Procedia Computer Science

Procedia Computer Science 101, 2016, Pages 135 – 142

YSC 2016. 5th International Young Scientist Conference on Computational Science

# Machine Learning Models of Text Categorization by Author Gender Using Topic-Independent Features

Aleksandr Sboev[1,2,3], Tatiana Litvinova[1,4,5], Dmitry Gudovskikh[1], Roman Rybka[1], Ivan Moloshnikov[1]

[1]*NRC "Kurchatov Institute", Moscow, Russia*
[2]*MEPhI National Research Nuclear University, Moscow, Russia*
[3]*Plekhanov Russian University of Economics, Moscow, Russia*
[4]*Voronezh State Pedagogical University, Voronezh, Russia*
[5]*Voronezh State University, Voronezh, Russia*
sag111@mail.ru, centr_rus_yaz@mail.ru, dvgudovskikh@gmail.com, rybkarb@gmail.com, ivan-rus@yandex.ru

**Abstract**

In the present article, we address the problem of automatic text classification according to the author's gender. We used a preexisting corpus of Russian-language texts RusPersonality labeled with information on their authors (gender, age, psychological testing and so on). We performed the comparative study of machine learning techniques for gender attribution in Russian-language texts after deliberately removing gender bias in topics and genre. The obtained models of classifying Russian texts by their authors' gender demonstrate accuracy close to the state-of-the-art and even higher (up to 0.86 +/-0.03 in Accuracy, 86% in F1-score).

*Keywords:* text classification, gender identification corpus stylometry, authorship profiling, CNN, LSTM

## 1 Introduction

The last 20 years have seen an explosion of research in automated retrieving information from the text, in particular, information about its author (authorship profiling). The automatic extraction of information from text related to the gender, age and other demographic characteristics of the author is essential in forensics, security, and marketing. For example, one would like to learn about the linguistic profile of the author of aggressive textual message, or companies may be interested to learn about the demographics of people who like or dislike their products, given blogs and on-line product reviews as analysis source.

Peer-review under responsibility of organizing committee of the scientific committee of the 5th International Young Scientist Conference on Computational Science

doi:10.1016/j.procs.2016.11.017

This area of research has seen a massive development. There have been contests to find most accurate techniques for categorizing texts according to their authors' personal information (Rangel 2015). Most studies are dealing with the binary classification of text according to the gender of their authors. The winners of PAN 2015 obtained models that identify gender with the accuracy as high as 0.97 for Danish and Spanish, 0.86 for English (Rangel F. et al., 2015).

There are still a lot of issues to be addressed and selecting the parameters to study seems most crucial. Different research (see review by Rangel F. et al. 2015) have used as parameters 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» (Sarawgi R. et al., 2011). 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, especially for Slavic languages. The objective of this paper is to explore the possibility of automatically classifying Russian written texts according to their authors' gender using the parameters that are rather context-independent.

# 2 Related work

Starting from 2003 «Automatically Categorizing Written Texts by Author Gender» (Koppel M. et al., 2003) scientists have tried to tackle the categorization of texts according to the gender of their authors. Different groups of text parameters which can be extracted with NLP tools were used, such as content-based features (bag of words, words n-grams, dictionary words, slang words, ironic words, sentiment words, emotional words), style-based features (frequency of punctuation marks, capital letters, quotations, together with POS tags), and feature selection along with a supervised learning algorithm (Argamon S., et al., 2009), (Burger J., et al., 2011), (Schler J., et al., 2006). For example, (Schler J., et al., 2006), (Francisco Rangel) used a set of stylistic features like non-dictionary words, parts-of-speech, function words and hyperlinks, combined with content features, such as word unigrams with the highest information gain, and obtained accuracy of about 80% for blog author gender identification. Only a few researchers used syntactic features as well; see, e.g., (Cheng N., et al., July, 2011).

However, nearly all models presented in the state-of-the-art works in the area still depend on the datasets they were trained and tested on, since they heavily draw on content features, mostly on a large number of recurrent words or combinations of words extracted from the training sets (Company J.S., et al., 2014).

As for Russian language, sociolinguists, specialists of gender linguistics, and criminalists have been investigating differences in speech between men and women for a long time (Goroshko E., 1996), (Oshepkova E., 2003), (Ermolova E., 2008), (Viazgina N., 2012), (Ryzhkova E., 2015), (Golev N.D., 2015), (Litvinova T.A., 2015). However these studies were more descriptive and analytical than diagnostic. During a few last years there appeared some articles, devoted to mathematical models for diagnostic of Russian texts author by gender see (Litvinova T. A., et al., 2015), (Litvinova T.A., et al., 2014), (Romanov A.S., et al., 2011), (Korshunov A., 2013). But this works does not contain comparative analysis of different text classification methods and based on limited number of text parameters. A comparison of existing methods of gender classification of author texts was made in (Litvinova T.A., 2016).

In present paper we perform the comparative study of machine learning techniques for author gender attribution in full-sized Russian-language texts after deliberately removing gender bias in topics and genre using corpus with much more number of authors.

# 3 Materials

For this study, we used the "RusPersonality Corpus" which consists of Russian-language texts of different genres, which are samples of a natural written speech (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). All the texts were written by respondents during experiments, so, all the texts are not edited, do not contain plagiarism, and all the characteristics of the author are known. As of 23.05.2016 the corpus RusPersonality contained 1 867 texts of 1 145 respondents (depending on the type of a task, they wrote one or two texts). Overall it contains about 300 000 word or about 1 800 000 symbols.

For the current research we chose 556 respondents. Each of them wrote two texts (a description of a picture and a letter to a friend). Two texts of the same author were combined and considered as one text. The average length of combined text was 350 words.

# 4 Methods

## 4.1 Feature description

We have used a set of morphological and syntactic features that were chosen solely on the basis of that these would be more or less topic-independent. In addition, we have selected as parameters frequencies of the words describing emotions.

Automated analysis of texts relies on the software using the neural network method for morpho-syntactic analysis of Russian texts (Rybka R.B., et al., 2015). A wide range of morphological and syntactical features was selected and homonyms were processed. Python libraries were used for learning the classification models: scikit-learn with machine learning methods and Keras for designing neural network models (scikit-learn), (keras).

Besides morphological and syntactical features, we used emotion-based features which, as suggested in some papers see, e.g., (Montero C. S., et al., 2014), have positive implications on classification of author's gender.

The total of 141 parameters were identified, which are as follows:

1. Morphological features (13) – POS tag features. These are 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.
2. Syntactical parameters (60) – syntactic relations of different types (groups from 1 to 5 from (RNC, 2003-2016).
3. Derivative coefficients which reflect different relationships of parts of speech (Treiger index, dynamics coefficient, 27 in total) (Litvinova T.A., et al., 2014), (Sboev A.G., et al., 2015). The number of exclamatory marks, the number of question marks; the number of dots; the number of emoticons (4).
4. The number of words pertaining to a particular group "Emotion" (e.g., "Anxiety", "Discontent", the total of 37 categories, see (Information Retrieval System "Emotions and feelings in lexicographical parameters: Dictionary emotive vocabulary of the Russian language.") for details).

## 4.2 Algorithms

We used:

Approach 1. The set of different machine learning algorithms including extensively used in text classification tasks: 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).

Approach 2. Complicated topologies of ANN – a Convolution Neural Network (CNN) that applies convolutional filters to successive windows for a given sequence to extract global features by max-pooling (Collobert R., et al., 2011)composed with Long Short Term Memory neuronet (LSTM) (Hochreiter S., et al., 1997).

## 4.3 Experiments setup and evaluation

The model for gender classification using mentioned algorithms (Approach 1) was designed based on a specified set of features which were selected according to following stages.

1. Quantitative parameters were transferred into a numerical vector with min-max normalization from 0 to 1.
2. The most frequent significant features were identified using the following methods:
    - common – regular data with no preliminary processing;
    - PCA – selecting significant features using the method of major components, the number of the major components was 10, 30, 50 and 100;
    - impX – the number of the features are 10,20, quarter and half of total number of samples, selected by following algorithm:
        a. the train and test data were split into 4 folds using stratified k-fold; (Kohavi, 1995)
        b. - for each fold 1000 classification trees are computed using the Gini split index;
        c. - for each tree the f1-score is calculated which characterizes the accuracy of predicting each of the trees;
        d. - the largest Gini index for the parameter in each particular tree means that it is significant for the classification;
        e. - for each parameter a sum of the products of the f1-score and Gini index was calculated in 1000 trees in splits on 4 folds and the resulting value which is the sum of the values of the Gini index calculated for the parameter in each tree as assumed indicates the importance of the parameter.
3. Total sample set is split into subsets of: learning (80%), cross validation (10%) and trial (10%).
4. A set of the classification models was learned using a series of trials (10) followed by cross-validation by means of the stratified K-fold method.
5. The evaluation of the model. The average weighed F1-score was used.

The input data for ANN (Approach 2) was based on only grammatical features without word-forms and lemmas.

Grammatical signs included: part of speech, noun case, verb form, gender, number and etc. We allocated 43 morphological features in accordance with the Russian National Corpus format (RNC,

2003-2016). To represent them as numbers typically one converts each categorical feature using "one-hot encoding". Let $X$ be a set of documents, where $\{x_1,\ldots,x_i\} \in X$. $Y$ is a set of labels. Every text was represented by a matrix A($n*k*m$), where n — number of texts, k — the dimension of the feature space, $m$ – length of document ($m = L\text{-}l$, where $L$ — maximum permissible length of the document (300 words), $l$ — the actual length of the document).

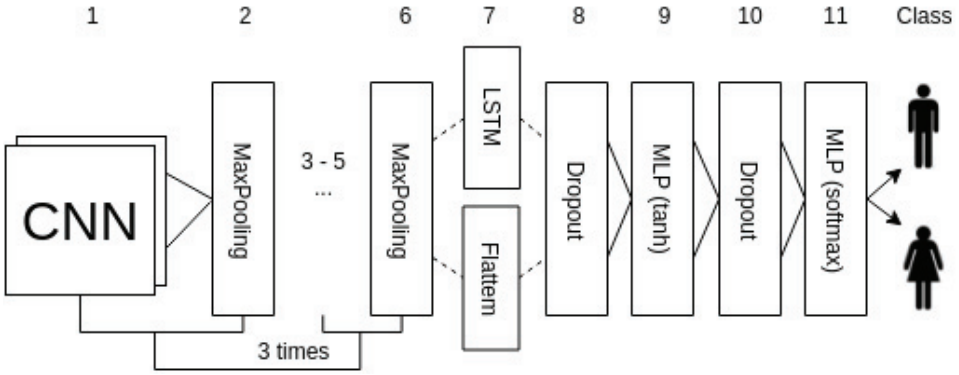


**Fig. 1.** Neural network structure with combination of CNN, MLP and LSTM:

- A complicated neural network combining CNN, MLP and LSTM includes:
- 1st, 3rd, 5th CNN layers: Number of convolution kernels to use = 30, the extension of each filter = 3, activation function is ReLU
- 2nd, 4th, 6th layers: MaxPooling (pool length = 2)
- *First variant 7th layer: flatten layer to transform matrix to vector*
  *Second variant 7th layer: Long-Short Term Memory (output dimension = 30)*
- 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 = tanh)
- 10th layer: dropout layer (Fraction of the input units to drop = 0.5)
- 11th layer: fully connected NN layer (Number of hidden neurons = 2, activation function = softmax)

Learning parameters:

RMSprop based on gradient descent optimization algorithm was used. The learning process takes place with the cross-validation. The number of permutation and split iterations = 10, 80% of samples for training, 20% for test)

The results of testing the models for gender classification are shown in Table 1.

| Model | Feature selection techniques | Mean F1-score | Mean Accuracy |
|---|---|---|---|
| Gradient Boosting | imp_quarter | 0.72+/-0.03 | 0.74+/-0.03 |
| adaBoosting | imp20 | 0.71+/-0.05 | 0.71+/-0.04 |
| ExtraTrees | imp10 | 0.71+/-0.04 | 0.72+/-0.03 |
| Random Forest | imp10 | 0.7+/-0.03 | 0.72+/-0.02 |
| PNN (sigma=0.1) | imp10 | 0.69+/-0.05 | 0.69+/-0.05 |
| SVM | PCA 10 | 0.72+/-0.04 | 0.72+/-0.04 |
| ReLU (1 Hidden Layer) | imp10 | 0.73+/-0.05 | 0.74+/-0.05 |
| CNN+MLP | Grammatical information | 0.82+/-0.05 | 0.83+/-0.05 |
| CNN+LSTM | Grammatical information | 0.86+/-0.05 | 0.86+/-0.03 |

**Table** 1: Results of gender identification

# 5 Conclusions and future work

The study performed for full-sized texts of Russian-language confirms previously reported for English and some other languages findings that gender can be traced in texts beyond topic and genre. It is also 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 among well studied and widely used algorithms ReLU is the most efficient classification algorithm with the accuracy 0,74+/-0,05 that seems to be state-of-arts for Russian language. Note that the result of work (Korshunov A., 2013) with the accuracy 0.86 was reached under other conditions on base of short texts from Twitter (up to 140 symbols) and it cannot be compared directly with above mentioned results. However, complicated neural network models (CNN+LSTM) demonstrate higher level of accuracy (86%+/-0.03) even on the base of very limited set of grammatical features. The disadvantage of such models is a difficulty in evaluating the features, because these models extract and process the input information nonlinearly. So, there are plans to create the procedure of extraction of the meaningful combinations of features from internal neural network parameters and to test obtained models on Russian social media texts.

# 6 ACKNOWLEDGEMENTS

This research is 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

Argamon S., et al. (2009, 2). Automatically profiling the author of an anonymous text. *Communications of the ACM 52 (2)*, pp. 119-123.

Burger J., et al. (2011). Discriminating gender on Twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1301-1309). Stroudsburg: Association for Computational Linguistics.

Cheng N., et al. (July, 2011). Author gender identification from text. *Digital Investigation: The International Journal of Digital Forensics & Incident Response, Volume 8 Issue 1*, Pages 78-88.

Collobert R., et al. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research, Volume 12*, 2493-2537.

Company J.S., et al. (2014). How to use less features and reach better performance in author gender identification. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, (pp. 1315-1319). Reykjavik, Iceland.

Ermolova E. (2008). A problem of determining the gender and the age of anonymous document author, basing on features of written language (in Russian). *Expert-Forensic №4*, 16-18.

Francisco Rangel, P. R. (n.d.). Overview of the Author Profiling Task at PAN 2013. *Forner P., Navigli R., Tufis D. (Eds.)Notebook Papers of CLEF 2013 LABs and Workshops. CEUR-WS.org, vol. 1179.*

Golev N.D., R. D. (2015). On gender-marked units in Russian-language texts as the statistic sharacteristics of the service parts of speech (in Russian). *Bulletin of Kemerovo state university, №3-1.*, pp. 148-153.

Goroshko E. (1996). Differentiation in male and female speech styles (psycholinguistic analysis). In *abstract of dissertation, Phd of philological Sciences;*. Moscow: The Institute of Linguistics, RAS.

Hochreiter S., et al. (1997). Long short-term memory. *Neural computation, Volume 9(8)*, 1735-1780.

*Information Retrieval System "Emotions and feelings in lexicographical parameters: Dictionary emotive vocabulary of the Russian language."*. (n.d.). Retrieved from http://lexrus.ru/default.aspx?p=2876

keras. (n.d.). *Keras Library*. Retrieved from http://keras.io/

Kohavi, R. (1995). A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Proc. of the 14th Int. Joint Conf. on A.I., Vol. 2*.

Koppel M. et al. (2003, 4 17). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, pp. 401-412.

Korshunov A., B. I. (2013). Detection of demographic attributes of microblog users. *Proceedings of the Institute for System Programming Volume 25*.

Litvinova T. A., et al. (2015). Using Part-of-Speech Sequences Frequencies in a Text to Predict Author Personality: a Corpus Study. *Indian Journal of Science and Technology, vol. 8(9)*, 93-97.

Litvinova T.A., e. a. (2015). *On identifying and diagnosis of author's personality of written text (in Russian).* Voronezh: “Gramota” Publishers.

Litvinova T.A., e. a. (2016). Gender Prediction for Authors of Russian Texts Using Regression And Classification Techniques. *Proceedings of the Third International Workshop on Concept Discovery in Unstructured Data (CDUD 2016)*, pp. 44-54.

Litvinova T.A., et al. (2014). Profiling the Author of a Written Text in Russian. *Journal of Language and Literature 5(4)*, 210-216.

Montero C. S., et al. (2014). Investigating the role of emotion-based features in author gender classification of text. *15th International Conference on Intelligent Text Processing and Computational Linguistics*, (pp. 2:98-114). Kathmandu, Nepal.

Oshepkova E. (2003). Identification of author's sex in written text (lexical and grammatical aspect) (in Russian). In *abstract of dissertation, Phd of philological Sciences.* Moscow: MSLU.

Rangel F. et al. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. *CLEF 2015 Labs and Workshops, Notebook Papers*, (p. France). Toulouse.

RNC. (2003-2016). *Russian National Corpus*. Retrieved from Russian language text corpus with syntactic markup: http://www.ruscorpora.ru/en/index.html

Romanov A.S., et al. (2011). Gender identification of the author of a short message. *Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference “Dialogue” (2011), Issue 10 (17)* (pp. 556-561). Moscow: Russian State University for the Humanities.

Rybka R.B., et al. (2015). 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). Saint-Petersburg, Russia.

Ryzhkova E. (2015). Problems of authorship classification examination of texts in internet communication (in Russian). *Philological Sciences. Issues of Theory and Practice №6. Part 1.*, pp. 130-133.

Sarawgi R. et al. (2011). Gender attribution: tracing stylometric evidence beyond topic and genre. *Proceedings of the 15th Conference on Computational Natural Language Learning* (p. USA). Stroudsburg: Association for Computational Linguistics.

Sboev A.G., et al. (2015). A Quantitative Method of Text Emotiveness Evaluation on Base of the Psycholinguistic Markers Founded on Morphological Features, Volum 66. *4th International*

*Young Scientist Conference on Computational Science* (pp. 307-316). Procedia Computer Science.

Schler J., et al. (2006). Effects of Age and Gender on Blogging. , *Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03*. Stanford, California, USA: Computational Approaches to Analyzing Weblogs.

scikit-learn. (n.d.). *scikit-learn*. Retrieved from http://scikit-learn.org/

Viazgina N. (2012). *Diagnosis of author's gender as task of authorship recognition (in Russian).* Retrieved from siberia-expert: http://siberia-expert.com/publ/konferencii/konferencija_2012/diagnostika_pola_avtora_kak_zadacha_avtorovedcheskoj_ehkspertizy_vjazigina_n_v/10-1-0-282