The Winning Approach to Cross-Genre Gender Identification in Russian at RUSProfiling 2017

Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, Alexander Gelbukh
CIC, Instituto Politécnico Nacional
Mexico City, Mexico
imarkov@nlp.cic.ipn.mx, helena.adorno@gmail.com, sidorov@cic.ipn.mx, www.gelbukh.com

ABSTRACT

We present the CIC systems submitted to the 2017 PAN shared task on Cross-Genre Gender Identification in Russian texts (RUSProfiling). We submitted five systems. One of them was based on a statistical approach using only lexical features, and other four on machine-learning techniques using some combinations of gender-specific Russian grammatical features, word and character *n*-grams, and suffix *n*-grams. Our systems achieved the highest weighted accuracy across all the test datasets, occupying the first four places in the ranking.

KEYWORDS

Author Profiling, Gender Identification, Cross-Genre, Social Media, Russian, Machine Learning, Computational Linguistics

1 INTRODUCTION

Author profiling (AP) is the task of identifying the author's demographics, such as age, gender, personality traits, or native language, basing on a sample of his or her writing. This task has numerous practical applications in forensics, security, and marketing, to name just a few. For example, in forensics and terrorism prevention applications, knowing the characteristics of the suspect can narrow down the search space for the author of a written threat; in marketing applications, this information can be important to predict a customer's shopping preferences or develop new targeted products.

The rapid growth of social media data available on the Internet has significantly contributed to the increased interest in this task. This interest led to establishing of the annual PAN evaluation campaign[1], which is considered one of the main fora on AP, authorship attribution, plagiarism detection, and other tasks related to the study of authorship and characteristics of the author of a text.

Recent trends in the field include cross-genre AP scenario [17], that is, the setting when the training corpus consists of texts of one genre, while the test set consists of texts of another genre. Cross-genre AP conditions better match the requirements of a real-life scenario of forensic applications, when the available texts by the candidate authors can belong to genre and thematic area different from the texts under investigation.

Following the recent trends in the field, the 2017 PAN shared task on Gender Identification in Russian texts (RUSProfiling) [7] provided cross-genre AP scenario: the training corpus was composed of tweets, while the provided test datasets covered five different genres: offline texts (such as a letter to a friend or a picture descriptions), Facebook posts, tweets, product and service online reviews, and gender imitation texts.

Machine-learning methods are commonly used for the AP task. From the machine-learning perspective, the task is viewed as a multi-class, single-label classification problem, in which automatic methods are to assign class labels (e.g., `male` or `female`) to the text samples. Recently, deep-learning techniques [19], such as character-, word-, and document-embedding approaches [10], have been used for the task; however, linear models still perform better, since they seem to be more robust in capturing stylistic information in the author's writing. Therefore, we employ the commonly-used linear machine-learning approaches, as well as propose a novel statistical approach aiming to identify the gender of an author basing on statistical analysis of lexical information.

The paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we provide some characteristics of the datasets used in the RUSProfiling shared task 2017. In Section 4, we describe the conducted experiments, providing the experimental settings for the submitted systems. In Section 5, we give the obtained results and their evaluation. Finally, in Section 6 we draw some conclusions and point to possible directions of future work.

2 RELATED WORK

The PAN evaluation campaign has become one of the main platforms for evaluation of AP approaches and methodologies. There have been various profiling aspects covered by PAN since 2013 [15], including age, gender, personality traits, and language variety identification, under both single- and cross-genre AP conditions.

PAN 2017 [16] attracted 22 submissions. Most of the teams (including the top three systems) used traditional machine-learning algorithms, such as SVM [9, 11, 20] or logistic regression [4, 13]. This edition can be characterized by the increased use of deep-learning techniques [5, 18], in particular word and character embeddings [2, 4, 19], which are gaining popularity and achieving competitive, but still lower than the linear models, results for the AP task.

Content-based and style-based features have been extensively used in the previous editions of PAN. As content-based features, bag of words, word *n*-grams, slang words, locations, brand names, topic words, among others, were used by several teams. As style-based features, character *n*-grams are the most popular feature type for AP, other feature types include ratio of links, character flooding, typed character *n*-grams, emoticons, hashtags, and user mentions.

Due to the scarcity of available training data, AP research in the Russian language has been limited. The first corpus in the Russian

[1] http://pan.webis.de

language annotated with the authors' metadata information—the Ruspersonality corpus—was introduced by Litvinova *et al.* [6]. The corpus is composed of texts labeled with the author gender, age, personality traits, native language, neuropsycological testing data, and educational level. The corpus also contains a subset of truthful and deceptive texts. At the time publication of [6], the corpus contained over 1,850 documents.

Several experiments were carried out in order to illustrate the usefulness of the Ruspersonality corpus [6, 8]. For gender identification, Litvinova *et al.* [6] used a range of context-independent features such as part-of-speech (POS) tags, syntactic relations, ratios of POS tags, punctuation marks, and emotion words. They also evaluated different machine-learning algorithms: gradient boosting, adaBoosting, random forest, SVM, ReLU, among others. The best performance was obtained by ReLU (mean F1-score of 74%).

3 DATASETS

The focus of the RUSProfiling shared task 2017 is on cross-genre gender identification. The organizers provided a training dataset composed of tweets and five different test datasets on the following genres:

- **Test 1:** Offline texts (such as picture descriptions or letter to a friend) from the Ruspersonality Corpus [6].
- **Test 2:** Facebook posts.
- **Test 3:** Twitter messeges.
- **Test 4:** Product and service online reviews.
- **Test 5:** Gender imitation corpus, that is, women imitating men and vice versa.

Table 1 presents general statistics of the training and five test datasets. In the table, *No. of docs* stands for the number of documents in each dataset. The statistics of the average number (*Avg.*) of words and characters per document, as well as standard deviation (*Std.*), were calculated after applying pre-processing steps, which included lowercasing and removal of all non-cyrillic characters (punctuation marks were also removed). In terms of average number of words and characters, the Test 2 dataset is the most similar to the training corpus. The main difference between the two datasets is the standard deviation, which is larger in the training corpus. The Test 3 dataset is on the same genre as the training corpus, but it contains shorter documents, of 729.22 words on average. The Test 1 and Test 5 datasets have similar statistics in terms of the number of words and characters, but differ in the number of documents (370 and 94, respectively). Finally, the Test 4 dataset contains the shortest documents, of 54.40 words on average.

Table 1: RUSProfiling datasets statistics

Dataset	No. of docs	Words Avg.	Words Std.	Chars Avg.	Chars Std.
Training	600	1,216.16	731.61	7,736.20	4,674.29
Test 1	370	277.75	109.83	1,650.54	639.44
Test 2	228	1,096.60	164.41	6,900.66	1,106.27
Test 3	400	729.22	686.23	4,672.79	4,426.54
Test 4	776	54.40	44.39	354.18	276.56
Test 5	94	272.92	157.07	1,685.84	945.69

4 EXPERIMENTAL SETTINGS

To evaluate our systems, we conducted experiments both on the provided training dataset under 10-fold cross-validation and using 80%–20% dataset splitting, that is, we used 80% (480 documents) of the training dataset for training and 20% (120 documents) for evaluation. The splitting was balanced across the genders. Following the official evaluation metrics of the shared task, we measured the performance in terms of classification accuracy.

We applied several pre-processing steps before feature extraction. Pre-processing has proved to be a useful strategy for author profiling [3, 11] and related tasks, such as authorship attribution [12]. Keeping in mind that test datasets are in another genre, we kept only cyrillic characters (non-cyrillic characters along with punctuation marks were removed). We also performed lowercasing, which yielded slight improvement in accuracy. These pre-processing steps were applied in all our runs (in the context of this shared task, systems are officially called runs).

In all the runs based on machine-learning techniques, we used Support Vector Machines (SVM) algorithm, which is considered among the best-performing classification algorithms for text categorization tasks, including cross-genre AP scenario [17]. We used the liblinear scikit-learn [14] implementation of SVM with the OvR multi-class strategy. We set the penalty hyper-parameter C to 100 basing on the evaluation results. In our experiments on the training dataset, SVM showed higher performance than other classification algorithms we tried, such as random forest, logistic regression, multinomial Naïve Bayes, LDA, and ensemble classifier.

In our machine-learning approaches, we used two different implementations of the term frequency–inverse document frequency (tf-idf) weighting: the default scikit-learn implementation and tf-idf with sublinear tf scaling, i.e., tf was replaced with 1 + log(tf). In our experiments on the training dataset, tf-idf systematically outperformed other examined weighting schemes, such as binary, tf, and log entropy.

The configurations of the five runs of the CIC team are described below.

4.1 Run CIC-1 (machine learning)

Features Since in the Russian language singular forms of the past tense verbs change by gender (singular masculine forms have the ending *-л* "*-l*", while an indicator of singular feminine forms is the ending *-ла* "*-la*"), we used "word ending in *-ла*" as a feature. Moreover, since the past tense reflexive verbs maintain the reflexive ending *-сь* "*-s'* ", we also used the feature "word ending in *-лась*" "*-las'* ". We employed the features *-ла* "*-la*" and *-лась* "*-las'* " in isolation, as well as in combination with the subject of the sentence if the subject was the first-person singular pronoun *я* "*ya*" and if this subject was within the window of 6 words after, or 3 words before, the verb. This gave four additional composite features: "*я -ла*", "*я -лась*", "*-ла я*", and "*-лась я*" with the meaning such as "I -ed$_{\text{feminine}}$ myself", as in *I dressed myself in a skirt.* The window size (+6/−3) was selected based on grid search.

In addition, since Russian adjectives agree with the pronouns in gender, we used the ending *-ая* "*-aya*" (nominative feminine

singular form) in combination with the first person singular pronoun *я* "*ya*" as feature if the pronoun was within the same +6/−3 window as above. This gave two more features: "*я -ая*" and "*-ая я*", with the meaning such as "I -feminine-singular-adjective", as in *I am a professor emerita*.

Additionally, we used the last three (cyrillic) characters of each word as features (suffix *n*-grams, $n = 3$), which, in particular, indirectly accounted for other grammatically meaningful endings such as "*ный*" (hinting at masculine adjective, as in *I am a professor emeritus*).

Frequency threshold Fine-tuning the size of the feature set has proved to be of a great importance in AP [11]. It allows to reduce significantly the size of the feature set and at the same time to improve the results in most cases. In this run, we selected only those features that occurred in at least two documents in the training corpus and occurred at least five times in the entire training corpus (min_df = 2; threshold = 5).

Weighting scheme Tf-idf weighting with sublinear tf scaling.

4.2 Run CIC-2 (machine learning)

Features Word features represent the lexical choice of a writer. These features have proved to be indicative of author's gender in other languages, such as English, Spanish, Portuguese, and Arabic [16]. In this run, we used word unigram features (bag-of-words approach) in combination with the last three characters of each word (suffix 3-grams).

Frequency threshold The threshold was the same as in the CIC-1 run.

Weighting scheme Tf-idf weighting without sublinear tf scaling.

4.3 Run CIC-3 (statistical)

First, we labeled the words that occur in the training corpus as male's or female's, depending on whether the word was used (not counting repetitions) more frequently in male's or female's documents, except when the difference was less than 2.

Next, for each document we calculated the ratio of such male's to female's words (not counting repetitions). We labeled a document as male's if this ratio was above a threshold; otherwise, as female's. Since the dataset was balanced, as the threshold we used the median of the distribution of this ratio.

We also experimented with taking repetitions of words into account, thresholds other than 2 for classifying words, as well as with some formulas other than ratio for classifying documents; however, we observed a lower performance.

4.4 Run CIC-4 (machine learning)

Features Combination of word and character *n*-gram features usually provides good results for AP, for instance, a combination of word and character *n*-grams was used by the best performing system [1] at this year's PAN shared task [16]. In this run, we used a combination of word unigrams with character *n*-grams ($n = 2$–3).

Frequency threshold We selected only those features that occurred in at least two documents in the training corpus and occurred at least four times in the entire training corpus (min_df = 2; threshold = 4).

Table 2: 10-fold cross-validation and 80%–20% train-test split results (accuracy)

Run	10FCV acc.	No. of features	80%–20% acc.	No. of features
CIC-1	**0.8833**	3,136	**0.8583**	2,922
CIC-2	0.8550	19,139	**0.8583**	16,155
CIC-3	0.7400	22,847	0.7417	19,353
CIC-4	0.8683	31,045	0.8500	27,222
CIC-5	0.8683	22,625	**0.8583**	20,003

Weighting scheme We used tf-idf weighting with sublinear tf scaling.

4.5 Run CIC-5 (machine learning)

Features Word unigrams, word 3-grams, and character *n*-grams ($n = 2$–4).

Frequency threshold In this run, we set a hight frequency threshold value: we selected only those features that occurred in at least two documents in the training corpus and occurred at least 50 times in the entire training corpus (min_df = 2; threshold = 50). However, setting this high frequency threshold values only marginally affected 10-fold cross-validation and 80%–20% accuracy, making it very slightly higher or very slightly lower.

Weighting scheme Tf-idf with sublinear tf scaling.

5 RESULTS

The 10-fold cross-validation results, in terms of classification accuracy (*acc.*) for each run, as well as the results under 80%–20% dataset splitting, are shown in Table 2. For each experiment, the results for 10-fold cross-validation (*10FCV*) and 80%–20% splitting, as well as the number of features (*No. of features*), are provided. The best results for each evaluation procedure is highlighted in bold typeface.

Our first run, which included gender-specific Russian grammatical features, showed the highest 10-fold cross-validation accuracy with the smallest number of features. Three out of five of our runs (CIC-1, CIC-2, and CIC-5) showed the same accuracy under 80%–20% splitting, probably due to small size of the dataset. Statistical approach (run CIC-3) showed the lowest accuracy under both 10-fold cross-validation and 80%–20% setting, though, surprisingly, it showed the best results on several of the final test datasets, as shown in Table 3. We attribute this, again, to the small size of the datasets available for development.

A comparison of the participating systems, including the official ranking, is presented in [7]. We show the detailed results of our five runs on the five test datasets, along with the highest result achieved on each test set among all participating systems and the system that showed this result, in Table 3. The best result on each test dataset is highlighted in bold typeface. *Avg.* stands for the average accuracy of each run across the five test datasets; if a system was not tested on some test set, we counted its accuracy on this test set as zero. *Weighted* stands for the accuracy weighted by the number of documents in each test set (again, counting as zero if a system

Table 3: Results for the five runs of the CIC team on the five test sets

System	Test 1	Test 2	Test 3	Test 4	Test 5	Avg.	**Weighted**	Norm.
Best result	**0.7838**	**0.9342**	**0.6825**	**0.6186**	**0.6596**	**0.6580**	**0.6456**	**0.9258**
Best system	Bits_Pilani-4	CIC-2	CIC-3	CIC-3	Bits_Pilani-5	CIC-1	CIC-3	CIC-3
CIC-1	0.5865	0.9211	0.6525	0.5979	0.5319	**0.6580**	0.6435	0.9154
CIC-2	0.5838	**0.9342**	0.6650	0.5709	0.5213	0.6550	0.6354	0.9014
CIC-3	0.6027	0.7851	**0.6825**	**0.6186**	0.5426	0.6463	**0.6456**	**0.9258**
CIC-4	0.4676	0.8860	0.5975	0.5116	0.5213	0.5968	0.5675	0.8047
CIC-5	0.4973	0.8991	0.6275	0.5258	0.5000	0.6099	0.5862	0.8313
CIC best rank	4^{th}	$\mathbf{1^{st}}$	$\mathbf{1^{st}}$	$\mathbf{1^{st}}$	4^{th}	$\mathbf{1^{st}}$	$\mathbf{1^{st}}$	$\mathbf{1^{st}}$

was not evaluated on a test set); this was the measure used for the official ranking. *Norm.* is similar to *Weighted*, but is normalized by the highest accuracy on each test set (note that this is not accuracy; it is the average closeness of the given system to the best system).

As one can see from Table 3, none of the runs consistently outperformed other runs across all the test datasets. The Test 3 set consisted of documents that were collections of various tweets of the same author, similarly to the training corpus, so it was not exactly cross-genre scenario, but the documents in the Test 3 set contained fewer tweets than those of the training corpus. On this dataset, as well as on Test 4 with the shortest documents (online reviews), of our runs, the best performance was achieved by run CIC-3, which was based on the statistical approach. Test 2 (Facebook posts) was the only test set, on which our statistical approach (CIC-3) failed to produce good result.

Surprisingly, on the gender imitation corpus (Test 5), CIC-1 was our second-best run (after CIC-3), even though CIC-1 was based on gender-specific Russian grammatical (morphological) features, such as the grammatical gender of verbs and adjectives, which in imitated text follow the patterns of the gender being imitated.

Runs CIC-4 and CIC-5, in spite of showing similar 10-fold cross-validation and 80%–20% accuracy, performed worse on the test datasets than our first three runs. This can be due to the inclusion of character *n*-grams, which probably caused overfitting. Another reason for the relatively poor performance of CIC-5 could be the too high frequency threshold value set for this run.

For more in-depth analysis of the obtained results, the access to the golden standard for the test datasets would be required.

6 CONCLUSIONS

We have presented the description of the five systems submitted by the CIC team to the 2017 PAN shared task on Gender Identification in Russian texts (RUSProfiling), four of them occupying the first four places in the official ranking [7]. The task focused on cross-genre author profiling (AP) scenario: the training corpus was composed of tweets, while the provided test datasets were composed of offline texts, Facebook posts, tweets, online reviews, and gender imitation texts.

Our systems, which were not tuned for a specific genre, showed the highest accuracy on three out of five test datasets: Facebook posts, tweets, product and service online reviews, performing worse on two test datasets than more genre-specific systems, which were used only for some of the genres. Our first run based on a machine-learning approach using gender-specific Russian grammatical features showed the highest average accuracy across all the test datasets, while our statistical approach based on lexical features showed the best performance according to the weighted (official) and normalized evaluation.

One of the directions for future work would be to examine in more detail the importance of morphological features for gender identification in Russian texts, as well as to improve our statistical approach by automatically tuning the threshold value according to the size and genre of the test data.

ACKNOWLEDGMENTS

This work was partially supported by the Mexican Government (CONACYT projects 240844, SNI, COFAA-IPN, SIP-IPN 20171813, 20172008, and 20172044).

REFERENCES

[1] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2017. N-GrAM: New Groningen Author-profiling Model. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[2] Marc Franco-Salvador, Nataliia Plotnikova, Neha Pawar, and Yassine Benajiba. 2017. Subword-based Deep Averaging Networks for Author Profiling in Social Media. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[3] Helena Gómez-Adorno, Ilia Markov, Grigori Sidorov, Juan-Pablo Posadas-Durán, Miguel A. Sanchez-Perez, and Liliana Chanona-Hernandez. 2016. Improving Feature Representation Based on a Neural Network for Author Profiling in Social Media Texts. *Computational Intelligence and Neuroscience* 2016 (October 2016), 13 pages. https://doi.org/10.1155/2016/1638936

[4] Andrey Ignatov, Liliya Akhtyamova, and John Cardiff. 2017. Twitter Author Profiling Using Word Embeddings and Logistic Regression. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[5] Don Kodiyan, Florin Hardegger, Stephan Neuhaus, and Mark Cieliebak. 2017. Author Profiling with Bidirectional RNNs using Attention with GRUs. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[6] Tatiana Litvinova, Olga Litvinlova, Olga Zagorovskaya, Pavel Seredin, Aleksandr Sboev, and Olga Romanchenko. 2016. "Ruspersonality": A Russian Corpus for Authorship Profiling and Deception Detection. In *Proceedings of the 2016 International FRUCT Conference on Intelligence, Social Media and Web, ISMW-FRUCT 2016*. IEEE, St. Petersburg, Russia, 1–7.

[7] Tatiana Litvinova, Francisco Rangel, Paolo Rosso, Pavel Seredin, and Olga Litvinova. 2017. Overview of the RUSProfiling PAN at FIRE Track on Cross-genre Gender Identification in Russian. In *Notebook Papers of FIRE 2017, FIRE 2017 (CEUR Workshop Proceedings)*. CEUR-WS.org, Bangalore, India.

[8] Tatiana Litvinova, Pavel Seredin, Olga Litvinova, Olga Zagorovskaya, Aleksandr Sboev, Dmitry Gudovskih, Ivan Moloshnikov, and Roman Rybka. 2016. Gender

Prediction for Authors of Russian Texts Using Regression and Classification Techniques. In *Proceedings of the 3rd Workshop on Concept Discovery in Unstructured Data co-located with the 13th International Conference on Concept Lattices and Their Applications, CDUD@CLA*, Vol. 1625. CEUR-WS.org, 44–53.

[9] A. Pastor Lopez-Monroy, Manuel Montes-y-Gomez, Hugo Jair-Escalante, Luis Villasenõr Pineda, and Thamar Solorio. 2017. Social-Media Users can be Profiled by their Similarity with other Users. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[10] Ilia Markov, Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and Alexander Gelbukh. 2017. Author Profiling with Doc2vec Neural Network-Based Document Embeddings. In *Proceedings of the 15th Mexican International Conference on Artificial Intelligence, MICAI 2016*, Vol. 10062. Part II, LNAI, Springer, Cancún, Mexico, 117–131.

[11] Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov. 2017. Language- and Subtask-Dependent Feature Selection and Classifier Parameter Tuning for Author Profiling. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[12] Ilia Markov, Efstathios Stamatatos, and Grigori Sidorov. 2017. Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2017*. Springer, Budapest, Hungary.

[13] Matej Martinc, Iza Škrjanec, Katja Zupan, and Senja Pollak. 2017. PAN 2017: Author Profiling - Gender and Language Variety Prediction. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (November 2011), 2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195

[15] Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the Author Profiling Task at PAN 2013. In *Working Notes Papers of the CLEF 2013 Evaluation Labs (CEUR Workshop Proceedings)*. CLEF and CEUR-WS.org, Valencia, Spain, 23–26.

[16] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein. 2017. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*. CLEF and CEUR-WS.org, Dublin, Ireland.

[17] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. 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, Évora, Portugal.

[18] Nils Schaetti. 2017. UniNE at CLEF 2017: TF-IDF and Deep-Learning for Author Profiling. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[19] Sebastian Sierra, Manuel Montes-y-Gómez, Thamar Solorio, and Fabio A. González. 2017. Convolutional Neural Networks for Author Profiling. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.

[20] Eric S. Tellez, Sabino Miranda-Jiménez, Mario Graff, and Daniela Moctezuma. 2017. Gender and Language-Variety Identification with microTC. In *Working Notes Papers of the CLEF 2017 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1866. CLEF and CEUR-WS.org, Dublin, Ireland.