Overview of the RusProfiling PAN at FIRE Track on Cross-genre Gender Identification in Russian

Tatiana Litvinova
RusProfiling Lab
Russia
centr_rus_yaz@mail.ru

Francisco Rangel
Autoritas Consulting
Valencia, Spain
francisco.rangel@autoritas.es

Paolo Rosso
PRHLT Research Center
Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Pavel Seredin
RusProfiling Lab &
Kurchatov Institute
Russia
paul@phys.vsu.ru

Olga Litvinova
RusProfiling Lab &
Kurchatov Institute
Russia
olga_litvinova_teacher@mail.ru

ABSTRACT

Author profiling consists of predicting some author's traits (e.g. age, gender, personality) from her writing. After addressing at PAN@CLEF[1] mainly age and gender identification, in this RusProfiling PAN@FIRE track we have addressed the problem of predicting author's gender in Russian from a cross-genre perspective: given a training set on Twitter, the systems have been evaluated on five different genres (essays, Facebook, Twitter, reviews and texts where the authors imitated the other gender, where the users change their idiostyle). In this paper, we analyse the 22 runs sent by 5 participant teams. The best results (although also the most sparse ones) have been obtained on Facebook.

Keywords

author profiling; gender identification; cross-genre profiling; Russian;

1. INTRODUCTION

Author profiling involves predicting an author's demographics, personality traits, education and so on from her writing, with gender identification being the most popular task [10, 8, 12, 13, 11, 2, 5, 6, 15, 16, 4]. Author profiling tasks are popular among participants of PAN which is a series of scientific events and shared tasks on digital text forensics.[2] Slavic languages, however, are less investigated from an author profiling standpoint and have never been addressed at PAN.

This year at FIRE we have introduced a PAN shared task on Cross-genre Gender Identification in Russian texts (RusProfiling shared task) where we provided tweets as a training dataset and Facebook posts, online reviews, texts describing images or letters to a friend, as well as tweets as test datasets. The focus is especially on cross-genre gender profiling.

The rest of the overview paper is structured as follows. In Section 2, we describe the construction of the corpus and the evaluation metrics. In Section 3, participants' approaches are presented, and in Section 4 the obtained results are discussed. Finally, in Section 5 we draw some conclusions.

2. EVALUATION FRAMEWORK

In this section we describe the construction of the corpus, covering particular properties, challenges and novelties. Moreover, the evaluation measures are described.

2.1 Corpus

In this section, we describe the datasets that have been released for the tasks described in the previous section. We have designed these datasets using manual and automated techniques and made them available to participants through the task web page.[3]

Twitter dataset: (500 users per gender) was split into training (300 users per gender) and testing datasets (200 users per gender). Annotating social media texts is what makes designing such corpora particularly challenging. Some researchers automatically built Twitter corpora while others have solved this problem by using labor-intensive methods. For example, Rao et al. [14] use a focused search methodology followed by manual annotation to produce a dataset of 500 English users labeled with gender. The gender tag was ascribed based on the screen name, profile picture, self-description ('bio') and –in the few cases this was not sufficient– the use of gender markings when referring to themselves. For this research we used the same approach with manual labeling for tweet author gender. For those cases where the gender information was not clear we discarded the user. Retweets were removed.

The number of tweets from one user varied from 1 to 200 (depending on how active the users were at the time the data was collected – September 2016). All tweets from one user were merged together and considered as one text. As the analysis suggests, the tweets contain a lot of non-original information (hashtags, hidden citations (e.g., newsfeeds that are copied, etc.), hyperlinks, etc.), which makes it extremely challenging for them to be analyzed.

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

[2] http://pan.webis.de/index.html

[3] http://en.rusprofilinglab.ru/rusprofiling-at-pan/korpus/

Facebook dataset: 228 users (114 authors per gender) of different age groups (20+, 30+, 40+) from different Russian cities were randomly chosen (to get minimum mutual friendships). We used the same principals for gender labeling as were used for Twitter. All posts from one user were merged into one text with average length of 1000 words.

As well as for collecting data from Twitter, Facebook pages of famous people involved in administration or government or accounts of heads of major companies were not employed for the study. As the analysis show, in Russian Facebook texts there is less non-original information than on Twitter.

Essays dataset: 185 authors per gender, one or two texts per author (in case of two texts they were merged together and considered as one text). The texts were taken randomly from manually collected RusPersonality corpus [5]. RusPersonality is the first Russian-language corpus of written texts labeled with data on their authors. A unique aspect of the corpus is the breadth of the metadata (gender, age, personality, neuropsychological testing data, education level, etc). The texts were written by respondents especially for this corpus, do not contain any borrowings and are not edited. Topics of the texts were letter to a friend, picture description, letter to an employee trying to convince her to hire the respondent. The average text length in this dataset was 150 words.

Reviews dataset: 388 authors per gender, one text per author. The texts were collected from Trustpilot[4], the author's gender was identified based on the profile information. The average text length was 80 words.

Gender-imitated dataset: 47 authors per gender, three texts from each author that were merged together and considered as one text. The texts were randomly selected from the existing corpus we have collected called Gender Imitation Corpus. The Gender Imitation Corpus is the first Russian corpus for studies of stylistic deception. Each respondent (n=142) was instructed to write three texts on the same topic (from a list). Let us provide an example of the task: "Last summer you bought a package tour from a travel agency, but you were not at all pleased with your experience with that company and the trip was not worth the price. You are about to ask for a refund. Write three texts describing your negative experience providing a detailed account of it. Give a warning that you are intending to sue the company". The first text is supposed to be written in a way usual for whoever writes it (without any deception), the second one should be written as if by someone of the opposite gender ("imitation"); the third one should be as if one by another individual of the same gender so that her personal writing style will not be recognized (what is referred to as "obfuscation"). Most of the texts are 80-150 words long. All of the respondents are students of Russian universities. Besides the texts, the corpus includes metadata with the authors' characteristics: gender, age, native language, handedness, psychological gender (femininity/masculinity). Therefore, the corpus provides countless opportunities for investigating problems arising in imitating properties of the written speech in different aspects as well as gender (biological and psychological) imitation in texts. To the best of our knowledge, this is the first corpus of this kind. Presently, the corpus is being prepared to be made available on the RusProfiling Lab website.

In Table 1 a summary on the number of authors per dataset is shown.

Table 1: Distribution of authors per dataset (half per gender).

Dataset	Genre	Number of authors
Training	Twitter	600
Test	Essays	370
	Facebook	228
	Twitter	400
	Reviews	776
	Gender-imitated	94

2.2 Performance measures

For evaluating what done in the previous approaches we have used accuracy, following author profiling tasks at PAN. In the RusProfiling shared task, we have calculated the accuracy per dataset as the number of authors correctly identified divided by the total number of authors in this dataset. The global ranking has been obtained by calculating the average accuracy among all the datasets weighted by the number of documents in each dataset:

$$global\_acc = \frac{\sum_{ds} accuracy(ds) * size(ds))}{\sum_{ds} size(ds)} \tag{1}$$

2.3 Baselines

To understand the complexity of the task per genre and with the aim to compare the performances of the participants approaches, we propose the following baselines, as well as we did at PAN at CLEF in 2017 [11]:

- *majority*. A statistical baseline that emulates random choice. The baseline depends on the number of classes: two in case of gender identification.
- *bow*. This method represents documents as a bag-of-words with the 5,000 most common words in the training set, weighted by absolute frequency of occurrence, and it uses SVM as machine learning algorithm. The texts are preprocessed as follows: lowercase words, removal of punctuation signs and numbers, and removal of stop words for the corresponding language.
- *LDR* [9]. This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes. The key concept of LDR is a weight, representing the probability of a term to belong to one of the different categories (e.g. female vs. male). The distribution of weights for a given document should be closer to the weights of its corresponding category. LDR takes advantage of the whole vocabulary.

[4] https://ru.trustpilot.com/

3. OVERVIEW OF THE SUBMITTED APPROACHES

Following, we briefly describe the systems submitted by the five participants of the task, from three perspectives: preprocessing, features to represent the authors' texts and classification approaches. In Table 3 the teams and the corresponding references are presented.

Table 2: Participating teams and their references.

Team	Author
AmritaNLP	[18]
BITS_Pilani	[1]
CIC	[7]
DUBL	[17]
RBG	[3]

Preprocessing. Preprocessing was carried out to obtain plain text [1]. Various participants removed stopwords [1, 17], short words [17] and Twitter specific elements (user mentions, hashtags and links) [1, 17]. Some of them also removed punctuation marks [7, 1] as well as numbers [1], and the authors in [7] removed non-cyrillic characters. Finally, lemmatisation has been performed by the authors in [17].

Features. Traditionally, author profiling tasks have been approached with content and style-based features. In this vein, the authors in [18] extracted features such as the number of user mentions, hashtags and urls, emoticons, punctuation marks, and average word length, combined with tf-idf bag-of-words. Similarly, the authors in [7] combined different kinds of features in their systems such as word and character n-grams, words most frequently used per gender, linguistic patterns such as word endings or the use of first person singular pronouns within a distance to a verb in past tense. The mentioned linguistic rule has been combined with deep learning techniques in [1]. Finally, the authors in [17] performed topic modelling and the authors in [3] developed a representation scheme based on the texts belonging to the corresponding target classes.

Classification Approaches. Traditional features have been used with machine learning methods such as Support Vector Machines (SVM) [18, 7, 3], Random Forest [18] and AdaBoost [18]. The authors in [17] used Additive Regularization for Topic Modelling. Finally, the authors in [1], who combined a rule-based approach with deep learning, have used variations of Long-Short Term Memory networks.

4. EVALUATION AND DISCUSSION OF THE SUBMITTED APPROACHES

Due to the cross-genre perspective of the task, five datasets were provided. Five teams submitted a total of 22 runs, whose distribution per dataset is shown in Table 3. As can be seen, a total of 93 runs have been analysed, with 18-19 runs per dataset.

Table 3: Number of participants' runs per dataset.

Dataset	Number of runs
Essays	18
Facebook	19
Twitter	18
Reviews	19
Imitated	19
Total	93

The distribution of the results per dataset is shown in Figure 1. It is noteworthy the highest accuracy obtained on Facebook, with the median value about 75% and the highest one over 90%. However, results on this genre are the most sparse ones, with a standard deviation of 0.16. On the other hand, results on the gender-imitated corpus are the lowest ones, with most of the participants obtaining accuracies close to 50%, that would correspond to the majority class baseline. However, there were two participants who obtained results about 65%. In the following subsections we analyse the results per dataset more in depth.

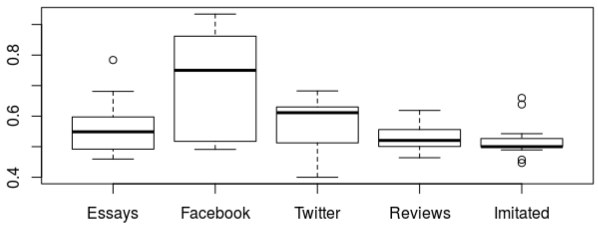

Figure 1: Distribution of results for gender identification in the different datasets.

4.1 Essays

Results on the essays dataset (Table 4) set forth an average accuracy of 55.39%, a median of 54.86% and a total of seven runs below the majority class and bow baselines. Apart from these low results, there are four runs improving in more than 10% this baseline, with accuracies between 60.27% and 78.38%.

The best result (78.38%) has been obtained by Bits_Pilani, who combined linguistic rules with deep learning techniques. The second best result (68.11%) has been obtained by AmritaNLP, who used stylistic features with traditional machine learning algorithms. As can be seen, the first result is more than 10% higher than the second one, and about 23% higher than the average, showing the power of deep learning in this task when training on Twitter and evaluating on essays. However, none of these systems overcame the LDR baseline (81.41%), that obtained a performance that was 3% and 13% higher, respectively.

Table 4: Accuracy in gender identification in essays.

Ranking	Team	Run	Accuracy
	LDR		0.8141
1	Bits_Pilani	4	0.7838
2	AmritaNLP	3	0.6811
3	dubl	4	0.6297
4	CIC	3	0.6027
5	AmritaNLP	2	0.5973
6	CIC	1	0.5865
7	CIC	2	0.5838
8	dubl	1	0.5486
9	dubl	2	0.5486
10	dubl	3	0.5486
11	AmritaNLP	1	0.5243
	bow		0.5027
	majority		0.5000
12	RBG	4	0.5000
13	CIC	5	0.4973
14	RBG	2	0.4919
15	CIC	4	0.4676
16	RBG	1	0.4595
17	RBG	3	0.4595
18	RBG	5	0.4595
	Min		0.4595
	Q1		0.4933
	Median		0.5486
	Mean		0.5539
	SDev		0.0861
	Q3		0.5946
	Max		0.7838

4.2 Facebook

Table 5: Accuracy in gender identification in Facebook.

Ranking	Team	Run	Accuracy
1	CIC	2	0.9342
2	CIC	1	0.9211
3	CIC	5	0.8991
4	CIC	4	0.8860
5	Bits_Pilani	5	0.8728
	LDR		0.8596
6	Bits_Pilani	3	0.8509
7	CIC	3	0.7851
	bow		0.7632
8	dubl	3	0.7588
9	dubl	2	0.7544
10	dubl	4	0.7500
11	AmritaNLP	1	0.7456
12	AmritaNLP	2	0.7237
13	AmritaNLP	3	0.6228
14	RBG	2	0.5351
	majority		0.5000
15	RBG	3	0.5000
16	RBG	4	0.5000
17	RBG	5	0.5000
18	RBG	1	0.4956
19	Bits_Pilani	2	0.4912
	Min		0.4912
	Q1		0.5175
	Median		0.7500
	Mean		0.7119
	SDev		0.1642
	Q3		0.8619
	Max		0.9342

In Table 5 the results on the Facebook dataset are shown. Both the average value (71.19%), the median (75%), the Q3 (86.19%) and the best value (93.42%) are the highest of all datasets. Indeed, they are even higher than the obtained on the Twitter dataset (shown in Table 6). However, the systems behaved in a heterogeneous way among datasets, obtaining the most sparse results with an inter-quartile range of 34.44%. The reason is due to five runs equal or below the majority baseline, and another run from the same participant very close to 50%. Furthermore, 12 systems performed worst than the bow baseline, that obtained an accuracy of 76.32%, even higher than the mean (71.19%) and the median (75%).

The four best results have been obtained by CIC, that trained SVMs with combinations of n-grams and linguistic rules, among others. The fifth and sixth best results have been obtained by BITS_Pilani with linguistic rules combined with deep learning. The best runs obtained a better performance than the LDR baseline of 2% and 12%, respectively. In this case, although the deep learning techniques obtained good results, they are more than 5% lower than traditional approaches.

4.3 Twitter

Table 6: Accuracy in gender identification in Twitter.

Ranking	Team	Run	Accuracy
1	CIC	3	0.6825
	LDR		0.6759
2	CIC	2	0.6650
3	Bits_Pilani	4	0.6525
4	CIC	1	0.6525
5	dubl	3	0.6300
6	CIC	5	0.6275
7	dubl	4	0.6275
8	AmritaNLP	3	0.6175
9	dubl	2	0.6125
10	AmritaNLP	2	0.6100
11	CIC	4	0.5975
12	AmritaNLP	1	0.5700
13	Bits_Pilani	2	0.5400
14	RBG	2	0.5125
	majority		0.5000
15	RBG	4	0.5000
	bow		0.4937
16	RBG	1	0.4650
17	RBG	3	0.4550
18	RBG	5	0.4000
	Min		0.4000
	Q1		0.5194
	Median		0.6112
	Mean		0.5787
	SDev		0.0815
	Q3		0.6294
	Max		0.6825

The results obtained on the Twitter dataset are shown in Table 6. The two best results (68.25%, 66.50%) have been obtained by CIC team, with the next result tied with BITS_Pilani (65.25%). These results are very similar to the one obtained by the LDR baseline (67.59%). The average result falls down to 57.87%, below the median of 61.12%, due to the low results obtained by most of the runs sent by

RBG team. In this vein, it is noteworthy to see that the results are below the majority baseline obtained by the bow baseline (49.37%).

Although the results on the Twitter dataset were expected to be the highest ones, they are much lower than the obtained on the Facebook dataset. In Facebook, besides maintaining the spontaneity of Twitter, posts use to be longer and grammatically richer, with fewer syntactic errors and misspellings. This may be the cause of the increase in accuracy. Furthermore, although the mean is higher, the best result in Twitter (68.25%) is 10% lower than the obtained in the essays dataset (78.38%).

4.4 Reviews

Results on the reviews dataset (Table 7) are lower than on the previous datasets although with lowest sparsity: most of the participants obtained results close to the average and median (52.87% and 52.06% respectively). As can be observed, these results are very close to the majority class (50%) and the bow baseline (50%), with five runs equal or below, and nine runs with less than a 5% of improvement. These low results expose the difficulty of the task on this genre when the training data comes from Twitter.

The best results have been achieved by CIC (61.86% and 59.79%) and Bits_Pilani (57.86% and 57.73%) teams, such as in the previous datasets (although about 4% lower than the 65.81% obtained by the LDR baseline). However, the difference is more than 7% in case of Twitter, 17% in case of essays and 30% in case of Facebook.

Table 7: Accuracy in gender identification in reviews.

Ranking	Team	Run	Accuracy
	LDR		0.6581
1	CIC	3	0.6186
2	CIC	1	0.5979
3	Bits_Pilani	5	0.5786
4	Bits_Pilani	4	0.5773
5	CIC	2	0.5709
6	AmritaNLP	1	0.5412
7	AmritaNLP	3	0.5296
8	CIC	5	0.5258
9	RBG	2	0.5232
10	RBG	4	0.5206
11	AmritaNLP	2	0.5155
12	Bits_Pilani	2	0.5142
13	CIC	4	0.5116
14	RBG	3	0.5013
	majority		0.5000
	bow		0.5000
15	RBG	1	0.5000
16	RBG	5	0.5000
17	dubl	3	0.4794
18	dubl	2	0.4755
19	dubl	4	0.4639
	Min		0.4639
	Q1		0.5007
	Median		0.5206
	Mean		0.5287
	SDev		0.0424
	Q3		0.5561
	Max		0.6186

4.5 Gender Imitation

In the gender-imitated corpus, the authors were asked to write the texts as if they were of the other gender or obfuscating their style, besides texts without imitation. In Table 8 the results of the gender identification task on this genre are shown. The average and median accuracies obtained by the systems on this dataset are the lowest (51.90% and 50% respectively). Most participants obtained accuracies close to the majority class and the bow baseline: 11 teams with an accuracy equal or lower than 50% and 6 teams with less than 5% of improvement. Only two runs of Bits_Pilani team obtained a significant improvement of 13% and 15% over the majority class. This team combined linguistic rules with deep learning techniques, showing the robustness of these techniques when the authors imitate the other gender and style. In this vein, we should highlight that LDR baseline (55.32%), AmritaNLP (54.26%) and CIC (54.26%), that obtained similar results among them, performed about 10% worst than the aforementioned deep learning techniques.

Table 8: Accuracy in gender identification in gender-imitated texts.

Ranking	Team	Run	Accuracy
1	Bits_Pilani	5	0.6596
2	Bits_Pilani	3	0.6383
	LDR		0.5532
3	AmritaNLP	1	0.5426
4	CIC	3	0.5426
5	CIC	1	0.5319
6	CIC	2	0.5213
7	CIC	4	0.5213
8	Bits_Pilani	1	0.5106
	majority		0.5000
	bow		0.5000
9	CIC	5	0.5000
10	dubl	2	0.5000
11	dubl	3	0.5000
12	dubl	4	0.5000
13	RBG	1	0.5000
14	RBG	3	0.5000
15	RBG	4	0.5000
16	RBG	5	0.5000
17	RBG	2	0.4894
18	AmritaNLP	2	0.4574
19	AmritaNLP	3	0.4468
	Min		0.4468
	Q1		0.5000
	Median		0.5000
	Mean		0.5190
	SDev		0.0517
	Q3		0.5266
	Max		0.6596

4.6 Global Ranking

The global ranking shown in Table 9 has been calculated following Formula 1. It is noteworthy that most participants obtained a weighted accuracy between 47% and 57%, with a median of 54.42%. That means that most of the participants obtained results close to the majority class (50%) and the bow baseline (53.13%). There are also three runs that obtained results much lower than the majority class due to their participation only on some datasets.

At the top of the ranking, we can highlight that the CIC team obtained the best first four results, with accuracies ranging from 58.62% to 64.56%, showing the robustness and homogeneity of their approach. However, it should be highlighted that, as Bits_Pilani runs different systems on the different datasets, although they obtained one of the bests results in each of them, a fair comparison has not been possible. For example, run 4 obtained 78.38% accuracy on essays (more than 10% than the next one), was not run neither on Facebook nor on gender-imitated sets, where the overall accuracy was lower. It is worth to mention that none of the systems outperformed the LDR baseline (71.21%), that obtained a 6.65% better performance with respect to the best system.

Table 9: Global ranking by averaging the accuracies on the different datasets, weighting by the size of the dataset.

Ranking	Team	Run	Accuracy
	LDR		0.7121
1	CIC	3	0.6456
2	CIC	1	0.6435
3	CIC	2	0.6354
4	CIC	5	0.5862
5	AmritaNLP	3	0.5857
6	AmritaNLP	2	0.5744
7	AmritaNLP	1	0.5691
8	dubl	4	0.5685
9	CIC	4	0.5675
10	dubl	3	0.5605
11	dubl	2	0.5546
12	Bits_Pilani	4	0.5337
	bow		0.5313
13	RBG	2	0.5145
	majority		0.5000
14	RBG	4	0.5086
15	RBG	1	0.4839
16	RBG	3	0.4829
17	RBG	5	0.4706
18	Bits_Pilani	2	0.3881
19	Bits_Pilani	5	0.3790
20	Bits_Pilani	3	0.1344
21	dubl	1	0.1065
22	Bits_Pilani	1	0.0236
	Min		0.0236
	Q1		0.4737
	Median		0.5442
	Mean		0.4780
	SDev		0.1740
	Q3		0.5731
	Max		0.6456

5. CONCLUSION

This paper describes the 22 systems sent by 5 participants to the RusProfiling shared task at PAN-FIRE 2017. Participants submitted a total of 93 runs on the five different datasets, with 18-19 runs per each dataset. They had to address the identification of the author's gender from a cross-genre perspective: given a training set of Twitter data, the systems have been evaluated on five different sets (essays, Facebook, Twitter, reviews and gender-imitated texts).

Participants have used different kinds of approaches, from traditional ones based on hand-crafted features and machine learning techniques such as Support Vector Machines, to the nowadays fashionable deep learning techniques. Depending on the genre, these approaches performed the best, such as in case of essays or the gender-imitated texts where they obtained more than 10% of improvement over the traditional ones.

Contrary to what was expected, the best results have not been achieved in Twitter but in Facebook. The reason may be that, although Facebook maintains the spontaneity of Twitter, their posts use to be longer and grammatically richer, with fewer syntactic errors and misspellings. On the other hand, almost the worst results have been obtained on reviews. Similar cross-genre effects were also observed at PAN-2014 [8].

In case of the gender-imitated texts, most systems failed, with 11 runs equal or below the majority baseline, and 6 runs with less than 5% of improvement. Only two systems of Bits_Pilani obtained results with more than 10% of improvement over the baseline. In this more difficult scenario, the deep learning approaches showed their superiority over traditional approaches.

6. ACKNOWLEDGMENTS

This work was supported in part of creation of Gender Imitation Corpus by the Russian Science Foundation, project No. 16-18-10050 ”Identifying the Gender and Age of Online Chatters Using Formal Parameters of their Texts”. Texts with style obfuscation were collected in the framework of the project ”Lie Detection in a Written Text: A Corpus Study” supported by the Russian Foundation for Basic Research N 15-34-01221. The third author acknowledges the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

7. REFERENCES

[1] R. Bhargava, G. Goel, A. Shah, and Y. Sharma. Gender identification in russian texts. In *Working Notes for PAN-RUSProfiling at FIRE'17. Workshops Proceedings of the 9th International Forum for Information Retrieval Evaluation (Fire'17), Bangalore, India*. CEUR-WS.org, 2017.

[2] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi, and F. Pianesi. The workshop on computational personality recognition 2014. In *Proceedings of the ACM International Conference on Multimedia*, pages 1245–1246. ACM, 2014.

[3] B. Ganesh HB, A. Kumar M, and S. KP. Representation of target classes for text classification - amrita_cen_nlp@rusprofiling pan 2017. In *Working Notes for PAN-RUSProfiling at FIRE'17. Workshops Proceedings of the 9th International Forum for Information Retrieval Evaluation (Fire'17), Bangalore, India*. CEUR-WS.org, 2017.

[4] T. Litvinova, D. Gudovskikh, A. Sboev, P. Seredin, O. Litvinova, D. Pisarevskaya, and P. Rosso. Author gender prediction in russian social media texts. In *Conf. on Analysis of Images, Social networks, and Texts, AIST-2017*.

[5] T. Litvinova, O. Litvinlova, O. Zagorovskaya, P. Seredin, A. Sboev, and O. Romanchenko. ” ruspersonality”: A russian corpus for authorship profiling and deception detection. In *Intelligence, Social Media and Web (ISMW FRUCT), 2016 International FRUCT Conference on*, pages 1–7. IEEE, 2016.

[6] T. Litvinova, P. Seredin, O. Litvinova, O. Zagorovskaya, A. Sboev, D. Gudovskih, I. Moloshnikov, and R. Rybka. Gender prediction for authors of russian texts using regression and classification techniques. In *CDUD@ CLA*, pages 44–53, 2016.

[7] I. Markov, H. Gomez-Adorno, G. Sidorov, and A. Gelbukh. The winning approach to cross-genre gender identification in russian at rusprofiling 2017. In *Working Notes for PAN-RUSProfiling at FIRE'17. Workshops Proceedings of the 9th International Forum for Information Retrieval Evaluation (Fire'17), Bangalore, India*. CEUR-WS.org, 2017.

[8] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans. Overview of the 2nd author profiling task at pan 2014. In *Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org, vol. 1180*, 2014.

[9] F. Rangel, P. Rosso, and M. Franco-Salvador. A low dimensionality representation for language variety identification. In *17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*. Springer-Verlag, LNCS, arXiv:1705.10754, 2016.

[10] F. Rangel, P. Rosso, M. Moshe Koppel, E. Stamatatos, and G. Inches. Overview of the author profiling task at pan 2013. In *Forner P., Navigli R., Tufis D. (Eds.), CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org, vol. 1179*, 2013.

[11] F. Rangel, P. Rosso, M. Potthast, and B. Stein. 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, Sept. 2017.

[12] F. Rangel, P. Rosso, M. Potthast, B. Stein, and W. Daelemans. Overview of the 3rd author profiling task at pan 2015. In *Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391*, 2015.

[13] 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 *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, Sept. 2016.

[14] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[15] A. Sboev, T. Litvinova, D. Gudovskikh, R. Rybka, and I. Moloshnikov. Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science*, 101:135–142, 2016.

[16] A. Sboev, T. Litvinova, I. Voronina, D. Gudovskikh, and R. Rybka. Deep learning network models to categorize texts according to author's gender and to identify text sentiment. In *Computational Science and Computational Intelligence (CSCI), 2016 International Conference on*, pages 1101–1106. IEEE, 2016.

[17] G. Skitalinskaya, L. Akhtyamova, and J. Cardiff. Cross-genre gender identification in russian texts using topic modeling working note: Team dubl. In *Working Notes for PAN-RUSProfiling at FIRE'17. Workshops Proceedings of the 9th International Forum for Information Retrieval Evaluation (Fire'17), Bangalore, India*. CEUR-WS.org, 2017.

[18] V. Vinayan, N. J.R., H. NB, A. Kumar M, and S. K P. Amritanlp@pan-rusprofiling: Author profiling using machine learning techniques. In *Working Notes for PAN-RUSProfiling at FIRE'17. Workshops Proceedings of the 9th International Forum for Information Retrieval Evaluation (Fire'17), Bangalore, India*. CEUR-WS.org, 2017.