To the question of data-driven identification of author’s age for Russian texts with age deceptions using machine learning

A Sboev¹², R Rybka¹, I Moloshnikov¹, D Gudovskikh¹ and T Litvinova³

¹ MEPhI National Research Nuclear University, Moscow, Russian Federation
² National research centre “Kurchatov institute”, Moscow, Russian Federation
³ Voronezskij Gosudarstvennyj Pedagogiceskij Universitet, Voronezh, Russian Federation

E-mail: sag111@mail.ru

Abstract. In this work we compare data-driven approaches to the task of author’s age identification for Russian texts with age deception. The data corpus has been specially gathered with crowdsourcing for this task. Two ways to determine age deception in author texts are considered and compared: the first is a traditional task of identification of age group of a text author, the second is identification of the occurrence of age imitation in the text with its type (imitating higher age or imitating lower age). The best results obtained by LinearSVC model with vector of TF-IDF features of character n-grams as input data demonstrate the F1-score of about 80% for the second task, and for the first one it is about 44%.

1. Introduction

Nowadays a great amount of information circulates in the Internet, but only part of it is truthful. Distribution of deceitful information on the internet may mislead people, in some cases leading to serious consequences [1, 2]. Thus for many practical fields, such as forensic, security, and others, it is important to have truthful information. In our previous works we investigated the task of gender identification of Russian text author in texts with gender deception. Here we consider the task of age identification in texts with age deception, i.e. when the authors tried to distort their ages. While the case of author age identification in texts without the age deception is widely represented in scientific literature for different languages [3, 4, 5, 6, 7, 8], sources for the case with deceptions are absent. The result overview of the age identification tasks at the PAN 2016 competition [4] demonstrates the state-of-the-arts for the case of author age identification in texts without age deception. There a Twitter corpus was used for training, and different corpora from social media, blogs, essays, and reviews were used 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 each 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 [7]. 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 [9, 5] 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 used the simplest classifiers: SVM [9] and logistic regression [5]. Results of age group determination gave the accuracy levels of 0.38 for English on Social Media and 0.58 on Blogs, 0.35 for Spanish on Social Media and 0.51 on Blogs. We too investigated the task of author age identification without deception in our works [8, 10] on base of the Gender-Imitation-Crowdsource (‘GI cs’) corpus, collected using the crowdsourcing platform, and a corpus [6] of blogs from LiveJournal. The best result for ‘GI cs’ is F1 = 0.48, it was obtained using Gradient Boosting Classifier with LDR [11], LIWC features [12]. In this case 3 age groups to classify texts: 18–23, 24–29, 30+ have been specified. Regarding blogs, the best result is similar, F1 = 0.49, it was obtained using compositions of selected features: lemma, unigrams, bigrams; POS unigrams and bigrams; QUITA features [13], tenses, aspects. The formulation, closest to the task observed in this paper, was issued in [14], where the task of presence of any deception in men and women texts was investigated, but it was also far from ours.

2. Corpus Description

The corpus for this task, Age-Imitation-Crowdsource corpus (AI cs), has been collected with a crowdsourcing platform. It expands the Gender-Imitation-Crowdsource corpus (GI cs [10]) we used in our previous works, and so keeps the same layout and style. This time, the job the crowdsourcing respondents have been given is the following: Each respondent is asked to fill a questionnaire including their gender, age, ethnicity, native city, native language, profession, educational level, and are they left-handed or right-handed. The respondents are to choose a topic out of a few suggested, and write three texts on it.

- Text A in their natural manner.
- Text B imitating the style of someone younger.
- Text C imitating the style of someone older.

The topics to choose from are the following.

- An attempt to persuade some arbitrary listener to meet the respondent at their place;
- A story about some memorable event/acquisition/rumour or whatever else the imaginary listener is supposed to enjoy;
- A story about oneself or about someone else, aiming to please the listener and win their favour.

The task does not pass checking and is considered improper work if it contains:

- Irrelevant answers to the questionnaire,
- Incoherent jumble of words,
- Chunks of text borrowed from somewhere else,
- Texts not conforming to the above list of topics.

Texts checking is performed firstly by automated search for borrowings (by an anti-plagiarism website), and then by manual review of compliance to the task. The corpus statistic is presented in Table 1. The distribution of examples by gender feature is shown in Figure 1.
3. Model

To obtain a base estimation we use a popular classification method based on support vector machine with linear kernel (Linear Support Vector Classifier – SVC). Due to the fact that SVC solves binary classification tasks, we use the one-vs-rest strategy: a separate classifier is created for each class of the task and is trained to determine its own class from all the remaining data. The SVC configuration is $C = 1$, with hinge squared loss function. For coding text data we use text representation based on TF-IDF vector of symbol n-grams, where $n$ ranges from 3 to 8.

The frequency vector characterizing document $d$ is calculated with:

$$\text{TF}(t, d) = \frac{n_t}{\sum_k (n_k)}$$

$$\text{IDF}(t, D) = \log \left( \frac{|D|}{|\{d_i \in D, t \in d_i\}|} \right)$$

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D),$$

where $t$ is an n-gram of 3-to-8-character-long sequence, $\text{TF}(t, d)$ is the number of occurrences of $t$ in the document $d$. $|D|$ is the total number of documents in the collection. $\sum_k (n_k)$ is the number of characters in document $d$. $|\{d_i \in D, t \in d_i\}|$ is the number of documents in $D$ that contain $t$. 

---

**Figure 1.** The distribution of texts by men/women in the age groups considered.

**Table 1.** The main properties of the AI–cs corpus employed.

| Unique authors | Number of texts | Text length (characters) |
|----------------|-----------------|--------------------------|
|                | Total           | By age group             | Min | Max | Median |
|                | 3302            | 12180 5049 3339 1680 2112 | 400 | 4678 | 492    |

---
Table 2. F1-score (mean ± std) of age determination after training on texts with age imitation, detailed by age groups and the presence of age imitation in the testing texts.

| Age group | Classifying all texts | Classifying age-deceptive texts | Classifying non-deceptive texts |
|-----------|-----------------------|---------------------------------|---------------------------------|
|           |                       | Imitating higher age            | Imitating lower age             |
| 20–30     | 0.48 ± 0.02           | 0.45 ± 0.03                     | 0.48 ± 0.04                     |
| 30–40     | 0.35 ± 0.03           | 0.34 ± 0.04                     | 0.35 ± 0.04                     |
| 40–50     | 0.49 ± 0.03           | 0.50 ± 0.03                     | 0.49 ± 0.04                     |
| Mean      | 0.44                  | 0.43                            | 0.44                            |

4. Experiments

4.1. Determination of age group

In the first step, the original body (AI_{cs}) is filtered keeping only texts of authors whose age ranges from 20 to 50 years. The resulting set is randomly split into 10 folds. In order to avoid unevenly represented classes affecting classification results, all the age groups – 20–30, 30–40, and 40–50 – are balanced by excluding samples from over-represented groups to make all the groups having similar number of samples. Then, each balanced age group is divided into training and testing sets, so that no author has their texts in both sets. Finally, we create a classifier for each balanced group and perform its training and testing on the corresponding sets. For quality assessment we use F1-score of classifying texts with and without age deception, detailed by age groups, in Table 2 – after training on age-deceptive texts, and in Table 3 – after training on non-age-deceptive texts.

To obtain the estimated values for Table 2, the training stage is accomplished using the training part of the balanced dataset AI_{cs} (which is 60% of it, 2982 texts), and for testing we use about 345 texts for each age group (about the 1035 texts, i.e. 20% of the balanced AI_{cs}).

For Table 3, the training set contained a part of the texts without imitation selected from the original balanced dataset AI_{cs} (about the 1725 texts), the test samples also as in the previous case contained examples of both with imitation and without it (about the 345 texts for each age group and about the 115 for each type of imitation (without, older, younger) in each group).

Table 3. F1-score (mean ± std) of age determination after training on texts without age imitation, detailed by age groups and the presence of age imitation in the testing texts.

| Age group | Classifying all texts | Classifying age-deceptive texts | Classifying non-deceptive texts |
|-----------|-----------------------|---------------------------------|---------------------------------|
|           |                       | Imitating higher age            | Imitating lower age             |
| 20–30     | 0.43 ± 0.01           | 0.33 ± 0.04                     | 0.47 ± 0.03                     |
| 30–40     | 0.33 ± 0.03           | 0.31 ± 0.05                     | 0.30 ± 0.04                     |
| 40–50     | 0.49 ± 0.02           | 0.51 ± 0.03                     | 0.47 ± 0.03                     |
| Mean      | 0.42                  | 0.38                            | 0.41                            | 0.45
Table 4. Age imitation type determination F1-scores, detailed by age groups.

| Imitation type | Whole set, mean ± std | Age Groups |
|----------------|------------------------|------------|
|                |                        | 20–30      | 30–40      | 40–50      |
| No imitation   | 0.74 ± 0.01            | 0.77 ± 0.02| 0.76 ± 0.01| 0.73 ± 0.05|
| Older          | 0.81 ± 0.01            | 0.83 ± 0.01| 0.82 ± 0.01| 0.81 ± 0.02|
| Younger        | 0.86 ± 0.01            | 0.86 ± 0.01| 0.88 ± 0.01| 0.86 ± 0.02|
| Mean           | 0.80                   | 0.82       | 0.82       | 0.80       |

4.2. Determination of the imitation type

This task is to classify texts into 3 classes:

- Texts that do not imitate other age.
- Texts that imitate age higher than the author’s.
- Texts that imitate the style of a lower age.

Like the previous experiment, texts from authors 20 to 50 year old have been used. The general experiment outline is the same. The main difference from the age group determination task is that the sets here do not need to be balanced by the number of samples from different classes. That is ensured by the crowdsourcing platform’s job setting itself: each respondent will have written equal number of texts for each of the three imitation types. The classification quality in this experiment is assessed by F1-scores of classifying the three types of imitation (no imitation, pretending to be older, pretending to be younger) averaged over the whole testing set (“Whole set” column in Table 4), and specifically for certain age groups (“Age groups” column in Table 4).

For Table 4, we used 60% (about the 7340 documents) of the original AI_cs dataset for training and 20% (about the 2420 documents) for testing phases.

5. Discussion

An analysis of results of Tables 2 and 3 shows that the two approaches used to determine age deception in author texts demonstrate essentially different accuracies. The accuracy of determining the deception occurrence with identification of the type of imitation is far higher and reaches the F1-score of 0.80. The accuracy of determining the age group is only about 0.44 by F1-score, which is just about 11% higher than the baseline. Notably, the latter result practically coincidences with the result of the case without imitation (F1-score of 0.42); in other words, our age determination model is independent on the age imitation occurrence in texts.

6. Conclusion

As a result of this work, for the first time a Russian-language Age-Imitation-Crowdsourcing corpus (AI_cs) has been created, to solve the task of identifying the author age when the author tries to deceive it. Two approaches to the classifying solution have been presented: the traditional one, to identify the age group of a text author, and the second one, to identify the occurrence of the deception in the text with its type (overstating or lowering). The accuracies obtained demonstrate the second approach to be more efficient, achieving the F1-score of 80%. The future experiments will be focused on analyzing cross-genre text corpora and their effect on deceived age determination accuracy.
7. Acknowledgments
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] Kocheshev P A 2016 Problemy nauki 6 24–25
[2] Morozova A P 2017 Nauchnye trudy Severo-Zapadnogo instituta upravleniya 8 154–163
[3] van de Loo J, De Pan G and Daelemans W 2016 Comput. Linguistics Netherlands 5 46–60
[4] Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M and Stein B 2016 Overview of the 4th author profiling task at pan 2016: cross-genre evaluations Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al. pp 750–784
[5] Modaresi P, Liebeck M and Conrad S 2016 Exploring the effects of cross-genre machine learning for author profiling in pan 2016. CLEF (Working Notes) pp 970–977
[6] Litvinova T Sboev A P P 2018 Artificial Intelligence and Natural Language. 7th International Conference, AINL 2018, St. Petersburg, Russia, October 1719, 2018, Proceedings 167–177
[7] Bougiatiotis K and Krithara A 2016 Author profiling using complementary second order attributes and stylometric features. CLEF (Working Notes) pp 836–845
[8] Sboev A, Moloshnikov I, Gudovskikh D and Rybka R 2018 DEStech Transactions on Computer Science and Engineering
[9] Alvarez-Carmona M A, López-Monroy A P, Montes-y Gómez M, Villasenor-Pineda L and Escalante H J 2015 Working Notes Papers of the CLEF
[10] Sboev A, Moloshnikov I, Gudovskikh D, Selivanov A, Rybka R and Litvinova T 2018 Procedia Computer Science 123 417–423
[11] Rangel F, Franco-Salvador M and Rosso P 2016 A low dimensionality representation for language variety identification International Conference on Intelligent Text Processing and Computational Linguistics (Springer) pp 156–169
[12] Tausczik Y R and Pennebaker J W 2010 Journal of language and social psychology 29 24–54
[13] Kubát M, Matlach V and Čech R 2014 Studies in Quantitative Linguistics 18: QUITA-Quantitative Index Text Analyzer (RAM-Verlag)
[14] Hong J, Mattmann C A and Ramirez P 2017 Ensemble maximum entropy classification and linear regression for author age prediction 2017 IEEE International Conference on Information Reuse and Integration (IRI) (IEEE) pp 509–516