Gender Prediction for Authors of Russian Texts Using Regression And Classification Techniques

Tatiana Litvinova[1,3], Pavel Seredin[2,3], Olga Litvinova[1,3], Olga Zagorovskaya[1,3],
Alexandr Sboev[3], Dmitry Gudovskih[3], Ivan Moloshnikov[3], Roman Rybka[3]

[1]Voronezh State Pedagogical University, Voronezh, Russia
centr_rus_yaz@mail.ru
[2]Voronezh State University, Voronezh, Russia
paul@phys.vsu.ru
[3]Kurchatov Institute, Moscow, Russia
sag111@mail.ru

Abstract. Automatic extraction of information about authors of texts (gender, age, psychological type, etc.) based on the analysis of linguistic parameters has gained a particular significance as there are more online texts whose authors either avoid providing any personal data or make it intentionally deceptive despite of it being of practical importance in marketing, forensics, sociology. These studies have been performed over the last 10 years and mainly for English. The paper presents the results of the study of a corpus of Russian-language texts *RusPersonality* that addressed automatic identification of the gender of the author of a Russian text using mostly topic-independent text parameters. The identification of the gender of authors of texts was addressed as a classification as well as regression task. For the first time for Russian texts we have obtained the models classifying authors of texts according to their gender with the accuracy identical to the state-of-the-art one.

Keywords: authorship profiling · corpus · stylometry · text classification · regression · gender attribution

1 Introduction

In recent years, exponential increase in textual information has sparked interest in automatically predicting users' personal information (gender, age, personality traits and so on). This field of research is often referred to as authorship profiling. Automatic prediction of such information has various applications in the fields of forensics, business intelligence and security.

The general algorithm for solving this problem is as follows:

1) Collecting a corpus of texts with metadata containing information about their authors;

2) Designing a list of text parameters, linguistic labelling of a corpus and extraction of numerical values of selected text parameters;

3) Designing a mathematical model to detect a certain personality trait based on qualitative values of texts and evaluation of their accuracy.

This area of research has been rapidly developing. There have been contests to find most accurate techniques for categorizing texts according to their authors' personal information [10].

One of the most important characteristics of authors of texts is their gender, i.e. there are a lot of papers on automatic detection of personality traits using texts. Research in identifying author's gender started with extensions of the earlier work on categorization and classification of text [7]. Using the various methods and features, researchers have automated prediction of an author's gender with accuracies ranging from 80% to 90%. [1;2;4;5;12]. For instance, the winners of PAN 2015 obtained models to classify texts according to the gender of their authors with the accuracy as high as 0.97 for Danish and Spanish and 0.86 for English [10].

There are still a lot of issues to be addressed and selecting the parameters to study seems most crucial. Different groups of text parameters were used which can be extracted using NLP tools such as content-based features (bag of words, words n-grams, dictionary words, slang words, ironic words, sentiment words, emotional words) and style-based features (frequency of punctuation marks, capital letters, quotations, together with POS tags) as well as feature selection along with a supervised learning algorithm (see [10] for review). Different research including the one mentioned above have used the parameters of the frequency of words of different topics but it is obvious that the resulting models might not be appropriate to use for corpora of texts of other genres. We are also cautious about the fact that «the reported performance might be overly optimistic due to non-stylistic factors such as topic bias in gender that can make the gender detection task easier» [12, p. 78]. Therefore it is essential that the high-frequency parameters less dependent on a particular topic and genre are used.

Most studies of the classification of texts according to the gender of their authors have been conducted using English texts and there have been only a few studies dealing with other languages (see [3] for details), especially Slavic ones.

The author's gender is known to be explicitly expressed in Russian texts if a verb in a sentence is in the past form and the subject is a singular first-person pronoun "я". Compare: "Прошлой зимой я **ездила** в Альпы" (a female speaker); "Прошлой зимой я **ездил** в Альпы " (a male speaker). If the subject is not the pronoun "я" or if the verb is not in the past form, the gender of the speaker is not explicit. Compare: "Я **поеду** в Альпы" (the gender of the speaker is not explicit). It is worth emphasizing that the existence of grammatical forms which reflect the speaker's gender does not automatically make gender identification in Russian texts a trivial task. In Russian "the gender of the speaker" is explicit in a statistically insignificant number of statements. Any non-first-person narrative does not indicate the gender of its author. Besides, it is easy for the author to imitate the speech of an individual of the other gender using the above forms. Therefore it is only by relying on these parameters that the gender of the author can be identified particularly in a forensic context.

In our lab, we focus on identifying the gender of authors of Russian texts using different methods of data analysis and different sets of text parameters. The basic assumptions of our research rely on issues facing forensic analysis, i.e. we use relatively short texts (150-300 words) as material for training and testing models and the text parameters were relatively topic-independent and cannot be consciously controlled and imitated and quantifiable and extracted by means of different NLP tools.

2 Design of the study

2.1 Dataset

For this study, we used corpus “RusPersonality” which consists of Russian-language texts of different genres (e.g. description of a picture, essays on different topics, etc.) labelled with information on their authors (gender, age, results of psychological tests, and so on). As of now, the corpus “RusPersonality” contains 1 867 texts by 1 145 respondents (depending on the type of a task, they wrote one or two texts). Overall corpus contains about 300 000 words. The average length of texts was 230 words. Most of the respondents were students of Russian universities. For experimental studies of automatic identification of an author’s gender we selected only students’ texts so that other factors (age, education level, etc.) do not have any influence on gender and linguistic text parameters. Selections in all of the experiments were balanced by gender. The selections in all the experiments were balanced by gender.

2.2 Feature set

We employed different text parameters that are relatively topic-independent.

Morphological features:

- POS tag features, which mainly represent a particular part of speech for every word in a given text: the number of nouns; the number of numerals; the number of adjectives; the number of prepositions; the number of verbs; the number of pronouns; the number of interjections; the number of adverbs; the number of particles, the number of conjunctions, the number of participles, the number of infinitives, the number of finite verbs (were extracted in different experiments using a morphological parser by XEROX, pymorphy 2 library script, morphosyntactic parser [11]);
- Derivatives of the coefficients which were different relationships of parts of speech: Treiger index, dynamics coefficient, 27 in total [9], [13];
- POS bigrams extracted using a morphological parser by XEROX [8];

1. Syntactical features (60):

- Synto – frequencies of different types of syntactic relationships between heads and dependents. Syntactic structure of sentences was analyzed as a dependency tree and extracted using a morphosyntactic parser [11];
- number of sentences of different types: compound and complex, etc. (extracted manually);

2. *Punctuation features* – the number of commas, exclamatory marks, the number of question marks; the number of dots; the number of emoticons etc. extracted by means of a specially designed script;
3. *Lexical features* – lexical diversity indices extracted using online service istio.com and EmoDicts – frequencies of words denoting different types of emotions (e.g., “Anxiety”, “Discontent”, the total of 37 categories, see [6] for details).

2.3 Methods

We have addressed automatic detection of an author’s gender as a regression and text classification task. Logistic regression was designed using IBM SPSS Statistics software.

Basically, the prediction of gender and age of the author of a text document is made by machine learning algorithms. Independent of the classifier used (see Section IV-D), the input consists of a large list of features with appropriate values and a corresponding classification class. The class is used to train the algorithms if the document is part of the training set, as well as for evaluating if the document is part of the test set. To determine the best working algorithm for this approach, several commonly used methods have been tested, which are well studied and have been used extensively in several text classification tasks. In particular, we used Gradient Boosting Classifier, Adaptive Boosting Classifier (adaBoosting), ExtraTrees, Random Forest, PNN (sigma = 0.1), Support Vector Machine with linear kernel (SVMs), ReLU (1 Hidden Layer with 26 neurons). Python libraries were used for learning the classification models: scikit-learn fitted with machine learning methods and keras for designing neural network models (http://scikit-learn.org/, https://pypi.python.org/pypi/Keras).

3 Results

3.1 Regression models

1. First experiment. For a pilot study, 150 texts from 75 participants (26 males, 49 females) were selected with the average number of words being 166. There was a total of 75 text parameters all of which are relatives values that is correlations of numerical values of different text parameters (part-of-speech correlations, e.g. (vfin+vinf)/noun, adj/(adv+pronadv), correlations of the number of types of various syntactic structures and so on) [9]. Ratios, i.e. relative frequencies, were used as the parameters in order to refrain from the dependence on the length of the text.

In order to estimate the closeness and direction of the linkage between the parameters of the text and personality and to establish the analytical expression (form), correlation and regression analysis was used based on modern statistical data visualization software. The main aim of the study was to establish a function dependence of a conditional mean of the result property (Y) (gender) on the factor properties (x_1, x_2, …, x_k), which are the parameters of the text. Therefore the initial regression equation, or a statistical model of the relationship between the author’s gender and quantitative parameters of the text is given by the function

$$Y(x) = f(x_1, x_2, \ldots, x_n), \tag{1}$$

where n is a number of factors included in the model; x_i are the factors that influence the result Y.

In order to determine the characteristics of and type of connection between the text parameters and individual characteristics of the author, a correlation analysis was performed ($p < 0.05$) using the software IBM SPSS Statistics. We established a number of correlations between the text parameters and the author's gender (0 – woman, 1 – man). The following correlations ($p < 0.05$) are found: the number of content words / the number of function words (0.258); the number of nouns / the total of words (0.252), the number of function words / the number of nouns (0.297), (pronouns of all types + prepositions + pronominal adverbs) / the total of words (-0.269) and so forth.

The accuracy of the model assessed on test corpus is ~60%.

2. Second experiment. The study [8] is a follow-up of the search for the text parameters independent of its subject matter and consciously uncontrolled by the author and therefore impossible to imitate. As the analysis of scientific literature suggests, these are frequencies of sequences (bigrams) of parts of speech. The research using English-language materials has proved the analysis of the frequencies of different bigrams in texts to be efficient in authorship profiling [14].

96 texts were used for the study which were randomly selected from RusPersonality corpus. The frequencies of POS bigrams in each text (227 types of bigrams were overall identified) were calculated, bigrams were then selected which are found in no less than 75% of the analyzed texts.

The only bigram found to have a significant correlation with the gender of the author of the text is **prep_noun bigram**. Its Pearson's correlation coefficient is 0.215. Therefore, it can be stated that there is weak linear connection between the proportions of prep_noun bigrams in the text and the gender of its author, males typically score more on this parameter.

Selecting different types of linear functions revealed that this dependence is most accurately described by a four-parameter linear regression.

The model was tested on test set (texts not used for designing the model, 10 written by males, 10 written by females, mean length = 161 word). The model was found to be 65% accurate. It also should be noted that the model was considerably better at distinguishing females than males.

3. Third experiment. A number of the text parameters correlated with gender of their authors allowed us to design a regression models [8;9]. However, testing of the quality of the models showed that this type of approximation yields a low level of accuracy as the parameters of texts by individuals of different gender are usually in overlapping ranges. This makes it impossible to design a functional model as part of a multiparameter regression. Therefore, it was decided to design a few regression models instead. In order to design regression models, 1090 texts by 545 authors were randomly selected (two texts by each respondent) from *RusPersonality*.

The text parameters were only those that were not consciously controlled: indicators of lexical diversity of a text, proportions of parts of speech, and different correlations of parts of speech (a total of 78 parameters).

For each text parameter a regression model was designed based on an optimal selection considering the sign of a correlation coefficient and exclusion of statistical outliers. Let us show the suggested approach using an example of 5 text parameters correlated with the gender of an author ($p<0.05$): TTR (type-token ratio, $r = 0.390$), formality ($r = 0.315$), a proportion of prepositions and pronoun-like adjectives ($r = 0.243$), proportion of the 100 most frequent Russian words in a text ($r = -0.322$); a ratio of function words to content words in a text ($r= -0.295$).

In order to properly estimate the obtained result, let us determine the average arithmetic values from the solution of the five equations:

$$GENDER_1 = -0.669 + \left(2.622 TTR\right), \tag{2}$$

$$GENDER_2 = -0.637 + \left(0.971\, Formality\right), \tag{3}$$

$$GENDER_3 = -0.188 + \left(0.0432\, preposition + pronoun - like\ adjective\right), \tag{4}$$

$$GENDER_4 = 1.500 - \left(0.0303\, Frequent\right), \tag{5}$$

$$GENDER_5 = 1.392 - \left(0.0229\, Function\right) \tag{6}$$

In order to properly estimate the obtained result, let us determine the average arithmetic values from the solution of the five equations.

In order to estimate the suggested approach, we used a corpus of texts with contributions from 553 individuals (368 women, and 185 men, while two texts from each respondent were considered as one text). Their topic and length were identical to those used to design the regression models. Gender was correctly identified in 65% of women and 63% of men. Thus, the accuracy of the approach was 64%.

4 Classification models

For the current research we have chosen 556 respondents and each of them wrote two texts (a description of a picture and a letter to a friend). Each of two texts were joined and considered as one text with average length of 350 words. All the texts were split into the learning (80%), cross-validation (10%) and trial (10%) samples. We used different groups of features, in total 141 features:

1) *Emomarkers* – psycholinguistic markers of emotiveness based on morphological features [8];

2) *EmoDicts* – frequencies of emotional words (e.g., “Anxiety”, “Discontent”, the total of 37 categories [6]);

3) *Litvinova* – a set of parameters used in [9] which are ratios of PoS frequencies, number of sentences in a text, number of clauses, number of exclamation marks etc.

4) *PoS* – frequencies of part-of-speech (nouns, adjectives, adverbs, pronouns, numerals, particles, prepositions, verbs, conjunctions) [11];

5) *Sinto* – frequencies of different types of syntactic relationships between head and dependents [11].

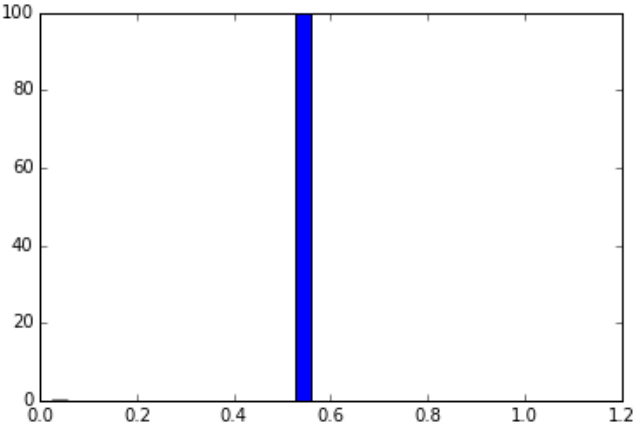
Comparative analysis of different machine learning algorithms has shown that ReLu is the most efficient classification algorithm with the F-score of 0.74 (see Table 1).

Table 1. F-scores for different classification algorithm

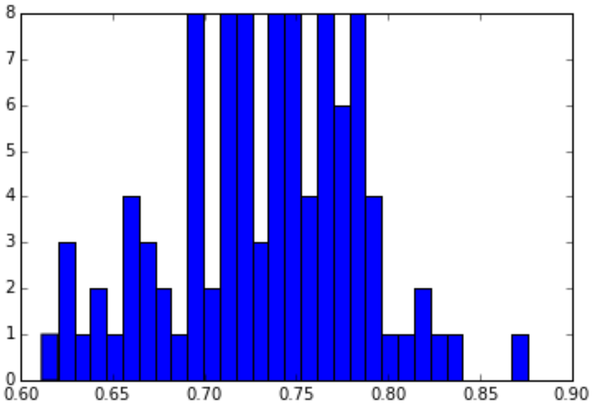
Model	**Feature selection techniques**	**Mean F1-score (25 cycles)**
Gradient Boosting	imp_quarter	0.72
adaBoosting	imp20	0.71
ExtraTrees	imp10	0.7
adaBoosting	common	0.7
Random Forest	imp10	0.7
PNN(sigma = 0.1)	imp10	0.68
SVM	PCA (30)	0.66
ReLU (1 Hidden Layer with 26 neurons)	imp10	0.74

In order to understand which groups of parameters yield the most accurate result, we designed graphs of f1-score distribution of the trained models (for SVM with a linear core) with different sets of parameters, each model was trained 100 times with a new combination of example sets. The selection was divided into the training (80%) and testing (20%) samples each time (see Fig. 1).

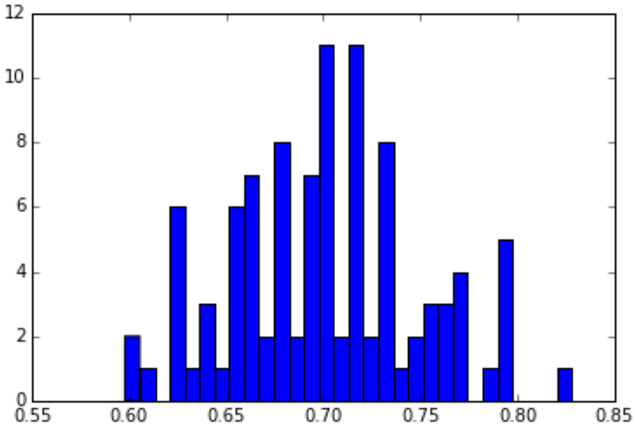
Fig. 1. F1-scores of the trained models (for SVM with a linear core) with different sets of parameters (the values of f1-score are plotted along the axis X, the number of models with a specified accuracy is plotted along the axis Y)



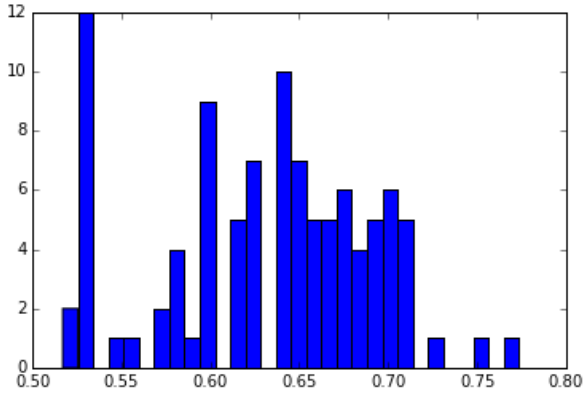
a) EmoMarkers



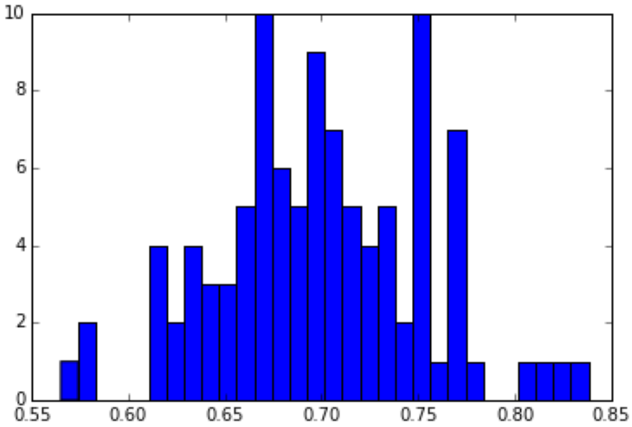
b) EmoDicts



c) Litvinova



d) PoS



e) Sinto

As Fig. 1 shows, the most informative parameters for identifying the gender of text author are different ratios of parts of speech, syntactic parameters, frequencies of various emotional words.

5 Conclusion and future work

The study performed for the first time for Russian-language texts confirms previously reported for English and some other languages findings that gender can be traced in texts beyond topic and genre using both regression and classification approach. It is shown that the author's gender is conveyed through specific syntactical and morphological patterns and use of emotion words. Comparative analysis of different machine learning algorithms has shown that ReLu is the most efficient classification algorithm with the F-score of 0.74. There are plans to expand the list of the text parameters and to test the obtained models on a Russian corpus of tweets and online chats.

Acknowledgements. This research is financially supported by the Russian Science Foundation, project No 16-18-10050 “Identifying the Gender and Age of Online Chatters Using Formal Parameters of their Texts”.

References

1. Burger, J. D., Henderson, J., Kim, G., Zarella, G.: Discriminating Gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Edinburgh, 27-31 July 2011, pp. 1301-1309.
2. Cheng, Na, Chandramouli, R., Subbalakshmi, K.P.: Author gender identification from text. Digital Investigation 8(1), 78-88 (2011)
3. Ciot, M., Sonderegger, M, Ruths, D.: Gender Inference of Twitter Users in Non-English Contexts. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2013).
4. Corney, M., Vel, O. de, Anderson, A., Mohay, G.: Gender-preferential text mining of e-mail discourse. In: Computer Security Applications Conference, 2002. Proceedings. 18th Annual, 2002, pp. 282-289.
5. Deitrick, W., Miller Z., Valyou B., Dickinson B., Munson T., Hu W.: Author Gender Prediction in an Email Stream Using Neural Networks. Journal of Intelligent Learning Systems and Applications 4(3) (2012)
6. Information Retrieval System "Emotions and feelings in lexicographical parameters: Dictionary emotive vocabulary of the Russian language.", Web: http://lexrus.ru/default.aspx?p=2876
7. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically categorizing written texts by author gender. Literary Linguist. Comput. 17(4), 401–412 (2002)
8. Litvinova, T. A., Seredin, P. V. Litvinova, O. A.: Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study. Indian Journal of Science and Technology 8(9) [S. l.], 93—97 (2015)
9. Litvinova, T. A.: Profiling the Author of a Written Text in Russian. Journal of Language and Literature 5(4): 210—216 (2014)
10. Rangel, F., Fabio, C., Rosso, P., Potthast, M., Stein, B., Daelemans W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Linda Cappellato and Nicola Ferro and Gareth Jones and Eric San Juan (eds.): CEUR Workshop Proceedings. Toulouse, France (2015) http://www.sensei-conversation.eu/wp-content/uploads/2015/09/15-pan@clef.pdf
11. Rybka, R., Sboev, A., Moloshnikov, I. Gudovskikh, D.: Morpho-syntactic parsing based on neural networks and corpus data. In: Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), pp. 89 –95. IEEE, St. Petersburg (2015)
12. Sarawgi, R., Gajulapalli, K., Choi, Y.: Gender attribution: tracing stylometric evidence beyond topic and genre. In: EMNLP '11 Proceedings of the 15th Conference on Computational Natural Language Learning, pp. 78–86. Association for Computational Linguistics, Stroudsburg (2011)
13. Sboev, A., Gudovskikh, D., Rybka, R., Moloshnikov, I.: A Quantitative Method of Text Emotiveness Evaluation on Base of the Psycholinguistic Markers Founded on Morphological Features. Procedia Computer Science vol. 66, 307-316 (2015)
14. Wright, W. R., Chin D. N., Personality Profiling from Text: Introducing Part-of-Speech N-Grams, User Modeling, Adaptation, and Personalization. Lecture Notes in Computer Science, 8538:502-507 (2014).