# COVID-19 related communication on Twitter: analysis of the Croatian and Polish attitudes

Karlo Babić[1,2] https://orcid.org/0000-0001-6343-0938, Milan Petrović[1] https://orcid.org/0000-0001-5302-9366, Slobodan Beliga[1,2] https://orcid.org/0000-0003-1407-6156, Sanda Martinčić-Ipšić[1,2] https://orcid.org/0000-0002-1900-5333, Andrzej Jarynowski[3](✉) https://orcid.org/0000-0003-0949-6674, and Ana Meštrović[1,2](✉) https://orcid.org/0000-0001-9513-9467

[1] University of Rijeka, Rijeka 51000, Croatia,
{karlo.babic,milan.petrovic,sbeliga,smarti,amestrovic}@uniri.hr,
WWW home page: https://infocov.uniri.hr/
[2] Center for Artificial Intelligence and Cybersecurity, University of Rijeka
[3] Interdisciplinary Research Institute
Wroclaw, Poland,
ajarynowski@interdisciplinary-research.eu

**Abstract.** In this paper, we analyze and compare Croatian and Polish Twitter datasets. After collecting tweets related to COVID-19 in the period from 20.01.2020 until 01.07.2020, we automatically annotated positive, negative, and neutral tweets with a simple method, and then used a classifier to annotate the dataset again. To interpret the data, the total number as well as the numbers of positive and negative tweets are plotted through time for Croatian and Polish tweets. The positive/negative fluctuations in the visualizations are explained in the context of certain events, such as the lockdowns, Easter, parliamentary elections, etc. In the last step, we analyze tokens by extracting the most frequently occurring tokens in positive or negative tweets and calculating the positive to negative (and reverse) ratios.

**Keywords:** COVID-19 · Twitter · Social media · Sentiment analysis · NLP

## 1 Introduction

Social media accelerates information spreading and may cause an infodemic, especially during a crisis. As stated by the 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. In this light, automatic classification of positive, neutral, and negative attitudes in social media and automatic detection and prediction of fake news spreading play an important role and may improve various aspects of crisis communication.

Sentiment analysis refers to the set of NLP based techniques used to identify, extract, or characterize opinions, emotions, and subjective information expressed using text [1]. Its main purpose is to classify attitudes related to various topics into positive, negative, or neutral. There are many possible applications of sentiment analysis [4, 12]. One very important application could be during natural disasters and emergencies when knowledge about positive and negative attitudes and sentiment, in general, could be used to improve situational awareness and crisis management. In the field of NLP, there have already been defined numerous approaches and techniques for sentiment analysis and opinion mining. However, the COVID-19 crisis brings a new set of challenges in terms of large communication volumes (massive datasets, new terminology, new aspects, and new specific topics that have come into focus).

As one of the most popular and used social networks, Twitter is one of the most studied networks in the domain of natural language processing (NLP) and social network analysis (SNA). There is a variety of tasks that ranges from tweet classification [17], hate speech detection [9], to link prediction [10, 11], etc.

Our work aims to deploy a multidisciplinary approach with a focus on quantitative analysis of empirical social media data from Twitter, allowing practical applications. Thus the methods of Digital (computational) Epidemiology [14] make it possible to analyze a huge amount of data (Big Data) at low cost which allows for the optimization of decision-making processes in public health in the context of COVID-19. We choose a comparative approach [15] between Poland and Croatia (both are post-communistic European countries), as political transformation and second demographic transition have common patterns. Moreover, the sociolinguistics perspective is more or less similar too (e.g., Polish and Croatian are in the same Slavic language group). In both countries, there was an election campaign in June 2020, which has impacted communications on Tweetosphere.

In this research, we deployed techniques from the field of NLP to analyze and compare Croatian and Polish COVID-19 tweets. First, we collected COVID-19 related tweets in the Croatian and Polish languages and performed automatic annotation of tweets using sentiment lexicons and a classifier. Next, we analyzed the distribution of positive and negative tweets over time. We found that there are some peaks of positive and negative tweets and try to explain them in the context of important events in Croatia and Poland. After that, we analyzed the most frequent tokens in positive and negative tweets in both datasets.

## 2 Background and Related Work

There are plenty of attempts around the challenge of media research in the COVID-19 era.

One of the most detailed and relevant studies about tweets' sentiment in the era of the pandemic is described in [19]. Xue et al. perform sentiment analysis of COVID-19 tweets by classifying each tweet into eight pairwise emotions: joy-sadness, trust-disgust, fear-anger, and surprise-anticipation. Their research is

based on the 1.9 million tweets related to coronavirus tweeted at the beginning of the pandemic. Moreover, the authors identify 11 topics using the LDA method and show that fear is the dominant emotion in all tweets. In [8] authors aim to examine trends of four emotions: fear, anger, sadness, and joy during the COVID-19 pandemic. Their results show that public emotions shifted from fear to anger during the pandemic. In [2] authors examine key themes and topics of COVID-19 related tweets in English at the beginning of the pandemic. All of these studies are focused mostly on the English language domain. Some research examines other languages and states as well, such as: [6], a study which analyzes the social dimension of the public health in Poland; [18] which examines irony on Twitter as the first response to the pandemic in Italy; and [13] which analyses sentiment of the tweets in Nepal.

## 3 Methodology

### 3.1 Datasets

We collect 3,945 tweets in the Croatian language and 989,004 tweets in the Polish language during the period of the first five months of the pandemic, from the 20th January to the 1st July of 2020. We filter relevant Croatian tweets by selecting the tweets that include COVID-19 or other related terms. For Polish tweets [5], we filter the relevant tweets with a selection criteria: tag #Koronawirus and language Polish.

The annotation method we use follows two main steps for both Croatian and Polish dataset: automatic annotation and trained annotation.

Automatic annotation uses a sentiment lexicon [3] (which contains positive and negative words) to score each tweet by calculating the number of positive words minus the number of negative words in a tweet. By doing so a tweet with more positive words than negative will have a positive score, and vice versa. A tweet is negative if the calculated score is smaller than -5, positive if the score is larger than 3 for Croatian tweets and larger than 8 for Polish tweets, and neutral otherwise. Words in the lexicons are first shortened by removing the last 20% of letters in words that have more than 4 letters. By doing so those words have a higher chance of matching the words in tweets.

Annotations produced with the automatic annotator are rough and the number of negative and positive tweets is very low. Because of the strict automatic annotator, tweets annotated as negative or positive are very likely in the right class. All tweets for which the automatic annotator is uncertain (a majority) are annotated as neutral.

Trained annotation is an iterative process of training a classifier:

i Initially, the classifier is trained on the annotated data produced with the automatic annotator.
ii After the initial training, it classifies the neutral tweets (as annotated by the automatic annotator) into positive, negative, and neutral.

iii After that step, the total number of positive and negative tweets is higher, and the classifier is trained again on the newly annotated data.
iv All tweets are then again annotated by the classifier.

For the classifier in the trained annotation step, we use Naive Bayes. We vectorized the data by using a bag-of-words approach, where singular words and word-pairs (two-grams) are used as units. After the two steps of annotation, the resulting dataset is described in Table 1.

**Table 1.** Annotated dataset statistics.

| | Total Tweets | Negative | Neutral | Positive |
|---|---|---|---|---|
| Croatian | 3,945 | 32.2% | 12.1% | 55.7% |
| Polish | 989,004 | 51.9% | 31.4% | 16.6% |

The annotation method is evaluated on datasets that are manually annotated. They have 99 tweets and 618 tweets for the Croatian and Polish dataset respectively. The number of negative, neutral, and positive tweets in those datasets is equal. Tweets with the same content as some of the tweets in the evaluation datasets are removed from the training datasets. The Croatian annotator achieves 50.5% accuracy, and the Polish annotator achieves 37.4% accuracy.

### 3.2 Data Analysis

We first analyze the number of COVID-19 related tweets per day during the period of 163 days for the Croatian and Polish language separately.

Next, we analyze the sentiment over time by using mean values and the normalized ratio of positive and negative tweets for both the Croatian and Polish dataset. We explain positive and negative peaks for certain days in the context of the social and political events related to the pandemic.

In the last step, we represent each tweet as a list of tokens. Tokens are sequences of characters separated by spaces or punctuation marks. As lemmatization or stemming did not improve annotations, tokens are not further processed. Then we analyze the most frequent COVID-19 related tokens that appear in positive and negative tweets separately.

## 4 Results

First, by showing the number of tweets per day, we can see when tweets about COVID-19 were trending (Figure 1). The first peak in the number of tweets about COVID-19 happened at the same time in Croatia and Poland, on the 35th day (25.02.2020). The biggest peak happened first in Poland, on the 50th day (11.03.2020), and Croatia followed on the 60th day (21.03.2020). Interest in coronavirus in Poland is similar to pan-European peak around 11-16.03.2020 [7],

however, Croatia is an exceptional case. In Poland, there was a sharp and steep increase in interest, while in Croatia attention was build week by week until it reaches the peak. The number of tweets about COVID-19 gradually got smaller, until the 150th day (19.06.2020) when in Croatia the number of COVID-19 related tweets begins to climb again. All the mentioned peaks correlate with the COVID-19 infection waves.

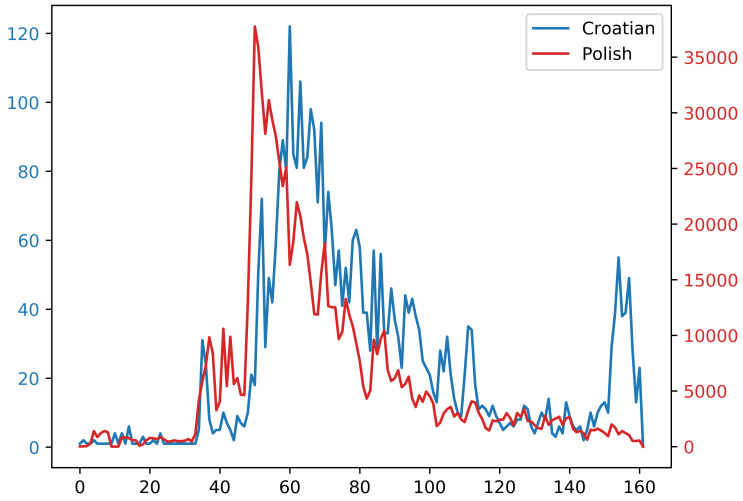


**Fig. 1.** The number of Croatian and Polish tweets about COVID-19 (blue and red respectively) per day calculated from 20.01.2020 until 01.07.2020.

### 4.1 Trends of Positive and Negative Attitudes in Tweets

By setting the values for negative, neutral, and positive sentiment to -1, 0, and 1, it is possible to calculate and visualize sentiment's behavior through time by using different methods.

As is visible in Figure 1, the number of tweets at the beginning of the pandemic and the end (except the very end which correlates with the second wave) is very low. Because of that, sentiment data is noisy at those periods.

For Figure 2, sentiment values are first averaged per day. That step normalizes the dataset so the differences in the number of tweets per day do not affect the visualizations. The values that are plotted are calculated by taking the average value of the sliding window (which is 7 days long). In this plot correlation of trend in sentiment change is visible. Both in Croatia and Poland sentiment was at first at its lowest, after which it begins to climb until the 45th day (06.03.2020), where it stabilizes.

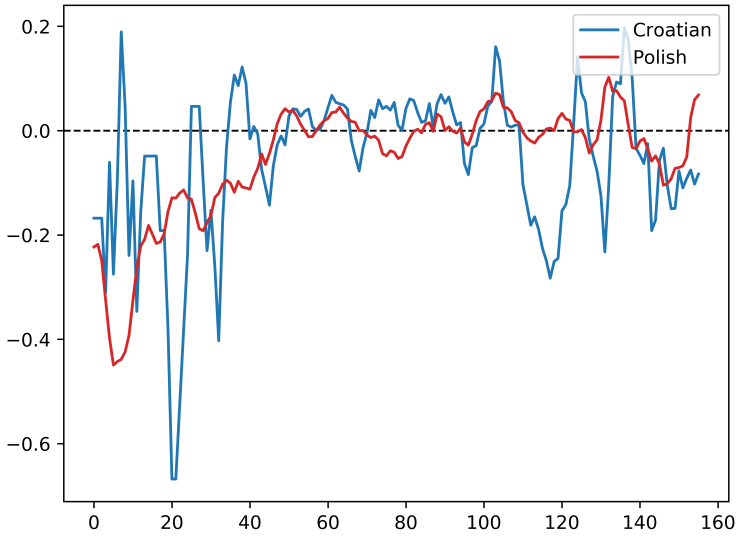


**Fig. 2.** Here we plotted the mean value for Croatian and Polish tweets (blue and red respectively) through time, aggregated by seven days (sliding window with length equal to seven days).

Figure 3 plots the ratio between the normalized number of positive and negative tweets (the values are first averaged with a seven-day long sliding window). By doing so, we can more clearly see the local correlations between Croatia and Poland.

The visible features in Figure 3 are:

- Day 33 (23.02. - 01.03. average): A positive peak, especially in Croatian tweets. At that time first COVID-19 cases were detected in Croatia, but it seems that people are ironic and in denial.
- Day 41 (02.03. - 09.03.): A negative peak. Croatians began to realize that COVID-19 is a serious threat and events started to cancel. First COVID-19 cases in Poland.
- Day 50 (11.03. - 18.03.): After the initial reaction, this positive peak in sentiment corresponds with the acceptance of the current status in both countries. Probably due to consolidation of the society and forming support networks among Internet users (more visible in the relative change in Poland).
- Day 57 (18.03. - 25.03.): A negative peak that correlates with the worst day in Italy so far. In Croatia, the number of infections rises and a destructive earthquake hits the capital of Croatia.
- Day 62 (23.03. - 30.03.): Sentiment has a positive peak as people started to post tweets supporting the healthcare system, at-risk groups, and respect measures.
- Day 69 (30.03. - 06.04.): A negative peak in Croatia that could be explained by the depression caused by isolation and the government closing the borders.

Polish tweets show a downward trend due to the lockdown restrictions being tightened and the consolidation effect has probably ended.
- Day 78 (08.04. - 15.04.): A negative peak, probably due to isolation at the time of Easter.
- Day 96 (26.04. - 03.05.): A negative peak in Poland potentially corresponding to decisions in schooling.
- Day 103 (03.05. - 10.05.): A positive peak correlating with almost no new COVID-19 cases in Croatia and restrictions being released (e.g., borders being opened in Poland).
- Day 115 (15.05. - 22.05.): A negative peak that was probably due to the Croatian parliamentary election and Polish presidential election campaigns.
- Day 122 (22.05. - 29.05.): A positive peak in Croatia as the number of COVID-19 cases is extremely low and it seems as if the pandemic ended.
- Day 129 (29.05. - 05.06.): A negative peak after a short period of optimism due to restriction releasing in Poland.
- Day 135 (04.06. - 11.06.): A positive peak due to another branch of restriction releasing in Poland.
- Day 146 (15.06. - 22.06.): A negative trend correlating with the second wave in Croatia. It could be related to the political campaign and polarization of societies before the Polish presidential elections on 28.06. and Croatian parliamentary election on 05.07.

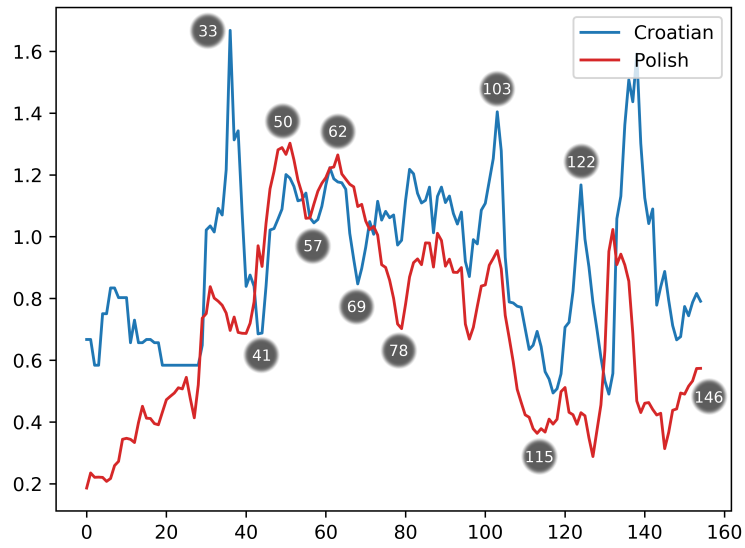


**Fig. 3.** Normalized ratio of number of positive to number of negative tweets for Croatian and Polish tweets (blue and red respectively) through time, aggregated by seven days (mean value from the sliding window with length equal to seven days).

### 4.2 COVID-19 Related Tokens in Positive and Negative Tweets

We analyze token frequencies in negative and positive tweets. In Table 2 and Table 3 tokens are sorted by their frequency of occurrence in positive and negative tweets for Croatian and Polish tweets respectively. Along with the frequency of tokens in positive and negative tweets, ratios are shown as well (e.g., token "masks" has a ratio of 1.43:1, which means that it occurs in 1.43 times more positive than negative tweets).

In Croatia, tweets directly mentioning COVID-19 or the pandemic are mostly negative. Political tweets are often negative as well. Tweets mentioning measures such as quarantine and masks are mostly positive.

**Table 2.** The most frequent tokens in positive (left) and negative (right) Croatian tweets translated to English. Ratios of positive to negative (left) and ratios of negative to positive (right) token occurrences. Ratios are normalized w.r.t. the total number of positive and negative tweets.

| N | Ratio | Token | N | Ratio | Token |
|---|---|---|---|---|---|
| 1028 | 0.73 | coronavirus | 818 | 1.38 | coronavirus |
| 294 | 0.78 | virus | 231 | 1.43 | virus |
| 283 | 1.1 | corona | 189 | 2.14 | covid19 |
| 270 | 1.2 | stayhome | 168 | 1.94 | corona |
| 255 | 0.86 | covid19 | 149 | 1.5 | self-isolation |
| 209 | 1.04 | measures | 130 | 0.83 | stayhome |
| 208 | 1.81 | quarantine | 118 | 1.37 | viliberos |
| 182 | 2.38 | croatia | 112 | 1.9 | croatia |
| 178 | 0.72 | self-isolation | 107 | 0.99 | measures |
| 153 | 2.18 | isolation | 105 | 1.28 | ravnateljstvocz |
| 149 | 0.73 | viliberos | 93 | 1.4 | people |
| 142 | 0.78 | ravnateljstvocz | 80 | 1.28 | vladarh |
| 108 | 0.78 | vladarh | 69 | 0.6 | quarantine |
| 94 | 1.43 | masks | 61 | 1.53 | thanks |
| 85 | 0.95 | headquarters | 60 | 1.12 | time |
| 83 | 1.1 | zagreb | 59 | 9.11 | pandemic |
| 79 | 6.95 | work | 53 | 13.74 | county |
| 69 | 0.65 | thanks | 52 | 1.06 | headquarters |
| 65 | 1.04 | home | 47 | 1.52 | zagreb |
| 57 | 1.27 | help | 46 | 1.66 | mup |
| 54 | 0.89 | pandemic | 45 | 3.71 | hdz |
| 53 | 1.8 | protection | 44 | 1.58 | question |
| 52 | 0.51 | infected | 42 | 1.51 | davorbozinovic |
| 49 | 1.35 | crisis | 42 | 9.1 | test |
| 48 | 0.6 | mup | 40 | 1.82 | dnevnikhr |

In Poland, negative tweets are frequently associated with reporting cases, deficiency in the healthcare system, economic crisis, or elections. Positive tweets are frequently associated with anti-crisis, social mobilization, and support.

**Table 3.** The most frequent tokens in positive (left) and negative (right) Polish tweets translated to English. Ratios of positive to negative (left) and ratios of negative to positive (right) token occurrences. Ratios are normalized w.r.t. the total number of positive and negative tweets.

| N | Ratio | Token | N | Ratio | Token |
|---|---|---|---|---|---|
| 162037 | 1.01 | coronavirus | 56784 | 0.86 | government |
| 42291 | 5.47 | people | 52595 | 1.77 | coronavirus in poland |
| 39266 | 235.18 | voivodeship | 502491 | 0.99 | coronavirus |
| 36822 | 16.18 | cases | 47429 | 0.94 | poland |
| 32346 | 30752.57 | confirmed | 39473 | 0.9 | epidemic |
| 31776 | 5.35 | tests | 39363 | 0.92 | law&justice |
| 28351 | 1.38 | government | 38549 | 1.74 | covid19 |
| 23936 | 19.21 | infections | 30077 | 1.39 | minister |
| 23757 | 8.94 | new | 27182 | 0.88 | today |
| 21487 | 67179.23 | lab | 25536 | 0.53 | health |
| 18524 | 57915.4 | result | 24163 | 0.18 | people |
| 18491 | 57812.22 | positive | 22308 | 1.43 | time |
| 15424 | 1.89 | health | 18556 | 0.19 | tests |
| 13739 | 1.09 | law&justice | 18340 | 1.83 | polish |
| 11438 | 0.93 | just | 18152 | 2.43 | know better |
| 9830 | 1.13 | day | 17820 | 0.9 | may |
| 9390 | 13.66 | issue | 17795 | 1.16 | morawiecki |
| 8860 | 27700.84 | related | 16433 | 1.13 | lukasz wzumowski |
| 16370 | 1.09 | poland | 15751 | 1.25 | pandemic |
| 7645 | 0.99 | stay at home | 15428 | 1.64 | million |
| 7236 | 1.24 | epidemic | 14844 | 1.44 | days |
| 6480 | 20259.76 | lower silesia | 14586 | 0.86 | relationship |
| 6422 | 1.54 | poles | 14266 | 2.0 | thousand |
| 6420 | 3.7 | life | 27644 | 0.81 | fight |
| 6401 | 23.41 | safety | 13696 | 1.26 | andrzej duda |

## 5 Discussion and Conclusion

This is a comparative approach to present Twitter discourse on COVID-19, in Poland and Croatia, in the first wave of COVID-19 (from 20.01. to 01.07.).

The interest in the COVID-19 topic on Twitter in Europe is primarily of a social, not medical, dimension and the peaks of interest and changes in sentiment are mainly driven by social or political issues. Tweets about COVID-19 in Poland and Croatia have the same intensity patterns (Figure 1) as in many other

countries [7], however peak of interest in Croatia is a week later than in most of the European countries. Our study is confirming that interest in COVID-19 on Twitter is unrelated to the actual physical risk of acquiring COVID-19 in Poland (official disease prevalence) which could be mostly the case corresponding to Google Trend Studies [7,16], but it is slightly correlated with the incidence in Croatia. Sentiment analysis (Figure 3 revealed positive attitudes during Stay at Home consolidation and the support phase during the second part of March and negative attitudes during political campaigns in late June. Positive/negative tweet's ratio (Figure 3) is a more informative factor than average sentiment (Figure 2).

For future work, we plan to expand the datasets by using a longer range of dates, and by training a classifier that can filter more precisely tweets related to COVID-19. Comparison of sentiment trends for COVID-19 related tweets and tweets that are not related to COVID-19 could provide us with a new understanding. Further, we can examine semantic content by analyzing the distribution of topics through time.

## 6 Acknowledgment

This work has been supported in part by the COST Action CA15109 COSTNET and 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).

## References

1. Beigi, G., Hu, X., Maciejewski, R., Liu, H.: An overview of sentiment analysis in social media and its applications in disaster relief. In: Sentiment analysis and ontology engineering, pp. 313–340. Springer (2016). DOI 10.1007/978-3-319-30319-2_13
2. Chandrasekaran, R., Mehta, V., Valkunde, T., Moustakas, E.: Topics, trends, and sentiments of tweets about the covid-19 pandemic: Temporal infoveillance study. Journal of medical Internet research **22**(10), e22,624 (2020). DOI 10.2196/22624
3. Chen, Y., Skiena, S.: Building sentiment lexicons for all major languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 383–389 (2014). DOI 10.3115/v1/P14-2063
4. Jakopović, H., Mikelić Preradović, N.: Identifikacija online imidža organizacija temeljem analize sentimenata korisnički generiranog sadržaja na hrvatskim portalima. Medijska istraživanja: znanstveno-stručni časopis za novinarstvo i medije **22**(2), 63–82 (2016). DOI 10.22572/mi.22.2.4
5. Jarynowski, A.: A dataset of media releases (twitter, news and comments, youtube, facebook) form poland related to covid-19 for open research
6. Jarynowski, A., Wójta-Kempa, M., Płatek, D., Czopek, K.: Attempt to understand public health relevant social dimensions of covid-19 outbreak in poland. Available at SSRN 3570609 (2020). DOI 10.2139/ssrn.3570609
7. Lampos, V., Moura, S., Yom-Tov, E., Cox, I.J., McKendry, R., Edelstein, M.: Tracking covid-19 using online search. arXiv preprint arXiv:2003.08086 (2020)

8. Lwin, M.O., Lu, J., Sheldenkar, A., Schulz, P.J., Shin, W., Gupta, R., Yang, Y.: Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. JMIR public health and surveillance **6**(2), e19,447 (2020). DOI 10.2196/19447
9. Markoski, F., Zdravevski, E., Ljubešić, N., Gievska, S.: 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). DOI 20.500.12188/8269
10. Martinčić-Ipšić, S., Močibob, E., Meštrović, A.: Link prediction on tweets' content. In: International Conference on Information and Software Technologies, pp. 559–567. Springer, Berlin, Germany (2016). DOI 10.1007/978-3-319-46254-7_45
11. Martinčić-Ipšić, S., Močibob, E., Perc, M.: Link prediction on twitter. PloS one **12**(7), e0181,079 (2017). DOI 10.1371/journal.pone.0181079
12. Načinović, L., Perak, B., Meštrović, A., Martinčić-Ipšić, S.: Identifying fear related content in croatian texts. In: Proceedings of the Eighth Language Technologies Conference, pp. 153–156 (2012)
13. Pokharel, B.P.: Twitter sentiment analysis during covid-19 outbreak in nepal. Available at SSRN 3624719 (2020). DOI 10.2139/ssrn.3624719
14. Salathé, M.: Digital epidemiology: what is it, and where is it going? Life sciences, society and policy **14**(1), 1 (2018). DOI 10.1186/s40504-017-0065-7
15. Strzelecki, A., Azevedo, A., Albuquerque, A.: Correlation between the spread of covid-19 and the interest in personal protective measures in poland and portugal. In: Healthcare, p. 203. Multidisciplinary Digital Publishing Institute (2020). DOI 10.3390/healthcare8030203
16. Szmuda, T., Ali, S., Hetzger, T.V., Rosvall, P., Słoniewski, P.: Are online searches for the novel coronavirus (covid-19) related to media or epidemiology? a cross-sectional study. International Journal of Infectious Diseases (2020). DOI 10.1016/j.ijid.2020.06.028
17. Tutek, M., Sekulić, I., Gombar, P., Paljak, I., Čulinović, F., Boltužić, F., Karan, M., Alagić, D., Šnajder, J.: 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), pp. 464–468 (2016). DOI 10.18653/v1/S16-1075
18. Vicari, S., Murru, M.F.: One platform, a thousand worlds: On twitter irony in the early response to the covid-19 pandemic in italy. Social Media+ Society **6**(3), 2056305120948,254 (2020). DOI 10.1177/2056305120948254
19. Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., Zhu, T.: Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. PloS one **15**(9), e0239,441 (2020). DOI 10.1371/journal.pone.0239441