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

A Sboev\textsuperscript{1,2}, I Moloshnikov\textsuperscript{1}, D Gudovskikh\textsuperscript{1}, and R Rybka\textsuperscript{1}

\textsuperscript{1} National Research Centre “Kurchatov Institute”, Moscow, Russian Federation
\textsuperscript{2} MEPhI National Research Nuclear University, Moscow, Russian Federation

E-mail: sag111@mail.ru, ivan-rus@yandex.ru, dvgudovskikh@gmail.com, rybkarb@gmail.com

Abstract. In this work we compare several data-driven approaches to the task of author’s gender identification for texts with or without gender imitation. The data corpus has been specially gathered with crowdsourcing for this task. The best models are convolutional neural network with input of morphological data (f1-measure: 88\%\pm3) for texts without imitation, and gradient boosting model with vector of character n-grams frequencies as input data (f1-measure: 64\%\pm3) for texts with gender imitation. The method to filter the crowdsourced corpus using limited reference sample of texts to increase the accuracy of result is discussed.

1. Introduction

Nowadays a great amount of information circulates in the Internet, but only part of it is truthful. However, for many practical fields, such as forensic, security, and others, to have truthful information is important. There are different formulations of the task of information deception identification. In this paper we investigate the task of gender identification of Russian text author in texts with gender deception (i. e. when the authors tried to mimic other gender) in comparison with the same task for texts without deception. While the latter case is widely represented in scientific literature, see for example \cite{1, 2, 3, 4, 5, 6, 7, 8}, sources for the former case are absent. The closest formulation is issued in \cite{9}, where the task of presence of any deception in men and women texts is investigated.

The gender identification task was a part of the annual competition PAN 2017 \cite{1} which was devoted to text author profiling for such languages as English, Arabic, Spanish, and Portuguese. That task was to determine author gender by Twitter messages. The corpus contained 500 tweets from 100 authors. Training was performed on 60\% of the samples, and the remaining 40\% were used for testing according to the competition rules. Classical methods of machine learning showed the best results for the English and Spanish model based on SVM \cite{2}, for the Arabic language the classification model was constructed on the basis of logistic regression \cite{3}. Deep neural networks showed the best result for the Portuguese language, their topologies were described in the work \cite{4}. The authors used recurrent neural networks (RNN) for words and convolutional layers (CNN) for symbols. In the network topologies, full-connected...
words, maxpooling and attention mechanism were used. That approach combined different text representations to classify the author’s gender. Formally, these results show that today classical methods demonstrate the best result for the problem of gender identification. However, the approaches based on neural networks demonstrated their applicability for this task even on small data sets.

Regarding the gender deception task, the closest example is the work [9], where the research was concentrated on any lie in texts from men and women. The accuracy of the best model was just about 5% above baseline, based on Support Vector Machine (SVM) and the combination of semantic features (number of words and word sequences belonging to certain semantic classes) and word forms. The standard deviation was not shown in the article. Training data was collected using Amazon Mechanical Turk [10], the size of corpus was 7168 sentences (3584 truths, 3584 lies) from 512 users.

Moreover, the PAN 2017 competition results, where the Russian language track based on RusProfiling and Gender Imitation corpora took place, are presented in [6]. Unfortunately, the test using the Gender Imitation of this track had the special formulation: the learning sample consisted of Twitter texts, but the testing texts were formed by uniting in one text both the textual part without deception and the textual part with deception from the Gender Imitation corpus. It did not allow to evaluate correctly the accuracy of identifying text author gender only on texts with gender deception.

All the datasets used in our work are described in section 2. The section 4 describes the used methods of machine learning with different sets of input features described in section 3. The description of the experiments without gender imitation (subsection 5.1), with gender imitation (5.2) and with filter algorithm (5.3) and the accuracy obtained in them is presented in the section 5. Finally, we present the discussion of the results in 6 and conclusion in 7.

2. Used corpora
We used a crowdsourcing platform to collect texts imitating other gender. In addition, various corpora with texts without imitation of gender were used, and manually collected Russian Gender Imitation corpus [5], [6], which contains 125 documents from men and 269 documents from women, as the test case. The main crowdsourcing corpus GI\textsubscript{cs} contains 1161 men and 2043 women. The users who made markup on the crowdsourcing platform used the methodology used to create the Gender Imitation corpus: each author wrote 3 texts on a theme chosen from the following list: a description of oneself and one’s potential partner for a dating site, an attempt to persuade an unfamiliar person to come, or a negative tour review. The task was that these three texts needed to be written in certain different styles: Text A in the authors natural style, Text B imitating other gender style, and Text C in a different style but without gender imitation. Subcorpus GI\textsubscript{cs\_a\_b} was obtained from the main corpus GI\textsubscript{cs}, by deleting C texts.

In addition, we used the RusPersonality corpus (RusPer) [11], the RusProfiling (RusProf) [12] corpus containing texts from various social networks Twitter (further denoted as Tw), Facebook (FB) and LiveJournal (LJ) and reviews. To increase the number of examples in these subcorpuses, the combined texts of each user were divided into examples of 15 sentences: Tw\textsubscript{s} (5062 documents from men, 2450 documents from women), LJ\textsubscript{s} (1632 documents from men, 1624 documents from women), FB\textsubscript{s} (868 documents from men, 749 documents from women).

3. Features
The following sets of features were used in combination with models of Section Models:

- **TF-IDF for n-grams (TF-IDF):** The vector of n-gram frequencies, characterizing a document \(d\), is obtained by applying by the TF-IDF formula
  \[
  tf - idf(t, d) = tf(t, d) \times idf(t)
  \]
  to each n-gram \(t\) of the collection \(D\). Here \(t\) is a character n-gram from 3 to 8 characters.
long, \( tf(t, d) \) is the number of times \( t \) occurs in \( d \); \( idf(t) = \log \frac{1 + n_d}{1 + df(d, t)} + 1 \). Here \( n_d \) is the total number of texts, and \( df(d, t) \) is the number of texts that contain \( t \). We tried different values of minimal \( df(d, t) \) in the range from 1 to 100, the best is 1.

- **Word2vec**: This is an already existing model that was preliminarily trained on a collection of random Russian web pages crawled in December 2014, contains 9 million documents in total. Corpus size is 660,628,738 tokens. Model was trained using the Continuous Skip-Gram algorithm. Vector dimensionality was set to 500, window size 2. Lemmas occurring less than 30 times were ignored [13].

- **Sequences feature (Seq. feat)**: Texts are represented as sequences of words with a full set of morphological tags (person, gender, part of speech, etc.), all one-hot encoded.

- **Linguistic Inquiry and Word Count (LIWC)**: a psycho-social dictionary [14]. As said in [15], “It reveals the links between the personality characteristics of the author and quantitative parameters of the text (number of the words of some of the parts of speech, of some lexical-thematic groups, the frequency of punctuation marks, etc.)”

4. Models

4.1. Conventional machine learning methods

- **Support Vector Machine (SVM)**. We used the classifier based on a support vector machine with linear kernel. The following hyperparameters were used: regularization parameter \( C = 1 \), \( L_2 \)-norm used in the penalization and squared hinge-loss function.

- **Gradient boosting classifier (GB)** trained with the following parameters: learning rate is 0.05, the number of boosting stages to perform is 300, the minimum number of samples to split an internal node is 19, the maximum tree depth is 12.

- **Random Forest classifier (RF)**. The parameters of this model are selected using a genetic algorithm, the following are set: the Gini index, the number of counts taken in calculating the best division is 55%, the minimum number of edges on the node 8, the minimum number of examples for the division of a node 5, the number of trees to be calculated 100.

Standard models were trained using 10-fold cross-validation, dividing the set into 80% training and 20% unused data, as testing was conducted on a separate corpus.

4.2. Deep neural network

**Model 1** is based on convolution and fully connected layers:

(i) **Convolutional Neural Network (CNN)**: 128 neurons, window size = 2, activation function = Relu

(ii) **Maxpooling layer**: window = 4, step = 4

(iii) **CNN**: 128 neurons, window = 2, activation function = Relu

(iv) **Maxpooling layer**: window = 4, step = 4

(v) **CNN**: 128 neurons, window = 2, activation function = Relu

(vi) **Maxpooling layer**: window = 4, step = 4

(vii) **CNN**: 128 neurons, window = 2, activation function = Relu

(viii) **GlobalMaxPooling**

(ix) **Dropout 0.5**

(x) **Fully-connected MLP layer**: 128 neurons, activation function = Relu

(xi) **Dropout 0.5**

(xii) **Output layer - MLP**: activation function = softmax
5. Experiments

In all experiments, the models are estimated by F1-score (denoted as F1 in tables) and standard deviation (F1_std in tables), obtained from the test data. The test and training samples are balanced by the number of examples in each of the classes, so the baseline (bl in tables) is 0.5. As the test set we used a balanced set of "A" and "B" of corpus Gender Imitation. Since separate corpora are used as learning and test sets, the intersection of the authors of texts is avoided. A number of experiments with different algorithms of machine learning based on combinations of training samples are described below.

5.1. Experiments 1

The goal of these calculations is to state the level of accuracy of gender identification of text author in texts without gender deception. The model learning has been made, on base of both separate parts corpora and the extended variant of its combinations. The model testing has been made only on texts without gender deception, that is, on part ‘A’ of the Gender Imitation corpus.

| #  | Train dataset     | Model | Features  | F1  | F1_std | Delta(bl) |
|----|------------------|-------|-----------|-----|--------|-----------|
| 1  | GI_cs_a_b        | GB    | TF-IDF    | 0.65| 0.03   | 0.15      |
| 2  | GI_cs_a          | GB    | TF-IDF    | 0.78| 0.02   | 0.28      |
| 3  | GI_cs_a          | SVM   | TF-IDF    | 0.73| 0.02   | 0.23      |
| 4  | GI_cs; RusPer; Reviews; Tw_s; LJ_s; FB_s | Model 1 | Seq. feat | 0.88| 0.03   | 0.38      |
| 5  | RusPer; Reviews ;Tw_s; LJ_s; FB_s; | Model 1 | Seq. feat | 0.84| 0.02   | 0.34      |
| 6  | GI_cs            | Model 1 | Seq. feat | 0.87| 0.03   | 0.37      |
| 7  | Tw_s; LJ_s; FB_s; | Model 1 | Seq. feat | 0.82| 0.02   | 0.32      |
| 8  | Tw; LJ; FB       | Model 1 | Seq. feat | 0.77| 0.04   | 0.27      |
| 9  | GI_cs; RusPer; Reviews | Model 1 | Seq. feat | 0.88| 0.03   | 0.38      |
| 10 | RusPer; Reviews  | Model 1 | Seq. feat | 0.79| 0.03   | 0.29      |
| 11 | GI_cs_a_b        | Model 1 | Seq. feat | 0.86| 0.03   | 0.36      |

5.2. Experiment 2.

In this case we exploited the approach of the previous experiment based on learning on different sets of corpora but with testing on part ‘B’ of the Gender Imitation corpus, with gender deception.

5.3. Experiment 3.

Calculations based on models, analogous to the ones used in second experiment 5.2, have been made on base of texts filtered from the crowdsourcing corpus. For filtering the special model is used, learned on different text samples of the reference corpus Gender Imitation.
Table 2. The results of experiment 2 on subset ‘B’ of GI

| #  | Train dataset | Model      | Features | F1 | F1_std | Delta(bl) |
|----|---------------|------------|----------|----|--------|-----------|
| 1  | GI\textsubscript{cs\_a\_b} | SVM        | TF-IDF   | 0.54 | -      | 0.04      |
| 2  | GI\textsubscript{cs\_a\_b} | GB         | TF-IDF   | 0.64 | 0.03   | 0.14      |
| 3  | GI\textsubscript{cs} | Model 1    | Word2vec | 0.45 | -      | -0.05     |
| 4  | GI\textsubscript{cs\_a\_b} | SVM        | LIWC     | 0.53 | 0.2    | 0.03      |
| 5  | GI\textsubscript{cs\_a\_b} | DT         | LIWC     | 0.51 | 0.04   | 0.01      |
| 6  | GI\textsubscript{cs\_a\_b} | GD         | LIWC     | 0.46 | 0.07   | -0.04     |
| 7  | GI\textsubscript{cs\_a\_b} | RF         | LIWC     | 0.63 | 0.02   | 0.13      |

(i) Learning Model 1 (SVM) on Gender Imitation corpus Gender Imitation using TF-IDF features. The ‘GI dataset’ group of columns in Table 3 shows the parts of the corpus used to learn the filtering model.

(ii) Using the model learned on the previous step, the crowdsourcing corpus has been filtered, taking only correctly predicted examples.

(iii) Training the model 2 (SVM) on the filtered data of ‘GI\textsubscript{cs} dataset’ of Table 3 using the TF-IDF features.

(iv) The testing was conducted on the texts from Gender Imitation. The results of testing the model 2 on the full body of Gender Imitation are presented in Table 3.

Table 3. The results of experiment 3 using the “data filtering” algorithm.

| #  | GI dataset | GI\textsubscript{cs\_a\_b} dataset | F1 (a) | F1 (b) | F1 (c) |
|----|------------|-------------------------------------|--------|--------|--------|
| 1  | a          | a, b, c                             | 0.87   | 0.57   | 0.67   |
| 2  | b          | a, b, c                             | 0.48   | 0.87   | 0.55   |
| 3  | c          | a, b, c                             | 0.67   | 0.49   | 0.95   |
| 4  | a, b       | a, b, c                             | 0.84   | 0.93   | 0.66   |
| 5  | a, c       | a, b, c                             | 0.81   | 0.59   | 0.86   |
| 6  | b, c       | a, b, c                             | 0.61   | 0.83   | 0.82   |
| 7  | a, b, c    | a, b, c                             | 0.84   | 0.83   | 0.84   |
| 8  | a          | a, b                                | 0.79   | 0.49   | 0.60   |
| 9  | b          | a, b                                | 0.45   | 0.83   | 0.49   |
| 10 | c          | a, b                                | 0.61   | 0.73   | 0.93   |
| 11 | a, b       | a, b                                | 0.80   | 0.90   | 0.60   |
| 12 | a, c       | a, b                                | 0.77   | 0.58   | 0.82   |
| 13 | b, c       | a, b                                | 0.61   | 0.86   | 0.78   |
| 14 | a, b, c    | a, b                                | 0.80   | 0.81   | 0.80   |
| 15 | a          | a                                   | 0.85   | 0.45   | 0.64   |
| 16 | b          | a                                   | 0.52   | 0.71   | 0.65   |
| 17 | c          | a                                   | 0.65   | 0.59   | 0.91   |
| 18 | a, b       | a                                   | 0.86   | 0.71   | 0.67   |
| 19 | a, c       | a                                   | 0.81   | 0.53   | 0.79   |
| 20 | b, c       | a                                   | 0.67   | 0.76   | 0.78   |
6. Results
Results of the first experiment demonstrate an accuracy level for the task of author gender identification on texts without gender imitation. It follows from Table 1, that the best model for this task is the convolutional neural network with input data represented as a morphological features sequence, and it gives an F1-score equal to 0.88 with standard deviation of 0.03 by cross-validation.

Results of the second experiment point out that complicated neural networks with morphological data are ineffective for texts with gender imitation, and in this case the best model is gradient boosting with TF-IDF character n-gram vectors as input data, which demonstrates the F1-score, equal to 0.64 with standard deviation of 0.03. It is 0.22 lower than in case of texts without the gender imitation. One possible reason of significant accuracy falling is the difference in the features distribution between Gender Imitation and crowdsourcing corpora. Section 5.3 gives partial answer on this. As could be seen, training on filtered crowdsourcing data gives a significant growth in F1-score on texts with gender imitation. Moreover, accuracy of the gender identification could reach 0.93 depending on training sample composition and subsample selection from A, B, and C.

7. Conclusion
For tasks with and without gender imitation, different models of author gender identification are effective. Further accuracy growth requires sufficiently large training samples. The most perspective way to obtain it is to use crowdsourcing, but manual filtering and checking of these texts is a labour-consuming task. The proposed approach to solve the task is a way to get a larger corpus by crowdsourcing with the same features distribution as a comparatively small reference sample and use it for training the new model of increased accuracy.

8. Acknowledgements
This research is supported by the Russian Science Foundation, project No 16-18-10050. 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
[1] F Rangel, P Rosso, M Potthast, and B Stein 2017 Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter Working Notes Papers of the CLEF
[2] A Basile, G Dwyer, M Medvedeva, J Rawee, H Haagsma and Malvina Nissim 2017 N-gram: New groningen author-profiling model arXiv preprint arXiv:1707.03764
[3] M Martinc, IzasKrianc, K Zupan, and S Pollak 2017 Pan 2017: Author profiling-gender and language variety prediction Notebook for PAN at CLEF
[4] Y Miura, T Taniguchi, M Taniguchi, and T Ohkuma 2017 Author profiling with word+ character neural attention network Notebook for PAN at CLEF
[5] T Litvinova, P Seredin, O Litvinova and O Zagorovskaya 2017 Differences in type-token ratio and part-of-speech frequencies in male and female russian written texts In Proceedings of the Workshop on Stylistic Variation 73
[6] Markov I et al 2017 The Winning Approach to Cross-Genre Gender Identification in Russian at RUSProfiling 2017 Training 600
[7] Sboev A et al 2016 Deep Learning Network Models to Categorize Texts According to Author’s Gender and to Identify Text Sentiment Computational Science and Computational Intelligence (CSCI) International Conference pp 1101-1106
[8] Sboev A et al 2016 Machine Learning Models of Text Categorization by Author Gender Using Topic-independent Features Procedia Computer Science 101 pp 135-142
[9] V Perez-Rosas and R Mihalcea 2015 Experiments in open domain deception detection EMNLP 11201125
[10] Amazon Mechanical Turk 2012 Amazon mechanical turok. Retrieved August 17
[11] T Litvinova, O Litvinlova, O Zagorovskaya, P Seredin, A Sboev and O Romanchenko 2016 "Ruspersonality": A russian corpus for authorship profiling and deception detection In Intelligence, Social Media and Web (ISMW FRUCT) IEEE 7
[12] RusProfiling Lab 2017 Rusprofiling corpus of russian texts [online] http://rusprofilinglab.ru/rusprofiling-at-pan/corpus/
[13] A Kutuzov and E Kuzmenko 2017 Building web-interfaces for vector semantic models with the webvectors toolkit EACL 2017 99
[14] Yla R Tausczik and James W Pennebaker 2010 The psychological meaning of words: Liwc and computerized text analysis methods Journal of language and social psychology 29(1) 54
[15] O Litvinova, P Seredin, T Litvinova and J Lyell 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 52
[16] Lin Wu, Chunhua Shen and Anton van den Hengel 2016 Personnet: Person re-identification with deep convolutional neural networks arXiv preprint arXiv:1601.07255