Problems of Analyzing Socio-Political Content of Internet Resources Based on Neural Network Technologies

A F Rogachev$^{1, 2}$

$^1$Department of Mathematical Modeling and Informatics, Volgograd State Agrarian University, 26 Universitetskii Prospekt, Volgograd, Russian Federation, 400002
$^2$Laboratory of Economic Analysis, All-Russian Research Institute of Irrigated Agriculture, 9 Timiryazeva, Volgograd, Russian Federation, 400002

E-mail: Rafr@mail.ru

Abstract. Ensuring information security requires identifying undesirable information content of Internet resources. Semantic and lexicological diversity of Internet content requires improvement of methods of neural network analysis of natural language texts (NLP). The problem is complicated by the presence of "information garbage", which is a specific information noise that complicates the task of classifying texts. Well-known NLP technologies using artificial neural networks (ANN) include substantiation of the structure and construction of a subject-oriented database of text data bodies, frequency analysis and construction of dictionaries. To identify semantic content and latent threats, a dense vector representation of the analyzed texts in a multidimensional space (embedding) is justified. The authors substantiate a modified NLP approach to identifying sociocultural and cyber threats, contained in the information content of Internet resources. To justify and research the ANN architecture and hyperparameters focused on the socio-political content under study, the ANN family was built in Python 3. The ANN architecture included combinations of fully connected, convolutional, and/or recurrent layers. The number of neurons of the recognizing fully connected layer with the "softmax" activation function (or sigmoid in multiple classification) was taken by the number of classes marked in the text corpus.

1. Introduction
Opposition to ideological extremism requires identifying the socio-cultural and cyber threats that may be contained in the online content. The semantic and lexicological diversity of the exponentially growing content of electronic sources makes it necessary to create new methods for its computer analysis in order to identify undesirable content using artificial intelligence (AI) methods.

In the field of political technologies, the use of artificial neural network (ANN) models is promising. Their use includes analyzing the results and generalizing opinion polls, predicting the dynamics of ratings, identifying significant factors, objective clustering of the electorate, and visualizing the social dynamics of the population. The use of neural network technologies provides justification of the structure and construction of a subject-oriented database of problem-oriented text data, preliminary frequency preprocessing of text arrays, and vector representation of generated dictionaries of terms. It also requires justification and experimental research of the architecture, including hybrid ones, as well as ANN hyperparameters [1].
One of the directions of the strategy of scientific and technological development of the Russian Federation is "Countering man-made, socio-cultural threats, and ideological extremism, as well as cyber threats and other sources of danger to society, the economy and the state". However, according to a number of scientists, some of its methodological provisions are controversial [2]. The task of categorizing content, including identifying its belonging to the target category in an exponentially increasing amount of information, is relevant for various socio-political spheres [3]. To substantiate the approaches and methods of studying the content of Internet content, publications of domestic and foreign scientists-political scientists, sociologists, philologists [4], who have created a number of ML-methods for analyzing the emotional color and tone of texts in the media on the Internet, including cognitive and interpretive decoding [5][6].

The fundamental principles of technologies for creating and teaching ANN, including the formation of case papers, preprocessing of source data, architecture and macro parameters of Ann, are systematically described in [7][8], where methods and technologies for analyzing text information are considered. Researchers and experts in the field of ANN note the special capabilities of AI methods and neural network approaches Natural Language Processing (NLP) to identify the desired content [9][10][11][12].

One of the specific tasks of identifying undesirable information containing socio-cultural and cyber threats is to search for sets of terms using the "bag of words" method based on thematic thesauri. To find a contextual sequence of indicator terms, a preliminary vector representation of terms in a multidimensional space is promising to improve the efficiency of their computer analysis. Embeddings are a variant of this approach. The dimension of the space of such a vector representation, its optimal volume, and the method of obtaining it, taking into account the selected ins architecture, are a separate scientific and methodological problem [13][14][15][16]. Other specific areas of NLP technology improvement are also known [17][18][19][20][21][22]. A separate problem is the classification of texts in Russian [18][21]. Therefore, taking into account the specifics of the socio-political orientation of the discourse of the problem to be solved, it is necessary to conduct research in the field of substantiation and construction of corpora of typical Internet content texts, methods of their frequency preprocessing and analysis using ANN based on embeddings built on the basis of selected dictionaries of terms.

2. Methods and Materials
The purpose of this study is to develop a methodology for building ANN for analyzing the socio-political content of Internet resources using hybrid neural network technologies. ANN ensembles were written in the Google Colaboratory environment in Python 3 using specialized libraries - Keras, ScikitLearn, NLTK, Gensim, spaCy, NetworkX [9]. Research on the functioning of neural networks was carried out using specially formed corpora of thematic texts in the field of content containing cyber threats. Thematic cases, except for those downloaded from the Internet, were specially prepared for emulation of the subject area.

Basic NLP technologies for artificial neural networks include the following key procedures: (1) structural justification and construction of a domain-specific structure of text data, (2) frequency analysis, (3) construction of domain-specific dictionaries, (4) tokenization, and (5) multidimensional text vectorization.

2.1. Language Model
The construction of a dictionary based on a statistical approach that normalizes the frequency of lexemes in the document was carried out taking into account the content detected in the "term frequency–inverse document frequency" (TF–IDF) coordinates in the rest of the text corpus. The TF-IDF encoding method determines the relevance of document tokens by the scaled frequency of their appearance, normalized by the inverse scaled frequency of the appearance of the token in the entire corpus.

To program and get embeddings in Python, the "word2vec" algorithm was used using the Google Colaboratory cloud environment. A vector representation of words in a multidimensional space, in which similar words appear next to each other in a vector space with an a priori dimension of the order of 10...150, which was selected experimentally.

Construction, configuration and selection of ANN hyperparameters for natural language processing, including using the Russian-language "word2vec", was performed on the basis of methods and techniques.
of system analysis [14][15]. The formation of a subject-oriented database of texts was carried out by parsing Internet resources. RSS was also used, which is a type of web channels that publish news in a format that is convenient for computer download. The markup of the corpus base for training the studied ins architectures was carried out using the "one-hot-encoding" (OHE) format).

2.2. Architectural Solutions for Artificial Neural Networks

Combined architectures based on layers of various types (recursive, LSTM with long-term short-term memory, convolutional, etc.) were used to justify various ANN architectures, including hybrid ones, and to build neural networks focused on text processing [14]. The concatenation of these layers was implemented in Python using the functional programming method.

Justification and choice of architecture and hyperparameters for constructing an ANN that includes recursive (RN/RM/LSTM), convolutional (Conv1D), and fully connected (Dense) layers of neurons and their hybrid sequential and/or parallel combinations.

Experimental numerical study of training and functioning indicators of the developed ins architectures was carried out according to the selected criteria of accuracy (loss, accuracy) and quality indicators (stability, lack of retraining, etc.). To ensure stability and prevent over-training of ins, we used methods of regularization and normalization of data, as well as genetic algorithms for their structural and parametric optimization.

3. Results and Discussion

An essential component of the problem under consideration is the presence and identification of "information garbage" and "information trash" in the conditions of exponential growth of the volume of information. According to the authors' conceptual approach, the concept of "information garbage" is proposed to be classified as content components that are unsuitable for objective and/or subjective reasons for use for their intended purpose (fragments of texts and programs and/or texts and programs in general).

In contrast to "information garbage", information trash is content components (fragments of texts and/or texts) that are usable, but are withdrawn (or subject to withdrawal) from use for objective (unreliability, irrelevance, incorrectness) and/or subjective (rejection, misunderstanding, underestimation) reasons. Information garbage and information trash (table 1) pose a real threat to the intellectual sphere of society.

Urgent measures and effective mechanisms are required to identify the content containing them and then neutralize (block) the sources of such content, primarily on the global Internet.

3.1 Working Hypothesis

In the course of the research, a hypothesis was formulated about the possibility of selecting the architecture and hyperparameters of the neural network to solve the problem of identifying latent socio-cultural threats that are not always clearly contained in the studied text corpora. The object of the study was a corpus of socio-political texts downloaded from available Internet resources, including RSS feeds containing information about explicit or latent socio-cultural and cyber threats, as well as about ideological extremism.

| Table 1. Classification Criteria for Detecting Unwanted Content |
|---------------------------------------------------------------|
| **Unwanted content** | A concise definition of | Distinctive features |
| Intimidation | Dramatized threat of destructive influence on a person, people close to them, their property; | Promises to harm a person, organization, or their property; |
| | Wishes for an unfavorable result of the activity; | Demonstration of cruelty |
| Affront (indignity) | Deliberate humiliation of the honor and dignity of a | Obscene expressions; |
| | | Negative assessment of the qualities of a person; |
The main stages of the study that provide verification of the formulated hypotheses were the following.

1. Downloading and/or forming a subject-oriented database of thematic text corpora with their markup for teaching the studied ANNS.

2. Tokenization of the generated text corpora and compilation of a frequency dictionary (thesaurus) of terms.

3. Building a dictionary based on a statistical approach that normalizes the frequency of occurrence of tokens in the document, taking into account the content in the rest of the corpus in the coordinates "term frequency–inverse document frequency" (TF-IDF). The TF-IDF encoding method allows you to measure the relevance of document tokens by the scaled frequency of their appearance, normalized by the inverse scaled frequency of the appearance of the token in the entire corpus.

4. Building embeddings based on the "word2vec" algorithm implemented in the Google Colaboratory cloud environment. Vector representation of words in a multi-dimensional space, in which similar words appear next to each other in a vector space with an a priori dimension of the order of 10...150, which is selected experimentally.

3.2 Architectural and Software Solutions for Artificial Neural Networks

A combined architecture based on fully connected, convolutional, and recursive layers was adopted for the design of a subject-oriented ins. The latter were implemented using the LSTM architecture, the structure of the base element of which, the combination of which forms the intermediate layers [14], is shown in figure 1.

Each of the LSTM cells of the intermediate layer contains interconnected input, intermediate, and output filters, as well as several neurons and controlled adders that provide switching of transmitted signals. The activation functions of these neurons in the LSTM layer are sigmoid and hyperbolic tangent. With all the advantages of the described architecture—the ability to train temporal data, the sequence of which is typical for NLP, its training takes a long time, which, depending on the volume of generated training and verification samples, can be tens of hours. To eliminate these shortcomings, a modified version of the recursive ANN with memory can be used, as shown in figure 1-b.
The algorithm of functioning of such a structure, which contains three inputs and outputs, is mathematically described by a system of (1) equations.

\[
\begin{align*}
    z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\
    \tau_t &= \sigma(W_\tau \cdot [h_{t-1}, x_t]) \\
    \tilde{h}_t &= \tanh(W \cdot [\tau_t \cdot h_{t-1}, x_t]) \\
    h_t &= (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t,
\end{align*}
\]

where \(x_t\) - the input signal of each layer cell; 
\(h_t\) and \(h_{t-1}\) - intermediate signals transmitted between identical elements-cells of the hidden layer; 
\(\sigma\) - the sigmoidal activation function of neurons; 
\(\tanh\) - activation function in the form of a hyperbolic tangent.

Training ANN of this architecture with modified recursive layers takes significantly less time than with LSTM layers, which is especially important for automated selection of hyperparameters, for example, using genetic algorithms. After selecting the optimal architecture, building and compiling the ANN in the Google Collaboratory environment, a series of numerical studies of the performance indicators of the developed ins for ML-classification of texts will be conducted.

The authors justify 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 target Internet content, dictionaries of terms used for multi-class text analysis are pre-formed. After that, the text fragments are marked up according to their belonging to the corresponding six classes.

The software implementation of the ANN family was implemented using the built-in tools and libraries of Google Colaboratory. Technically, Google Colaboratory is a dedicated service hosted in the cloud. This service provides high-performance access to computing resources via the browser and does not require additional configuration for use. The developed ensemble of neural networks had the following structure. It included two types of blocks: invariant blocks designed for preprocessing texts that were identical for all the analyzed architectures, and variable blocks containing layers of neurons of different architectures. Text preprocessing units provided (1) loading text data, (2) splitting it into words, (3) creating dictionaries, (4) converting data to indexes, and (5) generating training and test samples [5].
To justify the choice of variable ANN blocks architectures, that are directly focused on the analysis of terms, phrases, and sequences of analyzed texts, layers of different types of neurons were analyzed. First of all, we studied modifications of recurrent layers that allows to remember the analyzed token sequences (CNN, LSTM, GRU). In addition, we considered one-dimensional convolutional layers (Conv1D) with variable window sizes and the number of convolution cores. Combinations of such powerful layers of neurons can contribute to the appearance of an undesirable phenomenon of retraining neural networks, so, for example, dropout layers were provided as a means of regularization.

Architecture of variable blocks developed by ANN that provide generalization and preliminary identification of various features and their combinations in the analyzed texts. This architecture included a combination of fully connected, convolutional, and/or recurrent layers.

Invariant blocks are modules that implement text classification directly. They contained data regularization and normalization layers, as well as a fully connected output layer. The number of output layer neurons corresponds to the number of recognized classes.

Activation functions of the recognition layer had the form "softmax" or "sigmoid" depending on the network architecture and the variant of the problem being solved. Recommendations are given for choosing hyperparameters for "loss", as well as the number of neurons and the activation function of the last recognition layer. When solving the problem of multiclass NLP-analysis for various architectures of hidden layers of hybrid ins, these parameters were invariant. In the course of the research, callback layers were used to ensure that the network learning process of neural networks is interrupted when the quality of training deteriorates. This approach made it possible to preserve the optimal values of its trained weight coefficients of neurons selected in the process of training the neural network.

3.3 Analysis of the Quality of Built ANN
The following describes the multi-class categorization-oriented ANN architecture, using the example of 6 text classes. The ANN architecture is based on fully connected layers with the "Relay" activation function and a 16-dimensional "embedding" word representation model with regularization via the "SpatialDropout1D" layer (Fig. 2).

![Figure 2. Fragment of the Embedding Source Code of the Program](image)

It is experimentally established that text models of the "bag of words" type are characterized by simplicity and relatively high learning speed. However, further research required for reliable detection of hidden threats required the use of more complex models that implement the representation of tokens in an n-dimensional dense vector space [2].

Figure 3 shows the results for the architecture described above, obtained during its training.
The diagrams in figure 3 for the test and test samples show a relatively high learning rate and classification accuracy of about 80%, even at 10...30 epochs, which can be considered acceptable in the subject area under study. This allows us to recommend such an architecture with the obtained values of hyperparameters for subsequent numerical experiments and the choice of a more optimal neural network architecture.

Research has also been done for ANN with an architecture based on LSTM layers with similar hyperparameters. Numerical experiments showed significantly longer learning time. This served as the basis for studying the architecture based on the one-dimensional convolution "Conv1D" with hyperparameters (20, 5, activation= 'relu') "in combination with the"MaxPooling1D" neuron layer.

Figure 3 shows the results of training ANN, whose architecture is based on one-dimensional convolution layers.

Analysis of the diagram in figure 3 shows a slightly lower learning rate compared to a network with only a fully connected architecture. The classification accuracy in the test sample is also about 80%. In this case, you can select an additional parameter 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 reduced hyperparameter values for further numerical experiments.

Below is a snippet of the Python 3 program code.

```python
modelECn = Sequential()
modelECn.add(Embedding(maxWrdsCunt, 10, inpt_lngth=xLn))
modelECn.add(SpatialDropout1D(0.2))
modelECn.add(BatchNormalization())
```
```python
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 4.** ANN Snippet Based on Fully Connected Layers

### 3.4. Scientific and Practical Significance of the Research Results

The scientific significance of the results of the research focused on identifying specific socio-political threats in Internet content is determined by the proposed scientific and methodological approaches to the formation of thematic text corpora and the construction of hybrid ins based on a combination of fully connected, convolutional and recursive layers.

![Graph](https://via.placeholder.com/150)

**Figure 5.** Results of ANN Training Based on the One-Dimensional Convolution Layer

The applied value of the results is in the software implementation of the proposed approach in Python, the formation of problem-oriented dictionaries of terms and the resulting set of neural network assessments of a number of Internet sources based on the criteria for the presence of obvious or latent signs of socio-cultural and cyber threats, as well as ideological extremism.

Thus, the modified approach is justified and the program implementation of neural network detection of texts containing explicit or latent socio-cultural and cyber threats, as well as signs of ideological extremism contained in the information content of Internet resources, based on thematic dictionaries (thesaurus), previously constructed by computer analysis of target Internet content, is performed.

### 4. Conclusions

The conducted research, including the methodology for forming the constructed thematic text corpora, as well as the construction of the ANN ensemble of combined architecture, allowed us to obtain the following conclusions.

1. The category of "information garbage" is introduced and classification features are justified as undesirable specific content, the identification and elimination of which helps to increase the level of relevance of information selected for further analysis.
2. For a preliminary experimental study, texts containing both the relevant information core and undesirable information, including socio-cultural and cyber threats and “information garbage”, were formed, which allowed us to get training, verification and testing samples for training the created ANNs.

3. An ensemble of combined ANN based on a combination of convolutional, fully connected and recursive layers is constructed, aimed at identifying undesirable text content containing sociotechnical and cyber threats, as well as "information garbage". The degree of recognition of thematic texts was achieved in the range of 0.79...0.85.

5. References
[1]. Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio 2014 Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. URL: https://arxiv.org/abs/1406.1078
[2]. Atamanov G 2015 Azbuka bezopasnosti. Methodology of information resource protection. Information Protection. Insider 2 (62), pp 8-13
[3]. Selva Prabhakaran. Topic Modeling with Gensim (Python) URL: https://www.machinlearningplus.com/nlp/topic-modeling-gensim-python (13.09.2020)
[4]. Olyanitch A, Khachmafova Z, Makerova S, Akhidzhakova M, Ostrovskaya T. 2019 Cognitive developing of semiotic data in computer-based communication (signs, concepts, discourse). Communications in Computer and Information Science. 1084. pp 109-121
[5]. Polyaakov P, Kalinina M, Pleshko V 2015 Automatic Object-oriented Sentiment Analysis by Means of Semantic Templates and Sentiment Lexicon Dictionaries. Proceedings of the 21st International Conference on Computational Linguistics Dialog–2015 2 pp 44–52
[6]. Surkova A, Chernobaev I 2019 Comparison of neural network architectures in the task of automatic text classification Modern informatization problems in the technological and telecommunication systems analysis and synthesis. MIP-2019' AS Proceedings of the XXIV-th International Open Science Conference pp 377-382
[7]. Bengforth B, Bilbro R, Ojeda T 2019 Applied text data analysis in Python. Machine learning and the creation of applications for natural language processing. (Saint Petersburg: Piter) 368 pp
[8]. Kyunghyun Ch 2014 Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation URL: https://arxiv.org/abs/1406.1078
[9]. Keras. The Python Deep Learning library. Official website of the Keras library. URL: https://www.keras.io/ (23.04.19).
[10]. Kim Y et al. 2015 Character-Aware Neural Language Models. ArXiv Prepr. ArXiv1508.06615
[11]. Tarasov D 2015 Deep Recurrent Neural Networks for Multiple Language Aspect-based Sentiment Analysis of User Reviews Proceedings of the 21st International Conference on Computational Linguistics Dialog. 2 pp 53–64
[12]. Gordeev D 2016 Detecting state of aggression in sentences using CNN. Lecture Notes in Computer Science 9811 pp 240-245
[13]. LeCun Y, Zhang X 2016 Text understanding from scratch. Computer Science Department arXiv:1509.01626
[14]. Smirnova O, Shishkov V 2016 Choosing the topology of neural networks and their application for classification of short texts. International Journal of Open Information Technologies 4 p 8
[15]. Chernobaev I, Skorynin S, Surkova A 2019 Application of recurrent neural networks in the task of detecting insincere messages in online services. System analysis in design and management. Collection of scientific papers of the XXIII International scientific and practical conference pp 403-412
[16]. Nikolenko S 2014 Text categorization and LDA model. URL: https://logic.pdmi.ras.ru/~sergey/teaching/mlkfu14/17-lda.pdf
[17]. Prokhorov S, Safronov V 2019 AI for AI: What NLP techniques help researchers find the right articles on NLP. Proceedings - 2019 International Conference on Artificial Intelligence: Applications and Innovations, IC-AIAI 2019. From Intentions to Concrete Results, from Discussion of Philosophical Questions "What is Truth? (Sorry, What is AI?) to Proposals for Innovative Projects pp 76-81

[18]. Rigin E, Fedorov S. 2020 Analysis of the quality of processing news articles in Russian by NLP systems. Modern technologies in the theory and practice of programming, proceedings of the conference. Peter the Great Saint Petersburg Polytechnic University; Dell Technologies; (EPAM Systems. Saint-Petersburg) pp 125-126

[19]. Timoshchuk E 2020 Assessing the quality of the requirements specification by applying gqm approach and using NLP tools. Proceedings of the Institute for System Programming of the RAS. T. 32 2 pp 15-28

[20]. Shu C, Dosyn D, Lytvyn V, Vysotska V, Sachenko A, Jun S 2019 Building of the predicate recognition system for the NLP ontology learning module. Proceedings of the 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2019.10 pp 802-808

[21]. Kutuzov A, Nikishina I 2019 Double-blind peer-reviewing and inclusiveness in Russian NLP conferences. Lecture Notes in Computer Science 11832 LNCS pp 3-8

[22]. Rogachev A 2019 Fuzzy Set Modeling of Regional Food Security. Advances in Intelligent Systems and Computing 726 pp 774-782. doi: 10.1007/978-3-319-90835-9_89

Acknowledgments

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