**IZDANJA FILOZOFSKOG FAKULTETA SVEUČILIŠTA U SPLITU**
**EDITIONES FACULTATIS PHILOSOPHICAE UNIVERSITATIS SPALATENSIS**

**SVEUČILIŠTE U RIJECI, FILOZOFSKI FAKULTET,**
**CENTAR ZA JEZIČNA ISTRAŽIVANJA**

**Nakladnik**
Sveučilište u Splitu, Filozofski fakultet, Poljička cesta 35, 21000 Split
Sveučilište u Rijeci, Filozofski fakultet, Sveučilišna avenija 4, 51000 Rijeka

**Odgovorne urednice**
izv. prof. dr. sc. Gloria Vickov, dekanica Filozofskoga fakulteta u Splitu
izv. prof. dr. sc. Ines Srdoč-Konestra, dekanica Filozofskoga fakulteta u Rijeci

**Recenzentice knjige**
prof. dr. sc. Aneta Stojić, Sveučilište u Rijeci
izv. prof. dr. sc. Maja Bezić, Sveučilište u Splitu

**Recenzenti pojedinačnih priloga**
Melita Aleksa Varga, Tatjana Balažic Bulc, Gorana Bandalović, Ivana Bašić, Nina Begičević Ređep, Branka Drljača Margić, Darko Hren, Cecilija Jurčić Katunar, Nejla Kalajdžisalihović, Danijela Marot Kiš, Nives Mikelić Preradović, Kristian Novak, Ivana Petrović, Anita Skelin Horvat, Anđel Starčević

**Korektura**
Trišnja Pejić (engleski), Anastazija Vlastelić (hrvatski)

**Grafičko oblikovanje i tisak**
Redak d. o. o., Split

**Naklada:** 30 primjeraka

**ISBN:** 978-953-352-071-1 (tisak), 978-953-352-072-8 (online)
**ISBN:** 978-953-361-040-5 (tisak), 978-953-361-041-2 (online)

Odobrilo Vijeće Filozofskoga fakulteta Sveučilišta u Splitu odlukom od 15. rujna 2021. (Klasa 003-08/21-06/0014; Ur. broj: 2181-190-00-21-0057).
Odobrilo Vijeće Filozofskoga fakulteta Sveučilišta u Rijeci odlukom od 24. rujna 2021. (Klasa 612-10/21-01/19; Ur. broj: 2170-24-01-03-21-2).
.

Urednice
Magdalena Nigoević ◆ Anastazija Vlastelić

# ODJECI SCIMETH-a

## (izazovi lingvističkih istraživanja)

Sveučilište u Splitu, Filozofski fakultet
Sveučilište u Rijeci, Filozofski fakultet, Centar za jezična istraživanja
Hrvatsko društvo za primijenjenu lingvistiku
Split – Rijeka, 2021.

# Sadržaj

**IV. Primijenjena istraživanja**

Izvorni znanstveni rad
UDK 811.163.42’27:616.2-036.21
004.738.5:811.163.42’27

Slobodan Beliga[1], Ana Meštrović[1, 2] i Mihaela Matešić[2, 3]
[1]University of Rijeka, Department of Informatics Rijeka, Croatia
[2]University of Rijeka, Center for Artificial Intelligence and Cybersecurity, Rijeka, Croatia
[3]University of Rijeka, Faculty of Humanities and Social Sciences, Rijeka, Croatia
sbeliga@inf.uniri.hr, amestrovic@inf.uniri.hr, mmatesic@ffri.uniri.hr

# NLP based framework for the comparison of the media coverage in Croatia during the first two waves of the COVID-19 pandemic

Online media play an important role in public health emergencies and serve as a communication platform. Infoveillance of online media during the COVID-19 pandemic is an important step towards a better understanding of crisis communication. The goal of this study was to perform a longitudinal analysis of the COVID-19-related content based on natural language processing methods. We present a possible framework for monitoring media coverage of crisis communication. For this purpose, we collected a dataset of news articles published by Croatian online media during the first 13 months of the pandemic. As the first step, we calculated the percentage of COVID-19-related articles in the total number of articles across eight online news media for different periods of the pandemic. The second step was to analyze the content by extracting the most frequent terms and applying the Jaccard similarity. Next, we compared the occurrence of the pandemic-related terms during the two waves of the pandemic. Finally, we applied named entity recognition to extract the most frequent entities and track the dynamics of changes during the observed period.

The results reveal that the online media have promptly responded to the pandemic with a large number of COVID-19-related articles. The total number of COVID-19-related articles in online media is rather high – even in the period between the two waves of the epidemic, when the number of new cases dropped to zero, the number of publications related to the

COVID-19 topics remained high. Furthermore, there are large overlaps in the terminology used in all articles published during the pandemic with a slight shift in the pandemic-related terms between the first and the second wave. Finally, our findings indicate that the most influential entities have lower overlaps for identified persons and higher overlaps for locations.

**Keywords:** online news media, infoveillance, crisis communication, natural language processing

## 1. Introduction

Media coverage plays an important role in responses to major crises such as the pandemic caused by the SARS-CoV-2 virus. During a crisis, the need for information exceeds normal levels and people try to gain access to information as soon as possible. As a result, online media serve as a key communication platform (Glik 2007). However, online media also have a negative side effect which is described as infodemic (Eysenbach 2002, 2009, 2011). The outbreak of the COVID-19 pandemic caused an information overload which was declared to be an infodemic and described as being potentially very dangerous (World Health Organisation, WHO 2020).

We have already analyzed media and social media content published in the Croatian language in our previous papers, but taking into concern some shorter time periods (Babić et al. 2020; Babić, Petrović, Beliga, Martinčić-Ipšić, Pranjić, et al. 2021; Babić, Petrović, Beliga, Martinčić-Ipšić, and Meštrović 2021; Bogović et al. 2021; Beliga et al. 2022 in press). In this paper, we focus exclusively on the news media space and explore the eight most representative online news media by scrutinizing their publications in the period from January 1st, 2020 to January 15th, 2021.

This research was motivated by the idea that the first step towards a better understanding of the situation with the COVID-19 infodemic is to perform a quantitative analysis of news articles published in online news media. Thus, the main goal of this study is to provide an overview of the trends concerning COVID-19-related articles published in online news media during the first 13 months of the pandemic in Croatia.

## 2. Related work

There is a large number of studies that examine communication in social media. Majority of these are based on Twitter datasets. However, there are only a few research papers that analyze more than one social network. In a study that addressed the problem of the COVID-19 infodemic in social media, Cinelli et al. (2020) analyzed the diffusion of information about COVID-19 on Twitter, Instagram, YouTube, Reddit and Gab and their findings revealed that there are no substantial differences between fake and true news spreading, only that the amount of fake news varies across platforms. Zarocostas (2020) elaborates that during a tsunami of information, with the amplification of spreading in social media, information spreads faster and further through the information space. He emphasizes that incorrect information is also spread through the traditional media. Cuomo et al. (2020) analyzed geospatial and longitudinal distributions of Twitter messages about the COVID-19 posted between March 3rd and April 13th 2020 and compared these results with the number of confirmed cases reported for sub-national levels in the United States, while Pulido et al. (2020) investigated the type of tweets that circulated on Twitter around the COVID-19 outbreak for two days, in order to analyze false and true information diffusion. According to Pulido et al.'s (2020) results, false information is tweeted more but retweeted less than science-based evidence or fact-checking tweets, while science-based evidence and fact checking tweets capture more engagement than mere "facts".

Generally, the infodemic concerning COVID-19 has been widely studied, primarily on Twitter, while studies of other platforms (such as Reddit, YouTube, Facebook, etc.) have not been carried out to such an extent. The majority of studies are focused on social networks and, as has already been mentioned, the most studied social network is Twitter. There are studies in which authors performed topic modelling of Tweets, sometimes combined with sentiment analysis (Jelodar et al. 2020; Lwin et al. 2020; Xue et al. 2020; Jang et al. 2021; Zhou et al. 2021; Rustam et al. 2021; Chakraborty et al. 2020; Satu et al. 2021; Kaur et al. 2021), and yet another direction in analysis of COVID-19 tweets is detection of fake news (Paka et al. 2021).

Some studies have focused on the issue of the COVID-19-related infodemic. Gallotti et al. (2020) assessed the infodemic risks from tweets in 64 languages collected during the period between January and March of 2020. The authors proposed a news reliability and infodemic risk index to monitor the exposure to the infodemic around the world. Their findings confirm that the epidemic and

infodemic co-evolve and that the infodemic risk exposure can be high regardless of the stage of the development of the country in question. They showed that the escalation of epidemic is accompanied by a shift of public attention toward more reliable sources. Mackey et al. (2020) conducted an infoveillance study of Twitter and Instagram posts on potential information about questionable sellers of COVID-19-related products such as "self-help remedies" and treatments. Their results obtained by topic modelling and classification are visualized on an interactive dashboard for better monitoring, removing and preventing of harmful content. One example of an extensive study of the infodemic in social media is that by Cinelli et al. (2020). In this study Cinelli et al. analyzed the diffusion of information related to the COVID-19 on different social media platforms and they tried to explain information spreading by using epidemic spreading models by identifying the R0 for each platform. Additionally, they show that there are no substantial differences between fake and true news spreading, only that the amount of fake news varies across platforms.

Ghafarian and Yazdi (2020) propose a model for understanding the effects of information seeking, information sources, information overload and the consequent information avoidance during the global health crisis caused by the COVID-19 disease. Their results reveal that among different information sources, social media exposure is significantly related to information overload. Bunker (2020) analyzed various platforms with the main goal of characterizing digital disruption and described the propagation of misinformation by recommender algorithms, bots and trusted individual platform users. The author concludes that the COVID-19 infodemic highlights the current problem confronting the information system discipline.

There is a somewhat smaller number of studies focused on the COVID-19 content published in the online news media. One of these describes a comparative linguistic analysis of headlines from serious and sensationalist newspapers in the UK (Almazán-Ruiz and Orrequia-Barea 2020). The authors constructed a corpus based on headlines from broadsheets and tabloids published over the period of one month at the beginning of the pandemic and performed an automatic analysis of the text and basic statistics in order to compare the headlines that appeared in broadsheets and tabloids. They concluded that tabloids included more instances of sensationalist features. Gozzi et al. (2020) studied the media coverage and collective internet response in the UK, USA, Canada and Italy in news articles, Wikipedia page views, Reddit posts and YouTube videos on official channels of major news organizations in the period between February and May 2020. Their findings revealed that the detected topics were aligned between news services and internet users, and that the collective attention was predominantly driven by media coverage rather than by the progression of the pandemic.

As can be seen, the majority of studies of the COVID-19-related content in media analyzed content published on social media, while content published in online news media tended to be studied to a lesser extent. Moreover, the analyzed datasets consisted of texts published at the beginning of the pandemic (the first wave of the pandemic), or the studies were based on datasets covering the time span of three to four months. Thus, although the COVID-19-related media content has been widely studied there is still room for improvement. Some possible extensions of the existing research directions are to include longitudinal data and/or to perform analyses over multiple internet-based data sources, which include both online news media and social media.

## 3. Methodological framework

The goal of this study is to analyze and compare the number and the content of the online media news articles published in Croatia during the first two waves[1] of the COVID-19 pandemic. Hence, we propose a framework for the comparison of media coverage of the pandemic in Croatia during the observed period based on natural language processing (NLP) techniques. In the first step we analyze the number of COVID-19 related articles. In the second, we analyze and compare the most frequent terms and entities and analyze how the main terminology has changed between the two waves. In the third step we analyze Google trends related to the COVID-19 terminology.

### 3.1. *Dataset*

The collected data cover the period of the first two waves of the pandemic (the period from January 1st, 2020 to January 15th, 2021). Data acquisition was carried out in such a way that all newspaper publications from eight online news media, covering the geographical and media space of the Republic of Croatia, were collected on daily basis. In the observed period, the number of collected

[1] An epidemic wave is a graph that tracks the number of people suffering from a disease over time. Epidemics usually begin with a sharp increase in the number of patients over a short time-period. That number then reaches a peak, after which it begins to decline until there are no new cases. Some epidemiological experts state that an end of an epidemic (epidemic wave) can be declared only if there are no new cases in a population for a certain number of days (e.g., 14 days). The definition of a second wave is that it occurs after the first wave has ended and after a certain period has passed in between. In the case of this study, there was no complete cessation for a period of fourteen days without a single case of infection in Croatia. However, there was a lull of 25 days in which the number of new cases was occasionally equal to zero, and occasionally one or two new cases occurred in the period from May 25th, 2020 to June 16th, 2020. The official date delimiting the two epidemic waves is not defined. Therefore, in this study we determined that the first wave lasted until May 15th, 2020. After that followed a period during which the number of new cases was greater than or equal to three. Due to the new sharp increase in the number of new cases, we define June 15th, 2020 as the beginning of the second wave of the epidemic.

articles from the selected eight online news media amounted to 270,359, while the number of COVID-19-related articles amounted to 121,095.

The filter used to determine the affiliation of a text to a class related to the COVID-19 was the mention of keywords related to the pandemic in the title, subtitle or the body of the text of the news article. Since the online media referred to COVID-19 disease in a range of terms (such as: *corona*, *korona*, *corona-virus*, *korona-virus*, *corona virus*, *korona virus*, *koronavirus*, *coronavirus*, *SARS-CoV-2*, *covid*, *COVID*, *COVID-19*, etc.) at the pandemic breakout, we used all of those variations as triggers to detect COVID-19 related news. For the purposes of further analysis of media content, the texts were pre-processed: (i) only the textual part of the news was retained (related images and videos were omitted) and (ii) all the titles, subtitles and body texts from all the collected news were lemmatized (to reduce morphological variations and make the text more suitable for analysis).

The selection of eight sources[2] used in the study was based on the following criteria: 1) well-known, most popular, viral online news media in Croatia; 2) covering the Republic of Croatia as a whole (i.e., sources that are narrowly oriented to regional news were excluded); 3) reporting on standard categories of news: basic daily news, business, economy, politics, crime, sports, entertainment, art, show business, health, religion, technology, etc. Additionally, it was important to cover various aspects of news publications in terms of (1) political orientation, (2) authorship, and (3) publication form. The first characteristic refers to the character of the online news media. What is meant by this is the attitudes about the world and about everything that surrounds us in a philosophical and practical sense through the lens of the editorial board and journalists of a particular online news medium. In that sense, we included online news media which are considered to belong to the yellow press or tabloids, media publishers that favor a variety of political stances (from the left through the center to the far-right political spectrum), independent online news media, but also those whose purpose is to entertain. In addition to the aspect of political orientation, authorship was also taken into account. Some of the selected online news media publish news written exclusively by the journalists they employ, but we have also covered four online news media in which editors select and publish texts written by their readers. In the case of such media, readers can send in photos and recorded video clips accompanied by a description of the current event, which the editorial office can officially select and publish, and sometimes even reward with a symbolic

[2] We omit the real names of the select online news media in accordance with the Electronic Media and the Copyright Law in Croatia.

cash prize. The third aspect is the form of publication. All of the eight online news media are published in electronic form. In addition, three of them are also published in print. Printed forms differ in the scope and frequency of publication from the electronic ones. Taking all this into consideration, we believe that such a data set, encompassing a large number of heterogeneous online news publications, constitutes a representative sample of news articles in the Croatian online media space during the observed epidemic period.

### 3.2. *Terminology extraction and named entity recognition*

In the second step we performed extraction of the most frequent terms and of the most frequent entities – persons (PER), locations (LOC), institutions (ORG) and miscellaneous (MISC) using Named Entity Recognition.

Named Entity Recognition (NER) is an NLP task aimed at extraction of named entities such as persons, locations, organizations, numeric expressions such as time, money, dates, etc. NER extraction can be modeled as a text sequence annotation problem. In this case Conditional Random Fields (CRF), as a non-directed graphical model, are trained to maximize the log likelihood, calculated from conditional probabilities of output labels sequence over the features of the input sequence and CRF states. In this study we used the NER system for Slovene, Croatian and Serbian (Fišer, Ljubešić, and Erjavec 2020; Ljubešić et al. 2013) trained on Slovene (ssj500k), Croatian (hr500k) and Serbian (SETimes.SR) corpora. The implemented NER is a slight modification of the CRF-based reldi-tagger with Brown clusters information added, capable of the recognition of entities.

We compared the trends concerned with changes in keywords during the time period covering 13 months and across the two waves by using the measure of Jaccard similarity that indicates overlapping keywords between two different time periods. Generally, the Jaccard index, also known as the Jaccard similarity coefficient, is a statistical measure used for expressing the similarity and diversity between sample sets. For two sets A and B, the Jaccard coefficient is defined as the size of the intersection divided by the size of the union of sets A and B:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

Note that by design, $0 \leq J(A,B) \leq 1$. If $A$ and $B$ are both empty, then we define $J(A,B) = 1$.

### 3.3. *Google trends in infoveillance*

Additionally, we checked the trends of searches on Google by using *Google trends*. *Google trends* enables analysis of the popularity of top search queries in *Google search* across various regions and languages. More precisely, *Google trends* uses graphs to compare the search volume of different queries over time. Furthermore, it is possible to compare the relative search volume of searches between two or more terms.

## 4. Results and discussion

### 4.1. *Description of the Online Newspaper Space*

The total number of COVID-19-related articles in online media is rather high. The ratio of COVID-19-related articles does not fall below 44% in any of the observed online news media. If we observe the average ratio across all explored online news media, COVID-19-related publications cover more than the half of the media space (about 57%). If we randomly read the news published in one of the eight observed media, this means that by tossing a coin on which the heads represents a topic related to the COVID-19, and the tails represents any other topic, we would have a 7% higher chance of getting to read a piece of news related to COVID-19.

Figure 1 shows the percentage of COVID-19 articles in the total number of published articles, summarized through all observed online news media, separately for the two pandemic waves and the period in which the epidemic subsided as well as the entire observed period from January 1st, 2020 to January 15th, 2021. In order to observe the pandemic picture during the entire 2020, data from January 1st to February 25th, 2020 is also observed despite the fact that there were no cases of coronavirus infection in Croatia at that time. Percentages of COVID-19-related articles were lower during this period (about 43 to 45%), i.e., very little was written about the COVID-19 topic in that period because, as already mentioned, there were no cases of infection in Croatia. However, in the period between the two waves of the epidemic, when the number of cases of infection dropped to zero, the number of publications related to the corona topics remained at a high 43%.

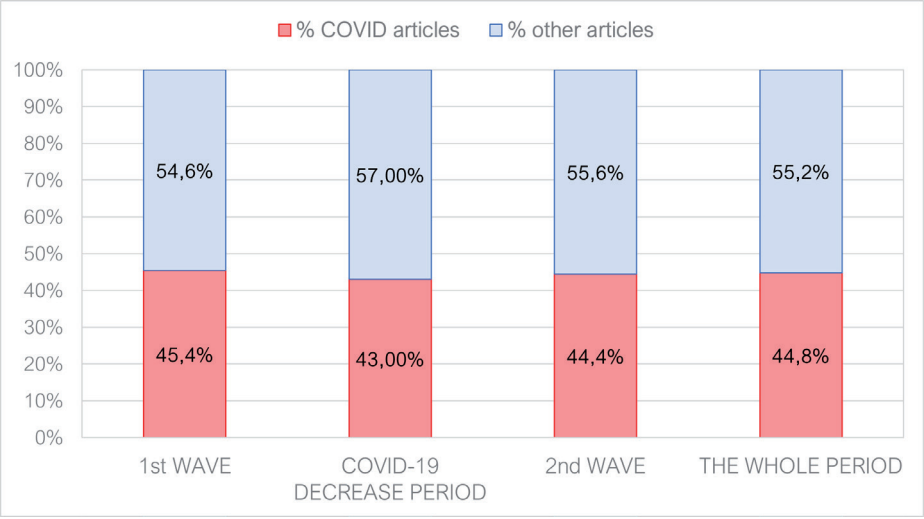


**Figure 1.** The percentage of COVID-19 related articles in the total number of articles summarized across eight online news media for different periods of the pandemic.

### 4.2. *Analysis of the most frequent terms and entities*

The analysis of the most frequent terms was performed in the first step at the level of pandemic waves. We analyzed the list of the top 250 most frequently used terms during the first and the second wave of the epidemic. Finally, the prevalence of pandemic terminology during the first and the second wave of the epidemic was quantified. Figure 2 displays the data for the nine most frequent terms related to the COVID-19 situation (the same terms were also explored in the next step of the proposed framework using *Google trends*).

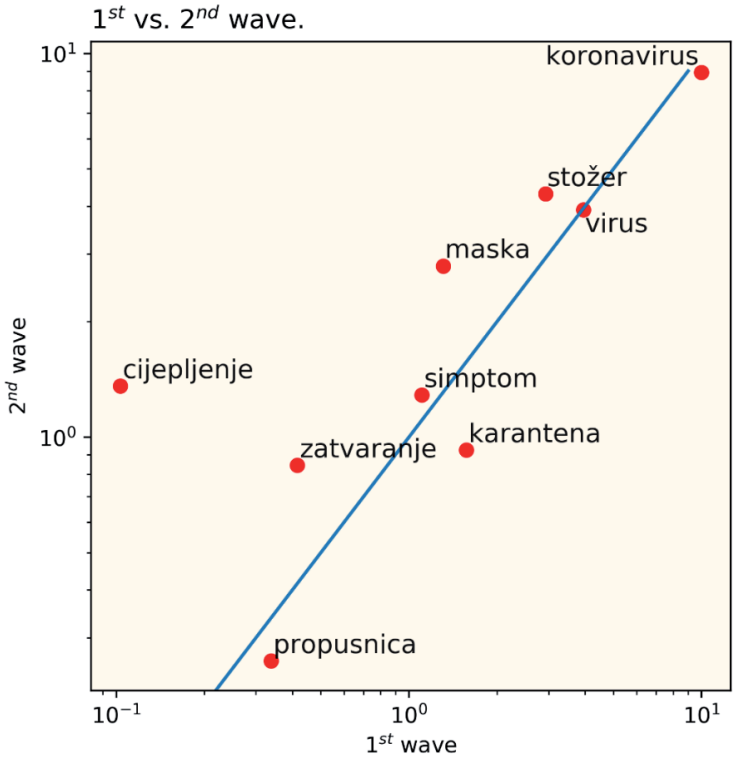


**Figure 2.** Comparison of the frequencies in the pandemic-related terminology across the two waves: virus (*virus*), coronavirus (*koronavirus*), symptoms (*simptomi*), vaccine (*cjepivo*), quarantine (*karantena*), lockdown (*zatvaranje*), headquarters (*stožer*), passes (*propusnice*), masks (*maske*)

The further away from the wave-dividing boundary a term is located, the more its frequency differs between the two waves. As can be seen in Figure 2, 'vaccination' (*cijepljenje*) was predominantly mentioned during the second wave because the expectations of an imminent discovery and successful production of a vaccine became more realistic (and finally real) during this period. The terms 'quarantine' (*karantena*) and 'pass' (*propusnica*) typically mark the first wave. During the first wave, quarantine was the main means of achieving the goal of reducing the spread of the epidemic, but during the second wave the same goal was approached differently, by selective lockdown, and this is the reason why the word 'lockdown' was more frequent during the second wave. The passes that ensured limited movement between municipalities during the first wave, were, during the second wave, at first mandatory for movement between counties, but were soon abolished (because of the severe earthquake which struck the area of Sisak and Petrinja on December 29th, 2020). 'Symptoms',

'virus' and 'coronavirus' were almost equally frequently mentioned during both waves, and 'mask' and 'headquarters' were more frequent during the second wave.

In order to better capture named entity recognition (NER) trends during the pandemic we have calculated the Jaccard similarity between the two pandemic waves. Here, we focused on the 250 most frequent entities per entity type: person (PER), location (LOC), institutions (ORG) and miscellaneous (MISC) and observed their overlap between the two epidemic waves. The results of the analysis are shown in Table 1.

**Table 1.** Jaccard similarity between the two epidemic waves for TOP 250 most frequent entities per type.

| ENTITY TYPE | Jaccard index |
|---|---|
| **PER** | 0,4045 |
| **LOC** | 0,5337 |
| **ORG** | 0,4793 |
| **MISC** | 0,3333 |

Jaccard similarity quantifies the similarities or differences between the presence of entities used in the news during different pandemic waves. In this case, the largest overlaps were found for the location entity type, slightly smaller ones for institutions, then for persons, and the smallest for the miscellaneous category. In daily news, locations were mostly constant during the pandemic waves (resulting in a low number of total locations). The results indicate that the focus was on a narrowed area restricted to Croatia, the neighboring countries, the EU and international locations connected with the COVID-19 crisis (Wuhan, Lombardy, etc.). This reflects the fact that the countries had closed their borders in order to limit the movement of people as a measure of precaution against spreading the disease. Predominant location entities during the whole period are Croatian cities and regions. When talking about persons mentioned in the media, it is interesting that the news is dispersed across many persons participating in daily events (interestingly, politicians were still more frequently mentioned than scientists in spite of the increased interest in the scientists' interpretations of the epidemic crisis). Among organizations, the focus was on the WHO, local infectious disease clinics, hospitals, etc. Focus was also placed on government entities (national headquarters, ministries, the Croatian parliament), and political parties.

The results of NER are presented in Figure 3, in a visualy evocative (easy-to-perceive) format: word-cloud visualization is useful for quickly noticing the

most prominent term as well as relative prominence of the terms (larger font indicates higher frequency.

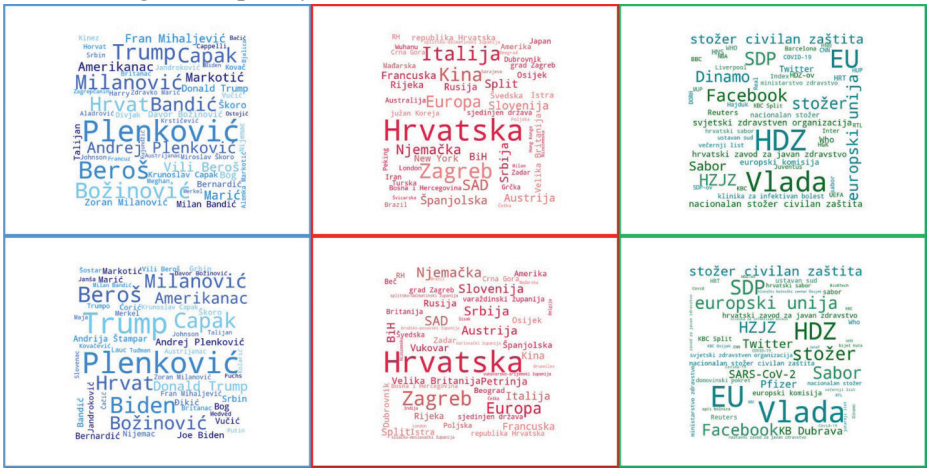


**Figure 3.** Visualization of summarised entities: persons (blue word-clouds), locations (red word-clouds) and institutions (green word-clouds) – during the 1st wave of the pandemic (in the first row) and during the 2nd wave of the pandemic (in the second row)

We also analyzed the total number of entities according to each category across the two waves. The results are shown in Figure 4.

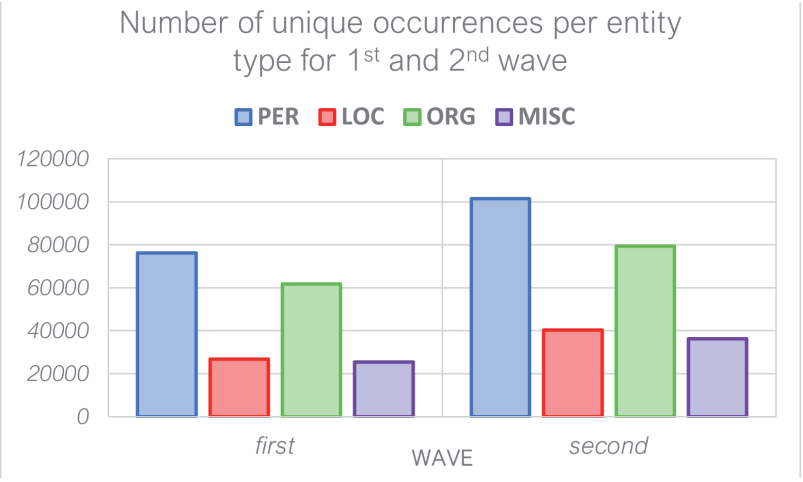


**Figure 4.** Number of unique occurrence per entity type by categories: persons (PER), locations (LOC), institutions (ORG), and miscellaneous (MISC), across two pandemic waves

### 4.3. *The results of the analysis based on the Google trends*

In the last step we chose a certain number of words specified as keywords belonging to the semantic field related to the COVID-19 pandemic in the Croatian language (these are the same frequent terms that we have explored in section 4.2). We checked how these keywords were represented in searches on Google (by using *Google trends*) from January 1st, 2020 until January 15th, 2021. Here we chose to start with the period that began a little bit earlier than the pandemic really started in Croatia in order to study the trends before the pandemic. In Figure 5 we show comparisons between some of the keywords that we found interesting.

The first comparison in Figure 5 is related to the four most frequently searched keywords from the medical domain: virus (*virus*), coronavirus (*koronavirus*), symptoms (*simptomi*), and vaccination (*cijepljenje*). It is obvious that the most frequently searched keyword was 'coronavirus' as this is the central word of the pandemic. In the beginning, the term 'virus' was also popular in searches on Google, but then during the lockdown period in March and April 'coronavirus' was the most dominant term in searches. The keyword 'symptoms' was also a popular search term during the first three months of the pandemic and a bit less so during the summer. That can be explained by the fact that by then we had already learnt what the symptoms of the COVID-19 disease are. In comparison to other keywords, 'vaccination' was the least popular search term during the first six months of the pandemic. This may indicate that, at the beginning, people were not so interested in vaccination, or that maybe they did not widely believe in the fast development of a vaccine against the COVID-19, etc.

Next, we singled out two terms that reflect the primary means officially recommended for preventing the spread of the epidemic (Figure 6): during the first wave it was masks, and during the second wave it was the vaccine (in addition to masks). During the whole observed period of 13 months, searches for the term 'mask' were significantly more predominant over those for the term 'vaccine'. Only at the end of the second wave the frequencies of the searches for both terms overlap, and this was followed by a short period of more frequent searches for the term 'vaccine'. This period coincided with the news about the production of the first vaccines, which consequently raised interest for this term. The same finding can be observed in the data presented in section 2.4. – online news media were writing more about the masks than the vaccines.

Figure 7 shows the comparison of the four keywords that are related to the social aspects of the pandemic: 'quarantine' (*karantena*), 'lockdown' (*zatvaranje*), 'headquarters' (*stožer*), and 'passes' (*propusnice*). If we compare the popularity of the keywords 'quarantine', 'lockdown' and 'headquarters' (Figure 7 (a)), we can see that 'headquarters' was the most popular term in the search queries,

especially during the lockdown period. This is probably due to the fact that headquarters plays an important role in all decisions about pandemic response measures. The keywords ‘quarantine’ and ‘lockdown’ show similar patterns in search queries. They were most popular at the beginning of the pandemic, similar to the keyword ‘headquarters’. Later, the term ‘lockdown’ seems to be more present in the search queries than the term ‘quarantine’. When we compare the keywords ‘passes’ and ‘headquarters’, it shows that the keyword ‘passes’ was a more popular search term during the lockdown. The pass refers to the document that was introduced to enable travelling from one municipality to another in Croatia during the closing and the lockdown. People were interested in searching for the instructions on how to get passes from the online service that was established (since Croatia is a country with a truly mosaic organization of municipalities and counties, passes were essential for running errands, performing daily tasks, etc.).

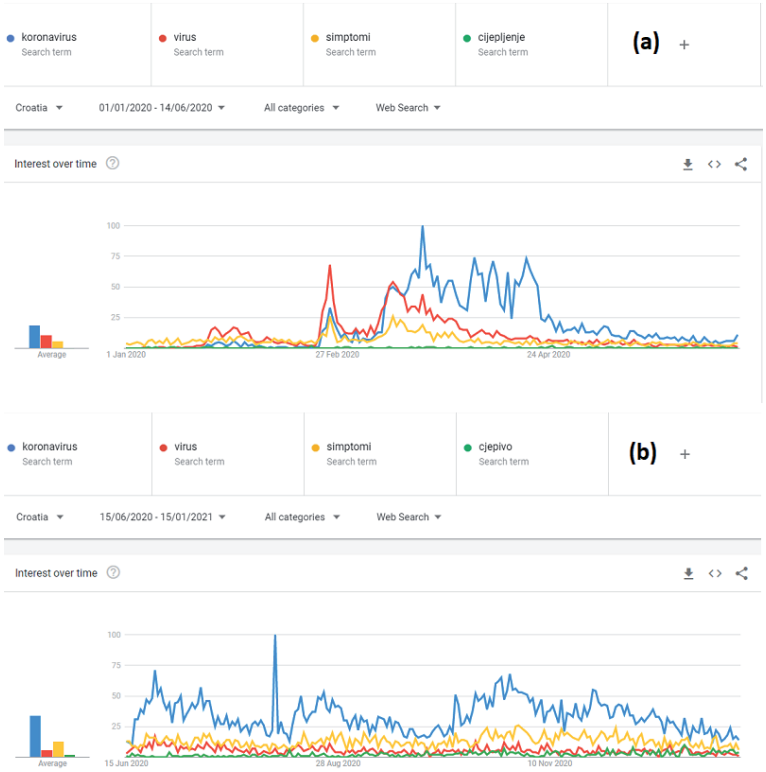


**Figure 5.** Comparison of four search terms: virus (*virus*), coronavirus (*koronavirus*), symptoms (*simptomi*), vaccination (*cijepljenje*) during the first wave (a) and the second wave (b), according to the *Google trends*

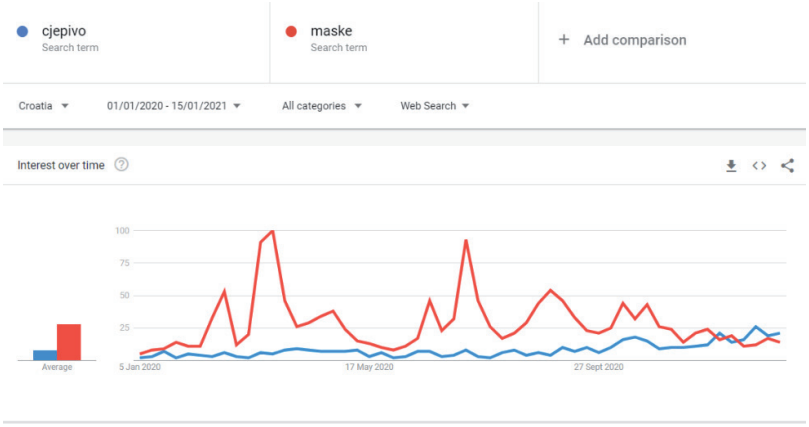


**Figure 6.** Comparison of two search terms during the whole period of 13 months: vaccine (*cjepivo*) and masks (*maske*)

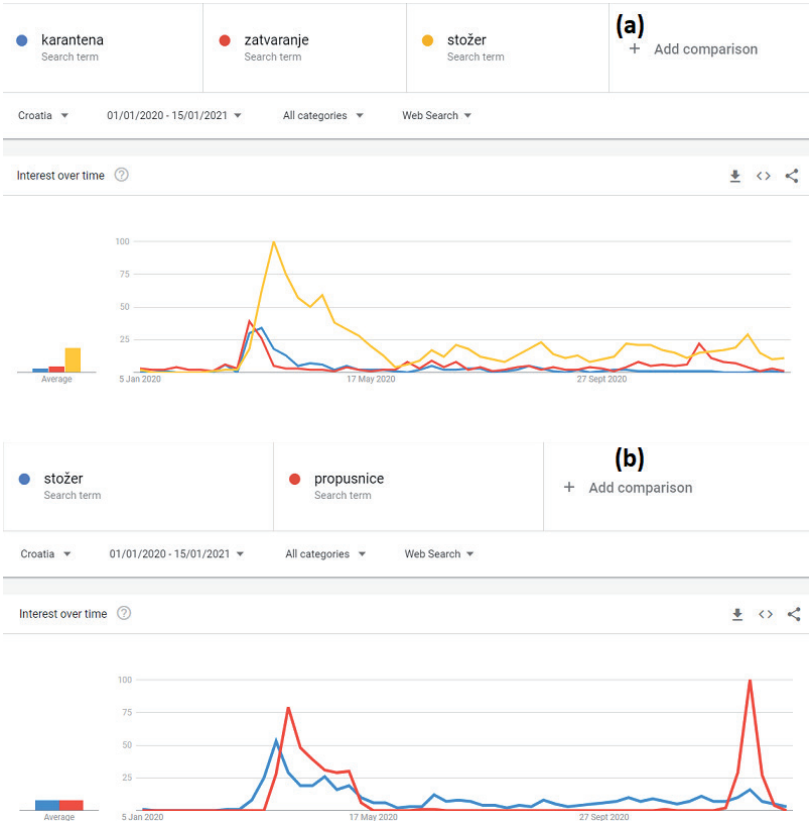


**Figure 7.** Comparison of different search terms during the whole period of 13 months: quarantine (*karantena*), lockdown (*zatvaranje*) and headquarters (*stožer*) (a), and headquarters (*stožer*) and passes (*propusnice*) (b)

## 5. Conclusion

The situation with the COVID-19 pandemic in comparison with other recent epidemics appears to be different in terms of both its spread and the strong impact on different fields of everyday life it has had. One of the consequences of the COVID-19 pandemic is an **information overload**, which was declared to be an infodemic by the WHO (Tangcharoensathien et al. 2020; WHO 2020). The COVID-19 **infodemic** has already been present for more than a year and it poses many challenges, such as large communication volumes, large number of users involved in the communication on the social media, massive datasets, new terminology related to the COVID-19, various domains, and topics (healthcare, economy, politics, education, etc.). Such demanding situations, which the society is currently experiencing, require **information monitoring** of online media content in order to gain insights into social media communication, crisis communication and human behavior during the pandemic. In an attempt to provide answers to the question of how media have responded to the COVID-19 outbreak, information monitoring has proven to be one of the key elements of infodemic management. Furthermore, new technologies have ensured progress in the task of **infoveillance** (*information surveillance,* Eysenbach, 2009). However, there is still room for improvement. In an attempt to further our understanding of this issue, in this study we have presented a possible framework for monitoring the media coverage of crisis communication.

This research provides a quantitative analysis of news articles published in online news media during the first two waves of the pandemic in Croatia. The analysis is based on natural language processing (NLP) techniques, and it is focused on the comparison of the features of the media coverage during the observed periods. In the first step we analyzed the number of the COVID-19-related articles. In the second step we analyzed and compared the most frequent terms and entities and how the main terminology has changed between the two waves. In the third step we analyzed *Google trends* related to the COVID-19 terminology. The results revealed that the total number of COVID-19-related articles in online media is rather high. The ratio of COVID-19-related articles does not fall below 44% in any of the observed eight most popular and viral online news media. By observing the average ratio across all explored online news media, we found that the COVID-19-related publications cover more than the half of the media space (about 57%). Interestingly, during the period between the two waves of the epidemic, when the number of new cases dropped to zero, the number of publications related to the COVID-19 topics remained at a high 43%. Next, by exploring the 250 most frequently used terms during the

first and the second wave of the epidemic, we have found that the trends were as follows: firstly, 'vaccination' (*cijepljenje*) was prevalently mentioned during the second wave, as scientists were on the verge of discovering a vaccine; 'quarantine' (*karantena*) and 'pass' (*propusnica*) typically marked the first wave; 'lockdown' was more frequent during the second wave as quarantine was replaced by a selective lockdown, a "softer" measure which was expected to lead to similar results as quarantine had done during the first wave. During the second wave the rules about passes were first softened (they were kept at the level of the counties and were not prescribed for movement between municipalities), but they were soon revoked due to the severe earthquake that had hit Croatia. 'Symptoms', 'virus' and 'coronavirus' were almost equally frequently mentioned during both waves, and 'mask' and 'headquarters' were more frequent during the second wave.

Next, we explored the mentions of persons, locations and institutions in COVID-19-related online news articles, which is a task termed named entity recognition (NER). We have calculated the Jaccard similarity between the two pandemic waves. The results revealed that during the 13 months the locations mentioned in media were mostly constant, which indicates that the focus was on a narrowed area restricted to Croatia, the neighboring countries, the EU and international locations connected with the COVID-19 crisis (Wuhan, Lombardy, etc.). This reflects the fact that countries closed their borders to prevent people from spreading the disease. The predominant location entities during the whole period were Croatian cities and regions. When talking about persons, it is interesting that news is dispersed across many persons participating in daily events, and that politicians were still more frequently mentioned than scientists in spite of the increased interest in the scientists' interpretations of the epidemic crisis. Among organizations, the focus was on the WHO, local infectious disease clinics, hospitals, government entities (national headquarters, ministries, the Croatian parliament), and political parties.

The presented framework encompasses a selection of NLP approaches which can be recommended for exploring crisis communication, since they can provide us with accurate and relevant comparisons of the data relating to different time periods.

## Acknowledgements

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 the University of Rijeka project number uniri-drustv-sp-20-58.

## Supplementary materials

All the colored figures presented in this work are also available in digital form on the website of the scientific project *InfoCoV* <https://infocov.uniri.hr/scimeth2021>.

### References

Almazán-Ruiz, Encarnación; Orrequia-Barea, Aroa (2020) "The British Press' Coverage of Coronavirus Threat: A Comparative Analysis Based on Corpus Linguistics." *Çankaya University Journal of Humanities and Social Sciences* 14: 1–22.

Babić, Karlo; Guerra, Francesco; Martinčić-Ipšić, Sanda; Meštrović, Ana (2020) "A Comparison of Approaches for Measuring the Semantic Similarity of Short Texts Based on Word Embeddings." *Journal of Information and Organizational Sciences* 44 (2): 231–246.

Babić, Karlo; Petrović, Milan; Beliga, Slobodan; Martinčić-Ipšić, Sanda; Meštrović, Ana (2021) "COVID-19 Related Communication on Twitter: Analysis of the Croatian and Polish Attitudes." In: *Sixth International Congress on Information and Communication Technology*. London: Springer. (in press)

Babić, Karlo; Petrović, Milan; Beliga, Slobodan; Martinčić-Ipšić, Sanda; Pranjić, Marko; Meštrović, Ana (2021) "Prediction of COVID-19 Related Information Spreading on Twitter." In *44th Proceedings of the IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics*. Opatija: IEEE, 424–428.

Beliga, Slobodan, Martinčić-Ipšić, Sanda; Matešić, Mihaela; Meštrović, Ana (2022). Natural Language Processing and Statistic: The First Six Months of the COVID-19 Infodemic in Croatia. In: *The Covid-19 Pandemic as a Challenge for Media and Communication Studies*. Routledge, Taylor & Francis Group. (in press)

Bogović, Petar Kristijan; Beliga, Slobodan; Martinčić-Ipšić, Sanda; Meštrović, Ana (2021) "Topic Modeling of Croatian News during COVID-19 Pandemic." In: *44th Proceedings of the IEEE International Convention on Information*

*and Communication Technology, Electronics and Microelectronics*. Opatija: IEEE, 1205–1212.

Bunker, Deborah (2020) "Who Do You Trust? The Digital Destruction of Shared Situational Awareness and the COVID-19 Infodemic." *International Journal of Information Management* 55: 102201.

Chakraborty, Koyel; Bhatia, Surbhi; Bhattacharyya, Siddartha; Platos, Jan; Bag, Rajib; Hassanien, Aboul Ella (2020) "Sentiment Analysis of COVID-19 Tweets by Deep Learning Classifiers – A Study to Show How Popularity Is Affecting Accuracy in Social Media." *Applied Soft Computing* 97: 106754.

Cinelli, Matteo; Quattrociocchi, Walter; Galeazzi, Alessandro; Valensise, Carlo Michele; Brugnoli, Emanuele; Schmidt, Ana Lucia; Zola, Paola; Zollo, Fabiana; Scala, Antonio (2020) "The Covid-19 Social Media Infodemic." *Scientific Reports* 10: 16598. [https://doi.org/10.1038/s41598-020-73510-5]

Cuomo, Raphaela E.; Purushothaman, Vidya; Li, Jiawei; Cai, Mingxiang; Mackey, Timothy K. (2020) "Sub-National Longitudinal and Geospatial Analysis of COVID-19 Tweets". *PLoS ONE* 15 (10): e0241330. [https://doi.org/10.1371/journal.pone.0241330]

Eysenbach, Gunther (2002) "Infodemiology: The Epidemiology of (Mis) Information." *The American Journal of Medicine* 113 (9): 763–765.

Eysenbach, Gunther (2009) "Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet." *Journal of Medical Internet Research* 11 (1): e11.

Eysenbach, Gunther (2011) "Infodemiology and Infoveillance: Tracking Online Health Information and Cyberbehavior for Public Health." *American Journal of Preventive Medicine* 40 (5): 154–158.

Fišer, Darja; Ljubešić, Nikola; Erjavec, Tomaž (2020) "The Janes Project: Language Resources and Tools for Slovene User Generated Content." *Language Resources and Evaluation*, 54 (1): 223–246. [https://doi.org/10.1007/s10579-018-9425-z]

Gallotti, Riccardo; Valle, Francesco; Castaldo, Nicola; Sacco, Pierluigi; De Domenico, Manlio (2020) "Assessing the Risks of 'Infodemics' in Response to COVID-19 Epidemics." *Nature Human Behaviour* 4 (12): 1285–1293.

Ghafarian, Seyed Hossein; Yazdi, Hadi Sadoghi (2020) "Identifying Crisis-Related Informative Tweets Using Learning on Distributions." *Information Processing & Management* 57 (2): 102145.

Glik, Deborah C. (2007) "Risk Communication for Public Health Emergencies." *Annu. Rev. Public Health* 28: 33–54.

Gozzi, Nicolò; Tizzani, Michele; Starnini, Michele; Ciulla, Fabio; Paolotti, Daniela; Panisson, André; Perra, Nicola (2020) "Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis." *Journal of Medical Internet Research* 22 (10): e21597.

Jang, Hyeju; Rempel, Emily; Roth, David; Carenini, Giuseppe; Janjua, Naveed Zafar (2021) "Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis." *Journal of Medical Internet Research* 23 (2): e25431.

Jelodar, Hamed; Wang, Yongli; Orji, Rita; Huang, Shucheng (2020) "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or Covid-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach." *IEEE Journal of Biomedical and Health Informatics* 24 (10): 2733–2742.

Kaur, Harleen; Ahsaan, Shafqat Ul; Alankar, Bhavya; Chang, Victor (2021) "A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets." *Information Systems Frontiers*, 1–13.

Ljubešić, Nikola; Stupar, Marija; Jurić, Tereza; Agić, Željko (2013) "Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene." *Slovenščina 2.0* 1 (2): 35–57.

Lwin, May Oo; Lu, Jiahui; Sheldenkar, Anita; Schulz, Peter Johannes; Shin, Wonsun; Gupta, Ray; Yang, Yinping (2020) "Global Sentiments Surrounding the COVID-19 Pandemic on Twitter: Analysis of Twitter Trends." *JMIR Public Health and Surveillance* 6 (2): e19447.

Mackey, Tim Ken; Li, Jiawei; Purushothaman, Vidya; Nali, M.atthew; Shah, Neal; Bardier, Cortni; Cai, Mingxiang; Liang, Bryan (2020) "Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales: Infoveillance Study on Twitter and Instagram." *JMIR Public Health and Surveillance* 6 (3): e20794.

Paka, William Scott; Rachit, Bansal; Abhay, Kaushik; Shubhashis, Sengupta; Chakraborty, Tanmoy (2021) "Cross-SEAN: A Cross-Stitch Semi-Supervised Neural Attention Model for COVID-19 Fake News Detection." *Applied Soft Computing* 107: 107393.

Pulido, Cristina M.; Villarejo-Carballido, Beatriz; Redondo-Sama, Gisela; Gómez, Aitor (2020) "COVID-19 Infodemic: More Retweets for Science-

Based Information on Coronavirus than for False Information." *International Sociology* 35(4) 377–392.

Rustam, Furqan; Khalid, Madiha; Aslam, Waqar; Rupapara, Vaibhav; Mehmood, Arif.; Choi, Gyu Sang (2021) "A Performance Comparison of Supervised Machine Learning Models for Covid-19 Tweets Sentiment Analysis." *PLoS ONE* 16 (2): e0245909. [https://doi.org/ 10.1371/journal.pone.0245909]

Satu, Md. Shahriare; Khan, Md. Imran; Mahmud, Mufti; Uddin, Shahadat; Summers, Matthew A.; Quinn, Julian M. W.; Moni, Mohamed Ali (2021) "TClustVID: A Novel Machine Learning Classification Model to Investigate Topics and Sentiment in COVID-19 Tweets." *Knowledge-Based Systems* 226: 107126.

Tangcharoensathien, Viroj; Calleja, Neville; Nguyen, Tim; Purnat, Tina; D'Agostino, Marcelo; Garcia-Saiso, Sebastian; Landry, Mark et al. (2020) "Framework for Managing the COVID-19 Infodemic: Methods and Results of an Online, Crowdsourced WHO Technical Consultation." *Journal of Medical Internet Research* 22 (6): e19659.

WHO – World Health Organisation (2020) "An Ad Hoc WHO Technical Consultation Managing the COVID-19 Infodemic: Call for Action, 7-8 April 2020."

Xue, Jie; Chen, Junxiang; Chen, Chen; Zheng, Chengda; Li, Sijia; Zhu, Tingshao (2020) "Public Discourse and Sentiment during the COVID 19 Pandemic: Using Latent Dirichlet Allocation for Topic Modeling on Twitter." *PLoS ONE*, 15 (9): e0239441. [https://doi.org/10.1371/journal. pone.0239441]

Zarocostas, John (2020) "How to Fight an Infodemic." *The Lancet* 395 (10225): 676.

Zhou, Jianlong; Yang, Shuiqiao; Xiao, Chun; Chen, Fang (2021) "Examination of Community Sentiment Dynamics Due to COVID-19 Pandemic: A Case Study from a State in Australia." *SN Computer Science* 2 (3): 1–11.

# Metodološki okvir za usporedbu medijskog praćenja prvih dvaju valova pandemije bolesti COVID-19 u Hrvatskoj zasnovan na obradi prirodnoga jezika (NLP)

*Online*-mediji kao komunikacijska platforma imaju važnu ulogu u javnozdravstvenoj zaštiti. Praćenje informiranja (engl. *infoveillance*) u *online*-medijima u pandemijskim okolnostima važan je korak prema boljem razumijevanju kriznih komunikacija. Cilj je ovoga istraživanja provesti longitudinalnu analizu onoga sadržaja u *online*-medijima koji se odnosi na izvješćivanje o pandemiji bolesti COVID-19, uz pomoć metoda obrade prirodnoga jezika (NLP), koja bi (analiza) poslužila kao mogući metodološki okvir za praćenje krizne komunikacije u medijima. U tu su svrhu prikupljeni podaci iz članaka s vijestima koje su objavljivali hrvatski *online*-mediji tijekom prvih 13 mjeseci trajanja pandemije (tj. tijekom prvih dvaju pandemijskih valova u Hrvatskoj). Metodologija obuhvaća: izračunavanje postotka objavljenih članaka na temu bolesti COVID-19 u osam najvažnijih *online*-medija u odnosu na ukupan broj objavljenih članaka u tim medijima u promatranom razdoblju, zatim analizu sadržaja vijesti s obzirom na najučestalije termine vezane uz pandemiju i izračunavanje podudarnosti sadržaja u praćenom razdoblju (engl. *Jaccard similarity*), potom usporedbu pojavnosti termina vezanih uz pandemiju tijekom prvih dvaju pandemijskih valova (tj. tijekom cijele prve godine krize) i naposljetku primjenu prepoznavanja imenovanih entiteta (NER) radi utvrđivanja najučestalijih entiteta i dinamike promjena u njihovoj zastupljenosti u promatranom razdoblju. Rezultati pokazuju da su mediji popratili pojavu pandemije velikim brojem vijesti – količina vijesti o bolesti COVID-19 bila je znatna u cijelom promatranom razdoblju, a nije opadala čak ni u razdoblju smirivanja širenja zaraze, tj. između dva vala, kad nije bilo novih slučaja zaraze. Nadalje, utvrđena je visoka podudarnost u terminima koji se upotrebljavaju u vezi s pandemijom tijekom prvoga i drugog vala, uz neke posebnosti vezane uz dinamiku proučavanja bolesti i preporučenih mjera za zaštitu od zaraze. Rezultati prepoznavanja imenovanih entiteta (NER) pokazuju veću razlikovnost kad je riječ o osobama spominjanim u kontekstu pandemije, a manju kad je riječ o lokacijama, što se tumači posljedicom ograničavanja kretanja kao mjere za zaštitu od pandemije.

**Ključne riječi:** *online*-mediji, praćenje informiranja, *infoveillance*, krizna komunikacija, obrada prirodnog jezika