2018 International Conference on Multidisciplinary Research
Track 1: STEM (Science, Technology, Engineering & Mathematics)
https://doi.org/10.26803/MyRes.2018.04

A Gender Identification of Russian Text Author on Base of Multigenre Data-Driven Approach using Machine Learning Models

Aleksandr Sboev
NRC "Kurchatov Institute", MEPhI National Research Nuclear University
Moscow, Russia

Ivan Moloshnikov, Dmitry Gudovskikh and Roman Rybka
NRC "Kurchatov Institute",
Moscow, Russia

Abstract
In this work data-driven approaches to identify the gender of author of Russian text are investigated with the purpose to clarify, to what extent the machine learning models trained on texts of a certain genre could give accurate results on texts of other genre. The set of data corpora includes: one collected by a crowdsourcing platform, essays of Russian students (RusPersonality), Gender Imitation corpus, and the corpora used at Forum for Information Retrieval Evaluation 2017 (FIRE), containing texts from Facebook, Twitter and Reviews. We present the analysis of numerical experiments based on different features(morphological data, vector of character n-gram frequencies, LIWC and others) of input texts along with various machine learning models (neural networks, gradient boosting methods, CNN, LSTM, SVM, Logistic Regression, Random Forest). Results of these experiments are compared with the results of FIRE competition to evaluate effects of multi-genre training. The presented results, obtained on a wide set of data-driven models, establish the accuracy level for the task to identify gender of an author of a Russian text in the multi-genre case. As shown, an average loss in F1 because of training on a set of genre other than the one used to test is about 11.7%.

Keywords: data-driving, machine learning, multigenre gender identification.

Introduction

A few last years the task of author's profiling identification with data driven approach is investigated greatly from different points of view because of its indisputable importance for practical implementations, commercial and governmental. Recently a new accent of this task has emerged: to what extent the

machine learning models trained on text of some certain genre could give accurate results on texts of other genres. This question was partly addressed at the PAN 2016 (Rangel F., 2016) competition where Twitter was used for training, and different corpora from social media, blogs, essays, and reviews were used for evaluation. A similar competition on Forum for Information Retrieval Evaluation 2017 (Litvinova T., 2017) took place for Russian texts. But the format of the competition was hardly appropriate to clarify this question of solving multi-genre task of profiling, because according to the competition rules only the single best run was taken into account without cross-validations. This is particularly discernible for such languages as Russian, for which the size of existing corpora is currently not so large (600 authors with 1 to 200 tweets per author, compared to 428 authors in English and 250 authors in Spanish with up to 1,000 tweets per author) and, consequently, the result of a single run is more stochastic. As a result, this makes it difficult to evaluate multi-genre effects. The purpose of this paper is to evaluate the accuracy of solving multi-genre profiling task more correctly with cross-validation, using an extended set of text features along with neural net and machine learning models.

Models, which are effective and largely independent of genre and external dictionaries, have been proposed in our previous work (Sboev A., 2017).The sets of used features included: morpho-syntactic, linguistic features (LIWC), a generalized dictionary approach of low displacement rank (LDR), along with different variants of vector representation of the text. The adaptations of these models and more complex models, based on convolutional neural networks (CNN), long short-term memories (LSTM), gradient boosting methods, are described in Section Materials and Methods, see Subsection Models. The experiments (described in the section Materials and Methods: Experiments) were carried out on the basis of various combinations of these corpora and models with various features. All used corpora and subcorpora are described in Section Materials and Methods: Datasets description. The results are summarized in Section Results and Conclusions.

Material and methods

Dataset description

The Gender Imitation-Crowdsource (CS) corpus was collected using the crowdsourcing platform. Each participant wrote 3 texts on one topic out of: a letter to a friend, a post for a dating site, a complaint letter to the boss or a negative review on the tour operator. The first text had to be written in the natural style (these texts are further referred to as `CS a' collection), the second in form to mimic opposite gender (`CS b'), and the third in form to imitate different style but not the gender (`CS c').

At the preprocessing stage, the texts were manually and automatically checked for borrowing on the Internet. After this stage, the corpus contained 5150 documents of, 1205 unique authors, 615 female and 588 male. Some authors wrote several texts for each collection, in which case we combined these texts into a single document. The average length of combined documents is about 300 words The number of authors

was balanced by excluding excessive authors, after balancing the resulting dataset contained about 800 authors on average.

We tried combining corpus documents into learning samples in several ways:
'CS' – contains parts A, B and C from each author as separate texts.
'CS combined' - in this sample parts A, B and C were combined into one document for each author.
'Cs ab' - in this sample parts A and B were combined into one document for each author.
The corpus 'Cs free theme' was collected according to the same exercise as Gender Imitation Croudsource but on a free topic. This corpus was also split to 'Cs free theme', 'Cs free theme A', 'Cs free theme B', 'Cs free theme C', and 'Cs free theme AB', analogously to 'CS', 'CS a', 'CS b', 'CS c', and 'CS ab' respectively.
In addition, we used the Gender Imitation corpus (GI) that was collected under the same rules as CS, but offline, in a fully controllable environment. This corpus also have parts 'GI a', 'GI b', 'GI c', 'GI combined' and 'GI ab' formed in the same way.

RusPersonality(RusPer) (Litvinova T. Z. O., 2016): is a corpus of Russian-language texts labeled with a large amount of metadata on their authors like gender, age, personalities, education level, neuropsychological testing data, etc. The corpus currently contains over 1850 documents, 230 words long in average, from 1145 respondents. This paper uses a part of the corpus, consisting of 1108 texts on two topics: letter to a friend and picture description.

RusProfiling (RusProfiling, 2017): texts collected from different social media platforms as Twitter and Facebook , along with reviews. In this paper, this data set was divided into subsamples, in order to study the influence of the corpus features on the quality of the model and to evaluate the possibility of genre-independent classification:
Reviews - manually collected reviews (641 men, 392 women).
Twitter - twitter messages (998 men, 543 women). 1000 messages were collected for each user, and then merged into one text.
Facebook - messages from walls, combined into one text (136 men, 114 women).

Features
Group 1. Linguistic Inquiry and Word Count (LIWC)
LIWC – is a set of psychosocial dictionaries (Tausczik, 2010) that describe linguistics categories (the number of words of certain parts of speech, some lexical-thematic groups, the frequency of punctuation marks, etc.) adapted for Russian language (Litvinova, 2017)

Group 2. LDR - Lower dimentional reduction
In this model the document representation is as a vector of dependencies on categories (class-dependent vector). It is calculated based on the matrix of TF-IDF terms and class-dependent term weights. Each document is represented as $F(c_1)$,

$F(c_2)$, ..., $F(c_n) \sim \forall c \in C$, where n – number of categories, $F(c_i)$ contains: the average weight of a document calculated as the sum of weights of its terms divided by the total number of vocabulary terms of the document, the standard deviation of weights of documents, the minimum weight among the term weights of document, the maximum weight found in the document, the sum of weights of the terms of the document divided by the total number of terms of the document, the proportion between the number of vocabulary terms of the document and the total number of terms of the document. The model is described in the work "A Low Dimensionality Representation for Language Variety Identification" (Rangel, 2017).

Group 3 Model Synt
As input to neural network each document is represented as a sequence of words, where each word is encoded by a syntactic pair "word-parent". The word and the parent are encoded by three types of features: unique index of lemma, Word2vec vector for word, (Kutuzov, 2017), binary encoded morphological tag.

Group 4. Model NN
As input for neural network the morphological characteristics of words are used: noun, verb, nominative case, masculine gender, feminine gender, etc.

Group 5. Symbol model.
In this model symbol-representation of text was used where each symbol convert to a one hot encoding vector. All English and Russian characters are used for encoding.

The features vector (groups 3, 4, 5) for the input document is formed based on the distribution of weights of the pre-trained neural networks described above.

Group 6. The vector of document is formed on the basis of the calculation of the term frequency-inverse document frequency (TF-IDF) of symbols n-grams in range from 3 to 5 letters.

Models
1) NNSyntMorph. Deep neural network based on morpho-syntactic features. The topology is shown on Fig. 1. Batch size is 32, early stopping after 5 epochs.

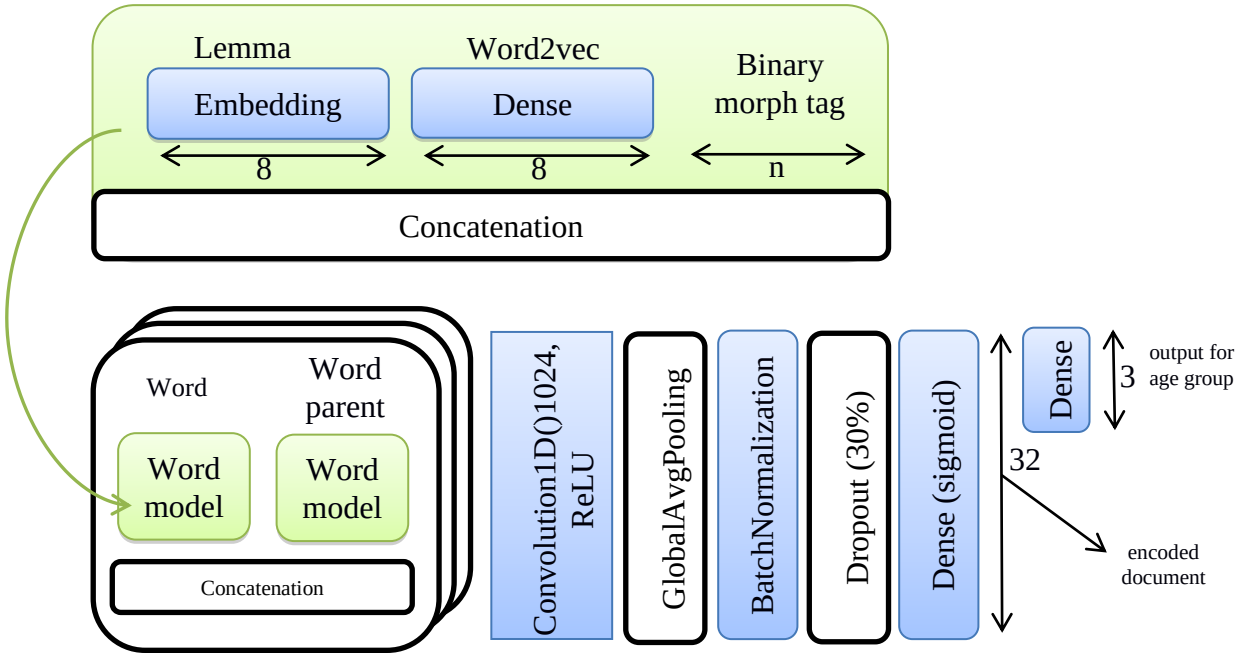

Figure 1 – Topology of morpho-syntactic neural network (NNSyntMorph)

2) Model NN. The topology of the network was taken from the work to determine the gender of the author, it based only on the morphological characteristics of words. (Sboev, 2016). A complicated neural network combining CNN, MLP and LSTM includes:

- 1st, 3rd, 5th CNN layers: Number of convolution kernels to use = 128, the extension of each filter = 2, activation function is ReLU
- 2nd, 4th, 6th layers: MaxPooling (pool length = 2)
- 7th layer: Long-Short Term Memory (output dimension = 128)
- 8th layer: dropout layer. (Fraction of the input units to drop = 0.5)
- 9th layer: fully connected NN layer (Number of hidden neurons = 10, activation function = softmax)

Learning parameters:
The learning process includes cross-validation with 10 permutation-and-split iterations, 80% of samples for training, 20% for testing.

3) Symbols model. This model uses symbol-representation of text where each symbol is converted to a one-hot encoded vector. The topology and parameters of network:

- 1st, 3rd, CNN layers: Number of convolution kernels to use = 8, the extension of each filter = 2, activation function is ReLU
- 2nd, 4th, 6th layers: MaxPooling (pool length = 2)
- 5th CNN layer: Number of convolution kernels to use = 32, the extension of each filter = 2, activation function is ReLU
- 7th layer: Global average pooling operation for temporal data.
- 8th layer: dropout layer. (Fraction of the input units to drop = 0.3)
- 9th layer: fully connected NN layer (Number of hidden neurons = 3, activation function = softmax)

The optimizer in models 1), 2), and 3) is the Adam algorithm (Kingma, 2014) based on gradient descent, with mean squared error as the optimization score function.

4) We used conventional models: RandomForestClassifier, LogisticRegression and Gradient Boosting Classifier (Gradient Boosting). These standard models were used with default hyperparameters of the sklearn library. Support vector machine(SVM) with linear kernel. As hyperparameters the following were used: regularization parameter C is 1.0, L2-norm used in the penalization and squared hinge-loss function. Each model uses features shown in Table 3.

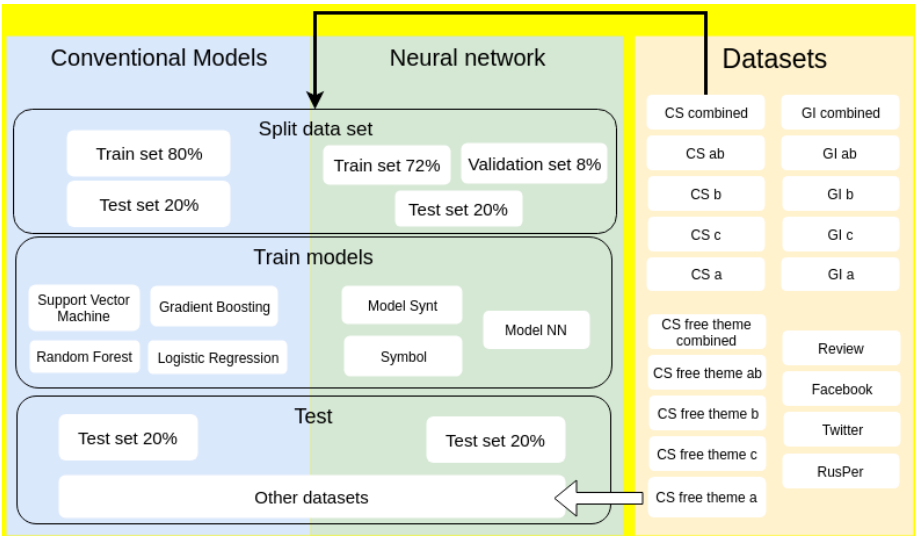

Figure 2. General flow diagram of experiments.

Experiments

We have performed a set of author identification experiments for identifying author gender by text features, employing various neural nets and machine learning models. The textual training sets are of different genres: Twitter (as at PAN@FIRE 2017, that allows to compare our results with its results); Facebook; different partitions of CS (a, b, c, ab, combined), RusPersonality, and Gender Imitation corpus. In case of conventional models the corpus is split to 20% for testing set, and 80% for training, or in case of neural network models 72% for training, 8% for validation, and 20% for testing. To investigate the relation of two genres, say, 1 and 2, we compare the results of two types of calculations, cross-genre and single-genre. The first type is to train (as well as validate when necessary) on texts of genre 1 with testing on texts of genre 2, and then vice versa: to train on genre 2 with testing on genre 1. The second type is to use texts of the same genre for training and testing. The general flow diagram of the experiments is shown in Figure 2. Each training set is balanced by excluding excessive examples, and 10-fold cross-validation is performed.

Table 1 – Cross-genre testing with non-deceptive corpora. Rows indicate training sets (number of examples, average document length), columns – testing sets (number of examples, average document length). F1-score(%) is presented, with standard deviation in brackets.

Corpus	Reviews (784, 62)	RusPer (1150, 173)	Cs a (1664, 84)	Cs free theme a (320, 102)	Facebook (228, 1389)	Twitter (1086, 1692)	GI a (90, 150)	Average	Std
Reviews (784, 62)	**79 (03)**	72 (02)	63 (02)	81 (02)	79 (03)	71 (01)	68 (04)	73	6
RusPer (1150, 173)	73 (01)	**82 (02)**	71 (01)	87 (01)	75 (05)	72 (01)	75 (04)	76	5
Cs a (1664, 84)	72 (02)	76 (03)	**78 (02)**	**85 (02)**	69 (11)	67 (01)	80 (03)	75	5
Cs free theme a (320, 102)	71 (01)	75 (02)	64 (04)	**82 (06)**	77 (04)	70 (03)	69 (06)	73	5
Facebook (228, 1389)	67 (02)	66 (01)	51 (03)	71 (02)	**84 (06)**	66 (01)	59 (04)	66	7
Twitter (1086, 1692)	72 (01)	73 (01)	62 (02)	79 (01)	**85 (02)**	**79 (02)**	69 (03)	74	6
GI a (90, 150)	57 (04)	58 (03)	56 (02)	58 (03)	49 (08)	51 (02)	**61 (06)**	56	3
Average	70	72	64	78	74	68	69		
Std	5	6	6	8	9	6	5		

Results

Table 1 demonstrates the F1-scores of the experiments mentioned above with corpora without other gender imitation or style distortion (so, for datasets containing A, B, and C partitions, only A subsets were selected). Rows correspond to datasets used for training, and columns correspond to testing datasets; so that on the diagonal (marked grey) are the non-cross-genre results, trained and tested on the same datasets. Each cell of the table shows the F1-score and standard deviation obtained by the model that is the best for that particular pair of training and testing set. Std denotes the standard deviations for average values in columns and rows.

These results show that, in general, in case of evaluating different genre corpora without gender deception the resulting F1 are in range 49%-87%, average 70.5%. The result accuracy mostly depends on the size of the training set: the smaller the set is (see the corpora sizes in row and column heads), the lower F1 is. For example see Table 1, lines with GI a and Facebook, which give lower result on every testing dataset.

Table 2 – Cross-gender testing, including corpora with gender imitation or style distortion. Before slash - F1 (standard deviation in brackets) of training on a dataset stated in the row and testing on a dataset stated in the column, after slash -vice versa, training on a dataset in the column, and testing on a dataset in the row. Names of dataset in columns and rows consists number of examples and average document length.

Corpus	Reviews (784, 62)	RusPer (1150, 173)	Facebook (228, 1389)	Twitter (1086, 1692)	Average	Std	Same-genre test
Cs (4996, 82)	58(02)/ 52(01)	65(10)/ 56(01)	57(03)/ 46(02)	57(03)/ 53(01)	59/ 52	3/ 3	**59 (01)**
Cs free theme (960, 104)	55(03)/ 59(02)	63(06)/ 62(01)	61(05)/ 52(02)	58(03)/ 58(01)	59/ 58	3/ 3	**61 (02)**
GI (250, 117)	54(02)/ 57(03)	50(04)/ 56(03)	45(07)/ 50(02)	48(02)/ 56(02)	49/ 55	3/ 2	**70 (04)**
Average	56 / 56	59 / 58	54 / 49	54 / 56			
Std	2 / 3	6 / 3	6 / 2	4 / 2			
Cs ab (1544, 77)	55(02)/ 49(02)	55(02)/ 51(01)	52(04)/ 45(01)	50(02)/ 49(02)	53/ 49	2/ 2	**58 (02)**
Cs free theme ab	49(01)/ 53(01)	53(03)/ 53(01)	54(03)/ 47(01)	50(04)/ 52(01)	52/ 51	2/ 2	**66 (05)**
GI ab	52(02)/ 54(02)	49(01)/ 52(02)	47(04)/ 48(02)	48(02)/ 54(02)	49/ 52	2/ 2	**62 (11)**
Average	52 / 52	52 / 52	51 / 47	49 / 52			
Std	2 / 2	2 / 0,7	3/ 1	0,9 / 2			
Cs b (1666, 79)	53(01)/ 46(03)	46(02)/ 44(01)	42(03)/ 41(03)	42(02)/ 42(01)	46/ 43	4/ 2	**78 (02)**
Cs free theme b (320, 103)	53(02)/ 50(02)	47(02)/ 38(01)	43(02)/ 39(02)	48(01)/ 37(01)	48/ 41	3/ 5	**82 (04)**
GI b (84, 100)	50(03)/ 42(05)	45(03)/ 40(02)	42(08)/ 43(04)	47(03)/ 46(03)	46/ 43	3/ 2	**66 (14)**
Average	52 / 46	46 / 41	42 / 41	46 / 42			
Std	1 / 3	0,7 / 2	0,4 / 1	2 / 3			
Same-genre test	**79(03)**	**82(02)**	**84(06)**	**79(02)**			

Training and testing on different genres cause an average loss in f1-measure is approximately 11.7%. As it follows from Table 1, the best result of testing on Crowdsourcing and Gender Imitation texts is shown by training on RusPersonality texts, which is not surprising, because the methodologies to form these corpora were similar, so these texts were written in a similar stylistic manner.

Table 3 – Cross-genre testing results for "combined" partitions of CS, CS-free-theme, and Gender Imitation datasets. The layout is same as in Table 2.

Corpus	Reviews (784, 62)	RusPer (1150, 173)	Facebook (228, 1389)	Twitter (1086, 1692)	Average	Std	Same-genre test
Cs combined	57(01)/ 54(01)	61(01)/ 57(01)	59(02)/ 51(01)	57(03)/ 54(02)	59/ 54	2/ 2	**60 (05)**
Cs free theme combined	56(13)/ 62(03)	62(03)/ 58(02)	62(06)/ 57(01)	57(03)/ 56(02)	59/ 58	3/ 2	**70 (21)**
Gender Imitation combined	51(01)/ 56(06)	50(04)/ 56(03)	49(04)/ 48(04)	49(02)/ 59(04)	50/ 55	0,8/ 3	**60 (11)**
Average	55 / 57	58 / 57	57 / 52	54 / 56			
Std	2 / 3	5 / 0,7	5 / 3	4 / 2			
Cs C (1666, 82)	72(01)/ 60(03)	72(01)/ 66(00)	65(03)/ 52(03)	65(04)/ 61(01)	69/ 60	4/ 4	**71 (01)**
Cs free theme C (320, 106)	68(02)/ 72(04)	73(03)/ 78(02)	67(05)/ 62(04)	67(02)/ 71(02)	69/ 71	2/ 4	**74 (05)**
GI C (76, 97)	52(02)/ 67(06)	49(05)/ 67(04)	41(11)/ 56(02)	48(05)/ 61(05)	4/ 63	3/ 4	**58 (12)**
Average	64 / 66	65 / 70	58 / 57	60 / 64			
Std	8 / 4	10 / 5	11 / 4	8 / 4			
Same-genre test	**79(03)**	**82(02)**	**84(06)**	**79(02)**			

Tables 2 and 3 present cross-genre testing results for datasets with gender imitation and style distortion. The F1-scores are averaged over all the models used, in contrast to best models' results in Table 1. In addition, average F1-score and standard deviation over similar partitions (B, C, AB, combined, and ABC contains parts A, B and C of all authors as separate texts) of different datasets are shown. For comparison, the rows and columns "same-genre test" present the same-genre testing results like the diagonal in Table 1: training and testing on the same dataset of the corresponding row or column.

Among Internet-texts training sets, the best F1 results for testing on GI and CS datasets are shown by training on Reviews (average result is 58%), then Twitter (56.6%), and slightly worse on Facebook (51%). Generally, the presence of texts with gender and style distortions in testing set increases the magnitude of loss in F1 up to 30%.

Table 3 represents ranked levels of efficiency of different input text features with various machine learning models, calculated over all test experiments. Formula 1 demonstrate how to evaluate the rank.

$$Rank = \frac{\sum_{j=1}^{n} i \in M}{n} (1)$$

Table 3 – Rank of models based on results of all experiments.

Model name	**Rank**
Gradient Boosting (Group 6. Char ngram)	2,44
SVM(with LIWC and LDR features)	3,21
Random Forest (based on LIWC and LDR features)	3,49
Logistic Regression(with LIWC and LDR features)	3,62
NNSyntMorph	4,6
Model NN	4,64
Symbols model	5,99

where i – a place in j-th experiment, place means the position in the F1 rating of the set of models M for j-th pair of training and testing datasets, n is the total number of such pairs. In Table 4 our results are compared to that of FIRE best runs, and these results, in general, are close.

Table 4 – Comparison with results of FIRE competition. Names of dataset in columns and rows contains number of examples and average document length. F1-score is presented with standard deviation consist in brackets.

Train\Test	**Reviews (784, 62)**	**RusPer (1150, 173)**	**RusPer (370, 150)**	**Facebook (228, 1389)**	**Twitter (1086, 1692)**	**GI combined**
Twitter (1086, 1692)	72 (01)	73 (01)	-	85 (02)	79 (02)	59 (04)
Twitter FIRE (600, 1600)	61	-	78	93	68	65

Conclusion

The presented results, obtained on a wide set of data-driven models, establish the accuracy level for author gender identification task on the Russian texts in the case of multi-genre. That is, we trained our models on texts of one genre and tested on texts of other genres.

The obtained loss in F1 compared to non-multi-genre case about 10% in average (for texts without gender imitation and style distortion) and depends on the size of training set. The presence of texts with gender deceptions and style distortions in testing set makes the magnitude of loss yet more in F1, up to 30%. The most effective model in our case is the Gradient Boosting model with symbolic n-gram representation of input text (Table 3). In future work we will continue collecting our crowdsourcing corpus, which would enlarge the size of training sets on cross-genre base to create highly effective model appropriate for the wide range of different genres, while at the same time effective for texts with gender deceptions.

Acknowledgements

This research is supported by the Russian Science Foundation, project No 16-1810050. This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC "Kurchatov Institute", http://ckp.nrcki.ru/

References

Kingma, D. P. (2014). *Adam: A method for stochastic optimization*. Retrieved from arXiv preprint arXiv:1412.6980

Kutuzov, A. &. (2017). Building web-interfaces for vector semantic models with the webvectors toolkit. *In Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. pp. 99-103.

Litvinova T., R. F. (2017). Overview of the RUSProfiling PAN at FIRE Track on Cross-genre Gender Identification in Russian. *Notebook Papers of FIRE*, pp. 8-10.

Litvinova T., Z. O. (2016). "Ruspersonality": A Russian corpus for authorship profiling and deception detection. *In Intelligence, Social Media and Web (ISMW FRUCT), 2016 International FRUCT Conference on. IEEE*, pp. pp. 1-7.

Litvinova, O. S. (2017). Deception detection in russian texts. *In Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. pp. 43-52.

Rangel F., R. P. (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/Balog, Krisztian [edit.]; et al.*, pp. pp. 750-784.

Rangel, F. F.-S. (2017). *A low dimensionality representation for language variety identification*. Retrieved from arXiv preprint arXiv:1705.10754

RusProfiling. (2017). *RusProfiling Lab 2017 Rusprofiling corpus of russian texts*. Retrieved from [online]: http://rusprofilinglab.ru/rusprofiling-at-pan/

Sboev A., M. I. (2017). A comparison of Data Driven models of solving the task of gender identification of author in Russian language texts for cases without and with the gender deception. *In Journal of Physics: Conference Series (Vol. 937, No. 1). IOP Publishing.*, p. p. 012046.

Sboev, A. L. (2016). Machine learning models of text categorization by author gender using topic-independent features. *Procedia Computer Science, 101*, pp. pp. 135-142.

Tausczik, Y. R. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology, 29(1)*, 24-54.