Building Artificial Neural Networks for NLP Analysis and Classification of Target Content

Aleksey Rogachev[1,*], Elena Melikhova[2], Gennady Atamanov[3]

[1] *Economic analysis laboratory, All-Russian research Institute of irrigated agriculture, Department of mathematical modeling, Volgograd state agrarian university, Volgograd, Russia*
[2] *Department of mathematical modeling, Volgograd state agrarian university, Volgograd, Russia*
[3] *Department of information security, Volgograd state university, Volgograd, Russia*
[*] *Corresponding author. Email: rafr @mail.ru*

ABSTRACT
The problems of analyzing texts in natural language (NLP) using artificial intelligence (AI) methods are caused by the semantic and lexicological diversity of texts. This circumstance causes the appearance of various machine learning (ML) metrics for neural network analysis. The problem of AI analysis is further complicated by the fact that the content under study often contains "information garbage", which is information noise, complicating the solution of a well-known problem of text classification. The lexicological diversity of Internet content requires improving the methods of neural network NLP analysis. The purpose of the research is to identify and solve problems that arise when analyzing information texts using artificial neural networks (ins), using the example of socio-political content. Well-known NLP technologies include substantiation of the structure and formation of a subject-oriented database of text data bodies, construction of dictionaries based on frequency analysis, and digital vectorization of texts. To identify the latent semantic content, the expediency of using a dense vector representation of terms in a multidimensional space (the embedding model) is justified. In order to justify the choice of basic architectures developed by ins to account for sequences and combinations of analyzed terms, modifications of convolutional (Conv1D) recurrent (CNN, LSTM, etc.) layers were selected that allow storing token sequences. Since such powerful layers contribute to the appearance of undesirable re-training of ins, effective means of regularization are necessary, for example, dropout layers. The authors substantiate a modified NLP approach to identifying sociocultural and cyber threats contained in the information content of Internet resources. Based on the frequency analysis of the target Internet content, dictionaries of terms used for multi-class text analysis are pre-formed, as well as their markup. To justify and study the ins architecture and hyperparameters focused on the content of the analyzed subject area, the ins family was built in Python 3 using specialized libraries - Keras, ScikitLearn, and others. The ANN architecture included combinations of fully connected, convolutional, and/or recurrent layers. When training ins in the Google Colaboratory environment, high-performance GPUs were used. Recommendations are given for selecting ins hyperparameters that are invariant for various architectures of hidden layers of hybrid ins that are focused on solving the problem of multiclass NLP analysis. The degree of correct text recognition in the test sample exceeded 80%. Recommendations for improving it are given.

Keywords: *artificial neural network, hybrid architecture, multiclass text analysis, cyber threat*

1. INTRODUCTION

The problems of analyzing texts in natural language (NLP) using artificial intelligence (AI) methods are caused by a significant semantic and lexicological diversity of texts, which causes the appearance of numerous machine learning (ML) and neural network analysis metrics [1, 2]. The problem of AI analysis is complicated by the fact that in most cases the content under study contains "information garbage" and "information trash", which are specific information noise, complicating the solution of a well-known problem of binary or multiple classification of texts. Since the main NLP methods [3, 4], primarily tokenization and vector representation of texts, are focused primarily on English, additional difficulties arise that are typical for other languages-Chinese, Russian, Arabic, etc. This should be taken into account when forming the structure and content of thematic text corpora for training and testing artificial neural networks (ANN). In particular, the detection of undesirable content that includes socio-political and cyber threats requires the inclusion of specific terms, phrases and phrases, as well as typical "information garbage" and phishing texts [5]. In addition, significant differences in thesauri, structure, language structures, and presentation styles of thematic material that are characteristic of different subject areas are reflected in the architecture, set of hyperparameters, and teaching methods of ANN [6-8]. These features require improvement of methods and techniques for neural network analysis of NLP, focused on various subject areas.

2. METHODOLOGY

Multi-class markup of texts containing terms specific to socio-cultural and cyber threats, as well as combinations of words and expressions related to "information garbage"was used in the formation of text corpora for ins training. To identify semantic content and latent threats, a dense vector representation of tokens of analyzed texts in a multidimensional space (embedding) is justified, the dimension of which is determined experimentally taking into account the specifics of text corpora and the capabilities of the computing platform. The texts included keywords containing socio-cultural and cybernetic threats, as well as signs of ideological extremism. In addition, the content included metered "information garbage".

To substantiate and study the ins architecture and hyperparameters focused on the analysis of socio-political content, the ins family was built in Python 3 using specialized libraries-Keras, ScikitLearn, and others.

The ins family was built and trained using high-performance GPUs provided by the cloud environment, which is a product of Google Research [https://research.google.com/colaboratory/faq.html#resource-limits]. Google Colaboratory allows you to write and execute interpreted Python code through the browser, which is very convenient for data analysis and ANALYSIS. The types of GPUs available in Colaboratory that significantly speed up ins training can vary in performance and are selected from the following list of Nvidia - K80s, T4s, P4s and P100s..

3. RESULTS AND DISCUSSION

3.1. Choosing the ANN family architecture

The purpose of the research is to solve the problems of NLP using neural network AI methods that arise when analyzing thematic content, using the example of identifying undesirable content of Internet resources of a socio-political profile. The results of an analytical review of the current state of neural network analysis, in relation to the problem of identifying thematic content on the example of identifying undesirable content and latent threats, allowed us to justify the main approaches to mathematical modeling [8-10], building ins and their training [11, 12].

Well-known NLP technologies using artificial neural networks (ANN) primarily include substantiation of the structure and construction of a subject-oriented database of text data bodies, frequency analysis and construction of subject-oriented dictionaries, as well as tokenization and digital vectorization of texts.

The authors substantiate a modified NLP approach to identifying sociocultural and cyber threats, including latent threats, contained in the information content of Internet resources. Based on the frequency analysis of targeted Internet content, dictionaries of terms used for multi-class text analysis are pre-formed, as well as their markup by belonging to the corresponding classes.

The software implementation of the ins family was carried out using the built-in tools and libraries of Google Colaboratory. Technically, Google Colaboratory is a Jupyter hosted service that does not require configuration for use, but provides high-performance access to computing resources.

The developed family of neural networks contained both invariant modules that are identical for all analyzed architectures, and variable blocks with ins layers of different architectures. Text preprocessing blocks were invariant : loading data and dividing it into words, creating dictionaries and converting data into indexes, forming training and test samples [5].

In order to justify the choice of architectures for variable blocks of the developed ins that are directly focused on the analysis of terms, phrases and sequences of analyzed texts, modifications of recurrent layers (CNN, LSTM, GRU) were selected that allow to remember the analyzed token sequences, as well as one-dimensional convolutional layers (Conv1D) with adjustable sizes of Windows and convolution cores. It is known that such powerful layers can contribute to the appearance of undesirable re-training of ins, so effective means of regularization were provided, for example, dropout layers.

The architecture of variable ins blocks also included a combination of fully connected, convolutional, and/or recurrent layers that provide generalization and preliminary identification of various features of the analyzed texts.

Invariant blocks were also modules that implement text classification. They included layers of data regularization and normalization, as well as a fully connected layer whose number of neurons corresponds to the number of recognized classes.

Activation functions of the recognition layer were accepted by "softmax"(or "sigmoid" depending on the network architecture and the task being solved. Recommendations on the choice of hyperparameters "loss" of ins, as well as the number and activation function of neurons of the last recognizing layer, which are invariant in solving the problem of multiclass NLP analysis for various architectures of hidden layers of hybrid ANN, are given.

In the course of the study, Callback layers were used to ensure that the network learning process is interrupted when the quality of learning deteriorates, while maintaining the optimal values of its weights.

3.2. Study of the quality of training built ANN

The following are indicators of the quality of ANN training, built in the course of the study.

Despite the relative simplicity and experimentally obtained relatively high speed of learning text models - " bag of words "("bag of words"), the need to identify latent threats led to further research with more complex models of" embedding", implementing the representation of tokens in an n-dimensional dense vector space [1].

The following is an ANN architecture for 6 text classes based on fully connected layers with the "ReLu" activation function and a 16-dimensional "embedding" word representation model with the "SpatialDropout1D(0.2)"layer regularization (Fig. 1).

```
modelEm = Sequential()

modelEm.add(Embedding(maxWrdsCnt, 16, inp_lngth = xLn))

modelEm.add(SpatialDropout1D(0.2))

modelEm.add(Flatten())

modelEm.add(BatchNormalization())

modelEm.add(Dense(200, activation='relu'))

modelEm.add(Dropout(0.2))

modelEm.add(BatchNormalization())

modelEm.add(Dense(6, activation='sigmoid'))
```

Figure 1 Architecture of a fully connected neural network for six recognized text classes

The results of ANN training for this architecture are shown in fig. 2.

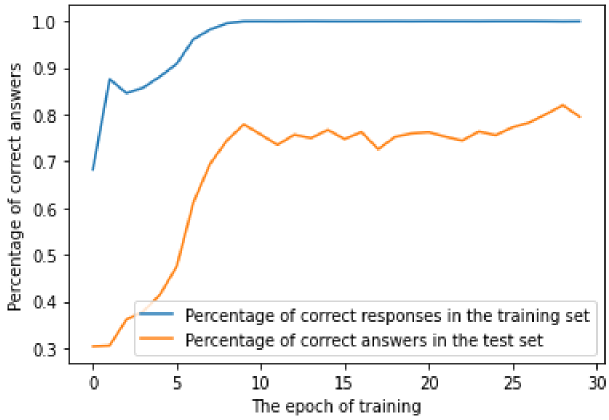

Figure 2 ANN learning results based on fully connected layers

Analysis of the diagram in Fig. 2 shows a high learning rate (10 ... 30 epochs) and an acceptable classification accuracy of about 80% in the test sample. This allows us to recommend such an architecture with the reduced values of hyperparameters for conducting search numerical experiments.

Numerical experiments with the LSTM architecture with similar parameters showed an excessively high learning time, which was the basis for the study of the architecture based on the one-dimensional convolution "Conv1D(20.5,activation='relu')" in combination with "MaxPooling1D(2)" (Fig. 3).

```
modelECn = Sequential()

modelECn.add(Embedding(maxWrdsCunt, 10, inpt_lngth=xLn))

modelECn.add(SpatialDropout1D(0.2))

modelECn.add(BatchNormalization())

modelECn.add(Conv1D(20,5,activation='relu'))

modelECn.add(MaxPooling1D(2))

modelECn.add(Dropout(0.2))

modelECn.add(BatchNormalization())

modelECn.add(Flatten())

modelECn.add(Dense(6,activation='softmax'))
```

Figure 3 The architecture of the neural network on the basis of one-dimensional convolutional hidden layer

The results of ANN training for this architecture based on one-dimensional convolution are shown in fig. 4.

Analysis of the diagram in Fig. 4 shows a slightly lower learning speed, compared to a network with only a fully connected architecture. The accuracy of classification in the test sample is also about 80%. At the same time, it becomes possible to select an additional hyperparameter for the size of the convolution window for texts of various contents, taking into account the presence and nature of "information garbage". This allows us to recommend such an architecture with the reduced values of hyperparameters for further numerical experiments.

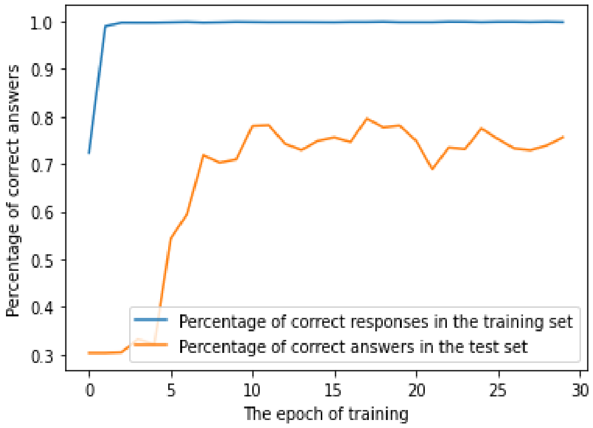

Figure 4 Results of ANN training based on the one-dimensional convolution layer

Thus, the degree of correct text recognition during experiments exceeded 80% in the test sample. Since this accuracy is still insufficient, the use of automatic optimization tools for hyperparameter values, for example, based on genetic algorithms or the KerasTuner tool, can be recommended to improve it in the course of continuing research.

4. CONCLUSION

As a result of the research, the authors justified a modified NLP approach to identifying sociocultural and cyber threats, including latent threats, contained in the information content of Internet resources.

Recommendations are given for choosing the architecture and hyperparameters that are invariant for hidden ins layers when solving the problem of multiclass NLP analysis.

A set of neural network assessments of typical RSS feeds based on the criteria of detected signs of explicit or latent socio-cultural and cyber threats and ideological extremism was obtained.

ACKNOWLEDGMENT

The reported study was funded by RFBR and EISR according to the research project № 20-011-31648.

REFERENCES

[1] B. Bengforth, R. Bilbro, T. Ojeda, Applied text data analysis in Python. Machine learning and the creation of applications for natural language processing, Piter, 2019.

[2] O.S. Smirnova, V.V. Shishkov, Choosing the topology of neural networks and their application for classification of short texts, International journal of open information technologies 4(8) (2016).

[3] D. Gordeev, Detecting state of aggression in sentences using CNN, Lecture Notes in Computer Science 9811 (2016) 240-245.

[4] X. Zhang, Y.Le Cun, Text understanding from scratch, Computer Science Department, arXiv:1509.01626, 2016.

[5] G.A. Atamanov, Azbuka bezopasnosti. methodology of information resource protection, Information Protection, Insider trading 2(62) (2015) 8-13.

[6] T. Mikolov, I. Sutskever, K. Chen, Distributed Representations of Words and Phrases and their

Compositionality, arXiv:1310.4546 [cs.CL], Avaliable at: https://arxiv.org/abs/1310.4546v1, 2013.

[7] P.Yu. Polyakov, M.V. Kalinina, V.V. Pleshko, Automatic Object-oriented Sentiment Analysis by Means of Semantic Templates and Sentiment Lexicon Dictionaries, in: Proceedings of the 21st International Conference on Computational Linguistics Dialog, 2015, vol. 2, pp. 44-52.

[8] A.S. Surkova, I.D. Chernobaev, Comparison of neural network architectures in the task of automatic text classification, Modern informatization problems in the technological and telecommunication systems analysis and synthesis, Proceedings of the XXIV-th International Open Science Conference, 2019, pp. 377-382.

[9] A.F. Rogachev, Computer Modeling of the Development of Russian Small Towns on the Basis of Cognitive Maps, in: Russia and the European Union: Development and perspectives, Contributions to Economics, 2017, pp. 113-118. DOI: 10.1007/978-3-319-55257-6_16

[10] A. Rogachev, E. Melikhova, Monitoring of agricultural land productivity using unmanned aerial vehicles and artificial neural networks, in: IOP Conference Series: Earth and Environmental Science, 12th International Scientific Conference on Agricultural Machinery Industry, INTERAGROMASH, 2019, p. 012175. DOI: 10.1088/1755-1315/403/1/012175

[11] D.S. Tarasov, Deep Recurrent Neural Networks for Multiple Language Aspect-based Sentiment Analysis of User Reviews, in: Proceedings of the 21st International Conference on Computational Linguistics Dialog 2 (2015) 53-64.

[12] K.E. Tokarev, Yu.A. Orlova, A.F. Rogachev, A.N. Chernyavsky, Yu.M. Tokareva, The intelligent analysis system and remote sensing images segmentation engineering by using methods of advanced machine learning and neural network modeling, in: IOP Conference Series: Materials Science and Engineering, Krasnoyarsk Science and Technology City Hall of the Russian Union of Scientific and Engineering Associations, Krasnoyarsk, 2020, p. 12124. DOI:10.1088/1757-899X/734/1/012124