Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 123 (2018) 417–423

Procedia Computer Science

www.elsevier.com/locate/procedia

8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017

Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception

Alexander Sboev[1,2], Ivan Moloshnikov[1], Dmitry Gudovskikh[1], Anton Selivanov[1], Roman Rybka[1], and Tatiana Litvinova[1,3]

[1] National Research Centre 'Kurchatov Institute', Moscow, Russian Federation
[2] MEPhI National Research Nuclear University, Moscow, Russian Federation
[3] Voronezh State Pedagogical University, Voronezh State University, Voronezh, Russian Federation
sag111@mail.ru, ivan-rus@yandex.ru, dvgudovskikh@gmail.com, aaselivanov.10.03@gmail.com, rybkarb@gmail.com, centr_rus_yaz@mail.ru

Abstract

We present the analysis of approaches to solve an author gender identification task for Russian-language texts with gender deception, using different Data-Driven models based on conventional machine learning (Support Vector Classifier, Decision Tree, Gradient Boosting) and neuronet algorithms (convolutional layers, long short-term memory layers, etc.) The source of training and testing data are collections of texts from the Gender Imitation corpus, expanded by crowd-sourcing and supplemented with files of RusProfiling and RusPersonality corpora. The reached accuracy of this task milestone is presented and discussed.

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures

Keywords: natural language processing, gender identification task, machine learning, deep neural networks

1 Introduction

The automatic identification of gender of text author is now in demand in many practical fields, especially for security and forensic tasks. The latter is a typical case when the author of the text to be identified may try to conceal his/her actual gender. So in this situation it is impossible to use the straightforward meaningful features, including such morphological ones as face, genus, etc. The literature devoted to such a formulation of the gender identification problem is practically absent. The task closest to ours is to identify the gender of the author of text with a deception of any type. As shown in [1], even this task is very difficult, and the best accuracy obtained by classifier based on Support Vector Machine(SVM) with the combination of word unigrams and semantic features (numbers of words in a sentence belonging to a specific semantic class) is only about 5% above baseline. The uncertainty of this result is not shown. The

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures
10.1016/j.procs.2018.01.064

dataset was specially collected using Amazon Mechanical Turk [2], consisting of 7168 sentences (3584 truths and 3584 lies) from 512 contributors. Considering texts with gender deception, the only corpus available for our task is the Russian Gender Imitation [3]. Since the size of the corpus had not been enough to form the training and testing datasets of appropriate sizes to reach an agreeable accuracy, it was extended with a corpus 'GI_cs' specially collected by crowdsourcing, and a few corpora without gender deception. These corpora are described in Section 2. In Section 3.1 we describe the sets of features used to encode texts. The learning techniques we used - Convolutional and Long Short-Term Memory networks (LSTM), Support Vector Machine (SVM), Decision Tree and Gradient Boosting Classifiers - are those which proved themselves well on the non-gender-deceptive text author identification task in our previous work. They are viewed in Section 3.2 with details of the use of selected features. Section 4 is devoted to methodology of calculation experiments and methods of results evaluation. In Section 5 we present the obtained results and discuss ways to their improvement. The today's accuracy milestone in the task along with the directions of future works are summarized in Section 6.

2 Datasets

Gender imitation (GI) is a Russian texts corpus manually collected with respondents. Each author chose a topic out like a letter to a friend, a self-description for a dating site, a complaint about the boss or about a tour, and then wrote three texts on the same topic: text A - in the author's natural style, text B as someone of the opposite sex, text C - as someone else of the same sex (125 men, 269 women).

GI_cs: texts collected by crowdsourcing to extend the Gender imitation corpus. It was collected the same way and has the same subsets as GI(1161 man, 2043 woman). **GI_cs_a_b** denotes the part of **GI_cs** without the C texts.

RusPersonality (RusPer) [4]: is a corpus of Russian-language texts labeled with a large amount of metadata on their authors like gender, age, personalities, education level, neuropsychological testing data, etc. The corpus currently contains over 1850 documents, 230 words long in average, from 1145 respondents. This paper uses a part of the corpus, consisting of 1108 texts on two topics: letter to a friend and picture description.

RusProfiling (RusProf) [5]: texts collected from different social media platforms as Twitte(Tw)r, Facebook(FB) and LiveJournal(LJ), along with reviews. In this paper, this data set was divided into subsamples, in order to study the influence of the corpus features on the quality of the model and to evaluate the possibility of genre-independent classification: **Reviews(R)** - manually collected review(641 men, 392 women). **Tw** - twitter messages(998 men, 543 women). 1000 messages were collected for each user, and then merged into one text. **LJ** - for each user his messages were combined into one text(6 men, 5 women). **FB** - messages from walls, combining into one text(136 men, 114 women). **Tw_split** - the same as Tw, all user messages were broken into separate documents of 15 sentences(5062 men, 2450 women). **LJ_split** - the same as LJ, splitting the combined texts into separate documents of 15 sentences(1632 men, 1624 women). **FB_split** - the same as FB, splitting the combined texts into separate documents of 15 sentences(868 men, 749 women).