Prediction of COVID-19 tweeting: classification based on graph neural networks

Milan Petrović[1, 2], Andrea Hrelja [1], Ana Meštrović[1,2]
[1] Faculty of Informatics and Digital Technologies, University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: {milan.petrovic, ahrelja, amestrovic}@uniri.hr
[2] Center for Artificial Intelligence and Cybersecurity, University of Rijeka

***Abstract*—This paper presents the application of graph neural networks (GNNs) to the task of node classification. GNNs have been shown to be useful in various classification tasks where data and the relationships between them can be represented using graphs. This research aims to develop a classifier that can identify two possible classes of Twitter nodes: COVID and nonCOVID. COVID nodes refer to Twitter users (nodes) that post tweets related to COVID-19 and nonCOVID are users (nodes) that do not post tweets about COVID-19. For that purpose, in the first step, we implement a pipeline that enables the automatic, continuous collection of data from Twitter and network construction. In the second step, we prepare the data and train a graph convolutional networks(GCN) classifier. We compare GCN and multilayer perceptron (MLP) in terms of standard measures: precision, recall, F1 and accuracy. The results show that GCN performs better than MLP in the task of node classification.**

I. Introduction

Social networks are a valuable source of information and may serve as an important communication platform in the global crises, such as the COVID-19 pandemic [1]–[3]. During the last two decades, social media have amplified the spread of information, as well as misinformation and disinformation which may lead to an infodemic as a negative side effect [4]. Thus, social network analysis plays an important role and may improve various aspects of crisis communication.

As a highly popular and used online social network, Twitter is one of the most studied networks in the domain of natural language processing (NLP) and social network analysis (SNA) in general. There is a wide set of tasks that ranges from general tweet classification [5], [6], sentiment analysis [7], [8], hate speech detection [9] to fake news detection [10], [11], etc.

During the COVID-19 pandemic, Twitter was again one of the most studied social networks and numerous studies analyzed various aspects of tweets related to the COVID-19. A large number of research papers reported the results of the fake news detection [1], [12] and sentiment analysis [13], [14], while some of the research was dedicated to the analysis of information spreading and infodemic in general [15]–[18].

One important task related to the analysis of information spreading is link prediction. There are many different link prediction algorithms, some are based on the local network properties and others on the global network properties [19]–[22]. In relation to this task, it may be of particular interest to identify which user will post about a certain topic. For example, in the context of the COVID-19 infodemic, it is useful to predict which users will tweet about COVID-19. This task is somewhat different from link prediction, but it also takes into account the properties of networks and can be formulated as a classification task.

Motivated by these considerations, this study aims to develop an approach that predicts whether a Twitter user will tweet about COVID-19. An important aspect in predicting tweet behaviour is the position of a node in the network of followers. Following this idea, we based our prediction on the structure of the network of Twitter followers. Thus, we chose graph neural networks (GNNs) for the classification task. Generally, graph neural networks are introduced in [23] more than a decade ago and have started to be widely used in many different tasks. There are successful applications of GNNs in research related to social network analysis such as fake news detection [24], sentiment analysis [25], social recommendation [26], user localisation [27], etc.

This work is an extension of our previous research focused on COVID-19 related tweets [28]–[30]. While in the previous approaches we analyzed how the content of the tweet influenced the spreading of tweet [29], [30], in this study, we utilize the network structure.

In the first part of the research, we implement a pipeline that enables automatic, continuous collecting of data from Twitter in the first step and automatic network construction in the second step. We collect the dataset of tweets in the Croatian language posted during the fourth wave of the COVID-19 pandemic in Croatia and construct the network of followers. The result is a network dataset of 8,808 users (Cro-USERS) which are represented as nodes and 319,845 links. The network of users followers is the subject of analysis in the next steps of the research. Additionally, we construct a dataset of 1,703,626 tweets.

In the second part of the research, we train neural network models for node representation and classification. Since tweeting depend on the node position in the network of followers it

This work has been supported in part by the Croatian Science Foundation under the project IP-CORONA-04-2061, “Multilayer Framework for the Information Spreading Characterization in Social Media during the COVID-19 Crisis” (InfoCoV), and by University of Rijeka project number uniri-drustv-18-38.

is necessary to capture the network structure. Thus, we choose node2vec model to represent each node as an embedding.

Next, we train models to classify the user into one of the two defined classes. We train and compare two models: graph convolutional network (GCN) [31] and multilayer perceptron (MLP). For that purpose, we divide the Cro-USERS dataset into two disjoint classes: (i) Cro-CoV-USERS - dataset that contains users that tweet about COVID-19 and (ii) Cro-nonCoV-USERS - dataset that contains users that do not tweet about COVID-19. We perform an evaluation in terms of precision, recall, F1-measure and accuracy. According to our preliminary results it seems that GCN outperforms MLP in the task of node classification.

The rest of the paper is organized as follows. In the next section we present the used methodology: (i) first we describe the methods and extensive procedure that we implement for collecting the data and (ii) second we describe the training procedure. In the third section, we report the results of the training. The last section is concluded with a discussion of the results and future work.

II. Methodology and Experiment Setup

A. Dataset

All the data is collected from the Twitter social network using a pipeline that we implemented for automatic, continuous collecting of data from Twitter. Data is structured that for each user we have its friends, followers, and a list of published tweets at a certain time. First, accounts location-based in Croatia are collected, second, all friends and followers are taken from accounts gathered in step one and the last step is collecting tweets from that profiles.

Fig. 1 represents the conceptual model of the pipeline implemented for collecting and preparing the data.

For this study, we collected the dataset of 8,808 Twitter users from the Republic of Croatia (Cro-USERS) and 1,703,626 of their tweets (Tweets dataset) in the Croatian language posted during the fourth wave of the COVID-19 pandemic.

After creating the Tweets dataset, the tweet messages are preprocessed as follows:

- All letters are converted to lowercase.
- Twitter links are removed from the tweets.
- Special characters are removed.
- Tweets without content are removed.

B. Network construction

A user following network is constructed from a sample of users who have more than 10 and less than 5000 followers and friends who posted at least 10 tweets from 31 June 2021. This date is considered to be the beginning of the "fourth wave" of the COVID-19 pandemic in Croatia. With this type of selection, the aim was to filter users who are real persons (not official or public profiles) and active users who share content with their friends and followers. This way we a get better insight into the public communication related to COVID-19 and behaviour of real people on online networks during the pandemic. Users are nodes and the directed link between two nodes is established if the first node follows the second one.

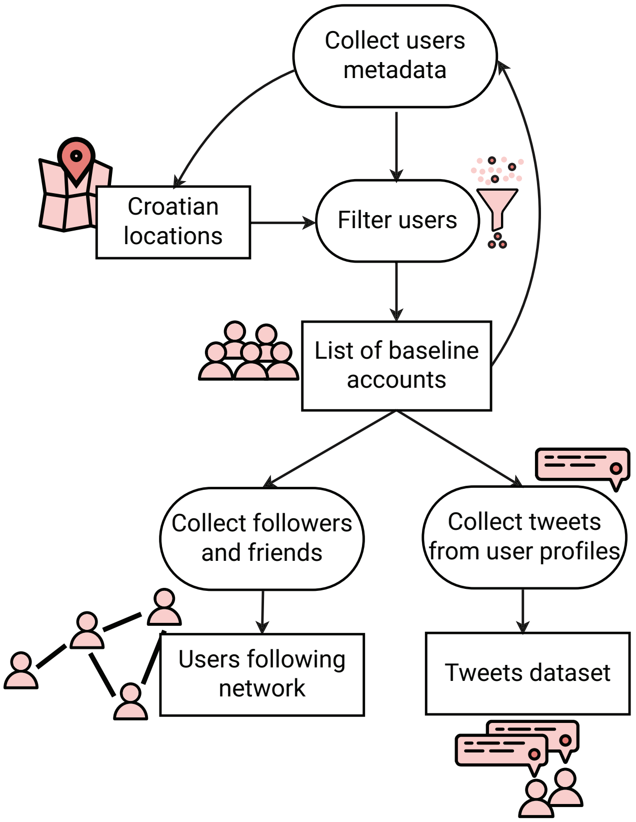

Fig. 1. Conceptual model of pipeline implemented for collecting and preparing the datasets and networks.

Fig. 2 shows the degree distribution in the constructed network. Due to the directed network of followers, we have in-degree and out-degree measures. It is obvious that degree distribution follows the power-law, which is usual distribution in the cases of online social networks. That indicates that the dataset that we selected for this experiment is a representative sample of nodes from the social network.

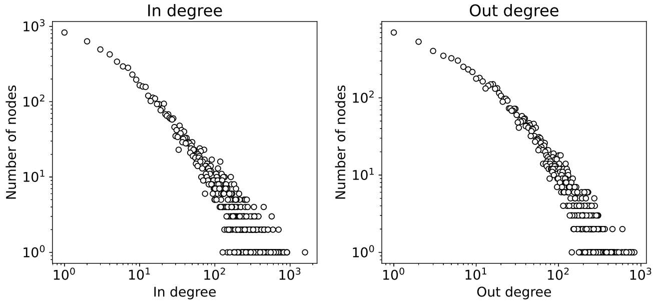

Fig. 2. Degree distribution within the constructed companion network. Due to the directed network of followers, we have in-degree and out-degree measures. Logarithmic scaling on the x and y axes was used.

In the next step each node is assigned a label: COVID or nonCOVID depending on whether the user has been posted or shared a COVID-19 related tweet. COVID-19 tweets are identified using the list of keywords which are related to the COVID-19 topic (the list of COVID-19 keywords in the Croatian language is given in the Appendix).

In Table I we report the total number of nodes and tweets and numbers of COVID-19 related data.

TABLE I
DATASETS

	All	COVID	nonCOVID
Number of nodes	8808	3670	5138
Number of tweets	1,703,626	70,394	1,633,232

Fig. 3 visualizes the user following network with colored COVID-19 related nodes. The network structure influence the action of tweeting about COVID-19. Thus, capturing the network structure as the node embedding using the node2vec can be used for the prediction of COVID-19 related tweeting.

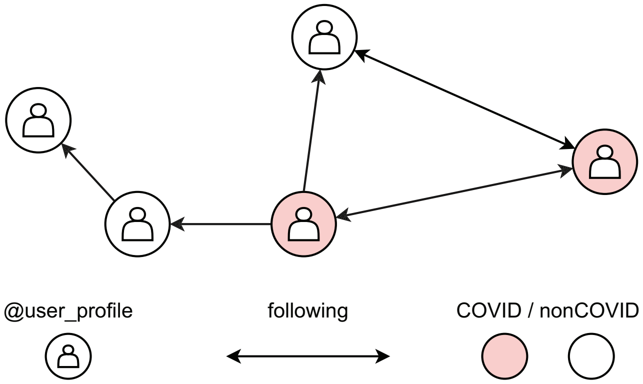

Fig. 3. Visual representation of Twitter following network, where node color represents COVID and nonCOVID label in dataset. Links between nodes represent following and friend relations between user nodes, where arrow direction corresponds if user is following or being followed by other node.

C. Model and training procedure

Model training consists of two steps. In the first step, node embeddings are created with node2vec method, with dimensions set on 16, 32, and 64. In the second steps these embeddings are being forwarded to GCN model. Each dimension was created with parameters where the batch size is 64, the learning rate is 0.01, walk length is set to 10, and number of walks is 3 and this was performed in 5 epochs.

We trained a model to classify the tweets into one of the two classes defined above. When creating the model architecture and setting the parameters, the ones that gave the best results were selected. Selected parameters are listed below.

The GCN's architecture is as follows:

- The input vector has the dimension of the embedding vector (16, 32 or 64).
- There are 3 layers with the number of hidden channels is 256 per layer.
- Dropout is set to 0.3 and learning rate to 0.005.
- Optimizer used for training is Adam.

In order to evaluate the performance of the GCN model, we used MLP as the baseline. Thus, we additionally train the MLP with the following architecture:

- The input vector has the dimension of the embedding vector (16, 32 or 64).
- There are 3 layers with the number of hidden channels is 64 per layer.
- Dropout is set to 0.3 and learning rate to 0.005.
- Optimizer used for training is Adam.

III. Results

Here we report the results of node classification task. We train GCN and MLP models using various dimensions of embeddings and compare their performance in terms of accuracy, precision, recall and F1-measure. The Cro-USERS dataset of 8,808 Twitter users was split with a ratio (55/30/15) for train, validation, and test. The model was trained on embeddings of different dimensions created using the node2vec algorithm. We experimented with three different embeddings dimensions: 16, 32 and 64. The models were trained on 200 epochs on 100 runs of which the best results are shown in the Table II.

TABLE II
MODEL PRECISION

	Precision	Recall	F1	Accuracy
MLP + n2v 16d	0.4643	0.2743	0.3449	0.5787
GCN + n2v 16d	0.7721	0.6029	0.6771	0.7676
MLP + n2v 32d	0.4825	0.3230	0.3869	0.5962
GCN + n2v 32d	0.7551	0.6985	0.7256	0.7865
MLP + n2v 64d	0.4736	0.6217	0.5376	0.5677
GCN + n2v 64d	**0.7916**	**0.7294**	**0.7592**	**0.813**

According the results GCN outperforms MLP in all three experiment setups. The best performance is achieved in the case of GCN combined with the embedding dimension set to 64. As expected, higher embedding dimensions provide better results in almost all cases.

The poor performance of the MLP can be explained by the fact that the training dataset was small. As shown in Fig. 4, the proportion of COVID-19 related tweets is very low in comparison to other tweets. Thus, we need models that can successfully deal with small training datasets. Since node2vec embeddings capture the network structure, graph neural networks are expected to be a better option than traditional neural networks models, such as MLP.

IV. Conclusion

In this paper, we described a classification of Twitter users into two classes: (i) users that post tweets about COVID-19 and (ii) users who do not post tweets about COVID-19. The representation of the user is based on graph neural networks. In particular, we used node2vec to model a user as the vector

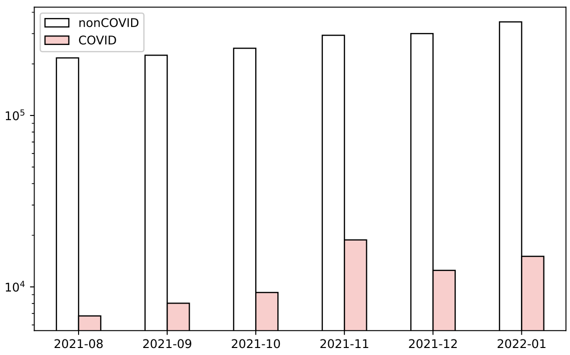

Fig. 4. The distribution of the number of collected tweets by months from the collection period is shown. For each month, two bars are displayed where one shows the number of tweets that do not contain COVID keywords while the other bar shows the number of tweets with COVID keywords. Logarithmic scaling was used on the y axis.

of properties extracted from the users followers network. In the next step, we trained GCN model and compare it with the MLP as the baseline using three different sizes of node2vec embeddings (16, 32 and 64). We evaluated both models in terms of precision, recall, F1-measure and accuracy. Our results that GCN performs better than MLP in all cases. The best performance is achieved for the largest size of embedding for which the F1-measure is 0.7592 and accuracy is 0.813.

The relatively poor performance of the MLP model can be explained by the fact that we have a small sample of COVID -19 related tweets. Moreover, the good results of the GCN model are achieved due to the network-based representation of the dataset. Thus, we conclude that a similar methodology and pipeline could be used for other datasets that have inherent network or graph-based structures.

In further work on this research, the goal is to try other types of embedding techniques, based natural language processing and deep learning. The goal is also to try predefined embeddings for nodes and/or links. These types of embedding can be extracted from the content shared by the user online for example his tweets. Some of the features can also be local node measures from the network or quantitative values taken from the user profile (number of friends, number of followers, etc.). We also plan to test other GNNs architectures. The architectures to be tested in the current research will be determined depending on the tasks. There are a variety of tasks that can be solved with GNN, such as node classification, binary and multiclass classification, link prediction, subgraph and community detection, and whole graph classification.

References

[1] D. Bunker, "Who do you trust? the digital destruction of shared situational awareness and the covid-19 infodemic," *International Journal of Information Management*, vol. 55, p. 102201, 2020.

[2] C. Cuello-Garcia, G. Pérez-Gaxiola, and L. van Amelsvoort, "Social media can have an impact on how we manage and investigate the covid-19 pandemic," *Journal of clinical epidemiology*, vol. 127, p. 198, 2020.

[3] D. C. Glik, "Risk communication for public health emergencies," *Annu. Rev. Public Health*, vol. 28, pp. 33–54, 2007.

[4] G. Eysenbach, "Infodemiology: The epidemiology of (mis) information," *The American journal of medicine*, vol. 113, no. 9, pp. 763–765, 2002.

[5] M. Tutek, I. Sekulić, P. Gombar, I. Paljak, F. Čulinović, F. Boltužić, M. Karan, D. Alagić, and J. Šnajder, "Takelab at semeval-2016 task 6: Stance classification in tweets using a genetic algorithm based ensemble," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 464–468.

[6] E. Mocibob, S. Martinčić-Ipšić, and A. Meštrović, "Revealing the structure of domain specific tweets via complex networks analysis," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2016, pp. 1623–1627.

[7] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the workshop on language in social media (LSM 2011)*, 2011, pp. 30–38.

[8] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine learning-based sentiment analysis for twitter accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 11, 2018.

[9] F. Markoski, E. Zdravevski, N. Ljubešić, and S. Gievska, "Evaluation of recurrent neural network architectures for abusive language detection in cyberbullying contexts," in *Proceedings of the 17th International Conference on Informatics and Information Technologies - CIIT 2020*, 2020.

[10] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 274–277.

[11] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular twitter threads," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2017, pp. 208–215.

[12] C. M. Pulido, B. Villarejo-Carballido, G. Redondo-Sama, and A. Gomez, "Covid-19 infodemic: More retweets for science-based information on coronavirus than for false information," *International Sociology*, p. 0268580920914755, 2020.

[13] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter," *PloS one*, vol. 15, no. 9, p. e0239441, 2020.

[14] M. O. Lwin, J. Lu, A. Sheldenkar, P. J. Schulz, W. Shin, R. Gupta, and Y. Yang, "Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends," *JMIR public health and surveillance*, vol. 6, no. 2, p. e19447, 2020.

[15] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," *Scientific Reports*, 2020.

[16] H. W. Park, S. Park, and M. Chong, "Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea," *Journal of Medical Internet Research*, vol. 22, no. 5, p. e18897, 2020.

[17] R. E. Cuomo, V. Purushothaman, J. Li, M. Cai, and T. K. Mackey, "Sub-national longitudinal and geospatial analysis of covid-19 tweets," *Plos one*, vol. 15, no. 10, p. e0241330, 2020.

[18] C. E. Lopez, M. Vasu, and C. Gallemore, "Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset," *arXiv preprint arXiv:2003.10359*, 2020.

[19] C. Ahmed, A. ElKorany, and R. Bahgat, "A supervised learning approach to link prediction in twitter," *Social Network Analysis and Mining*, vol. 6, no. 1, pp. 1–11, 2016.

[20] S. Martinčić-Ipšić, E. Močibob, and A. Meštrović, "Link prediction on tweets' content," in *International Conference on Information and Software Technologies*. Springer, 2016, pp. 559–567.

[21] S. Martinčić-Ipšić, E. Močibob, and M. Perc, "Link prediction on twitter," *PloS one*, vol. 12, no. 7, p. e0181079, 2017.

[22] K. Zhuang, H. Shen, and H. Zhang, "User spread influence measurement in microblog," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 3169–3185, 2017.

[23] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[24] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, 2019.

[25] M. Wang and G. Hu, “A novel method for twitter sentiment analysis based on attentional-graph neural network,” *Information*, vol. 11, no. 2, p. 92, 2020.

[26] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, “Graph neural networks for social recommendation,” in *The world wide web conference*, 2019, pp. 417–426.

[27] T. Zhong, T. Wang, J. Wang, J. Wu, and F. Zhou, “Multiple-aspect attentional graph neural networks for online social network user localization,” *IEEE Access*, vol. 8, pp. 95 223–95 234, 2020.

[28] K. Babić, M. Petrović, S. Beliga, S. Martinčić-Ipšić, M. Pranjić, and A. Meštrović, “Prediction of covid-19 related information spreading on twitter,” in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, 2021, pp. 395–399.

[29] K. Babić, M. Petrović, S. Beliga, S. Martinčić-Ipšić, M. Matešić, and A. Meštrović, “Characterisation of covid-19-related tweets in the croatian language: framework based on the cro-cov-csebert model,” *Applied Sciences*, vol. 11, no. 21, p. 10442, 2021.

[30] K. Babić, M. Petrović, S. Beliga, S. Martinčić-Ipšić, A. Jarynowski, and A. Meštrović, “Covid-19-related communication on twitter: analysis of the croatian and polish attitudes,” in *Proceedings of Sixth International Congress on Information and Communication Technology*. Springer, 2022, pp. 379–390.

[31] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv preprint arXiv:1609.02907*, 2016.

Appendix

List of words used to extract tweets relevant to COVID.

keywords = [’alemka’, ’markotic’, ’markotić’, ’beros’, ’beroš’, ’capak’, ’hzjz’, ’antigensk’, ’antimaskeri’, ’antivakseri’, ’cijep’, ’cijepiv’, ’cijeplj’, ’cijepljen’, ’cjep’, ’cjepiv’, ’cjepljen’, ’booster doza’, ’prva doza’, ’druga doza’, ’treca doza’, ’treća doza’, ’astra zeneca’, ’biontech’, ’curevac’, ’inovio’, ’janssen’, ’johnson’, ’novavax’, ’moderna’, ’pfizer’, ’vaxart’, ’sojevi koronavirusa’, ’brazilski’, ’britanski’, ’ceski soj’, ’delta’, ’indijski’, ’juznoafricki’, ’južnoafrički’, ’lambda’, ’njujorski’, ’njujorški’, ’omikorn’, ’omikron’, ’novi soj’, ’češki soj’, ’coron’, ’corona’, ’covid’, ’covid-19’, ’covid 19’, ’koron’, ’korona’, ’kovid’, ’ncov’, ’mutira’, ’mutaci’, ’n95’, ’sars-cov-2’, ’sarscov2’, ’sputnik’, ’inkubacij’, ’ljekov’, ’obolje’, ’novozaražen’, ’nuspoj’, ’patoge’, ’regeneron’, ’medicin’, ’infekc’, ’dezinf’, ’bolnic’, ’dijagnost’, ’doktor’, ’epidem’, ’respir’, ’respirator’, ’simpto’, ’rt pcr’, ’terapij’, ’viro’, ’virus’, ’slusaj struku’, ’slušaj struku’, ’propusnic’, ’ostani doma’, ’ostanimo doma’, ’zaraž’, ’festivala slobod’, ’pcr’, ’samoizola’,’samoizolacij’, ’testira’, ’zaraz’, ’distanc’, ’izolac’, ’karant’, ’lockd’, ’mask’, ’festival slobod’, ’ostanimo odgovorni’, ’pandem’, ’pandemij’, ’stozer’, ’stožer’,]