Solving the problem of determining the author of text data using a combined assessment

Vadim Moshkin[a], Ilya Andreev[a], Nadezhda Yarushkina[a]

[a] *Ulyanovsk State Technical university, Severny Venets str., Ulyanovsk, 432027, Russian Federation*

Abstract

The article describes the main approaches to the automated determination of authorship of texts based on the analysis of copyright styles. The architectures of a convolutional neural network, a multilayer perceptron, and LSTM neural network were proposed to solve this problem. Also, experiments were conducted in which the effectiveness of each of the proposed approaches was evaluated using the example of the task of determining the authorship of English-language poems.

Keywords 1

Machine learning, neural network, binary classification of texts, authorship definition

1. Introduction

Currently, the task of determining the authorship of a text is solved by analyzing the characteristic features of the language and copyright techniques using syntactic, lexical, phraseological, and stylistic analysis of the text. This is a time-consuming process, including an analysis of the author's texts by an expert to identify the features of his style. Hence the need for automation of this process.

In most cases, the following approaches are most often used to determine authorship of texts: mathematical statistics and probability theory, neural networks, cluster analysis, pattern recognition theories, etc.

The purpose of this work is to assess the applicability of machine learning methods to solve the problem of determining the authorship of the text and to compare models of neural networks in solving this problem.

Formal methods for determining authorship of texts are divided into two large groups: statistical methods and machine learning. The application of machine learning methods to determine the author of the text in the framework of the current study was investigated [1]. Machine learning algorithms include several categories:

- genetic algorithms
- neural networks,
- Bayesian classifier,
- decision trees, etc. [2]

The application of statistical methods for analyzing the authorship of the text was carried out in [3-5]. The authors of the study proposed a method based on taking into account statistics on the use of syntagmatic chains in the text. The accuracy of this method was less than 65%.

The "Linguoanalyzer" system [5] determines the author of the text by applying compression algorithms and Markov chains. This software system could determine the author with an accuracy of 70 to 89 percent depending on the length of the text.

Russian Advances in Fuzzy Systems and Soft Computing: selected contributions to the 8-th International Conference on Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT-2020), June 29 – July 1, 2020, Smolensk, Russia
EMAIL: v.moshkin@mail.com; ares-ilya@ya.ru; jng@ulstu.ru.
ORCID: 0000-0002-9258-4909; 0000-0002-6217-9566; 0000-0002-5718-8732.

The SMALT system [6-7] is based on dependency trees and types of relationships and statistical methods for analyzing a literary text. However, the authors were unable to achieve a sufficiently accurate determination of the author of the text using both 16 and 156 attributes.

The basis of the software system "Autologist" is the construction of a binary classifier. In this method, all texts from the training and test sets are expanded into a very large vector indexed by words. In this case, the texts are two sets of points from the training set in multidimensional space - some of them belong to the author, while others do not belong. This software system has achieved an average accuracy of 88% [8-9].

Despite the results obtained by researchers in the field of detecting authorship of texts of various styles, the use of machine learning methods to solve such problems will allow one to obtain better results, taking into account their effectiveness in working with natural language.

2. Models and implementation

The dataset for research is a collection of poems in English. This collection is publicly available in the Kaggle system [10].

The data set consists of 15638 poems by 3310 authors. Poems were chosen for the study, as the author's style is more expressed in them. English texts were selected.

The development was carried out in the online service Google Colab in Python [11]. The following libraries were used:

- Keras library [12] for building models of neural networks;
- pandas library for working with selections
- spaCy library for word processing
- matplotlib library for graphing.

We assume that there is a target author of the training and its reference texts (poems) and many other authors with their texts (poems). From here the binary classification problem is formed.

For machine learning, it is necessary to divide the data set into training and test sets. In this implementation, the division was made 80% to 20% in the training and test sets, respectively.

One of the most important components of the author's style is the syntactic construction of sentences in its texts. It is on this idea that this study is based.

The spaCy library is used for parsing text. spaCy encodes all strings into hash values to reduce memory usage and increase efficiency.

The schematic parsing of the proposal by the spaCy library is shown in Figure 1.

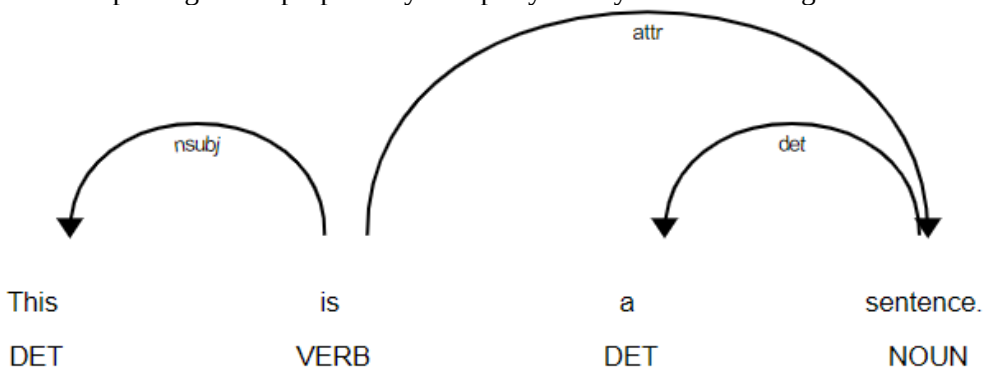

Figure 1: Parsing schema

A classifier was developed based on the k-nearest neighbors' algorithm [13]. Poems are presented in vector form. The algorithm for vectorizing poems includes the following actions:

- each word in the poem is represented in vector form using the Word2vec library (the pretrained GloVe model was used) [14, 15];
- vectors of all words are added;
- the resulting vector is normalized according to the length of the poem in words;
- an average vector of all words in the poem is obtained (presumably it expresses the theme of the poem).

The author's poems form his vector space. In the case of a binary classification, two sets of vectors will be obtained corresponding to the poems of the two authors, which are included in the training dataset. The belonging of poems from the test dataset is determined by the k-nearest neighbors method.

For experiments, the following neural network models were selected:

- multilayer perceptron;
- recurrent neural network LSTM;
- convolutional neural network (CNN).

The multilayer perceptron model consists of the following layers:

1. Layer Embedding;
2. Layer Flatten;
3. A fully connected Dense layer with 256 neurons and a relu activation function.
4. A fully connected Dense layer with 64 neurons and a relu activation function.
5. A fully connected Dense layer with one neuron and sigmoid activation function.

The neural network LSTM model consists of the following layers:

1. Layer Embedding;
2. LSTM layer with 128 neurons;
3. LSTM layer with 64 neurons;
4. A fully connected Dense layer with one neuron and sigmoid activation function.

The convolutional neural network (CNN) model consists of the following layers:

1. Layer Embedding;
2. Layer Conv1D with 256 neurons and relu activation function;
3. Layer Conv1D with 64 neurons and relu activation function;
4. Layer GlobalMaxPooling1D;
5. A fully connected Dense layer with 30 neurons and a relu activation function.
6. A fully connected Dense layer with one neuron and sigmoid activation function.

All models use the “adam” optimizer, the “binary_crossentropy” loss function, and “the accuracy” metric. In all models, the embedding layer is a fully connected layer consisting of 16 ordinary neurons. The last layer of each model has a sigmoid activation function, which is one of the best in solving the binary classification problem.

3. The results of the experiments

The 10 authors with the highest number of poems in the dataset were selected to conduct the experiments.

45 experiments were carried out. The experimental results are shown in Figure 2.

	William Shakespeare	Anonymous	Alfred, Lord Tennyson	Rae Armantrout	William Wordsworth	Emily Dickinson	William Butler Yeats	John Ashbery	Yusef Komunyakaa	Percy sshe Shelley
William Shakespeare	-	0,64	0,8	0,91	0,73	0,9	0,92	0,88	0,68	0,77
Anonymous	0,64	-	0,8	0,75	0,7	0,8	0,71	0,75	0,64	0,64
Alfred, Lord Tennyson	0,8	0,8	-	0,88	0,57	0,93	0,38	0,46	0,55	0,5
Rae Armantrout	0,91	0,75	0,88	-	0,93	0,77	0,88	0,83	0,68	0,73
William Wordsworth	0,73	0,7	0,57	0,93	-	1	0,58	0,79	0,59	0,64
Emily Dickinson	0,9	0,8	0,93	0,77	1	-	0,88	0,83	0,73	0,82
William Butler Yeats	0,92	0,71	0,38	0,88	0,58	0,88	-	0,71	0,59	0,64
John Ashbery	0,88	0,75	0,46	0,83	0,79	0,83	0,71	-	0,77	0,82
Yusef Komunyakaa	0,68	0,64	0,55	0,68	0,59	0,73	0,59	0,77	-	0,73
Percy sshe Shelley	0,77	0,64	0,5	0,73	0,64	0,82	0,64	0,82	0,73	-

Figure 2: Assessment of the accuracy of the developed classifier

The average classification accuracy was 73.8%.

The idea of the work is that we can get a higher classification accuracy by combining the estimates of the neural network classifier described in [16] and the classifier based on the k-nearest neighbors algorithm.

Thus, the assessment will take into account both the author's style of the poem and its theme.

The convolutional neural network model was chosen as a neural network classifier, since it showed the best results in the experiments of [16].

The combination of scores is done as follows:

- let $x_1 \in$ R[0:1] is the neural network classifier estimate;
- $x_2 \in$ R[0:1] is an estimate of the classifier based on the k-nearest neighbors algorithm;
- Then $x_3 = x_1^2 + x_2^2$ is the combined estimate;
- Combining by squaring numbers allows you to compensate for discrepancies in estimates.

For the experiment, poems by two authors (William Shakespeare, Alfred Lord Tennyson) were taken from the dataset since they have the largest number of poems in the dataset.

The training and test set were divided at a ratio of 75% / 25%.

3.1 Multilayer perceptron

The multilayer perceptron has been trained for 8 eras. Learning speed 2 seconds per era. The training schedule is shown in Figure 3.

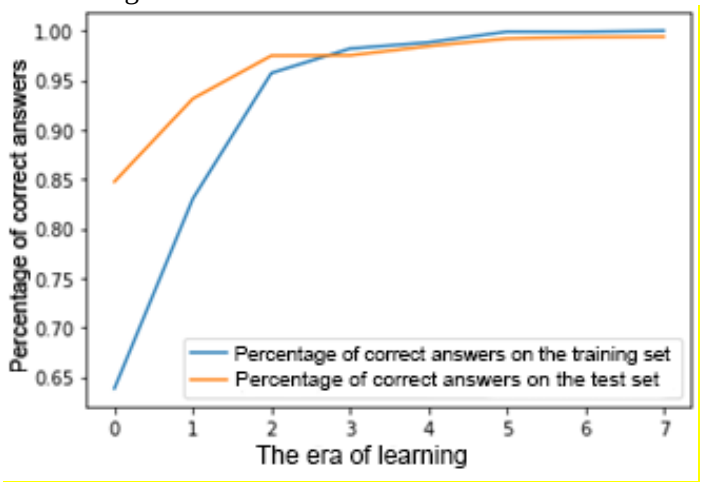

Figure 3: Perceptron training graph

This graph shows a correlation between accuracy in the training and test sets. This suggests that perhaps the model has extracted the correct features. Also on this graph, the accuracy of the training set is constantly growing, which means the model is successfully trained.

Figure 4 shows the result of evaluating the accuracy of a classifier based on a convolutional neural network. The accuracy was 77.5%. You can also see that this classifier is more inclined to assign poems to the second author (25 out of 40).

		Forecast		
		1	**2**	
Actual	**1**	**13** 32,50%	**7** 17,50%	65% 35%
	2	**2** 5%	**18** 45%	10% 90%
		86,67% 13,33%	72% 28%	77,50% 22,50%

Figure 4: Classification accuracy (convolutional neural network)

3.2 LSTM neural network

LSTM neural network trained 8 eras. The training time averaged 16 seconds per era. The training graph is shown in Figure 5. In this graph, there is a correlation between accuracy in the training and test sets, but it is much weaker, which means that the model found incorrect signs. Accuracy in the training set does not tend to grow continuously, therefore, the model does not learn. Thus, the model is incorrectly constructed, the hyperparameters of the model are incorrect, or the LSTM network is poorly suited for this form of data representation.

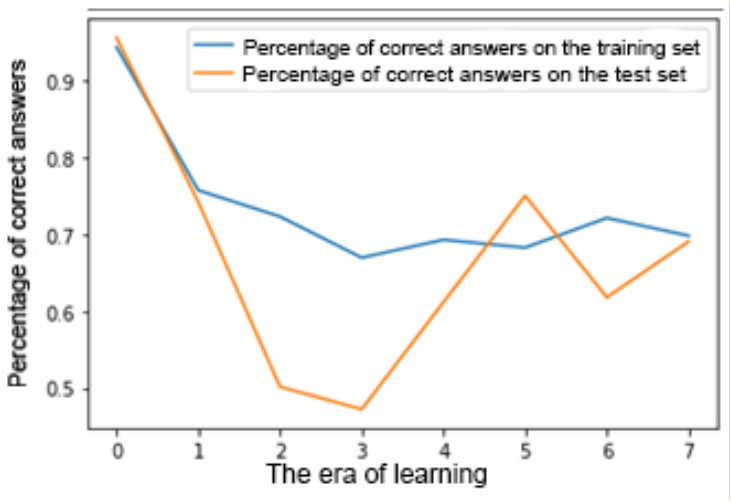

Figure 5: LSTM network training schedule

Figure 6 shows the result of evaluating the accuracy of the classifier based on the k-nearest neighbors algorithm. The accuracy was also 77.5%.

		Forecast		
		1	**2**	
Actual	**1**	**15** 37,50%	**5** 12,50%	75% 25%
	2	**4** 10%	**16** 40%	20% 80%
		78,95% 21,05%	76,19% 23,81%	77,50% 22,50%

Figure 6: Classification accuracy (developed algorithm)

3.3 Convolutional neural network

The convolutional neural network (CNN) model has been trained for 16 eras. Learning speed averaged 1 second per era. The training graph is shown in Figure 7. This graph shows a correlation between the accuracy of the training and test sets. This suggests that perhaps the model has extracted the correct features. Also on this graph, the accuracy of the training set is constantly growing, which means the model is successfully trained.

The convolutional neural network (CNN) training schedule is shown in Figure 7.

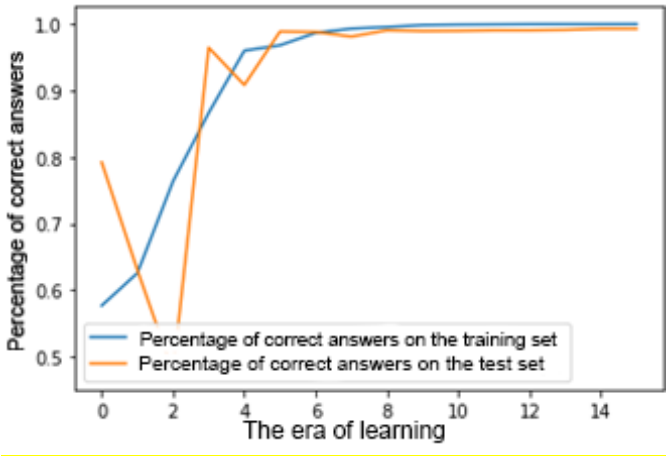

Figure 7: Schedule for convolutional network training (CNN)

Although classifiers have the same precision, they make different mistakes. Therefore, they can compensate for each other's inaccuracies. The result of assessing the accuracy of the combined classifier is presented in Figure 8.

		Forecast		
		1	**2**	
Actual	**1**	**16** 40%	**4** 10%	80% 20%
	2	**3** 7,50%	**17** 42,50%	15% 85%
		84,21% 15,79%	80,95% 19,05%	82,50% 17,50%

Figure 8: Accuracy of classification (combined classifier)

It was possible to improve the classification accuracy to 82.5% as a result of combining the estimates. The increase was 5% in absolute terms and 6.45% in relative terms.

4. Conclusion

As a result of the project, a text presentation method was proposed for computing neural networks with partial preservation of the author's style. In addition, models of neural networks were implemented such as a multilayer perceptron, LSTM, a convolutional neural network. An assessment was made of the quality of work of these models in the task of determining the authorship of poems and a comparison was made based on the results of which a rating of models can be made:

1. convolutional neural network;
2. multilayer perceptron;
3. LSTM neural network.

You can make an assessment and comparison based on the results of quality assessment of trained models. The multilayer perceptron showed good results, but its accuracy in determining the poems belonging to the author is small. The LSTM neural network better defines poems belonging to the author, however, the accuracy of determining non-author poems is too low. In general, the LSTM model of the neural network has been trained much longer and worse. The convolutional neural network (CNN) turned out to be the best, albeit slightly inferior in terms of final accuracy to the multilayer perceptron.

An algorithm for combining scores has been implemented.

The efficiency of the work of both a classifier based on the k-nearest neighbors algorithm and a combined classifier was experimentally proved.

We managed to improve the classification accuracy by 5% in absolute terms and by 6.45% in relative terms.

5. Acknowledgement

This study was supported by the Russian Foundation for Basic Research (Grants No. 18-47-732007, 18-47-730035 and 19-07-00999).

6. References

[1] A. V. Mukha, V. L. Rozaliev, Yu. A. Orlova, A. V. Zaboleeva-Zotova, An automated approach to determining the authorship of a text, Bulletin of the Volgograd State Technical University, 14 (2013) 51-54.

[2] P. Shrestha, S. Sierra, F. González, M. Montes, P. Rosso, T. Solorio, Convolutional Neural Networks for Authorship Attribution of Short Texts, 2017. doi:669-674. 10.18653/v1/E17-2106.

[3] Zh. Ge, Yu. Sun, M. Smith, Authorship Attribution Using a Neural Network Language Model, AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016, 4212-4213.

[4] O. V. Kukushkina, A. A. Polikarpov, D. V. Khmelev, Definition of authorship of the text using alphabetic and grammatical information, Problems of information transfer. 37, 2 (2001) 96-109.

[5] D. V. Khmelev, Recognition of the author of the text using chains of A. A. Markov, Tomsk State University Journal. Moscow State University. Ser. 9: Philology, 2 (2000) 115-126.

[6] A. A. Rogov, Yu. V. Sidorov, A. V. Korol, Automated system for processing and analysis of literary texts SMALT, Transactions and Materials of the II Intern. Congress of Russian Language Researchers “Russian Language: Historical Fates and the Present”. M: Moscow State University, (2004) 485-486.

[7] A. A. Rogov, G. B. Gurin, A. A. Kotov, Yu. V. Sidorov, T. G. Surovtsova, Software package SMALT, Digital Libraries: Advanced Methods and Technologies, Electronic Collections: Proceedings of X All-Russia. scientific conf. ”RCDL'2008.” Dubna, 2008, 155-160.

[8] A. S. Romanov, Technique of identification of the author of the text based on the apparatus of support vectors, Reports of Tomsk State University of Control Systems and Radioelectronics , 1-2 (19) (2009).

[9] A. S. Romanov, Methods and software for identifying the author of an unknown text: Abstract. Tomsk 26 (2010).

[10] Kaggle library URL: https://www.kaggle.com/johnhallman/complete-poetryfoundationorg-dataset.

[11] Python 3.8.1 documentation URL: https://docs.python.org/3.

[12] Keras Documentation URL: https://keras.io.

[13] T. Daniel, Larose. Discovering Knowledge in Data: An Introduction to Data Mining. Wiley-Interscience, USA, 2004.

[14] Algorithm Word2Vec URL: https://neurohive.io/ru.

[15] GloVe: Global Vectors for Word Representation URL: https://nlp.stanford.edu/projects/glove.

[16] V. S. Moshkin, I. A. Andreev, I. M. Shigabutdinov, Machine learning algorithms for attribution of text fragments, Proceedings of the VIII International Scientific and Practical Conference "Fuzzy Systems, Soft Computing and Intelligent Technologies" FSSCIT-2020, Smolensk: Universum 1 (2020) 173-181.