Data-driven Approaches to Author’s Profiling Identification for Russian Texts on Base of Complex Machine Learning Models in Combinations with Siamese Networks

Aleksandr Sboev\textsuperscript{1,2,*}, Ivan Moloshnikov\textsuperscript{1}, Dmitry Gudovskikh\textsuperscript{1} and Roman Rybka\textsuperscript{1}

\textsuperscript{1}NRC “Kurchatov Institute”, Moscow, Russia
\textsuperscript{2}MEPhI National Research Nuclear University, Moscow, Russian Federation
*Corresponding author

Keywords: Data-driving modeling, Author’s profiling, Age detection, Gender identification, Deep neural networks, Siamese networks.

Abstract. In this work data-driven approaches to author’s profiling identification for Russian texts are investigated on base of a united data corpus. This corpus has been specially collected by crowdsourcing, and currently contains texts from 1161 men and 2043 women. The adaptation of complicated models, based on convolutional neural networks, gradient boosting methods, LSTM, Siamese networks along with different input data and features (morphological data, vector of character n-grams frequencies, Linguistic Inquiry and Word Count and others) to form the vector of derived features in order to identify gender and age of the author of text is described. The method to improve the accuracy using coding by the Siamese network is presented and analyzed.

Introduction

The task of author’s profiling identification is very popular and relevant because of its possible use in practical implementations, like commercial and governmental. For such languages as English, Spanish, and Arabic, there are large text corpora to create data-driven models to identify author’s profile (see Section 3). A few last years the author’s profiling identification task has been rapidly developing for Russian, and it is strongly related to formation of similar Russian corpora, but the sizes of these corpora are currently smaller: in our work we use Gender-Imitation-Crowdsource (“GI cs”) corpus \cite{1}, and Gender Imitation corpus \cite{2} Naturally, it makes it more difficult to reach the high precision of task solution.

In this work to solve the problem of determining the age, we have specified 3 age groups to classify texts: 18-23, 24-29, 30+. The groups were chosen according to the representedness of the examples, so that numbers of profiles in groups were about the same. We investigate the effects of different types of features and their representations. The sets of features (see Section 4.1) include: morpho-syntactic, Linguistic Inquiry and Word Count (LIWC), a generalized dictionary approach of low displacement rank (LDR), derived features (see Section 4.2), along with different variants of vector representation of the text. The adaptation of complex models, based on convolutional neural networks, gradient boosting methods, LSTM, Siamese networks is described in Section 4.2. We then apply the same methods to the ‘GI cs’ datasets for the task of gender prediction (Section 7) and compare them to our previous results \cite{1,3}, that demonstrated the accuracy of 88\% ±3\%, which is about 30\% more than the baseline.

Corpus Description

The Gender-Imitation-Crowdsource (‘GI cs’) corpus was collected using the crowdsourcing platform. Each participant wrote 3 texts on one of the selected topics: a letter to a friend, a post for a dating site, a complaint letter to the boss or a negative review on the tour operator. The first text had to be written in the natural style (these texts are further referred to as ‘a’ collection), the second in...
form to mimic opposite gender (‘b’), and the third in form to imitate different style but not the gender (‘c’). At the preprocessing stage, the texts were manually and automatically checked for borrowing on the Internet. Some authors wrote several texts for each collection, in which case we combined these texts into a single document, except in the case of Siamese networks. The average length of combined documents is about 300 words and does not depend on age. The number of authors in each age group (18-23, 24-29, 30+) was balanced by excluding excessive authors. After balancing the resulting dataset contained about 800 authors on average. In addition to this ‘GI cs’, we used the Gender Imitation corpus that was collected by the same rules as described above, but off-line, in a fully controllable environment.

Related Work

There are different formulations of the task of author profiling, in particular for age and gender identification. In particular, in [4] binary age prediction was used: whether the age of a text author is under or over a definite boundary. Four values of boundaries were considered: 16, 18, 22, and 28. The balanced data set contained 23300 Dutch documents for each boundary and a total of 70060 documents for the task of gender identification. The input features were n-grams of symbols and tokens with the highest relative frequencies in the training set, and the classifier was a Support Vector Machine (SVM) with a linear kernel. The obtained accuracies for age identification ranged from 0.77 to 0.90 depending on document size, and about 0.70 for gender. The binary formulation of age identification in this paper essentially differs from ours. The closer formulation is presented in the results overview of the Author Profiling task at the PAN 2016 competition [5]. There a Twitter corpus was used for training, and different corpora from social media, blogs, essays, and reviews for evaluation. The following classes for age were considered: 18–24; 25–34; 35–49; 50–64; and 65+. The test samples contained two parts: texts from social media (at least 100 words per author) and blogs (25 posts of the author). Approaches based on the use of a wide range of stylistic features such as: various n-grams of symbols, words, word combinations and part of speech (PoS), measure term frequency–inverse document frequency (TF-IDF), punctuation, Out of Dictionary Words, Vocabulary Richness, Emoticons were presented in the works of winners [6]. Besides that, Second-Order Attributes were used, that represent terms and documents by vectors in a space of profiles (age, gender) to evaluate the relationship of a term with different profiles. The authors [7,8] drew attention to the number of grammatical errors in the text as the feature clearly reflecting the literacy of the author and indicating a possible age group. In most cases the track participants have used simplest classifiers: SVM [7] and logistic regression [8]. Results of competitions demonstrate the following level of accuracy for English: on Social Media for gender identification 0.55 and same on Blogs 0.75. As for Spanish, on Social Media for gender identification 0.70 and on Blogs 0.73. Results to define age group have given the accuracy levels for English on Social Media 0.38 and on Blogs 0.58, For Spanish on Social Media 0.35 and on Blogs 0.51. In work of [9] the classification in age groups under 17; 17–24; 24–34; 34–49; 49–64; 64+ on base of Entropy Maximization is investigated, using corpora of decoded calls from Fisher telephone conversation collection and forms of cancer patients. Different sets of attributes: n-grams of words, parts of speech, LIWC, average number of words, sentences, complexes of above mentioned features, and their aggregations in one vectors were tried. Authors got the age prediction accuracy of at best 52% (about 20% over the baseline), using content-based and stylistic features together.

Materials and Methods

Features

In this section the groups of features are described.

*Group 1.* Linguistic Inquiry and Word Count (LIWC). LIWC is a set of psychosocial dictionaries [10] with linguistics categories (the number of words of certain parts of speech, some lexical-
thematic groups, the frequency of punctuation marks, etc.) had been described and then they were adapted for Russian language [11].

**Group 2.** LDR (Lower-dimensional reduction). In this model [12] the document representation is as a vector of dependencies to categories (class-dependent vector). It is calculated based on the matrix of TF-IDF terms and class-dependent term weights. Each document is represented as \( \{ F(c_1), F(c_2), \ldots, F(c_n) \} \), where \( n \) is the number of categories, and \( F(c_i) \) is a vector characterizing the relationship of the document with the category \( c_i \). The components of \( F(c_i) \) are: the average weight of a document (calculated as the sum of weights of its terms divided by the total number of vocabulary terms of the document), the standard deviation of weights of documents, the minimum weight among the term weights of document, the maximum weight found in the document, the sum of weights of the terms of the document divided by the total number of terms of the document, the proportion between the number of vocabulary terms of the document and the total number of terms of the document.

**Group 3.** Model synt. As input for neural network each document is represented as a sequence of words where each word is encoded by a syntactic pair “word-parent”. The word and the parent are encoded by three types of features: unique index of lemma, Word2vec vector for word [13], binary encoded morphological tag.

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

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

**Pre-trained Models**

In this case the documents were presented by different encoders to form the vector of derived features. Depending on the selected set of encoders, the resulting vectors were concatenated (the dimension is approximately equal to 500) to train the Gradient Boosting Classifier. Next, we provide a description of the topology and learning parameters of pre-trained models.

**Model synt.** Model synt topology is shown on Fig. 3. Adam algorithm [14] based on gradient descent optimization is used. The optimization score function is mean squared error. Batch size is 32, early stopping after 5 epochs.

**Model NN.** The topology of network is taken from the work [15] to determine the sex of the author, it based only on the morphological characteristics of words. A complicated neural network combining CNN, MLP and LSTM includes: 1st, 3rd, 5th CNN layers: Number of convolution kernels to use = 128, the extension of each filter = 2, activation function is ReLU; 2nd, 4th, 6th layers: MaxPooling (pool length = 2); 7th layer: Long-Short Term Memory (output dimension = 128); 8th layer: dropout layer. (Fraction of the input units to drop = 0.5); 9th layer: fully connected NN layer (Number of hidden neurons = 10, activation function = softmax. Learning parameters: Adam algorithm based on gradient descent optimization was used. The optimization score function is mean squared error. Cross-validation is performed, with the number of permutation and split iterations = 10, 80% of samples for training, 20% for test.

**Siamese networks.** We have trained encoders based on Siamese neural networks for the representation of the document in the form of feature vector. Two encoders were trained for each task based on common topology. The first siamese encoder for both tasks was trained to determine the belonging of two documents to one author. The second encoder, depending on the task, was trained to identify the belonging of two documents to the same gender of author (male or female), but not necessarily the same author, or, respectively, to the same age for the age task. Classes are used in the siamese model at pre-training stage to calculate the documents’ similarity value. Thus, the documents of the same class (gender, author or age group) have less distances between them than documents of different classes. The encoder consists of two BiLSTM layers with 32 neurons and dense layer with 128 neurons and sigmoid activation function. A pair of documents are fed into the Siamese network, where each document passes through one of the two equal encoders with
common weights. Then the difference is calculated for the two encoded representations in a step-by-step manner. The difference vector is fed to one output neuron with sigmoidal activation function. The output network predicts the distance class (0 for close documents, 1 for distant documents). Further the resulting encoder is used to encode individual documents. During network training, pairs are randomly generated and the set is balanced by the distance class. Learning parameters for siamese models: Batch size is 256; the number of batches in the epoch is 16; optimizer: rmsprop (lr = 0.0005); an optimization score function function is the mean squared error; early stopping after 150 eras; the maximum number of epochs is 3000. The validation set is 256 * 4 pairs.

Classifiers

*Gradient Boosting Classifier.* We used default hyperparameters of sklearn library.

*Symbols model.* This model uses symbol-representation of text where each symbol is converted to a one-hot encoded vector. The topology and parameters of network: 1st, 3rd, CNN layers: Number of convolution kernels to use is 8, the extension of each filter is 2, activation function is ReLU; 2nd, 4th, 6th layers: MaxPooling (pool length is 2); 5th CNN layer: Number of convolution kernels to use is 32, the extension of each filter is 2, activation function is ReLU; 7th layer: Global average pooling operation for temporal data; 8th layer: dropout layer. (drop 0.3); 9th layer: fully connected NN layer with 3 hidden neurons and activation function is ‘softmax’. Adam algorithm based on gradient descent optimization was used. The optimization score function was mean squared error.

### Table 1. Aggregate results for Age and Gender.

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>f1</td>
<td>std</td>
</tr>
<tr>
<td>Best model</td>
<td>0.48</td>
<td>0.07</td>
</tr>
<tr>
<td>Mean all experiments</td>
<td>0.41</td>
<td>0.05</td>
</tr>
<tr>
<td>Mean with siames network</td>
<td>0.4</td>
<td>0.05</td>
</tr>
<tr>
<td>Mean without siames network</td>
<td>0.41</td>
<td>0.04</td>
</tr>
<tr>
<td>Symbol model</td>
<td>0.45</td>
<td>0.08</td>
</tr>
<tr>
<td>baseline</td>
<td>0.33</td>
<td>-</td>
</tr>
</tbody>
</table>

Experiments

We performed series of experiments for two tasks: determining the age group (18–23, 24–29, 30+) and identifying the gender of the author. For age identification the ‘GI cs’ corpus was divided into training set 72%, validation set 18% and 10% test set. In case of experiments with gender identification the test part was discarded, because in this case we used Gender Imitation corpus as an independent test set. The learning process takes place on balanced corpus with 10-fold cross-validation.

Results

Out of all feature selections we tried, both for age and gender tasks best feature groups with similar accuracies can be selected (see Tables 1): F1=0.48-0.45, Std=0.06-0.08 for the age task and F1=0.74-0.72, Std=0.04-0.06 for the gender task. In comparison with results for other languages cited in Section 3 these results in case of the age task are lower, F1 about 15% higher than the baseline versus about 19% (Social Media, mixed corpora), 39% (Blogs) for English; 16% (Social Media), 32% (Blogs) for Spanish. This shows an evident need to enlarge our corpus. As concerns the gender task, the obtained level of accuracy F1 is about 24% higher than the baseline versus about 2% (Social Media, mixed corpora), 19% (Blogs) for English; 17% (Social Media), 23% (Blogs) for Spanish, so our level is compatible with other languages. However, it is worse than in our previous works [12], because the latter used more big mixed sets to learn. The results based only on the symbols representation demonstrate the same accuracy like other models for the age task, and
lower accuracy for the gender task. Nonetheless, this “from scratch” learning is notable for not employing complicated input representations. Since such a model usually requires more data to train, with larger dataset it might compete with other models.

Conclusion

For the first time for Russian language we have evaluated the level of accuracy for the task to identify the age of a text author with the help of the current version of our crowdsourcing corpus. At the moment this level in terms of F1-score is about 24% higher than the baseline. Regarding the task to identify the gender of a text author the obtained accuracy is compatible with accuracies for other languages. In future work we plan to continue collecting our crowdsourcing corpus to enlarge the size of training sets, which may improve the accuracy, and to extend it with different genres.

Acknowledgements

This research is supported by the Russian Science Foundation, project No 16-1810050. This work has been carried out using computing resources of the federal collective usage center Complex for Simulation and Data Processing for Mega-science Facilities at NRC “Kurchatov Institute” (ministry subvention under agreement RFMEFI62117X0016), http://ckp.nrcki.ru/.

References


