Prediction of COVID-19 Related Information Spreading on Twitter

Karlo Babić[1,2], Milan Petrović[1,2], Slobodan Beliga[1,2], Sanda Martinčić-Ipšić[1,2], Marko Pranjić[3], Ana Meštrović[1,2]
[1] Department of Informatics, University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: {karlo.babic, milan.petrovic, sbeliga, smarti, amestrovic}@uniri.hr
[2] Center for Artificial Intelligence and Cybersecurity, University of Rijeka
[3] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
Email: marko.pranjic@styria.ai

***Abstract*—In this paper, we explore the influence of COVID-19 related content in tweets on their spreadability. The experiment is performed in two steps on the dataset of tweets in the Croatian language posted during the COVID-19 pandemics. In the first step, we train a feedforward neural network model to predict if a tweet is highly-spreadable or not. The trained model achieves 62.5% accuracy on the binary classification problem. In the second step, we use this model in a set of experiments for predicting the average spreadability of tweets. In these experiments, we separate the original dataset into two disjoint subsets: one composed of tweets filtered using COVID-19 related keywords and the other that contains the rest of the tweets. Additionally, we modified these two subsets by adding and removing tokens into tweets and thus making them artificially COVID-19 related or not related. Our preliminary results indicate that tweets that are semantically related to COVID-19 have on average higher spreadability than the tweets that are not semantically related to COVID-19.**

***Keywords*—information spreading; neural networks; NLP; Twitter; COVID-19**

I. Introduction

Social media accelerates information spreading and may cause an infodemic, especially during a crisis. As stated by the World Health Organization (WHO), the COVID-19 outbreak culminated with a massive infodemic, which is potentially dangerous because it makes it difficult for individuals to find reliable sources of information when they need it [1], [2]. In the context of infodemic, exploring information spreading plays an important role and may improve various aspects of crisis communication. Additionally, understanding information spreading patterns is the first step towards fake news detection.

As the 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 variety of tasks that ranges from tweet classification [3], [4], fake news detection [5], [6], sentiment analysis [7], [8] to hate speech detection [9], etc. Many of these studies deal with information spreading, link prediction, virality prediction, and other similar tasks [10]–[12].

Since retweeting is the key mechanism for information spreading on Twitter, retweet prediction has been the focus of many research that deal with information spreading. There is a large number of published papers that propose approaches for retweet prediction based on various features that may have an influence on retweeting. For example, Zhang et al. [13] use three sets of features: the author information, content information, and user interests, while in [14] authors consider features related to the time interval of retweets and the location of users in information cascades. In [15], authors study retweet cascades and try to predict cascade size at the exact moment of time. Some other approaches combine various manually constructed features, such as linguistic features, users' personal information, temporal features, etc. Recently, deep learning approaches have been used to automatically learn optimal features. In [16] authors propose a novel attention-based deep neural network to incorporate contextual and social information for the prediction of retweeting. According to recent studies, deep learning seems to be the most promising approach for information spreading prediction tasks.

During the COVID-19 pandemic, a lot of researchers have analyzed various aspects of tweets related to the coronavirus. Most of the research is dedicated to the infodemic and information spreading [17]–[20], while some studies are focused on the tasks of fake news detection [21], [22] and sentiment analysis [23], [24]. One important research in that domain is the study conducted by Cinelli et al. which addresses the problem of COVID-19 infodemic in social media [17]. Authors analyzed COVID-19 information spreading on Twitter, Instagram, YouTube, Reddit, and Gab. They fitted information spreading with epidemic models and determine the basic reproduction number for each social media platform. They show that there are no substantial differences between fake and true news spreading, only that the amount of fake news varies across platforms. In another similar study [18], authors analyze the Korean COVID-19 related Twitter dataset collected on February 29, 2020. Their results indicate that the spread of information was faster in the coronavirus network than in the other networks.

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.

This study aims to explore the influence of COVID-19 content in Croatian tweets on their spreadability. Hence, we are focused on retweet prediction based only on the tweet content. In our approach, we use neural network embeddings for text representation [25]. In the first part of the experiment, we collect the dataset (cro-tw) of tweets in the Croatian language posted during the first year of the COVID-19 pandemic. For the purpose of the second part of the experiment, we divide the cro-tw dataset into two disjoint subsets: (i) cro-CoV-tw - dataset that contains COVID-19 related tweets and (ii) cro-noCoV-tw - dataset that contains tweets not related to COVID-19. Further, we train a feedforward neural network model to predict if a tweet is highly spreadable. This task is framed as a classification task with two classes: low spreadability and high spreadability. We define the spreadability factor as the retweet count normalized by the number of followers. In the second part of the experiment, the trained model is used to perform predictions on different variants and modifications of the cro-CoV-tw and cro-noCoV-tw datasets in order to explore the influence of COVID-19 related words on the predicted spreadability of tweets.

The rest of the paper is organized as follows. In the next section, we describe the dataset and the training procedure. In the third section, we present the experimental setup in which we analyze spreadability and report the results. The last section is concluded with a discussion of the results and future work.

II. Materials and Methods

We train a feedforward neural network on a dataset of Croatian tweets to predict the spreadability of each tweet.

A. Training Dataset

The Twitter dataset we use for training is collected from Croatian accounts. The tweets are in the Croatian language and posted from 01.01.2020. to 15.01.2021.

Dataset preprocessing:

- All letters are transformed to lowercase letters.
- Tweets in the Croatian language are filtered by checking if they contain at least one of the characters: č, ć, ž, š, or đ.
- Twitter links are removed from the tweets.
- Special characters are removed.
- Tweets with zero retweets are removed.

Tweets with zero retweets are removed because such tweets introduce noise into the model and because such tweets are less interesting from the perspective of information spreading.

The final dataset has 93,587 tweets. The tweets are vectorized by averaging token embeddings extracted from the CLARIN.SI-embed.hr dataset [26] (i.e., centroid calculation as described in [27], [28]). The embeddings in the dataset are based on the fastText skip-gram model [29]. Of the 93,587 tweets, 143 have no tokens found in the embeddings. Among the rest of the tweets, 93.8% of tokens have embeddings. Further details about the dataset can be seen in Table I.

TABLE I: Dataset statistics

	cro-tw	cro-CoV-tw	cro-noCoV-tw
Number of tweets	93,444	9,050	84,394
Number of words	2,118,622	261,974	1,856,648
Number of unique words	187,678	38,276	178,578
Average number of words	22.67	28.95	22.00
Average number of retweets	2.794	2.853	2.788

B. Training Procedure

As the retweet count for each tweet depends on the number of followers the account has, we calculate the spreading factor by normalizing the number of retweets by dividing it by the number of followers. Further, the normalized value is passed through the log function ($spreadingFactor = log(retweets/followers + 1)$). Fig. 1 and 2 show histograms of the retweet count and the spreading factor.

After ordering the tweets by the spreading factor, two classes are made by dividing the tweets into two equally large sets (one with lower spreading factor and the other with higher spreading factor). The value of the spreading factor that borders the two classes is 0.0016.

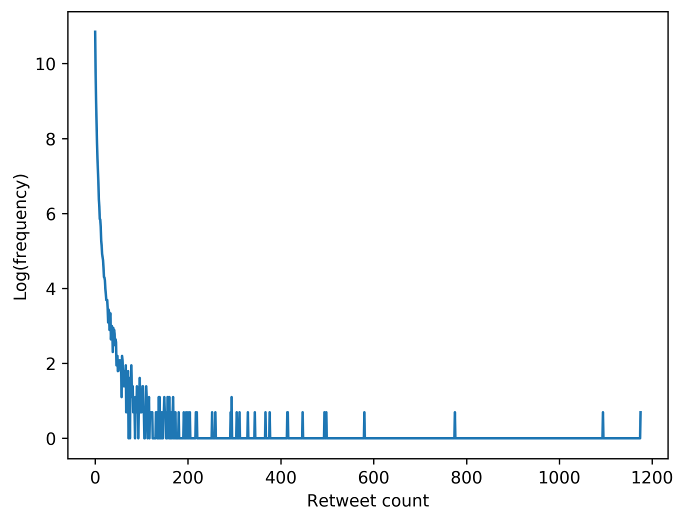

Figure 1: Histogram of the number of retweets each tweet has

We train a sequential neural network (in Keras) to classify the tweets into one of the two classes defined above.

The neural network's architecture is as follows:

- The input vector's length is 100.
- The first hidden layer has 100 nodes with linear activation.
- The next seven layers have 256, 128, 64, 32, 15, 8, and 4 nodes. These layers use rectified linear activation function. They are regularized with the L2 regularization technique.
- The last layer has one node with sigmoid as its activation function.
- Optimizer used for training is Adam with binary cross-entropy loss function.

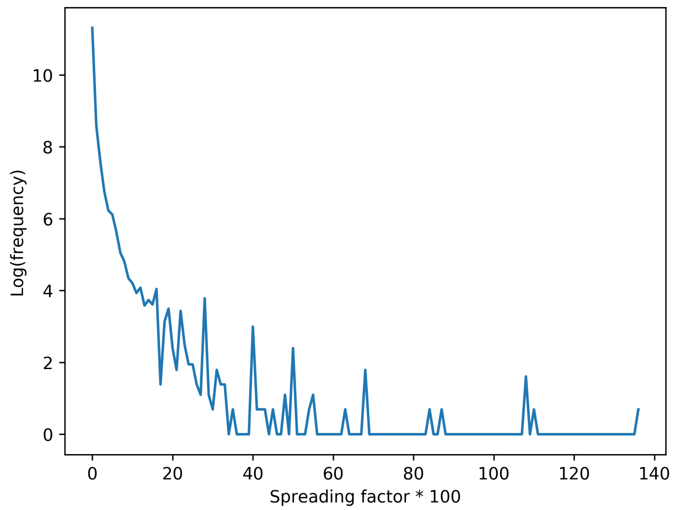

Figure 2: Histogram of the calculated spreading factors on a logarithmic scale

C. *Evaluations*

To evaluate the model we use 10% of the dataset (while the other 90% is used for training). It achieves 62.5% accuracy.

Fig. 3 shows the model's accuracy by comparing its predicted values with the true values. The predicted and true classes (which are denoted with zero for low spreadability and with one for high spreadability) are averaged by seven days and plotted through time.

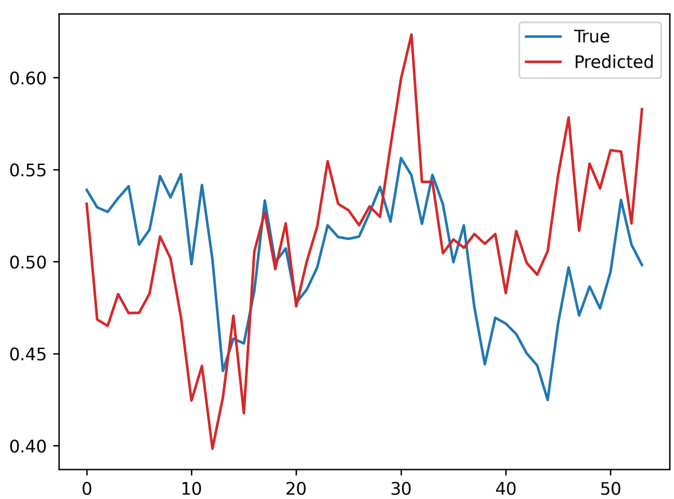

Figure 3: Comparison between true and predicted tweets' spreading factor, averaged by seven days (showing the model's performance). X-axis is time (aggregated by seven days), and Y-axis is the average class (which is zero for low spreadability and one for high spreadability) of all the tweets in the seven day window

III. Results

To explore the influence of COVID-19 related words in a tweet on its spreadability, we perform predictions with our model on seven subsets of the dataset. The main value used for the comparison is the number of tweets (and the relevant percentage) that our model classifies as highly-spreadable.

As our model is trained to predict the spreadability only from a tweet's text, any changes to the textual content of a tweet greatly influence the predicted class.

A. *Setup*

We use two lists of key tokens related to the COVID-19 pandemic (strong key tokens, Table II, and weak key tokens, Table III). Both lists are used when filtering tweets related to COVID-19. To detect if a tweet is related to COVID-19, the occurrence of only one token from the strong list is enough, but at least two occurrences are needed from the weak list. When removing words related to COVID-19 (to create the subsets of the dataset used in the experiments below), only the strong list is used.

TABLE II: The strong key tokens

Token	Word (Eng.)
korona	corona
virus	virus
covid	covid
kovid	covid
karant	quarantine
izolac	isolation
ostanidoma	stayhome
ostanimodoma	stayhome
slusajstruku	listenprofession
slušajstruku	listenprofession
ostanimoodgovorni	stayresponsible

TABLE III: The weak key tokens

Token	Word (Eng.)	Token	Word (Eng.)
stožer	headquarters	virol	virology
dezinf	disinfection	distanc	distance
epide	epidemic	zaraz	infection
pandem	pandemic	vizir	visor
odgovorni	responsible	sars	SARS
hzjz	HZJZ	who	WHO
infekc	infection	lockd	lockdown
inkubacij	incubation	simpto	symptom
mask	mask	alemka	Alemka
bolnic	hospital	markoti	Markoti
doktor	doctor	vili	Vili
ljuskav	pangolin	beroš	Beroš
terapij	therapy	beros	Beroš
patoge	pathogens	capak	Capak
mjer	measure	prosvjed	protest
dijagnost	diagnose	šveds	Sweden
obrana	defense	festival	festival
rad od	work from	slobode	freedom
ostanimo	stay	ostani	stay
doma	home	struk	profession
kući	home	liječ	treat
respir	respirator	starač	nursing
samoizol	self-isolation	dom	home

The seven subsets of the datasets we experiment on are:

- the whole dataset of all collected tweets: cro-tw;
- the subset of tweets that are related to COVID-19: cro-CoV-tw;
- modification of cro-CoV-tw by removing strong key tokens related to COVID-19: cro-CoV-tw-remTok;
- modification of the cro-CoV-tw by removing random words: cro-CoV-tw-remRnd;
- tweets that are not related to COVID-19: cro-noCoV-tw;
- modification of cro-noCoV-tw by adding the word ("koronavirus", eng. coronavirus): cro-noCoV-tw-addCov;
- modification of cro-noCoV-tw by adding a random word: cro-noCoV-tw-addRnd.

B. Results

The spreadability prediction results for all the seven subsets of the dataset are shown in Table IV.

TABLE IV: Spreadability prediction results for all the seven subsets of the dataset. #Total is the total number of tweets in each subset, #Spreading is the number of tweets classified as highly-spreadable, and %Spreading is the percentage of tweets that are classified as highly-spreadable

	#Total	#Spreading	%Spreading
cro-tw	93,444	40,498	43.3%
cro-noCoV-tw	84,394	33,707	39.9%
cro-noCoV-tw-addCov	84,394	47,818	56.7%
cro-noCoV-tw-addRnd	84,394	33,394	39.6%
cro-CoV-tw	9,050	6,791	75.0%
cro-CoV-tw-remTok	9,050	6,424	71.0%
cro-CoV-tw-remRnd	9,050	6,743	74.5%

Tweets in the subset cro-noCoV-tw have a lower percentage of highly-spreadable tweets (39.9%) than the subset cro-tw (43.3%). After artificially adding COVID-19 context to every tweet in the subset cro-noCoV-tw, the resultant subset cro-noCoV-tw-addCov has a higher percentage at 56.7%, which would suggest that the COVID-19 context makes a tweet more spreadable.

To further test the hypothesis that the COVID-19 context increases the predicted spreadability of a tweet, we add a random word to each of the tweets in the subset cro-noCoV-tw, instead of the word related to COVID-19. At 39.6%, the resultant subset cro-noCoV-tw-addRnd has a similar percentage of highly-spreadable tweets as the subset cro-noCoV-tw, as would be expected if the hypothesis is correct.

In contrast to the subset cro-noCoV-tw that has 39.9% highly-spreadable tweets, 75.0% of the tweets in the subset cro-CoV-tw are highly-spreadable. From the subset cro-CoV-tw, strong key tokens related to COVID-19 are removed, which produces the subset cro-CoV-tw-remTok. 71.0% of the tweets in the new subset are highly-spreadable, agreeing with the hypothesis, as the percentage is lower after removal of the strong key tokens related to COVID-19. Predicted spreadability is high even after key token removal, suggesting that there is more to the COVID-19 context than the key tokens related to COVID-19.

To test if the removal of words in general is responsible for the decrease in the percentage of highly-spreadable tweets (instead of specifically the removal of key tokens related to COVID-19), we create one more subset cro-CoV-tw-remRnd. In the new subset, we remove the same number of words as is removed in the subset cro-CoV-tw-remTok, but the words are chosen randomly. The resultant percentage is 74.5%, which is close to the percentage of the subset cro-CoV-tw.

We further test the influence of additional words (other than "koronavirus") on the predicted spreadability of tweets. The results are shown in Table V.

TABLE V: Spreadability prediction results for the subset cro-noCoV-tw with added words

Word (Cro.)	Word (Eng.)	%Spreading
koronavirus	coronavirus	56.7%
mjere	measures	51.8%
zaraženi	infected	51.7%
virus	virus	50.0%
stožer	headquarters	49.3%
izbori	elections	48.7%
broj	number	47.2%
vlada	government	45.8%
cjepivo	vaccine	43.9%
karantenta	quarantine	40.0%
None		39.9%
danas	today	38.9%
dijete	child	38.4%
hrana	food	36.6%
hrvatska	croatia	36.5%
kuća	house	35.6%
turizam	tourism	34.8%
maska	mask	32.0%
politika	politics	31.8%
auto	car	29.6%
muzika	music	29.4%

IV. Discussion and Conclusion

Our model proved its ability to predict a tweet's spreadability only from its text with a 62.5% accuracy on the binary classification problem, which could be further improved with a better embedding method and model.

COVID-19 context seems to be a large influence on a tweet's spreadability. By training a model on more text related to COVID-19 we will be able to explore in more detail the nature of its influence in social networks.

In future work, we plan to use BERT [30] for embedding instead of the fastText embeddings we used in this paper. We plan to fine-tune BERT with texts related to COVID-19 and continue analyzing different aspects of COVID-19 related communication in Croatian social media.

With embeddings, it is possible to produce helpful visualizations of the text space, which we plan to use with BERT on COVID-19 texts for exploration.

References

[1] W. H. Organization *et al.*, "An ad hoc who technical consultation managing the covid-19 infodemic: call for action, 7-8 april 2020," 2020.

[2] V. Tangcharoensathien, N. Calleja, T. Nguyen, T. Purnat, M. D'Agostino, S. Garcia-Saiso, M. Landry, A. Rashidian, C. Hamilton, A. AbdAllah *et al.*, "Framework for managing the covid-19 infodemic: methods and results of an online, crowdsourced who technical consultation," *Journal of medical Internet research*, vol. 22, no. 6, p. e19659, 2020.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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. 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.

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

[12] 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.

[13] Q. Zhang, Y. Gong, Y. Guo, and X. Huang, "Retweet behavior prediction using hierarchical dirichlet process," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[14] Z.-y. Ding, Y. Jia, B. Zhou, Y. Han, L. He, and J.-f. Zhang, "Measuring the spreadability of users in microblogs," *Journal of Zhejiang University SCIENCE C*, vol. 14, no. 9, pp. 701–710, 2013.

[22] 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.

[15] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 2335–2338.

[16] Q. Zhang, Y. Gong, J. Wu, H. Huang, and X. Huang, "Retweet prediction with attention-based deep neural network," in *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 75–84.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[23] 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.

[24] 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.

[25] K. Babić, S. Martinčić-Ipšić, and A. Meštrović, "Survey of neural text representation models," *Information*, vol. 11, no. 11, p. 511, 2020.

[26] N. Ljubešić, "Word embeddings CLARIN.SI-embed.hr 1.0," 2018, slovenian language resource repository CLARIN.SI. [Online]. Available: http://hdl.handle.net/11356/1205

[27] K. Babić, S. Martinšić-Ipšić, A. Meštrović, and F. Guerra, "Short texts semantic similarity based on word embeddings," in *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin, 2019, pp. 27–33.

[28] K. Babić, F. Guerra, S. Martinčić-Ipšić, and A. Meštrović, "A comparison of approaches for measuring the semantic similarity of short texts based on word embeddings," *Journal of Information and Organizational Sciences*, vol. 44, no. 2, pp. 231–246, 2020.

[29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.