Detecting Fake News About Covid-19 on Small Datasets with Machine Learning Algorithms

Elena Shushkevich
Technological University Dublin
Dublin, Ireland
elena.n.shushkevich@gmail.com

John Cardiff
Technological University Dublin
Dublin, Ireland
John.Cardiff@TUDublin.ie

***Abstract*—Nowadays the problem of fake news in social media is dramatically increasing, especially when it refers to fake news about Covid-19, as it is a recent and global problem. Because of this fact, it is important to have the ability to detect and delete such news immediately. In our research we concentrate our efforts on detecting fake news about Coronavirus on small datasets, using the Constraint-2021 corpus: the full dataset (10,700 messages) and the limited dataset (1,000 messages). We compare classical Machine Learning Algorithms (4 algorithms: Logistic Regression, Support Vectors Machine, Gradient Boosting, Random Forest) – algorithms of classification from the Scikit-learn library, GMDH-Shell tool (2 algorithms: Combi and Neuro), and Deep Neural Network (LSTM model). The results show that GMDH algorithms outperform traditional Machine Learning Algorithms and are comparable with Neural Networks model's results on the limited dataset.**

I. Introduction

Fake news in social media is an important problem because it may harm people mentally and physically, especially in the case when fake news is related to health areas. Nowadays when everybody, including children, has easy access to internet sources, and when there is no strict censorship of news and articles and anybody (not only experts or professional journalists) can be a news author, it is very easy to find misinformation in social media, which could lead to serious consequences.

In the last two years, news connected with Covid-19 has become fertile soil for such rumors and fake news. Fake news is information hoaxes designed to deliberately mislead the reader in order to gain a financial or political advantage [1]. In the context of Covid-19, such hoaxes could be targeted to accuse the government of wrongdoing, to increase sales of any medicines and many other actions, not to mention the anti-vaccine movement, which is becoming more widespread in connection with compulsory vaccination.

Due to the factors mentioned above, it has become an urgent task to create systems that can detect and classify such fake news in social media, and it is important to pay attention to the time from the fake news' publication to the fake news' detection and deletion – sometimes it is dramatically important to delete such news as soon as possible to prevent it spreading. In such time critical cases, there may not be enough time and a big enough dataset to create a complex system using neural networks which, as shown in [2], presented lower results than classical machine learning models on a small dataset. Therefore we are concentrating our efforts to create a system that will be able to detect and classify fake news connected with Covid-19 in the condition of a limited dataset. We are going to create a range of such systems and test them on the limited dataset to identify the best system. This system would be helpful for researchers, which do not have the opportunity, or have a time limit, to collect a big enough dataset for fake news classification.

The article is structured as follows: in the first section we describe the problem we deal with and the motivation of the research. Section 2 is devoted to the related works in the area of fake news detection. In section 3 we present the dataset we used for the experiments, and in section 4 we describe the preprocessing and modeling steps of our research. Section 5 contains the results we achieved and in section 6 we discuss these results and possible future steps.

II. Related Work

There is a lot of research devoted to the problem of fake news in the area of health. The research [3] explores ways in which social media messages focus on fake health information or misinformation and health evidence with real or potential social impact. The authors of [4] researched the impact of rumors and misinformation in social media in the context of the Covid-19 pandemic. The authors noted that the big volume of fake news in this area leads to a decrease in people's mental health and to an increase in the spread of Covid-19.

For the purpose of fake news detection and classification, researchers use both the classical machine learning approach and neural networks. For example, in [5] the detection of news connected with the pandemic was implemented using classical K-Nearest Neighbour Classifier, while the authors of [6] used an approach based on RoBERTa [7].

It should be noted that there are now some datasets devoted to fake news in social media related to Covid-19 being created. In the article we mentioned before [6] the authors collected 4,800 expert-annotated social media posts related to Covid-19 and 86 common misconceptions about the pandemic which help to evaluate the performance of fake news detection in this area. Also, there are some open challenges such as Constraint-2021 [8], which presented a dataset of fake and real news for binary classification. It should be noted that the best results on this dataset (more than 98% of weighted F1-score) were achieved by the winners of the challenge [9] who presented the neural networks based system for fake news detection.

The dataset we used for our research is from Constraint-2021, which we will describe below in more detail.

This article is a continuation of our research, initial results of which were presented in [10].

III. Dataset

For the experiments, we used the dataset Constraint-2021 which was created in the framework of the Constraint@AAAI2021 – COVID19 Fake News Detection challenge. The main goal of the open shared task was to create a system that allows the detection of fake news connected with Covid-19. There was a binary classification suggested in the challenge: the task being to classify if a message is fake news or real news. The challenge contained the same task for two languages: English and Hindi, and we took into account the English dataset only.

The full English dataset consists of 10,700 messages from social media posts and articles connected with Covid-19, where 5,100 of them are real news, collected from reliable sources such as the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), etc. The other 5,600 news are fake ones, and to collect them the authors used Facebook posts, Twitter tweets, Instagram posts and other types of social media. The dataset contains 37,503 unique words, and the full numeric features are presented in Table I.

TABLE I. Numeric features of the dataset

Attribute	Fake	Real	Combined
Unique words	19728	22916	37503
Avg words per post	22	32	27
Avg chars per post	143	218	183

Real news tends to be longer than fake by approximately 10 words (31.97 words in a real message vs. 21.65 words in a fake message, on average). Also, the number of characters per message (post) is higher in real news than in fake ones by nearly 75 characters.

Examples of fake and real news are presented in Table II below. All examples we have shown in this table were obtained from Twitter.

We used the Constraint-2021 dataset to create a limited dataset for the experiments in the condition of small data: we chose 500 real and 500 fake news and combined the news dataset, which we name 'limited dataset' as opposing to 'full dataset' of 10,700 messages.

IV. Modeling

We created the system for fake news detection using two steps: the preprocessing step, where we prepared messages for the experiments, and the modeling step, where we carried out the modeling itself. In this section, we will describe both steps in more detail.

A. Preprocessing

At the preprocessing stage we followed the below outlined steps:

- converting all characters to lowercase;
- removing all characters except those of the English alphabet and numbers;
- removing stopwords;
- removing words that occur too often (more than in 50% of messages);
- removing words that occur too rare (less than in 1% of messages);
- stemming;
- latent semantic indexing using PCA.

For all preprocessing steps we used Python, NLTK (Porter Stemmer and stopwords' library) and SciKit Learn (for converting texts to vectors and PCA) implemented on it. It is important to clarify the last-mentioned preprocessing step.

TABLE II. Examples of fake and real news

Label	Text
Fake	#Watch Italian Billionaire commits suicide by throwing himself from 20th Floor of his tower after his entire family was wiped out by #Coronavirus #Suicide has never been the way, may soul rest in peace May God deliver us all from this time
Fake	NEWS! New Government lockdown advice is either 'perfectly clear' or 'woefully confusing' depending on who you voted for
Fake	Trump announced that Roche Medical Company will launch the vaccine next Sunday and millions of doses are ready from it !!! The end of the play
Fake	China Muslims hidden at Bihari mosque has been taken to corona virus test by Bihari police. Erode police has caught Thailand Muslim mullahs infected with corona virus. Today Salem Police has caught 11 Indonesian Muslim mullahs at Salem mosque. This video shows that they are applying and putting saliva on spoons plates and utensils and also they are in the intention of spreading corona virus disease. Nobody knows what's happening in the Nation
Real	Almost 200 vaccines for #COVID19 are currently in clinical and pre-clinical testing. The history of vaccine development tells us that some will fail and some will succeed-@DrTedros #UNGA #UN75
Real	14 new cases of #COVID19 have been confirmed in Nigeria: 2 in FCT 12 in Lagos Of the 14 6 were detected on a vessel 3 are returning travellers into Nigeria; 1 is close contact of a confirmed case As at 7:35 pm 26th March there are 65 confirmed cases 3 discharged 1 death
Real	Currently most cases of #COVID19 in the US are in California and Washington State. However many other communities are also dealing with cases of COVID-19. See CDC recommendations for preventing spread of COVID-19 in communities.
Real	Solidarity is needed to provide a joint solution to the #COVID19 pandemic. The COVAX □ Vaccines Facility is the critical mechanism for joint procurement & pooling risk across multiple vaccines which is why I sent a □ to every Member State encouraging them to join-@DrTedros

We converted texts to vectors using the frequency of words as contained in the dataset, and we used Principal Component Analysis (PCA) to identify the most important parameters (tags), according to the idea that such tags are interconnected. Following this idea, we identified 260 parameters that cover 99% of dispersion. In more detail, the first parameter covers 7.5% of dispersion, 25% of dispersion are covered by the first 7 parameters, 50% of dispersion are covered by the first 34 parameters, and 75% of dispersion – by the first 97 parameters. The coverage of the dispersion by indexes is shown in Fig. 1.

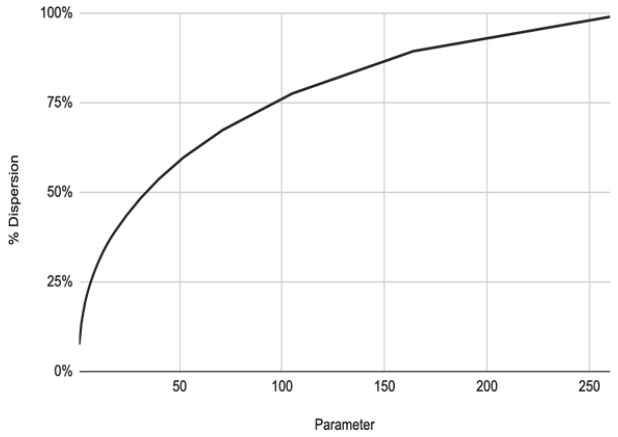

Fig. 1. The coverage of the dispersion by indexes

We used the Kaiser criterion [11] to choose the most important parameters: the criterion suggests choosing just the parameters with eigenvalues higher than the average value. In this way, we chose the 62 most important indexes.

B. Scikit-learn algorithms

After the preprocessing stage, we started modeling and we used four well-known classification algorithms for this purpose:

1) Logistic Regression (LR) – with the basic idea of the linear classifier, which can divide a space of features into two spaces by a hyperplane, where each half-space will reflect each class of a binary classification;

2) Support Vectors Machine (SVM) – a very popular algorithm for the classification tasks which creates a hyperplane, or set of hyperplanes, in multidimensional or infinite-dimensional space;

3) Gradient Boosting Classifier (GB) – the method which produces classification results using classification trees;

4) Random Forest Classifier (RF) – the classifier which also uses an ensemble of decision trees for prediction, but unlike Gradient Boosting, Random Forest builds each tree independently and combines results at the end of the process.

We implemented all these models using the Scikit-learn library, a highly respected free software machine learning library for Python, created for various tasks such as clustering and classification.

C. GMDH Shell

In our research we use two algorithms that showcase the technique of inductive modelling in the form of GMDH – Group Method of Data Handling. The idea and the perspectives of a Group Method of Data Handling are presented in [12], and the theoretical base of it are described in [13]. GMDH can be explained in the steps below:

1) The creation of the models' classes, in the frame of which the models are iterated from the simplest model to the most complex one, and the separation of these observations on the training and the testing datasets;

2) The tuning of parameters of the current model on the training dataset (using internal criteria) and the testing of this model on the testing dataset (using external criteria);

3) With the obtaining of the external criteria's extremum the counting is finished, in the opposite case the model's complexity is increased and the previous step is executed.

In our work we use the implementation of the Group Method of Data Handling in the form of GMDH Shell [14]. This implementation (GMDH Shell) targets the polynomial models class and includes a selection of classification algorithms. In our research we use two basic algorithms:

- The Combi algorithm, which observes all models on each complexity level;
- The Neuro algorithm, which selects the best models on each complexity level and forms a new set of models from the chosen ones based on the principles of evolution.

It should be noted that the complexity of the model is defined by the model's parameter quantity that needs to be assessed.

D. LSTM model

In our work we pay attention to the classification in conditions of a limited dataset. Despite the fact that, as we mentioned above, on a small dataset neural networks could not produce the highest performance, we decided to implement a Neural Network model, which allows us to make a fair comparison. The model we chose is LSTM (Long Short-Term Memory) [15], [16] with an embedding layer, which is a specific kind of recurrent neural network, and is capable of learning long-term dependencies.

V. Experiments

In this section, we describe the experiments we conducted on the full dataset (10,700 messages) and the limited dataset (1,000 messages) with the models we previously described (Logistic Regression, Support Vectors Machine, Gradient Boosting, Random Forest), GMDH Shell and LSTM.

The full options of the experiments are presented in Table III.

TABLE III. Options for experiments

Experiments	Number	Description
Algorithms	7	LR, GB, RF, SVM + Combi, Neuro+LSTM
Datasets	2	10700 and 1000 texts

A. Experiments with Scikit-learn

The results we achieved on the full dataset (10,700 texts) and on the limited dataset (1,000 texts), with the four classical machine learning algorithms we chose (Logistic Regression (LR), Gradient Boosting (GB), Random Forest (RF), and Support Vectors Machine (SVM)) are presented in Table IV.

TABLE IV. TESTING ALL METHODS ON THE FULL DATASET AND ON THE LIMITED DATASET (PRECISION (%) / RECALL(%) / MICRO F1-SCORE(%))

No	Method	10,700 texts	1,000 texts
1.	LR	75/75/75	69/57/64
2.	GB	76/76/76	70/68/68
3.	RF	76/76/76	72/69/69
4.	SVM	74/74/74	70/69/69

B. *Experiments with GMDH Shell*

The results we achieved on the full dataset (10,700 texts) and on the small dataset (1,000 texts), with the two classical GMDH-based algorithms we chose (Combinatorial (Combi) and NeuroNetwork (NN)) are presented in Table V.

TABLE V. TESTING BOTH METHODS ON THE FULL DATASET AND ON THE LIMITED DATASET (PRECISION (%) / RECALL(%) / MICRO F1-SCORE(%))

No	Method	10,700 texts	1,000 texts
1.	Combi	90/89/90	89/87/88
2.	Neuro	84/80/82	83/80/81

C. *Experiments with LSTM model*

The results we achieved with LSTM (Long Short-Term Memory) on the full dataset (10,700 texts) and on the limited dataset (1,000 texts) are presented in Table VI.

TABLE VI. TESTING LSTM ON THE FULL DATASET AND ON THE LIMITED DATASET (PRECISION (%) / RECALL(%) / MICRO F1-SCORE(%))

No	Method	10,700 texts	1,000 texts
1.	LSTM	92/91/92	91/86/88

VI. ANALYSIS AND CONCLUSIONS

In this section we describe briefly the experiments we conducted and analyse the results we achieved, and also we will propose some future steps for the improvement of the models we constructed and will mention the possible ways and the research areas which could be useful in case of limited dataset binary classification.

A. *Discussion*

In our work we present the three main steps of the research:

1) Preprocessing step: where we removed all punctuation marks, stopwords and also words that occurs in the dataset more often than in 50% of texts and rare than in 1% of texts, stemmed words, and also implemented PCA, choose the most important 62 parameters using Kaiser criterion and converted all texts to the numeric vectors based on these parameters.

2) Modeling step: using the Scikit-learn library, where we implemented four classical machine learning classification algorithms: Logistic Regression, Support Vectors Machine, Random Forest and Gradient Boosting.

3) Modeling step: using the GMDH Shell tool, using two methods from the tool: Combi and Neuro.

4) *Modeling step:* using Keras library, where we implemented the Neural Networks model – LSTM.

It is necessary to discuss the results we achieved with the experiments with the four classical machine learning algorithms, two GMDH Shell algorithms and LSTM model on the full dataset (10,700 messages) and on the limited dataset (1,000 messages).

As it is easy to see from Table 4, the highest result of the classification on the full dataset (10,700 messages) were achieved using Random Forest Classifier and Gradient Boosting Classifier, whereas on the limited dataset (1,000 messages) using Random Forest Classifier, but in both cases the difference between the highest and the lowest results is not significant.

To sum up the experiments with Scikit-learn library we conducted, despite the fact that the results obtained on the full dataset are better than the results we obtained on the limited dataset, the numbers we achieved are lower than the results of other researchers (in particular, the winners on the Constraint-2021 Challenge whose research we mentioned earlier) obtained on the same full dataset using a neural network approach. Therefore, neural networks look like a better decision in the case of binary classification with the Constraint-2021 dataset.

The interrelationships of the results between the classical machine learning algorithms we chose on the full and limited datasets are presented in Fig. 2.

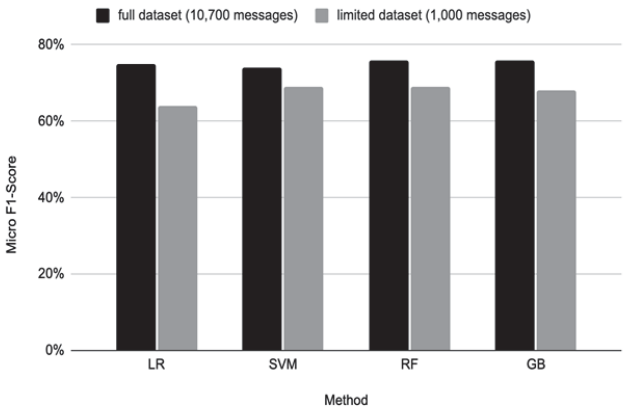

Fig. 2. The impact of the dataset size on the classification quality (Scikit-learn algorithms)

It is easy to see that in all cases the decrease of the dataset's size leads to the significant decrease in the quality of the classification. The greatest difference in micro F1-score we observe is in the case of Logistic Regression (11%), and the smallest difference between the result on the full and limited datasets is in the case of Support Vectors Machine (5%).

From the results we obtained with GMDH Shell algorithms (Combi and Neuro, as described previously) we can conclude

that GMDH approach performed better than the classical machine learning approach on both datasets (full and limited ones): the results on the full dataset is 90% micro F1-score for Combi method and 82% micro F1-score for Neuro method, while the best result with Scikit-learn algorithms is 76% micro F1-score only. As for the limited dataset, the results for GMDH Shell are 88% micro F1-score and 81% micro F1-score for Combi and Neuro method respectively, while the best results with the classical machine learning approach is 69% micro F1-score only. Figure 3 shows the impact of the dataset size on the quality of the classification of GMDH algorithms.

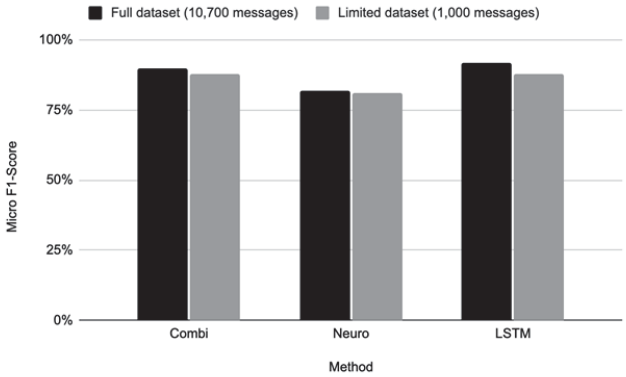

Fig. 3. The impact of the dataset size on the classification quality (GMDH Shell and LSTM Model)

The size of the dataset does not affected the quality of the classification as dramatically as it was in case of classical machine learning algorithms implemented via Scikit-learn – in case of GMDH Shell the difference between the results we achieved on the full dataset and on the limited dataset is only 2% micro F1-score for Combi method and 1% micro F1-score for Neuro method. Because of that, we can say that the GMDH Shell models are more stable than the models from Scikit-learn in case of the limitation of the dataset.

From the results obtained with the LSTM model we can make a conclusion that in the case of experiments on a full dataset the Neural Networks' model provides the best results – 92% micro F1-score in comparison with classical Machine learning algorithms and GMDH algorithms. This fact confirms our previous statement that on the full datasets Neural Networks achieved the highest results.

Figure 3 also shows the impact of the dataset size on the quality of the classification of the LSTM model. On the limited dataset the LSTM model achieved 88% micro F1-score, the same results we achieved with the Combi GMDH method. By decreasing the dataset's size, the performance of the LSTM model decreased by 4%, which is still high in comparison with the results we obtained with classical Machine Learning Algorithms and GMDH Shell.

B. Future Work

The appearance of a large amount of fake news on Covid-19 can be viewed as a manifestation of information wars. Models of information wars have already become the subject of consideration of specialists in mathematical modeling, see for example [17, 18]. In future it would be possible to consider such models to predict the dynamics of fake news and real news of Covid-19.

Furthermore, it looks useful to concentrate efforts to improve the models we have already created, for example, improve the Neural Networks model or to create an ensemble of the models which achieved the best results – the combination of such models can overperform the results we obtain with the models separately.

Also, during our research we used PCA and Kaiser criterion for the conversion of texts to vectors, and it would be interesting to implement smarter transformation methods which we believe would allow these algorithms to perform better.

Another possible way of improvement is to add an additional class of classification to the existing binary classification, and to collect in this class all messages we cannot mark as fake news or real news with a high probability, – this feature means that instead of binary classification we will have multi classification for 'fake news', 'real news' and 'I'm not sure which class is it' classes, and after this type of classification to add 'fake news' and 'real news' classes to the training dataset to expand it. This way looks promising, but it should be noted that for a significant improvement in the classification results the dataset should be bigger than our full dataset.

REFERENCES

[1] E. Hunt, "What is fake news? How to spot it and what you can do to stop it." *The Guardian*, Dec. 2016, Retrieved from https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate

[2] L. Akhtyamova, M. Alexandrov, J. Cardiff, O. Koshulko, "Opinion Mining on Small and Noisy Samples of Health-related Texts.", *Advances in Intelligent Systems and Computing III*. Springer, AISC book series, vol. 871, 2019, pp. 379–390; DOI: 10.1007/978-3-030-01069-0_31, 2018.

[3] C.M. Pulido, L. Ruiz-Eugenio, G. Redondo-Sama, B.Villarejo-Carballido, "A New Application of Social Impact in Social Media for Overcoming Fake News in Health". *Int. J. Environ. Res. Public Health* , 2020, 17, 2430.

[4] S. Tasnim, M. Hossain, H. Mazumder, "Impact of Rumors and Misinformation on COVID-19 in Social Media". *J Prev Med Public Health*, 2020 May, 53(3):171-174. doi: 10.3961/jpmph.20.094. Epub 2020 Apr 2. PMID: 32498140; PMCID: PMC7280809.

[5] S. Bandyopadhyay, S. Dutta, "Analysis of Fake News In Social Medias for Four Months during Lockdown in COVID-19.", 2020, https://doi.org/(10.20944/preprints202006.0243.v1).

[6] T. Hossain, R. Logan, A. Ugarte, Y. Matsubara, S. Young., S. Singh, "Detecting COVID-19 Misinformation on Social Media.", *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020,* 2020.

[7] Y. Liu, et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach.", 2019, CoRR abs/1907.11692.

[8] P. Patwa, S. Sharma., S. PYKL, V. Guptha, G. Kumari, M.S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, "Fighting an Infodemic: COVID-19 Fake News Dataset",.arXiv:2011.03327, 2020.

[9] A. Glazkova, M. Glazkov, T. Trifonov, "g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News Detection.", Combating Online Hostile Posts in Regional Languages during Emergency Situation, DOI:10.1007/978-3-030-73696-5_12, 2021, pp.116-127.

[10] E. Shushkevich, M. Alexandrov, J. Cardiff, "Detecting fake news about Covid-19 using classifiers from Scikit-learn", *International*

Workshop on Inductive Modeling IWIM'2021, in press [to be published in Sept. 2021]

[11] H. Kaiser, (April 1960). "The Application of Electronic Computers to Factor Analysis". *Educational and Psychological Measurement*. 20 (1): 141–151.

[12] V. Stepashko, "Developments and prospects of GMDH-based inductive modeling," Advances in Intelligent Systems and Computing II; Springer, *AISC book* series, vol. 689, 2017, pp. 346–360.

[13] V. Stepashko, "Method of critical variances as analytical tool of theory of inductive modeling," *J. Autom. Inf. Sci.*, vol. 40, no. 3, 2008, pp. 4–22.

[14] Platform GMDH Shell; http//www.gmdhshell.com

[15] S. Hochreiter, J. Schmidhuber, "Long short-term memory. Neural Computation.", 1997, 9(8): 1735–1780. PMID 9377276. https://doi.org/(10.1162/neco.1997.9.8.1735)

[16] F.A. Gers, J. Schmidhuber, "LSTM recurrent networks learn simple context-free and context-sensitive languages.", 2001, IEEE Transactions on Neural.

[17] A. Petrov, O. Proncheva, "Modeling propaganda battle: decision-making, homophily, and echo chambers", in *Proc. AINL-2018*, Springer, series CCIS,vol.930,, Web: //https://link.springer.com/chapter/10.1007%2F978-3-030-01204-5_19, 2018, pp.197-209.

[18] A. Petrov, O. Proncheva, "Modeling position selection by individuals during informational warfare with a two-component agenda", *J. Mathematical Models and Computer Simulations*, vol.12, No.2, https://link.springer.com/article/10.1134%2FS207004822002009X, 2020, pp.154–163.