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

Milan Petrović[1,2]; Andrea Hrelja[1]; Ana Meštrović, PhD[1,2]
[1]Faculty of Informatics and Digital Technologies, University of Rijeka, [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.

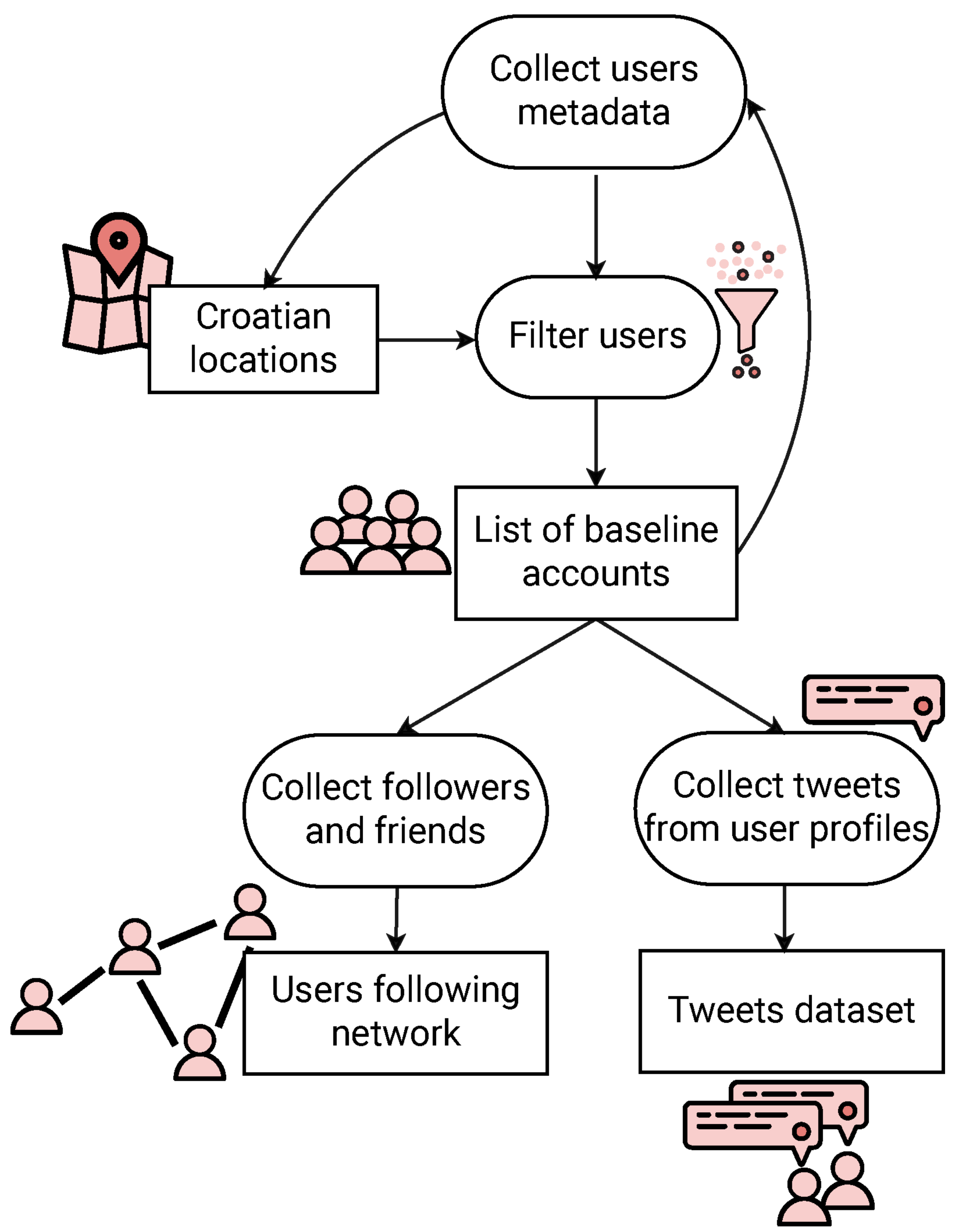

Figure 1. Conceptual model of pipeline implemented for collecting and preparing the datasets.

Introduction

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. Graph neural networks are widely used in many different tasks in research related to social network analysis such as fake news detection, sentiment analysis, social recommendation, user localization, etc.

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

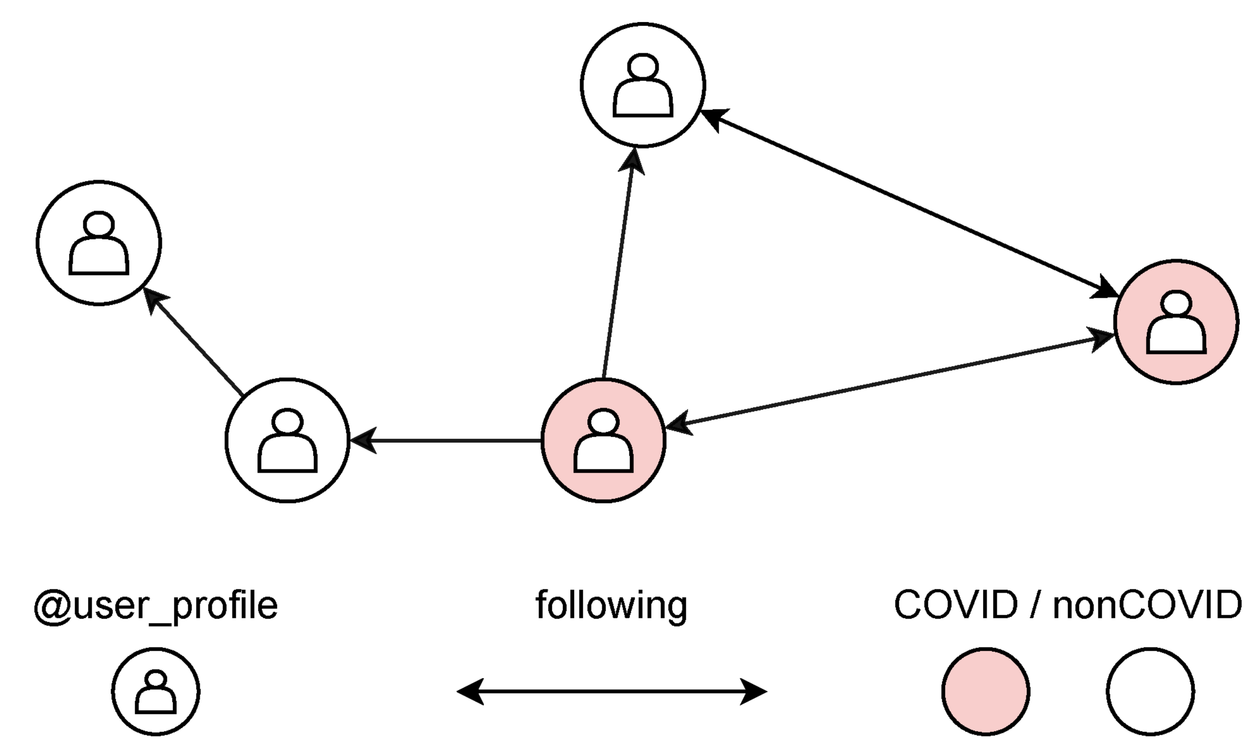

Figure 2. Visual representation of Twitter following network, where node color represents COVID and nonCOVID label in dataset. Links between nodes represent following and friend relations between user nodes, where arrow direction corresponds if user is following or being followed by another node.

Models

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. The selected parameters are listed below.

- 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 for MLP and 256 for GCN per layer.Dropout is set to 0.3 and learning rate to 0.005.
- Optimizer used for training is Adam.

Methods and Materials

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 presented in Figure 1. Data is structured so that for each user we have its friends, followers, and a list of published tweets at a certain time Figure 2.

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, distribution of data is shown in Chart 1.

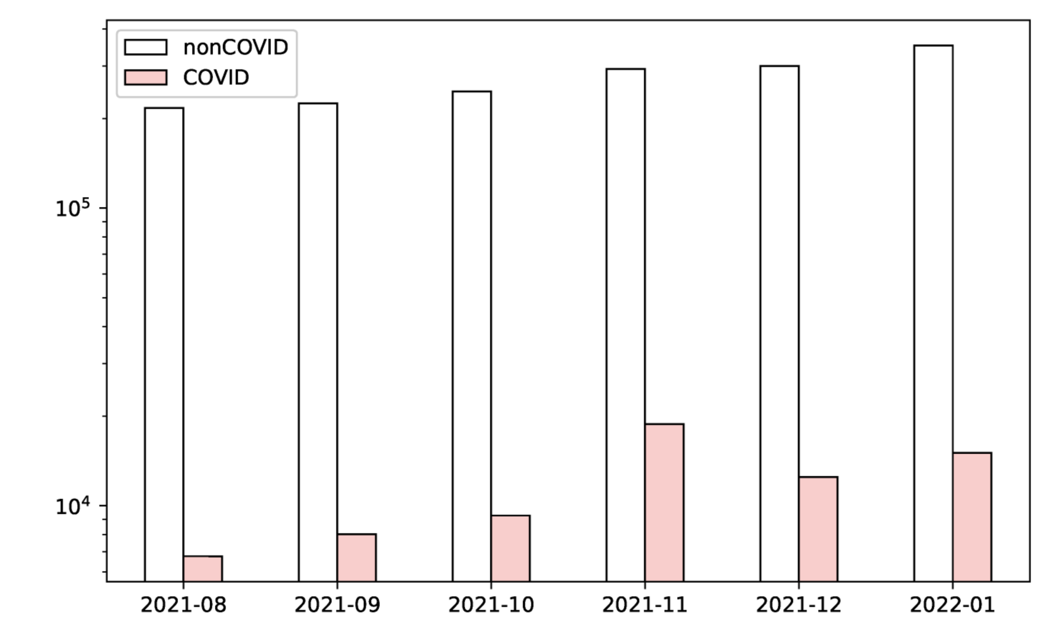

Chart 1. The distribution of the amount of collected tweets by months (Log scaling on the y axis).

Results

GCN and MLP models were trained 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 1.

According to 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.

	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

Table 1. Model precision.

Contact

Milan Petrović
Faculty of Informatics and Digital Technologies, University of Rijeka
Email: milan.petrovic@uniri.hr

References

1. Petrović, Milan, Andrea Hrelja, and Ana Meštrović. "Prediction of COVID-19 tweeting: classification based on graph neural networks." *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. IEEE, 2022.
2. 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.
3. 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
4. T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” arXiv preprint arXiv:1609.02907, 2016
5. 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