

# Gender Profiling in Social Network

Dmitry Badeev

December 2021

## Abstract

Topic of this work is *Robust (unbiased) methods of corpus research of social linguistic variability* on the example of the language of social networks.

The Task is *Single-genre and multi-genre gender profiling of authors by their social media texts*.

The proposed approach in problem solution is based on the use of a pretrained language model *Russian language SBERT*, with further fine-tuning for the training sample, in combination with stylometric methods.

Project code, Datasets are available here: [https://github.com/dbadeev/gender\\_profiling](https://github.com/dbadeev/gender_profiling)

## 1 Introduction

A goal of creating machine learning data driven models in computer linguistics is to construct the model most independent on genres, topics, and styles of text presented in training and testing sets. This fully concerns the task of gender identification of text author.

In this paper we'll investigate two questions: 1) in what degree the machine learning models trained on text of some certain genre could give accurate results on texts of other genres, and 2) the influence of style distortions and gender deceptions to the results of gender identification.

This theme was partly addressed at the PAN 2016 [F. Rangel and Stein, 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, 2017a] was held for Russian texts.

In the studies conducted the topic for different languages, the authors identified four main approaches in building gender profiling models for one- and multi-genre texts from social networks.

They can be summarized as follows:

1. Stylometric methods, the purpose of which is to capture the author's writing style using various statistical characteristics

- (a) lexical: the use of certain letters, words, emoticons, symbolic n-grams, the frequency of their use, etc.
  - (b) syntactic: use of punctuation, POS, punctuation frequency, etc.
  - (c) structural: how the author structures the text or uses certain patterns (greetings, signatures, etc.)
2. Content-based techniques that are designed to identify author profiles based on the content of the text (BOW, n-grams of words, slang words, etc.)

(a) *tf-idf*

The goal is to assess the importance of a word in the context of a document that is part of a document collection or corpus. This strategy has been suggested by numerous stylometric studies [Burrows, 1992], [Burrows, 2002], [Savoy, 2015].

However, after checking the *tf-idf* values, there were no significant differences between males and females. That is, it will not be possible to directly use this characteristic in order to distinguish the gender of the author.

Besides, the importance of the method is highlighting social media-specific features with high mention rates - for example, for Twitter, these are (@), hyperlinks (*urlink*), retweets (*rt*), and hashtags (#). For personal communication - frequent use of personal pronouns (for example, *I, it, you*), as well as some punctuation marks (. : ,) and functional terms (*the, to, of, is, and, in*) (for English)

(b) Highlight specific words for each of the categories

I.e., when terms are often used in one category and rarely in another. Then the difference in the probability of occurrence should detect all, or at least some of them.

$$D_{tf}(t_i) = p(t_i, M) - p(t_i, F) = \frac{tf_{i,M}}{n_M} - \frac{tf_{i,F}}{n_F},$$

where  $tf_{i,M}$  is the absolute frequency of the  $i$ -th word in the  $M$  category;  $n_M$  - length (in words) of all messages belonging to the  $M$  class (similar for the opposite  $F$  class).

(c) Definition of topics discussed in the text (*LDA, PLSA*)

3. Also, it is possible to apply the criteria for selecting characteristics that play the role of a filter. For each term, it calculates a distribution score across two categories. As criteria, you can use: odds ratio (*OR*), chi-square and correlation coefficient. Each of these functions usually finds different terms specific to this category.
4. Using deep learning methods (*CNN, LSTM, Bi-LSTM, GRU*) and ready-made pre-trained models (*GloVe, BERT, GPT, etc.*)

The presented work is based on the use of the pre-trained language model of the Russian language *SBERT*, with further fine-tuning for the training set, in combination with the stylometric methods.

## 1.1 Team

Research was done by **Dmitry Badeev**.

## 2 Related Work

Below are tables with the results of research on the following tasks:

- Determining the gender of the author in one-genre texts of social networks (all languages except Russian)
- Determining the gender of the author in one-genre Russian-language texts of social networks
- Determining the gender of the author in multi-genre Russian-language texts of social networks
- Determining the gender of the author in multi-genre texts of social networks (all languages, except Russian)

### **Determining the gender of the author in one-genre texts of social networks (all languages except Russian)**

Table 1 shows data on gender profiling in single-genre texts of social networks for different languages (except Russian), carried out in different years in which the top *accuracy* scores were achieved.

The columns shows: the Authors and the reference to the work, the Corpus and the reference, the Language of the corpus, the Features used, the Research Method, and the Accuracy Result.

The best result is from Ashraf and Nie [M.A. Ashraf and Nie, 2019], who used the 100-dimension Twitter-based GloVe pre-trained corpus and Bi-LSTM. The high result is probably due to good word representation embedding. In [Joo and Hwang, 2019], BERT was used for the English language. The result turned out to be comparable with other approaches considered by the authors and, according to them, the BERT model requires more fine tuning to obtain a better result.

According to the PAN'19 and PAN'17, results in works with traditional approaches have gained higher accuracy than deep learning methods: [Matej Martinc and Pollak, 2017], [Alex I. Valencia Valencia and Pineda, 2019], [Gishamer, 2019], [Joo and Hwang, 2019], [Pizarro, 2019], [Srinivasarao and Manu, 2019], [Régis Goubin and Fossi, 2019], [Angelo Basile and Nissim, 2017], [Yasuhide Miura and Ohkuma, 2017]. The top four teams used a combination of n-gramm and SVM, while the top team [Marco Polignano and Semeraro, 2019] using the deep learning model is 11-th.

The same result with traditional approach we can see in the old work in Turkish language: [B. Tayfun Kucukyilmaz and Can, 2006a].

In general, with the same datasets, the teams achieved the best results, mainly working with stylometric features (word, character n-gramms, POS, etc.), in combination with SVM.

Datasets, mentioned in Table 1:

[Zaghouani W, 2018], [B. Tayfun Kucukyilmaz and Can, 2006b],  
[Francisco Rangel and Stein, 2017], [Francisco Rangel, 2019].

Work	Corpus	Text languages	Features	Method used	Result
[M.A. Ashraf and Nie, 2019]	BT-AP-19 : [Zaghouani W, 2018]	English Roman-Urdu	GloVe	Bi-LSTM	0.88
[Alex I. Valencia Valencia and Pineda, 2019]	PAN '19 [Francisco Rangel, 2019]	English	char(3-4), word(1-3) n-grams, tf-idf	Logistic regression	0.8432
[Gishamer, 2019]	PAN '19 [Francisco Rangel, 2019]	English	char(3-5); word(1-3);POS; POS(2-3) -grams	SVM	0.84
[Angelo Basile and Nissim, 2017]	PAN '17 [Francisco Rangel and Stein, 2017]	Arabic	char(3-5); word(1-2) -grams; [tf_ = 1 + log(tf)] - idf;	SVM	0.8006
		English			0.8233
		Portuguese			0.8450
		Spanish			0.8321
[Joo and Hwang, 2019]	PAN '19 [Francisco Rangel, 2019]	English	lexical, tf-idf, Doc2Vec(DBOW+DM 1+2+3-gram), used words	features + BERT	0.8360
[Pizarro, 2019]	PAN '19 [Francisco Rangel, 2019]	English	char(1, 3), word(1, 3) n-grams, tf-idf	SVM	0.8356
		Spanish	char(3, 5), word(1, 3) n-grams, tf-idf		0.8172
		Spanish	char(3-5); word(1-2);POS(2-3) -grams;		0.712
[B. Tayfun Kucukyilmaz and Can, 2006]	from Heaven BBS [B. Tayfun Kucukyilmaz and Can, 2006b]	Turkish	stylistic features, CHI square	Naive Bayesian Style-based	0.819
[Matej Martinc and Pollak, 2017]	PAN '17 [Francisco Rangel and Stein, 2017]	Arabic	word(1, 2), char(4), suffix(4), punctuation(3) n-grams	Logistic regression	0.8031
		English			0.8071
		Portuguese			0.8600
		Spanish			0.8193
[Srinivasarao and Manu, 2019]	PAN '19 [Francisco Rangel, 2019]	English	char(3, 4), word(1, 3) n-grams, tf-idf	SVM	0.8398
		Spanish	char(3, 4), word(1, 2) n-grams, tf-idf		0.7967
[R'egis Goubin and Fossi, 2019]	PAN '19 [Francisco Rangel, 2019]	English	BOW, tf-idf	LSVC + stat calibration	0.8333
		Spanish			0.7917
[Yasuhide Miura and Ohkuma, 2017]	PAN '17 [Francisco Rangel and Stein, 2017]	Arabic	word & character embedding	RNN, CNN	0.7644
		English			0.8046
		Portuguese			0.8700
		Spanish			0.8118
[Marco Polignano and Semeraro, 2019]	PAN '19 [Francisco Rangel, 2019]	English		CNN, Word2Vec	0.7973
		Spanish			0.7417
		Spanish			0.7539

TABLE 1. One-genre Gender Detection (languages w/o Russian)

## Determining the gender of the author in one-genre Russian-language texts of social networks

Table 2 shows data on gender profiling in one-genre Russian-language texts of social networks, carried out in different years in which the top *accuracy* scores were achieved.

The columns show the Authors and the reference to the work, the Corpus and the reference, the Language of the corpus, the Features used, the Research Method, and the Accuracy Result.

The best result was achieved in 2013 [Korshunov, 2013]. However, as the same method was used in the works later (PAN FIRE '17), the achievement is possibly due to the specifics of the dataset and test cases on which the model's performance was evaluated.

Among the works where the PAN FIRE '17 [Litvinova T, 2017b] dataset was used, [Sboev, 2019] has the best result. Experiments were carried out with various models, including deep learning. The first place was taken by the Gradient Boosting model with symbolic n-gramm representation.

At the same time, in the work [C6oeb, 2020], the result was higher when using the model with GRU, CVAE. But in this work was not mentioned comparison with other models on the dataset used.

Work	Corpus	Features	Method used	Result
[Korshunov, 2013]	self-made	word (3-grams)	SVM	0.86
[Sboev, 2019]	[T. Litvinova and Romanchenko, 2016] [Litvinova T, 2017b]	char n-grams	Gradient Boosting	0.79
[C6oeb, 2020]	[Litvinova T.A., 2018]		GRU, CVAE	0.76
[I. Markov and Gelbukh, 2017] - CIC3	[Litvinova T, 2017b]	statistical		0.6825
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b]	probability distribution of occurrence of tdoc's words in the different classes.		0.6759
[I. Markov and Gelbukh, 2017] - CIC2	[Litvinova T, 2017b]	BOW, word (suffix 3-grams), tf-idf	SVM	0.6650
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b]	POS, rule-based classification	LSTM, Bi-LSTM	0.6525
[I. Markov and Gelbukh, 2017] - CIC1	[Litvinova T, 2017b]	POS combination, tf-idf	SVM	0.6525

TABLE 2. One-genre Gender Detection (Russian)

Datasets, mentioned in Table 2: [T. Litvinova and Romanchenko, 2016], [Litvinova T, 2017b], [Litvinova T.A., 2018].

## Determining the gender of the author in multi-genre Russian-language texts of social networks

Table 3 shows the top *accuracy* scores for gender profiling in multi-genre Russian-language texts of social networks, based on the results of RusProfiling PAN at FIRE Track, conducted in 2017.

The columns show the Authors and the reference to the work, the Training Corpus and reference, the Features used, the Research method, the Test Corpora (*Essays*, *Facebook*, *Reviews* and *Gender imitation* (texts in which the authors imitated a different gender, and users change their idiosyncrasy)) and the *accuracy* Result.

In each of the four test competitions the *LDR* (Low Dimensionality Representation) [F. Rangel and Franco-Salvador., 2016] method, chosen as the baseline, was in the lead. 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.

Among the *Essays*, the best result is shown in [R. Bhargava and Sharma, 2017], where a combination of rule-based classification and deep learning methods was applied.

Surprisingly, the results were quite high on a test sample from *Facebook*, which outperformed all other categories, including *Twitter*. According to the hypothesis of the authors of the review [Litvinova T, 2017b], this is due to the fact that *Facebook* posts are longer and grammatically richer, with fewer syntax errors and typos compared to *Twitter*.

According to the results, the models with the traditional approach turned out to be better than the models using deep learning methods.

Work	Corpus	Features	Method used	Test Corpus	Result
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b] (Twitter - training)	stylometric analysis		Essays	0.8141
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b] (Twitter - training)	POS, rule-based classification	LSTM, Bi-LSTM	Essays	0.7838
[V. Vinayan and Amrita, 2017]	[Litvinova T, 2017b] (Twitter - training)	exotic stat (average word length, URL usage, etc), tf-idf	SVM	Essays	0.6811
[I. Markov and Gelbukh, 2017] - CIC2	[Litvinova T, 2017b] (Twitter - training)	BOW, word (suffix 3-grams), tf-idf	SVM	Facebook	0.9342
[I. Markov and Gelbukh, 2017] - CIC1	[Litvinova T, 2017b] (Twitter - training)	POS combination, tf-idf	SVM	Facebook	0.9211
[I. Markov and Gelbukh, 2017] - CIC5	[Litvinova T, 2017b] (Twitter - training)	word(1), char(2-3) n-grams, tf-idf	SVM	Facebook	0.8991
[I. Markov and Gelbukh, 2017] - CIC4	[Litvinova T, 2017b] (Twitter - training)	word(1, 3), char(2-4) n-grams, tf-idf	SVM	Facebook	0.8860
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b] (Twitter - training)	POS, rule-based classification	LSTM, Bi-LSTM	Facebook	0.8728
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b] (Twitter - training)	stylometric analysis		Facebook	0.8596
[Sboev, 2019]	[Litvinova T, 2017b] (Twitter - training)	char n-grams	Gradient Boosting	Reviews	0.79
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b] (Twitter - training)	stylometric analysis		Reviews	0.72
I. Markov and Gelbukh, 2017] - CIC3	[Litvinova T, 2017b] (Twitter - training)	statistical		Reviews	0.6186
I. Markov and Gelbukh, 2017] - CIC1	[Litvinova T, 2017b] (Twitter - training)	POS combination, tf-idf	SVM	Reviews	0.5979
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b] (Twitter - training)	POS, rule-based classification	LSTM, Bi-LSTM	Reviews	0.5786
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b] (Twitter - training)	POS, rule-based classification	LSTM, Bi-LSTM	Gender imitation	0.6596
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b] (Twitter - training)	stylometric analysis		Gender imitation	0.6383

TABLE 3. Cross-genre Gender Detection (Russian)



### **Determining the gender of the author in multi-genre texts of social networks (all languages, except Russian)**

Table 4 shows the top *accuracy* scores for gender profiling in multi-genre social media texts (all languages except Russian), based on the results of the PAN CLEF conducted in 2016 and work [M.A. Ashraf and Nie, 2019].

The columns show the Authors and the link to the work, the Training Corpus and reference, the Features used, the Research Method, Test Corpuses (social networks, blogs, reviews) and the Accuracy Result.

As with single-genre text research, work [M.A. Ashraf and Nie, 2019]. scores best, using the 100-dimension Twitter-based GloVe pre-trained corpus and Bi-LSTM for testing Facebook texts, and CNN for SMS.

In PAN CLEF 2016, most of the participants considered the stylistic features of texts (frequency of use of function words, slang, use of capital letters, etc.) in combination with POS, n-grams, with not particularly impressive results.

Work	Corpus	Features	Method used	Test Corpus	Result
[M.A. Ashraf and Nie, 2019]	Zaghouni W, 2018] (Twitter - training)	GloVe	Bi-LSTM	RUEN-AP-17 [M. Fatima and R.M.A. Nawab, 2017], Facebook (English, Roman-Urdu)	0.833
[M.A. Ashraf and Nie, 2019]	Zaghouni W, 2018] (Twitter - training)	GloVe	CNN	SMS-AP-18 [M. Fatima and Masood, 2018], SMS (English, Roman-Urdu)	0.815
[Mart Busger op Vollenbroek and Nissim, 2016]	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	word(1-3),char(2-5) n-grams, POS	SVM	social media (English)	0.5575
Waser	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	not presented	not presented	social media (Spanish)	0.7031
Roman-Gomez	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	not presented	not presented	social media (Dutch)	0.6200
[Agrawal and Gon, calves, 2016]	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	tf-idf	SVM, NB	social media (English)	0.5431
[Pashutan Modaresi and Conrad, 2016]	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	word(1-2),char(4) n-grams, punctuation	Logistic regression	social media (Spanish)	0.6406
Waser	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	not presented	not presented	social media (Dutch)	0.6000
[Pashutan Modaresi and Conrad, 2016]	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	word(1-2),char(4) n-grams, punctuation	Logistic regression	blogs (English)	0.7564
Deneva	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	not presented	not presented	blogs (Spanish)	0.7321
[Bilan and Zhekova, 2016]	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	word(1-3) n-grams, tf-idf, CHI square, POS, stylistic	SVM(LinearSVC)	blogs (English)	0.7436
[Mart Busger op Vollenbroek and Nissim, 2016]	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	word(1-3),char(2-5) n-grams, POS	SVM	blogs (Spanish)	0.7143
Deneva	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	not presented	not presented	reviews (Dutch)	0.6180
[Bayot and Gon, calves, 2016]	[Linda Cappellato and Kraaij, 2014] [Verhoeven and Daelemans, 2014] [Ben Verhoeven and Plank, 2016] (Twitter - training)	CBOW	Word2Vec, SVM	reviews (Dutch)	0.5680

TABLE 4. Cross-genre Gender Detection (languages w/o Russian)

### 3 Model Description

Model containing three fully connected layers was built for an experiment with the Gender Profiling in Social Network task.

The *Russian SBERT NLU* model was used as the base *BertLayer*.

The *Russian SBERT NLU* model is a *BERT-large NLU* with 24 layers and 426.9 million parameters. The *Russian SBERT NLU* model is available for download for tensorflow models.

Averaging of token embeddings is taken from the last *BERT layer* (masked mean pooling is used).

Next layer - the hidden one (*Dense*) with 1024 neurons and the *relu* activation function.

For regularization, between the last hidden layer and the output layer, *Dropout* = 0.25 is used.

The last layer (*classifier*) consists of a fully connected layer with two output neurons and a *sigmoid* activation function.

Since there are two classes in the target (male and female), *Binary Cross Entropy Loss* is used as the loss function:

$$BCE = -\frac{1}{N} \sum_{i=0}^N (y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i))$$

Optimization algorithm - *Adam*, value *learning\_rate* =  $10^{-5}$ .

The stylometric feature *S* of gender definition in Russian is also used:

#### ***S* - the stylometric features of gender for the Russian language**

In Russian, the singular forms of the past tense verbs change by gender (the masculine singular forms have the ending “-л”, and the ending “-ла” is the indicator of the feminine singular forms). Also, reflexive verbs of the past tense differ in gender - for the feminine gender they have the ending “-лась”, and for the masculine - “-лся”. The described features of the verbs were used in combination with the subject of the sentence if the subject was the first person singular pronoun “я” (or “Я”), and if the subject came from the verb no more than within a 4-word window.

Thus, in total, four features are used: two for the masculine gender - “Я(я) [0-4] -л” and “Я (я) [0-4] -лся”, and two for the feminine gender - “Я (я) [0-4] -ла” and “Я (я) [0-4] -лась”.

## 4 Dataset

I planned to use data from the assignment for PAN at CLEF 2017, PAN at FIRE Track on Cross-genre Gender Identification in Russian. But links to datasets did not work, and a request to the organizers to give an opportunity to access the data remained unanswered.

Fortunately, I found the source - a common database of corpuses, from which datasets were formed for tasks at the conference. The structure of datasets, both training and test, is described detailly in [Litvinova T, 2017b]. I prepared datasets for my experiments similar to the data from the task on RusProfiling PAN '17 in accordance with that instructions.

### Twitter dataset

#### *Total*

- Messages from 1000 authors were used, 500 of each gender
- All messages from one author were combined and treated as one text
- Text length of one author - about 500 words or about 1500 words
- The entire dataset was splitted into two: training dataset and testing dataset

#### *Model training dataset (RuSb\_base)*

- Contains texts of 600 authors, 300 for each gender

#### *Dataset for testing single-genre texts*

- Contains texts of 400 authors, 200 of each gender

### Facebook dataset

#### *Dataset for testing single-genre texts*

- Contains the texts of 220 authors, 110 of each gender
- All messages from one author were combined and treated as one text
- Text size of one author is about 1000 words

The source for both Twitter and Facebook datasets was one text base, in which the source social network was not indicated. The division into different social networks was carried out according to the size of the text of one author, the only one criterion of difference mentioned in [Litvinova T, 2017b]. So, it is permissible to consider **all texts (Facebook and Twitter)** from the database **as one-genre**.

## Essays dataset

### *Dataset for testing multi-genre texts*

- Contains the texts of 370 authors, 185 of each gender
- One or two texts per author (in case of two texts they were merged together and considered as one text)
- The average text length in is 150 words.

## Reviews dataset

### *Dataset for testing multi-genre texts*

- Contains the texts of 776 authors, 388 of each gender
- One or two texts per author (in case of two texts they were merged together and considered as one text)
- The average text length in this dataset was 150 words.

## Gender imitation dataset

### *Dataset for testing multi-genre texts*

- Contains the texts of 94 authors, 47 of each gender
- Three texts from each author that were merged together and considered as one text
- All three texts have one theme (from a set of themes)
  - the first text is written in the usual way for the person who wrote it (without any deception)
  - the second is written from the person of the opposite sex ("imitation")
  - the third text should be as of author is the same gender, but personal writing style will not be recognized ("obfuscation")
- All texts from one author were combined and counted as one text
- Most of the texts are 80-150 words long

For experiments with models trained on datasets of different sizes, two additional single-genre datasets were formed.

*The Medium dataset* used the entire *Twitter dataset (training + test)* for training. *The Large Training Dataset* used the combined *Twitter Dataset (training + test)* with *the Facebook Dataset*.

## Medium training dataset

- Contains texts of 1000 authors, 500 for each gender
- All texts from one author were combined and counted as one text
- Text length of one author - about 500 words or about 1500 words

## Big training dataset

- Contains texts of 1220 authors, 610 for each gender
- All texts from one author were combined and counted as one text
- Text length of one author - evenly from about 500 words to 1500 words

All datasets are available at github.

# 5 Experiments

For the experiments, three models were prepared with the same parameters, but trained on datasets of different size:

- **RuSb\_base** - 600 Twitter posts, 300 for each gender (*Base training dataset*)
- **RuSb\_mid** - 1000 Twitter posts, 500 for each gender (*Average training dataset*)
- **RuSb\_big** - 1220 Twitter posts, 610 for each gender (*Large training dataset*)

Three more models (**RuSb\_base+S**, **RuSb\_mid+S**, **RuSb\_big+S**), additionally used *S* - the stylometric features of gender for the Russian language, described in 3<sup>1</sup>.

## 5.1 Metrics

The quality of model predictions were be assessed by two metrics - *accuracy* and *F<sub>1</sub>score*.

## 5.2 Experiment Setup

Training parameters:

- Basic BertLayer - Russian SBERT NLU model;
- batch size = 24;

---

<sup>1</sup>*S* - the stylometric features of gender for the Russian language

- seq\_len = 256;
- N\_tune\_lrs = 24;
- Optimizer - Adam;
- Learning rate = 1e-5;
- Metrics - F1, Accuracy;
- Number of epochs = 3.

Model summary is presented in Table 5:

Model:		
Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 1)]	0
bert_layer (BertLayer)	(None, 1024)	426908672
dense (Dense)	(None, 1024)	1049600
dropout (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 2)	2050
Total params: 427,960,322		
Trainable params: 303,361,026		
Non-trainable params: 124,599,296		

TABLE 5. Model summary

## 6 Results

### *Overall results*

The Table 6 shows the results (*accuracy*) obtained for all six models for test single- and multi-genre sets.

**RuSb\_mid** and **RuSb\_mid** tested only on **Facebook** data test, as models **RuSb\_big** and **RuSb\_big+S** are more informative in other data test categories, in comparison with first two.

It makes no sense to use models **RuSb\_base+S** and **RuSb\_big+S** on **Gender Imitation**, since accounting for stylistic features will lead to a deliberately incorrect result.

The result for **Facebook** can be considered one-genre, as mentioned earlier <sup>4</sup>.

	Twitter	Facebook	Essays	Reviews	Gender imitation
<b>RuSb_base</b>	0.87	0.85	0.86	0.80	0.93
<b>RuSb_base+S</b>	0.90	0.89	0.87	0.80	
<b>RuSb_mid</b>		0.85			
<b>RuSb_mid+S</b>		0.89			
<b>RuSb_big</b>			0.86	0.83	0.95
<b>RuSb_big+S</b>			0.86	0.83	

TABLE 6. Overall results

***Analysis:***

1. Stylistic features for *Reviews* did not affect the result, while in other cases, both single and multigenre, there is a slight increase. Perhaps this is due to the brevity of messages in Reviews (up to 80 words), where the forms of presentation considered by stylistic features are used rarely.
2. The lowest results in *Reviews* in comparison with other datasets - also, in my opinion, due to the brevity of messages.
3. Increasing the training dataset didn't have affect in two cases: *Facebook* and *Essays*. If in the one-genre *Facebook* case it can be assumed that the size of the base dataset was quite sufficient, then in the case of *Essays*, it seems logical to assume that the reason was the limited themes set as the basis for user texts.
4. The very high percentage of recognition shown in Gender Imitation is surprising. This phenomenon is beyond the scope of current work and deserves additional analysis.

More detailed information on the metrics of all experiments is given below in Table 7 and Table 8:

---

<sup>2</sup>The source for both Twitter and Facebook datasets was one text base



One-genre Gender Detection (Russian)

Twitter					
RuSb_base					
	precision	recall	f1-score	support	
0	0.88	0.85	0.87	200	
1	0.86	0.88	0.87	200	
accuracy			0.87	400	
macro avg	0.87	0.87	0.87	400	
weighted avg	0.87	0.87	0.87	400	
RuSb_base+S					
	precision	recall	f1-score	support	
0	0.89	0.92	0.90	200	
1	0.91	0.89	0.90	200	
accuracy			0.90	400	
macro avg	0.90	0.90	0.90	400	
weighted avg	0.90	0.90	0.90	400	

Facebook					
RuSb_base					
	precision	recall	f1-score	support	
0	0.86	0.85	0.85	110	
1	0.85	0.86	0.86	110	
accuracy			0.85	220	
macro avg	0.85	0.85	0.85	220	
weighted avg	0.85	0.85	0.85	220	
RuSb_base+S					
	precision	recall	f1-score	support	
0	0.88	0.90	0.89	110	
1	0.90	0.88	0.89	110	
accuracy			0.89	220	
macro avg	0.89	0.89	0.89	220	
weighted avg	0.89	0.89	0.89	220	
RuSb_mid					
	precision	recall	f1-score	support	
0	0.86	0.84	0.85	110	
1	0.84	0.86	0.85	110	
accuracy			0.85	220	
macro avg	0.85	0.85	0.85	220	
weighted avg	0.85	0.85	0.85	220	
RuSb_mid+S					
	precision	recall	f1-score	support	
0	0.89	0.89	0.89	110	
1	0.89	0.89	0.89	110	
accuracy			0.89	220	
macro avg	0.89	0.89	0.89	220	
weighted avg	0.89	0.89	0.89	220	

TABLE 7. Metrics on One-genre Gender Detection (Russian)

### Cross-genre Gender Detection (Russian)

<i>Essays</i>					<i>Reviews</i>				
<b>RuSb_base</b>					<b>RuSb_base</b>				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.85	0.88	0.86	185	0	0.95	0.64	0.77	388
1	0.88	0.84	0.86	185	1	0.73	0.97	0.83	388
accuracy			0.86	370	accuracy			0.80	776
macro avg	0.86	0.86	0.86	370	macro avg	0.84	0.80	0.80	776
weighted avg	0.86	0.86	0.86	370	weighted avg	0.84	0.80	0.80	776
<b>RuSb_base+S</b>					<b>RuSb_base+S</b>				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.86	0.88	0.87	185	0	0.95	0.64	0.77	388
1	0.88	0.86	0.87	185	1	0.73	0.97	0.83	388
accuracy			0.87	370	accuracy			0.80	776
macro avg	0.87	0.87	0.87	370	macro avg	0.84	0.80	0.80	776
weighted avg	0.87	0.87	0.87	370	weighted avg	0.84	0.80	0.80	776
<b>RuSb_big</b>					<b>RuSb_big</b>				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.86	0.85	0.86	185	0	0.95	0.70	0.81	388
1	0.85	0.86	0.86	185	1	0.76	0.97	0.85	388
accuracy			0.86	370	accuracy			0.83	776
macro avg	0.86	0.86	0.86	370	macro avg	0.86	0.83	0.83	776
weighted avg	0.86	0.86	0.86	370	weighted avg	0.86	0.83	0.83	776
<b>RuSb_big+S</b>					<b>RuSb_big+S</b>				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.87	0.86	0.86	185	0	0.95	0.70	0.81	388
1	0.86	0.87	0.87	185	1	0.76	0.97	0.85	388
accuracy			0.86	370	accuracy			0.83	776
macro avg	0.86	0.86	0.86	370	macro avg	0.86	0.83	0.83	776
weighted avg	0.86	0.86	0.86	370	weighted avg	0.86	0.83	0.83	776

TABLE 8. Metrics on Cross-genre Gender Detection (Russian)

#### *Analysis:*

1. Brings to notice the imbalance of *precision* and *recall* in all models on the *Reviews* dataset - high accuracy to distinguish class 0 (female) with a low percentage to detect the class itself, and vice versa for class 1 (male).

#### *Comparison with previous results*

The *Table 9* and *Table 10* shows the rating of **RuSb\_base**, **RuSb\_base+S**, **RuSb\_big+S**, **RuSb\_big+S** models in comparison with other models presented above, both in One-Genre 2<sup>3</sup> and Cross-Genre Gender Detection 2<sup>4</sup>. No comparison was made on Facebook, because a relevant test dataset for this category was not found.

<sup>3</sup>One-genre Gender Detection (Russian)

<sup>4</sup>Cross-genre Gender Detection (Russian)

Work	Corpus	Features	Method used	Result
<b>RuSb_base+S</b>	[Litvinova T, 2017b]	Stylometric features	Russian SBERT	<b>0.90</b>
<b>RuSb_base</b>	[Litvinova T, 2017b]		Russian SBERT	<b>0.87</b>
[Korshunov, 2013]	self-made	word (3-grams)	SVM	0.86
[Sboev, 2019]	[T. Litvinova and Romanchenko, 2016] [Litvinova T, 2017b]	char n-grams	Gradient Boosting	0.79
[C6oes, 2020]	[Litvinova T.A., 2018]		GRU, CVAE	0.76
[I. Markov and Gelbukh, 2017] - CIC3	[Litvinova T, 2017b]	statistical		0.6825
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b]	probability distribution of occurrence of tdoc's words in the different classes.		0.6759
[I. Markov and Gelbukh, 2017] - CIC2	[Litvinova T, 2017b]	BOW, word (suffix 3-grams), tf-idf	SVM	0.6650
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b]	POS, rule-based classification	LSTM, Bi-LSTM	0.6525
[I. Markov and Gelbukh, 2017] - CIC1	[Litvinova T, 2017b]	POS combination, tf-idf	SVM	0.6525

TABLE 9. Comparison on One-genre Gender Detection (Russian)

Work	Corpus	Features	Method used	Test Corpus	Result
<b>RuSb_base+S</b>	[Litvinova T, 2017b] (Twitter - training)	Stylometric features	Russian SBERT	Essays	<b>0.87</b>
<b>RuSb_base</b>	[Litvinova T, 2017b] (Twitter - training)		Russian SBERT	Essays	<b>0.86</b>
<b>RuSb_big</b>	[T. Litvinova and Romanchenko, 2016] [Litvinova T, 2017b]		Russian SBERT	Essays	<b>0.86</b>
<b>RuSb_big+S</b>	[T. Litvinova and Romanchenko, 2016], [Litvinova T, 2017b]	Stylometric features	Russian SBERT	Essays	<b>0.86</b>
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b] (Twitter - training)	stylometric analysis		Essays	0.8141
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b] (Twitter - training)	POS, rule-based classification	LSTM, Bi-LSTM	Essays	0.7838
[V. Vinayan and Amrita, 2017]	[Litvinova T, 2017b] (Twitter - training)	exotic stat (average word length, URL usage, etc), tf-idf	SVM	Essays	0.6811
<b>RuSb_big+S</b>	[T. Litvinova and Romanchenko, 2016], PAN FIRE '17 [d8]	Stylometric features	Russian SBERT	Reviews	<b>0.83</b>
<b>RuSb_big</b>	[T. Litvinova and Romanchenko, 2016], [Litvinova T, 2017b]		Russian SBERT	Reviews	<b>0.83</b>
<b>RuSb_base+S</b>	[Litvinova T, 2017b] (Twitter - training)	Stylometric features	Russian SBERT	Reviews	<b>0.80</b>
<b>RuSb_base</b>	[Litvinova T, 2017b] (Twitter - training)		Russian SBERT	Reviews	<b>0.80</b>
[Sboev, 2019]	[Litvinova T, 2017b] (Twitter - training)	char n-grams	Gradient Boosting	Reviews	0.79
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b] (Twitter - training)	stylometric analysis		Reviews	0.72
[I. Markov and Gelbukh, 2017] - CIC3	[Litvinova T, 2017b] (Twitter - training)	statistical		Reviews	0.6186
[I. Markov and Gelbukh, 2017] - CIC1	[Litvinova T, 2017b] (Twitter - training)	POS combination, tf-idf	SVM	Reviews	0.5979
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b] (Twitter - training)	POS, rule-based classification	LSTM, Bi-LSTM	Reviews	0.5786
<b>RuSb_base+S</b>	[Litvinova T, 2017b] (Twitter - training)	Stylometric features	Russian SBERT	Gender imitation	<b>0.95</b>
<b>RuSb_base</b>	[Litvinova T, 2017b] (Twitter - training)		Russian SBERT	Gender imitation	<b>0.93</b>
[R. Bhargava and Sharma, 2017]	[Litvinova T, 2017b] (Twitter - training)	POS, rule-based classification	LSTM, Bi-LSTM	Gender imitation	0.6596
[F. Rangel and Franco-Salvador., 2016]	[Litvinova T, 2017b] (Twitter - training)	stylometric analysis		Gender imitation	0.6383

TABLE 10. Comparison on Cross-genre Gender Detection (Russian)

Models, based on *Russian SBERT* with further tuning, as well as similar models that take into account stylometric features  $S$ , showed better results compared to models from other works built on similar datasets.

## 7 Conclusion

As part of the project, the following tasks were implemented:

- a review of works on similar topics for Russian and other languages was prepared;
- datasets for training models and testing in four different genres based on the database and description of datasets were formed;
- the basic model was built and five more models were trained on the basis of datasets of different size and stylistic features;
- the comparison of the results shown by the trained models with the results obtained in previous works was performed;
- the analysis of the results obtained was realized.

Based on the research results, the following conclusions can be drawn:

1. The *Russian SBERT NLU* model, as a base, with further fine-tuning for the gender classification task, gives a model that exceeds the results of the models discussed above in **Related Works 3** - both with a traditional approach and models using deep learning methods.
2. The results of prediction in gender profiling on single-genre and multi-genre datasets do not have significant differences, with the exception of specific requirements for texts if the genre of the tested datasets differs from the one on which the training took place (message size, limited topics, etc.).
3. An increase in the training dataset gives a slight increase in the accuracy of prediction on one-genre and some multi-genre datasets. If there is a requirement to limit the size of messages in the dataset to a small value (less than 100 words), or the genre of the tested sample differs from the one used for training, the accuracy does not change.
4. The use of stylistic features makes it possible to improve the prediction results of the base model built on the *Russian SBERT NLU* model in gender profiling on single and multi-genre datasets. However, the difference in the results is insignificant.

## References

- [Alex I. Valencia Valencia and Pineda, 2019] Alex I. Valencia Valencia, Helena Gomez Adorno, C. S. R. and Pineda, G. F. (2019). Bots and gender identification based on stylometry of tweet minimal structure and n-grams model. In *In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.*

- [Angelo Basile and Nissim, 2017] Angelo Basile, Gareth Dwyer, M. M. J. R. H. H. and Nissim, M. (2017). N-gram: New groningen author-profiling model—notebook for pan at clef 2017. In *In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland, September 2017. CEUR-WS.org.*
- [B. Tayfun Kucukyilmaz and Can, 2006a] B. Tayfun Kucukyilmaz, B. Barla Cambazoglu, C. A. and Can, F. (2006a). Arapâ tweet: a large multiâ dialect twitter corpus for gender, age and language variety identification. In *n International conference on advances in information systems. Springer, Berlin, Heidelberg*, pages 274–283.
- [B. Tayfun Kucukyilmaz and Can, 2006b] B. Tayfun Kucukyilmaz, B. Barla Cambazoglu, C. A. and Can, F. (2006b). Chat mining for gender prediction. In *In International conference on advances in information systems. Springer, Berlin, Heidelberg*, pages 1274–283.
- [Burrows, 1992] Burrows, J. (1992). Not unless you ask nicely: the interpretative nexus between analysis and information. In *Lit. Linguis. Comput.* 7(1), pages 91–109.
- [Burrows, 2002] Burrows, J. (2002). Delta: a measure of stylistic difference and a guide to likely authorship. In *Lit. Linguis. Comput.* 17(3), pages 267–286.
- [F. Rangel and Stein, 2016] F. Rangel, P. Rosso, B. V. W. D. M. P. and Stein, B. (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.],* pages 750–784.
- [F. Rangel and Franco-Salvador., 2016] F. Rangel, P. R. and Franco-Salvador., M. (2016). A low dimensionality representation for language variety identification. In *In 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing. Springer-Verlag, LNCS, arXiv:1705.10754.*
- [Francisco Rangel and Stein, 2017] Francisco Rangel, Paolo Rosso, M. P. and Stein, B. (2017). Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. In *In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland, September 2017. CEUR-WS.org.*
- [Francisco Rangel, 2019] Francisco Rangel, P. R. (2019). Overview of the 7th author profiling task at pan 2019: Bots and gender profiling in twitter.
- [Gishamer, 2019] Gishamer, F. (2019). Using hashtags and pos-tags for author profiling. In *In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.*

- [Joo and Hwang, 2019] Joo, Y. and Hwang, I. (2019). Steve martin at semeval-2019 task 4: Ensemble learning model for detecting hyperpartisan news. In *In Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 990—994.
- [Korshunov, 2013] Korshunov, A. (2013). Detection of demographic attributes of microblog users. In *Proceedings of the Institute for System Programming Volume 25*.
- [Litvinova T, 2017a] Litvinova T, Rangel F, R. P. S. P. L. O. (2017a). Overview of the rusprofiling pan at fire track on cross-genre gender identification in russian. In *Notebook papers of FIRE 2017, FIRE-2017, Bangalore, India, December 8–11, CEUR Workshop Proceedings. CEUR-WS.org, vol 2036*.
- [Litvinova T, 2017b] Litvinova T, Rangel F, R. P. S. P. L. O. (2017b). Overview of the rusprofiling pan at fire track on cross-genre gender identification in russian. In *In Notebook papers of FIRE 2017, FIRE-2017, Bangalore, India, December 8–11, CEUR Workshop Proceedings. CEUR-WS.org, vol 2036*, pages 1—7.
- [Litvinova T.A., 2018] Litvinova T.A., Sboev A.G., P. P. (2018). Profiling the age of russian bloggers. In *Conference on Artificial Intelligence and Natural Language, Springer, Cham*, pages 167—177.
- [M.A. Ashraf and Nie, 2019] M.A. Ashraf, R. N. and Nie, F. (2019). Author profiling on bi-lingual tweets. In *Journal of Intelligent Fuzzy Systems, (Preprint)*, pages 1—11.
- [Marco Polignano and Semeraro, 2019] Marco Polignano, Marco Giuseppe de Pinto, P. L. and Semeraro, G. (2019). Identification of bot accounts in twitter using 2d cnns on user-generated contents. In *In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org*.
- [Matej Martinc and Pollak, 2017] Matej Martinc, Iza Škrjanec, K. Z. and Pollak, S. (2017). Pan 2017: Author profiling - gender and language variety prediction—notebook for pan at clef 2017. In *In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland, September 2017. CEUR-WS.org*.
- [Pizarro, 2019] Pizarro, J. (2019). Using n-grams to detect bots on twitter. In *In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org*.
- [R. Bhargava and Sharma, 2017] R. Bhargava, G. Goel, A. S. and Sharma, Y. (2017). Gender identification in russian texts. In *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.*
- [Régis Goubin and Fossi, 2019] Régis Goubin, Dorian Lefeuvre, A. A. J. M. E. E.-Z. and Fossi, L. G. (2019). Bots and gender profiling using a multi-layer architecture. In *In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.*
- [Savoy, 2015] Savoy, J. (2015). Comparative evaluation of term selection functions for authorship attribution. In *Digit. Scholarsh. Humanit.* 30(2), pages 246–261.
- [Sboev, 2019] Sboev, D. Gudovskikh, . M. R. R. (2019). A gender identification of text author in mixture of russian multi-genre texts with distortions on base of data-driven approach using machine learning models. In *AIP Conference Proceedings 2116, 270006.*
- [Srinivasarao and Manu, 2019] Srinivasarao, M. and Manu, S. (2019). Bots and gender profiling using character and word n-grams. In *In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, CLEF 2019 Labs and Workshops, Notebook Papers, September 2019. CEUR-WS.org.*
- [T. Litvinova and Romanchenko, 2016] T. Litvinova, O. Litvinova, O. Z. P. S. A. S. and Romanchenko, O. (2016). “ruspersonality”: A russian corpus for authorship profiling and deception detection. In *in Intelligence, Social Media and Web (ISMW FRUCT), 2016 International FRUCT Conference on (IEEE, 2016)*, pages 167–177.
- [Yasuhide Miura and Ohkuma, 2017] Yasuhide Miura, Tomoki Taniguchi, M. T. and Ohkuma, T. (2017). Author profiling with word+character neural attention network—notebook for pan at clef 2017. In *In Linda Cappellato, Nicola Ferro, Lorraine Goeuriot, and Thomas Mandl, editors, CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland, September 2017. CEUR-WS.org.*
- [Zaghouani W, 2018] Zaghouani W, C. A. (2018). Arapâ tweet: a large multiâ dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the 11th international conference on language resources and evaluation (LREC), Miyazaki, Japan.*
- [Сбоев, 2020] Сбоев (2020). ГЕНЕРАТИВНО-ДИСКРИМИНАТИВНАЯ НЕЙРОСЕТЕВАЯ МОДЕЛЬ ДЛЯ ЗАДАЧИ АВТОРСКОГО ПРОФИЛИРОВАНИЯ. *Вестник Национального исследовательского ядерного университета МИФИ*, 9(1):50–57.