Infoveillance of the Croatian Online Media during the COVID-19 Pandemic: a One-Year Longitudinal NLP Study

Slobodan Beliga, Sanda Martinčić-Ipšić, Mihaela Matešić, Irena Petrijevčanin Vuksanović, Ana Meštrović

Submitted to: JMIR Public Health and Surveillance
on: June 24, 2021

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Infoveillance of the Croatian Online Media during the COVID-19 Pandemic: a One-Year Longitudinal NLP Study

Slobodan Beliga[1, 2] PhD; Sanda Martin?i?-Ipši?[1, 2] PhD; Mihaela Mateši?[3, 2] PhD; Irena Petrijev?anin Vuksanovi?[2] PhD; Ana Meštrovi?[1, 2] PhD

[1]University of Rijeka Department of Informatics Rijeka HR
[2]University of Rijeka Center for Artificial lntelligence and Cybersecurity Rijeka HR
[3]University of Rijeka Faculty of Humanities and Social Sciences Rijeka HR

Corresponding Author:
Slobodan Beliga PhD
University of Rijeka
Department of Informatics
Radmile Matej?i? 2
Rijeka
HR

Abstract

Background: Online media plays an important role in public health emergencies and serves as a communication platform. Infoveillance of online media during the COVID-19 pandemic is an important step toward a better understanding of crisis communication.

Objective: The goal of this study is to perform a longitudinal analysis of the COVID-19 related content based on natural language processing methods.

Methods: We collected a dataset of news articles published by Croatian online media during the first 13 months of the pandemic. Firstly, we test the correlations between the number of articles and the number of new daily COVID-19 cases. Secondly, we analyze the content by extracting the most frequent terms and apply the Jaccard similarity. Next, we compare the occurrence of the pandemic-related terms during the two waves of the pandemic. Finally, we apply named entity recognition to extract the most frequent entities and track the dynamics of changes during the observed period.

Results: The results show there is no significant correlation between the number of articles and the number of new daily COVID-19 cases. Furthermore, there are high 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, the findings indicate that the most influential entities have lower overlaps for the identified persons and higher overlaps for locations and institutions.

Conclusions: Our study shows that online media has a prompt response to the pandemic with a large number of COVID-19 related articles. There is a high overlap in the frequently used terms across the first 13 months, which may indicate the narrow focus of reporting in certain periods. However, the pandemic-related terminology is well covered.

(JMIR Preprints 24/06/2021:31540)
DOI: https://doi.org/10.2196/preprints.31540

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Infoveillance of the Croatian Online Media during the COVID-19 Pandemic: a One-Year Longitudinal NLP Study

Abstract

Background: Online media plays an important role in public health emergencies and serves as a communication platform. Infoveillance of online media during the COVID-19 pandemic is an important step toward a better understanding of crisis communication.
Objective: The goal of this study is to perform a longitudinal analysis of the COVID-19 related content based on natural language processing methods.
Methods: We collected a dataset of news articles published by Croatian online media during the first 13 months of the pandemic. Firstly, we test the correlations between the number of articles and the number of new daily COVID-19 cases. Secondly, we analyze the content by extracting the most frequent terms and apply the Jaccard similarity. Next, we compare the occurrence of the pandemic-related terms during the two waves of the pandemic. Finally, we apply named entity recognition to extract the most frequent entities and track the dynamics of changes during the observed period.
Results: The results show there is no significant correlation between the number of articles and the number of new daily COVID-19 cases. Furthermore, there are high 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, the findings indicate that the most influential entities have lower overlaps for the identified persons and higher overlaps for locations and institutions.
Conclusions: Our study shows that online media has a prompt response to the pandemic with a large number of COVID-19 related articles. There is a high overlap in the frequently used terms across the first 13 months, which may indicate the narrow focus of reporting in certain periods. However, the pandemic-related terminology is well covered.

Keywords: COVID-19; pandemic; online media; news coverage; infoveillance; infodemic; infodemiology; natural language processing; name entity recognition; longitudinal study.

Introduction

Background

Media coverage plays an important role in public health emergencies, such as the COVID-19 pandemic and serves as a key communication platform during global health crises [1]. The media represents a bridge between science and society and has great power in forming a collective opinion, attitudes, perspectives, and behaviors [2]. Recent studies proposed new disease-spreading models integrating media coverage as a serious factor that may influence human behavior in the context of disease transmission [3]–[5]. All of these studies confirm that the media may affect the spread and control of infectious diseases. Wang et al. explained that media coverage has an impact on the implementation of public intervention and control policies [4]. They pointed out that one of the measures is to educate people and explain how to prevent the disease through all available sources of information.

On the other side, the media, especially internet-based information sources may cause an infodemic, which is described as an overabundance of information, misinformation and disinformation. Coping with these phenomena created the discipline of infodemiology [6], [7]. Eysenbach [8] defined the four pillars of infodemic management and one of them is information monitoring – infoveillance that enables a better insight into how the media responds to the crisis.

The infodemic is one of the severe consequences of the COVID-19 pandemic [9], [10]. It raises many challenges for the task of infoveillance in terms of massive datasets, such as large communication volumes, new terminology related to the COVID-19, various topics and domains present in the media (healthcare, economy, politics, education, etc.) and the large number of users involved in communication in social media. Recently NLP technologies have ensured progress in dealing with the large amount of textual data [11] and thus are promising underlying methods as an integral part of infoveillance methodology.

Prior Work

The significance and impact of the media in the context of an epidemic was extensively studied for several epidemics before COVID-19, such as H5N1 influenza [12], SARS [13], MERS [14], H1N1 influenza [15], the Zika virus disease [16], etc. The outbreak of the COVID-19 pandemic resulted in research publications focused on the different aspects of public communication: the linguistic perspective of the online news media [17], the content analysis of global media framing of COVID-19 [18], the politicization and polarization in COVID-19 news coverage [19], the amount of media coverage in the context of the pandemic [2]. The studies related to the infoveillance are mostly focused on discovering topics [20], [21], sentiment [22], [23] or fake news detection [24], [25].

Most studies employed different NLP techniques for capturing specific aspects of the COVID-19 content published online. For discovering public perceptions, opinions and attitudes towards specific COVID-19 related topics authors mostly combine topic modelling and sentiment analysis [21], [26], [27], occasionally combined with the named entity recognition (NER) [28].

Although COVID-19 related media coverage has been widely studied, there are still some aspects of the task of infoveillance that can be improved. For example, existing studies are much often focused solely on the content of the texts, rather than on the amount of published texts. There are only a few exceptions in which the dynamics of publishing have been analyzed [2], [20]. Next, the majority of analyzed datasets consist of texts published at the beginning of the pandemic and they capture a short time span of three to four months. Given the lack of research that apply longitudinal data monitoring over larger time spans (i.e. the first year of COVID-19 pandemic), our study might be worth of attention.

In our research, we are following methodologies as described in the above-mentioned studies. However, in order to address the mentioned gaps, we propose extensions, contributing to the theoretical framework for the task of infoveillance. Firstly, we combine statistical methods and NLP techniques in order to perform the tracking of the amount of news articles and the content of news articles at the same time. Secondly, in the proposed approach we apply the Jaccard similarity coefficient for measuring the similarity of the most frequent terms and entities.

The Goal of this Study

In relation to the prior works, our study describes an approach for the task of infoveillance based on combining NLP methods and statistics, focused on the content from online news media.

By providing an analysis of the online media's response to the pandemic, we aim to contribute to the discipline of information monitoring, particularly to a better understanding of: 1) the role that Internet-based sources play in communication during the COVID-19 crisis and 2) the potential infodemic. Our goal is to provide an NLP based longitudinal tracking of the dynamics of changes in the coverage of Croatian online news space. Noting that the Croatian media are reported as being poorly trusted [29], gives us reason to explore how they have treated one of the most challenging situations in the everyday life of the country's citizens.

This study addresses the following research questions related to the period of the first 13 months of the pandemic:

RQ1: What is the amount of COVID-19 related news articles and is the number of COVID-19 related articles in correlation with the number of new COVID-19 cases?

RQ2: What are the main key terms, the most frequent pandemic related terms and the most frequent entities in the focus of the online news media?

RQ3: How COVID-19 related content (in terms of the most frequent words, most frequent pandemic-related terms, main entities related to the pandemic) has changed during the first 13 months of the pandemic?

To answer these questions, we perform the following analyses. First, we conduct an exploratory statistical analysis of online media to provide an overview of the trends of COVID-19 related articles published during the first year of the pandemic. Next, we propose a set of statistical and NLP-based methods for the task of infoveillance of the content published on online news media. More specifically, we apply named entity recognition (NER) for the automatic extraction of the entities that play a key role during pandemics. Next, we provide a simple visualization monitor enabling the longitudinal tracking of the change of the pandemic-related terms contrasted between the first and second waves of the pandemic. Finally, we quantify and visualize the changes of the most frequent terms and most frequent entities by using the Jaccard similarity over the 13 months.

Methods

Data Collection

In this longitudinal study, the collected data covers a period of more than a year, specifically the period from January 1, 2020 to January 15, 2021 covering the time of the first two pandemic waves in the Republic of Croatia (see Table 1). We include January and part of February 2020 although this is before the first reported COVID-19 case in Croatia. With the inclusion of this short period before the pandemic outbreak, the dataset contains the emergence of seed pandemic-related terminology. Moreover, the captured antecedent period serves as the control for the comparison with the official pandemic period. More details about the duration of the epidemic (pandemic) waves can be found in Section-A1 of Multimedia Appendix 1.

Table 1. Duration of pandemic waves.

period	start	end

1st pandemic wave	January 1, 2020 *February 25, 2020	May 22, 2020
the pandemic subsides	May 23, 2020	June 14, 2020
2nd pandemic wave	June 15, 2020	January 15, 2021

** the first COVID-9 case appears in Croatia*

The data acquisition includes publications from eight mainstream online news media, distributed to cover the geographical and media space of the Republic of Croatia. The articles were collected on a daily basis, resulting in 270,359 articles in total, while the number of COVID-19 related news articles is 121,095. Collected articles are a full sample of all articles published in these eight portals in defined period. We refer to the dataset of the COVID-19 related articles as Cro-CoV-texts2020 (In Multimedia Appendix 1, we provide a link to the publicly available lists of the word frequencies extracted from all news grouped by month). These eight portals included in the Cro-CoV-texts2020 dataset do not cover the entire online news media space of Croatia. Nevertheless, they form a representative sample for our longitudinal study. The criteria for their selection are described in detail in Section-A0 of Multimedia Appendix 1.

The filter used to determine the affiliation of an article to a COVID-19 class is the occurrence of keywords from the coronavirus thesaurus in the title, the subtitle or in the body of the text. The coronavirus thesaurus contains about twenty of the most important words describing the SARS-CoV-2 virus epidemic, as well as all their inflectional variations, and is available in the Multimedia Appendix 1 (Section-A2) of this paper. In addition to the general words (universal keywords related to the COVID-19 disease pandemic), the list has been expanded with additional terms specific to Croatia: the personal name of the public administration authorities (the Minister of Health, a leading state epidemiologist, the director of the National Civil Protection Headquarters, etc.).

The collected articles are preprocessed: (i) only the textual part of the news was retained (related images and videos are discarded) and (ii) all the titles, subtitles and body of the texts are lemmatized (to reduce the inflectional variations of the words as the standard NLP preprocessing procedure).

The epidemiological data related to COVID-19 disease, i.e. the number of newly-infected persons, were obtained from the official government portal. The data are available in Multimedia Appendix 1 (see Section-A0) for every day in the period from February 26, 2020 (when the first case of coronavirus infection was confirmed in Croatia) to January 15, 2021.

Statistical Analysis of the Amount of the Online Media Content

After filtering the collected content according to the defined thesaurus of coronavirus terms, we first determine the ratio of the COVID-19 related and remaining publications. We then perform an exploratory statistical analysis of the COVID-19 related online publications.

Specifically, the time series of COVID-19 daily cases is compared with daily-published COVID-19 related articles during the entire period from January 1, 2020 to January 15, 2021. Both time series have the same time resolution and the same length of 110 days in the first, and 215 days in the second pandemic wave. For the distribution of time series data that do not follow the Gaussian distribution, non-parametric tests were used. The standard Spearman's correlation coefficient (ρ) and Kendall's coefficient (τ) are used to measure the strength and direction of the association between the two variables – the number of cases and articles.

Additionally, the cross-correlation function (CCF) is applied to quantify a potential association, as well as the time lag between the two time series (see equation (1) in the Multimedia Appendix 1 Section-A3). The interpretation of the CCF dictates that larger absolute values of cross-correlation at the time lag indicate a stronger association between the two time series. The correlation is considered to be significant when the absolute value is greater than the threshold defined with equation (3) in Section-A3 in the Multimedia Appendix 1.

Another modality of the experiment aggregates the daily data into a one-week window for both time series resulting in the resolution of 15 weeks in the first, and 32 weeks in the second pandemic wave (46 weeks in total), and they are also suitable for calculating the CCF.

The autocorrelation function (ACF) calculates the strength of the relationship between a time series observation and observations at prior time steps, called lags. Because the correlation of the time series observations is calculated with values of the same series at previous times, this is called a serial correlation. A plot of the autocorrelation of a time series by a lag is often called the ACF, correlogram or an autocorrelation plot.

The graphs for the autocorrelation function (ACF) of the ARIMA (Autoregressive Integrated Moving Average) residuals include lines that represent the significance limits, and they are calculated by an equation (4). The values that extend beyond the significance limits are statistically significant at approximately α = .05, and show evidence that the autocorrelation does not equal zero [30].

The Mutual Information (MI) between the new COVID-19 case counts and the number of published articles related to COVID-19 during the period from February 26, 2020 to January 15, 2021 was quantified to further evaluate the mutual dependence of the two time series. The MI was calculated as the expected value of Pointwise MI (PMI) of the two time series. The calculations of PMI, MI, and NMI (Normalized Mutual Information) are defined by the equations (5), (6), and (7) defined in the Multimedia Appendix 1 Section-A3.

As suggested in [16], the CCF provided an overview of the association between real-world COVID-19 case counts and the published COVID-19 related articles over a period. In our case (for 325 observations and 28 lags), a CCF above 0.116 indicates a strong association between the two time series. However, the MI complemented the CCF and further quantified this association with an exact numerical value.

Analysis of the Most Frequent Terms and Dynamics of Changes

In the next step, we analyze the most frequent terms related to coronavirus and how the vocabulary trends are changing over time. Specifically, we calculate the frequencies of all the terms in the lemmatized dataset. We perform the same analysis in two different time spans: the first on a monthly level (13 months in total) and the second on two pandemic waves. In the analysis by months, the number of time units (days) depends on the total number of calendar days. In the second case, the duration of pandemic waves is 281 days in total: the first wave is shorter and lasts 166 days, while the second wave is longer and stretches over the remaining 215 days. Roughly speaking, we can say that the first wave lasts approximately six and the second seven months.

Being aware of the fact that other countries might not relate to the recognition and differentiation of pandemic waves, chunking for Croatia is justified by the collected data. The monthly level analysis is certainly appropriate for further comparison with other countries. In the analysis of coronavirus-related concepts, we compare the trends of how the most frequently used terms were changing during the 13 months and across two different pandemic waves by quantifying the Jaccard similarity that indicates the overlap of the terms between two different periods. There are many approaches for the extraction of key terms [31], however we decided to apply a simple approach based on the word frequencies.

Named Entity Recognition

Named Entity Recognition is an NLP task aimed at the extraction of named entities like persons, locations, organizations, numeric expressions (time, money, dates, etc.). NER extraction can be modeled as the text sequence annotation problem. In this case, Conditional Random Fields (CRF) as a non-directed graphical model is trained to maximize the log likelihood, calculated from the conditional probabilities of the output labels' sequence over the features of the input sequence and CRF states. The NER performance for Croatian is reported in [49] in the experiment with three named entity classes (ORG, PERS, LOC) and yielded an F1 score of 89.8%. In this work we use the NER system trained for the related Slavic languages Slovenian, Croatian and Serbian [32], [33] in order to automatically extract entities from large COVID-19 related dataset. The implemented NER is a slight modification of the CRF-based reldi-tagger with Brown clusters information added, capable of the recognition of person, person derivatives (adjectives derived from a person's name), locations, organizations and miscellaneous entities.

Results

Descriptive Analysis of the Online Newspaper Space

In our previous work, we analyzed isolated online and social media content published in the Croatian language in a shorter time period [34]–[37]. In this paper, we focus on the major eight representatives of online news media by scrutinizing their publications in the significantly longer period – from January 1, 2020 to January 15, 2021. The percentage of COVID-19 articles is quantified according to the coronavirus vocabulary.

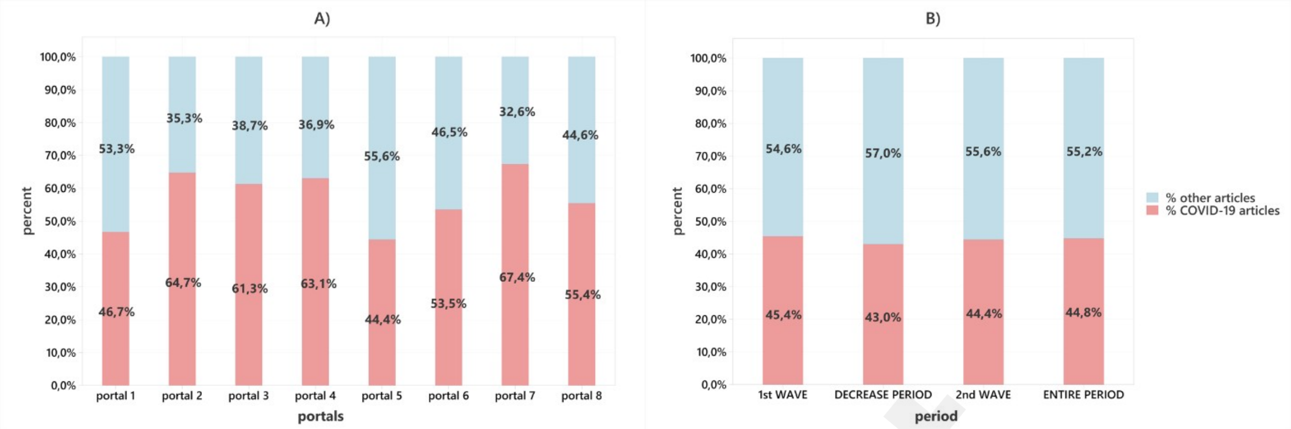


Figure 1. The percentage of COVID-19 related articles summarized per each of the eight online news media during the pandemic in Croatia (February 25, 2020 – January 15, 2021) (A), and the percentage of COVID-19 related articles in the total number of articles summarized across eight online news media for different periods during the pandemic (B).

The percentage of COVID-19 related articles does not fall below 44% on any of the eight observed online news media (A). The average ratio across all online news media of COVID-19 related publications occupies more than one half of the total media space (about 57%).

The right part (B) of Figure 1, shows the percentage of COVID-19 articles in the total number of published articles, summarized for all eight online news media, for different time periods: two pandemic waves, the period in which the pandemic subsided (marked as the decrease period) and for the entire period of 13 months. In order to observe the global picture in 2020, data from January 1 to February 25, 2020 are also observed despite the fact that there were no cases of COVID-19 in Croatia in that time. The percentage of COVID-19 related articles in the first wave would take a much higher value if the analysis did not include days in which there were no cases of infection in Croatia (the period from the beginning of the year to February 26, 2020) and would rise to a value of 57%. Surprisingly, in the period between the two pandemic waves, when the number of cases of infection drops to zero (the decrease period), the number of publications related to the COVID-19 remains at a high 43%, despite expectations that the media would write significantly less about COVID-19.

The Association between the COVID-19 Pandemic and Writing Dynamics in News Media

Many factors may influence the increased interest in COVID-19 issues in the media. For example, the number of patients on mechanical ventilation due to the deterioration of their condition, the number of people in self-isolation, the daily or total number of deaths from COVID-19, the number of companies and entrepreneurs who had to stop their regular business due to the pandemic, etc. The testing of all of these claims is impeded by the unavailability of reliable data. Nevertheless, below we examine an isolated variable that shows the potential to influence COVID-19 publications, and from which we can obtain reliable data. Hence, we aim to determine whether there is a correlation between the number of daily cases of newly-infected people with the SARS-CoV-2 virus and the number of published news articles related to the topic of COVID-19.

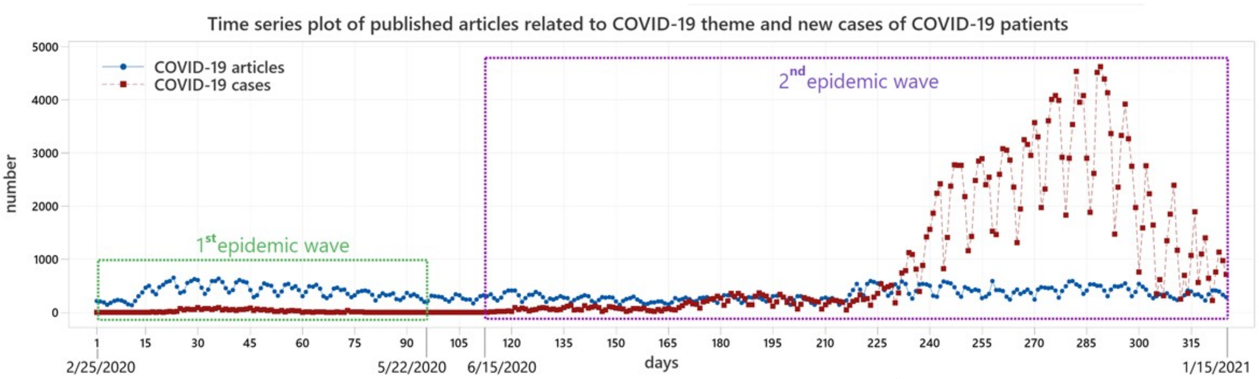


Figure 2. A time series plot comparing the number of published COVID-19 articles per day (blue) and the number of new COVID-19 cases (red) from February 25, 2020 to January 15, 2021

The time series plot in Figure 2 shows the number of new COVID-19 cases per day (red line) and the number of published COVID-19 related articles (blue line). The blue line has the same pattern of wavy repetition throughout the time of the observed period, regardless of the epidemic wave. While the red line has an elongated left tail and then a high ridge in the second epidemic wave. In addition to this, slight repetitive wave-like oscillations are also present along the time axis (days). Data distributions are shown with the histograms of the frequencies for both observed time series in Figure A4-1 (for details see Section-A4 in Multimedia Appendix 1).

Next, we examine whether there is a linear relationship between the number of new cases of COVID-19 per day and the number of publications of COVID-19 related news articles per day, and we calculate the Spearman rank correlation coefficient. We hypothesize H0: there is no correlation between the number of COVID-19 cases and the number of published articles related to COVID-19, (α = .05). A statistical test shows that we reject the null hypothesis and that there is a weak statistically significant correlation between the number of COVID-19 cases and the number of published COVID-19 related articles, ρ(325)=0.253, P<.001, which is additionally confirmed with Kendall's tau, τ(325)=.173, P<.001. More detailed results of 95% confidence intervals for a 2-tailed test are reported in Multimedia Appendix 1 (Section-A4).

Although statistically significant, the correlation is extremely weak. This is also confirmed by Kendall's tau coefficient. To obtain a direct interpretation of results, we will use Kendall's tau coefficient in terms of the probabilities for observing the agreeable (concordant) and non-agreeable (discordant) pairs. Therefore, we can say that the ratio of the occurrence of concordant to discordant pairs is 1:1.4 (resulted from 1+τ /1-τ), which means that the probability of occurrence of concordant pairs is 1.4 times higher than the occurrence of discordant pairs.

Realistically, it is to be expected that the number of publications on the topic of COVID-19 will not increase on the same day as the number of COVID-19 cases increases (or decreases), but the media will write about it subsequently (i.e. the next day or maybe a few days after). Next, we examine whether the correlation can be stronger if we observe the publication of COVID-19 related articles with a time delay compared to the daily number of COVID-19 cases.

Due to the fact that cycles can be seen in the time series data that repeat regularly over time in the form of a sine wave (see the time series graph in Figure 2), it may contain seasonal variations. However, a cycle structure in a time series may or may not be seasonal.

Correlograms (see Section-A4 in Multimedia Appendix 1) show a plot of the autocorrelation function (ACF) on a time series data of new COVID-19 cases (left plot), and published COVID-19 articles by lag (right plot). This autocorrelation measures the linear relationship between the lagged values of a time series. ACF for both, COVID-19 related articles, and new COVID-19 cases, show several significant peaks after a lag of seven days. This determines the cyclic behavior in the time series data in which the cycles are repeated every seven days. The reason for this is that on non-working days (Saturday and Sunday) less news is written and published (a minimum cycle value is achieved). During working days (usually in the middle of the week) a larger number of published news is reached (the maximum cycle value is reached). It is important to emphasize that there is no complete regularity in the cycles, i.e. there is no seasonality on a seven-day basis. The reason for this is that the maximum number of news is not always on the same day of the week. The peak is exchanged on Tuesdays, Wednesdays or Thursdays. The same holds for the number of new confirmed cases of COVID-19. At weekends, a smaller number of people are tested (just on the same days when less news is published), while on workdays more people are tested, so the number of confirmed infections is higher. The peak is reached again in the middle of the week, but not always on the same day, so the regularity in the form of seasonality cannot be credibly confirmed for the entire epidemic year. The results might be able to suggest the presence of a weekly seasonal component for certain shorter periods of the year. Finally, for the entire year, we observe with certainty a cyclical behavior on a weekly or seven-day basis.

According to these insights, we aggregate the data on the time series by week (seven days), i.e. we observe them in the one-week time window period. The Kolmogorov-Smirnov normality test shows that the data do not follow the Gaussian distribution (test details are available in the Multimedia Appendix 1 of Section-A4). Again, we hypothesize H0: that the correlation does not exist (α = .05). Spearman's correlation of ranks indicates the existence of a slight positive correlation ρ(47)=.277 (slightly higher than in the previous case per day), but it is not statistically significant (P=.060). Additionally, this is confirmed with Kendall’s tau τ(47)=.202, P=.05. 95% confidence intervals are reported in Section-A4 of Multimedia Appendix 1.

Due to the vague picture of the existence or non-existence of at least a weak positive correlation, we perform an additional cross-correlation test on the time series data measured on a daily basis.

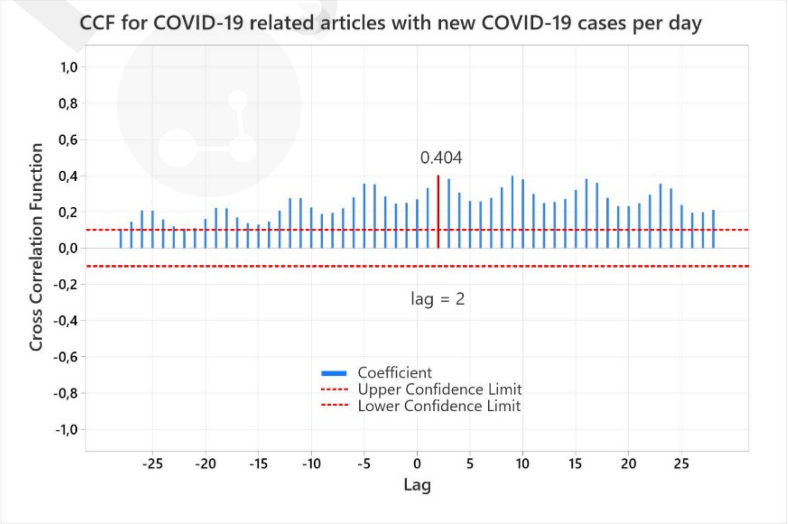


Figure 3. Cross-correlation function (CCF) between the published COVID-19 related articles

counts per day and the number of confirmed COVID-19 cases per day.
A significant cross-correlation between the published COVID-19 related articles' counts and the number of confirmed COVID-19 cases per day was observed for the pandemic in Croatia (Figure 3). The CCF was substantially above the threshold of statistical significance, while the strongest positive correlation in Figure 3 occurs at lag=2. This shows that the two variables are not contemporaneously correlated. However, the positive correlation at lag +2 suggests that higher numbers of COVID-19 infected cases lead to higher numbers of published articles related to COVID-19 themes two days later. Negative correlations are not present in the observed lag range.

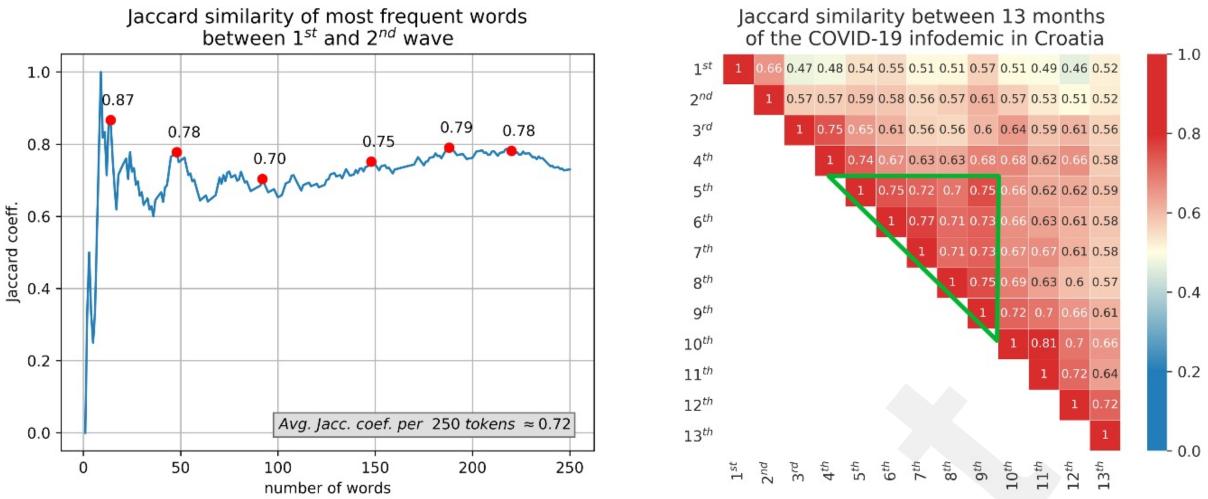
Cross-correlation tests indicate that publishing COVID-19 related news articles was not completely decoupled from the actual disease pandemic in the Republic of Croatia's online news space. This indicates the underlying COVID-19 pandemic effect on the writing about COVID-19. Finally, the strong dependence between the two time series is further quantified and confirmed by MI and the NMI measure (for details see Mutual Information results in Section 4 of Multimedia Appendix 1).

Next, we ask whether there is a linear relation among the eight major online news media considering the number of COVID-19 articles published per day. For all 28 possible cases, the correlations are statistically significant. In terms of the Spearman coefficient, all correlations are positive, and in only two cases the correlation is absent. Furthermore, in 12 cases, the positive correlation is weak, in the next 12 cases, it is substantial, and in two more cases, it is strong. The correlations are confirmed with a conservative coefficient – Kendall's tau (see Multimedia Appendix 1 – Section-A4).

Pandemic-Related Terminology Analysis

The analysis of the most frequent terms is performed at the granularity of pandemic waves. The top eight highly frequent terms in the first and in the second epidemic waves were found to be identical, according to their frequency in COVID-19 related media releases. This is an indication that throughout the pandemic year, regardless of the epidemic wave, journalists most often mention the terms: people, coronavirus, Croatia, year, measure, day, high/large and new. It is an extremely narrow vocabulary with a small set of three terms that constantly talks about the epidemic year in Croatia, and five more terms that are used day by day in the news describing the high daily number of newly-infected people.

If the monitored list is expanded to the top 250 most frequent terms during the first and second epidemic waves, the average value of the Jaccard similarity coefficient is 0.72 (see the curve oscillations in Figure 4 – left). This is an indication of a significant overlap of the most frequent terminology between the pandemic waves, and hence consistent content of pandemic related writing in online media.



	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th
1st	1	0.66	0.47	0.48	0.54	0.55	0.51	0.51	0.57	0.51	0.49	0.46	0.52
2nd		1	0.57	0.57	0.59	0.58	0.56	0.57	0.61	0.57	0.53	0.51	0.52
3rd			1	0.75	0.65	0.61	0.56	0.56	0.6	0.64	0.59	0.61	0.56
4th				1	0.74	0.67	0.63	0.63	0.68	0.68	0.62	0.66	0.58
5th					1	0.75	0.72	0.7	0.75	0.66	0.62	0.62	0.59
6th						1	0.77	0.71	0.73	0.66	0.63	0.61	0.58
7th							1	0.71	0.73	0.67	0.67	0.61	0.58
8th								1	0.75	0.69	0.63	0.6	0.57
9th									1	0.72	0.7	0.66	0.61
10th										1	0.81	0.7	0.66
11th											1	0.72	0.64
12th												1	0.72
13th													1

Figure 4. The Jaccard similarity of most frequent words (terms) between the first and second waves (left) and the Jaccard similarity between 13 months of the COVID-19 pandemic in Croatia (right).

Table 2, available in the Multimedia Appendix 1 Section-A5, lists the 50 most frequent terms used in news publications for the first and second epidemic waves, respectively.

In the second step, the terminology analysis was conducted at the month granularity. The Jaccard similarity is calculated for the 250 most frequent terms between every two months. The heat map in Figure 4 shows significant deviations in January and February (yellow-colored squares), followed by some high overlaps in terms of the Jaccard similarity (red hued squares). The green triangle on the heat map indicates the period with the highest overlap. In other words, the most used terminology in those months was the most similar. All the angles of the green triangle have a value of 0.75 and thus delimit the months in which the epidemic subsided and people lived with less pressure from infection. During the “green triangle” months the media virtually revolved around the same most frequent pandemic terms, and the reasons are to be found in two upcoming events: parliamentary elections and the tourist season. Moreover, the prime minister, running in the elections, announced that the income from tourism, always important for Croatian GDP, would be crucial during the pandemic.

The prevalence of pandemic terminology in the first and second epidemic waves was quantified and visualized in Figure 5. The terms below the blue diagonal line were more frequent in the media during the first wave and above the line during the second wave.

According to the results, the symptoms that are more common in the first wave are a cough, sore throat and respiratory symptoms, while in the second wave more was written about the lungs and breathing, taste, smell and a dry cough. It is important to note that the differences in frequencies between all these terms are small, and to conclude that they are spoken of almost equally in both waves. The symptoms of anosmia, ageusia, parosmia, etc., appear with the highest occurrence frequency.

The necessities for maintaining hygiene and preventing the spread of infection are predominantly mentioned in the first wave. These are: disinfectants, gloves, soap, visors, and even the pharmacies that trade in such supplies. The next important group of terms is related to drugs – azithromycin, Sumamed, paracetamol, and hydroxychloroquine are mentioned more in the first wave. In the second wave, when we gained more knowledge about the disease, Remdesivir is mentioned more, accompanied by the rise of the vaccination terminology - CureVac, Pfizer, AstraZeneca and Sputnik V. In addition to this, the word vitamins is frequent, as well as oxygen due to the intensification of the pandemic outbreak in the second wave.

Subjects from the political scene, such as the minister of the interior affairs (Božinović) or health (Beroš), director-general of the Institute of Public Health (Capak), director of the largest Clinic for Infectious Diseases in Croatia (Markotić) are more frequently mentioned in the second wave. Scientists (for example Lauc and Đikić) are mentioned more in the second wave because they made more media appearances at that time. Still, politicians are mentioned more often than scientists.

During the first wave, more attention was paid to the ways of spreading the disease and infection prevention. Therefore, terms such as spread (infection or disease), isolation, quarantine, infection, and disinfection are mentioned more often in this wave. Interestingly, the terms self-isolation and newly-infected have a significantly higher incidence in the second wave. Interestingly, the words self-isolation, newly-infected, infection, transmission, treatment, sample, positive test, testing, epidemiologist, social distance, to die, patient, mechanical ventilation (respirator) have a significantly higher incidence in the second wave. This might be a result of a significantly higher number of infections in the second epidemic wave, which was magnitudes higher than in the first.

Among the terms that name diseases, the plague and SARS prevail in the first wave, but influenza, SARS-CoV-2, and COVID-19 in the second wave.

General words used for describing COVID-19 infection and disease, such as: virus, coronavirus, infection, hospital and health care, pandemic, epidemic, life, and patient are immediately close to the wave-dividing boundary. Due to their generality, their frequency is magnitudes higher than the frequency of terms that describe or name symptoms, medications, public persons, names of medical institutions, etc.

We paid particular attention to drugs and vaccines that were most frequently mentioned at the time of the pandemic. The details of the observed word groups naming drugs and vaccines can be found in Table A5-2 (Section A5 of Multimedia Appendix 1). The representation of the groups in corpus is expressed as a percentage.
Results are reported separately for drug and vaccine terminology as normalized values per first and second wave. The group of drug-related words occupies 0.38% of the corpus from the first wave, and 0.61% of the second wave. The group of vaccine-related words occupies 0.24%, and 4.63% of the corpus in the first and second wave, respectively. The occurrence of words from both groups increases in the second epidemic wave: 0.23% more terms refer to drugs in the second wave than in the first wave, and as many as 4.02% more terms refer to vaccines in the second wave. In the first wave, existing drugs that could help treat COVID-19 were reported, but with the emergence of some new drugs (e. g., remdesivir), their mention in the second wave was relegated to the background. As the production of vaccines was announced mainly during the second wave, reporting on them became more exhaustive.

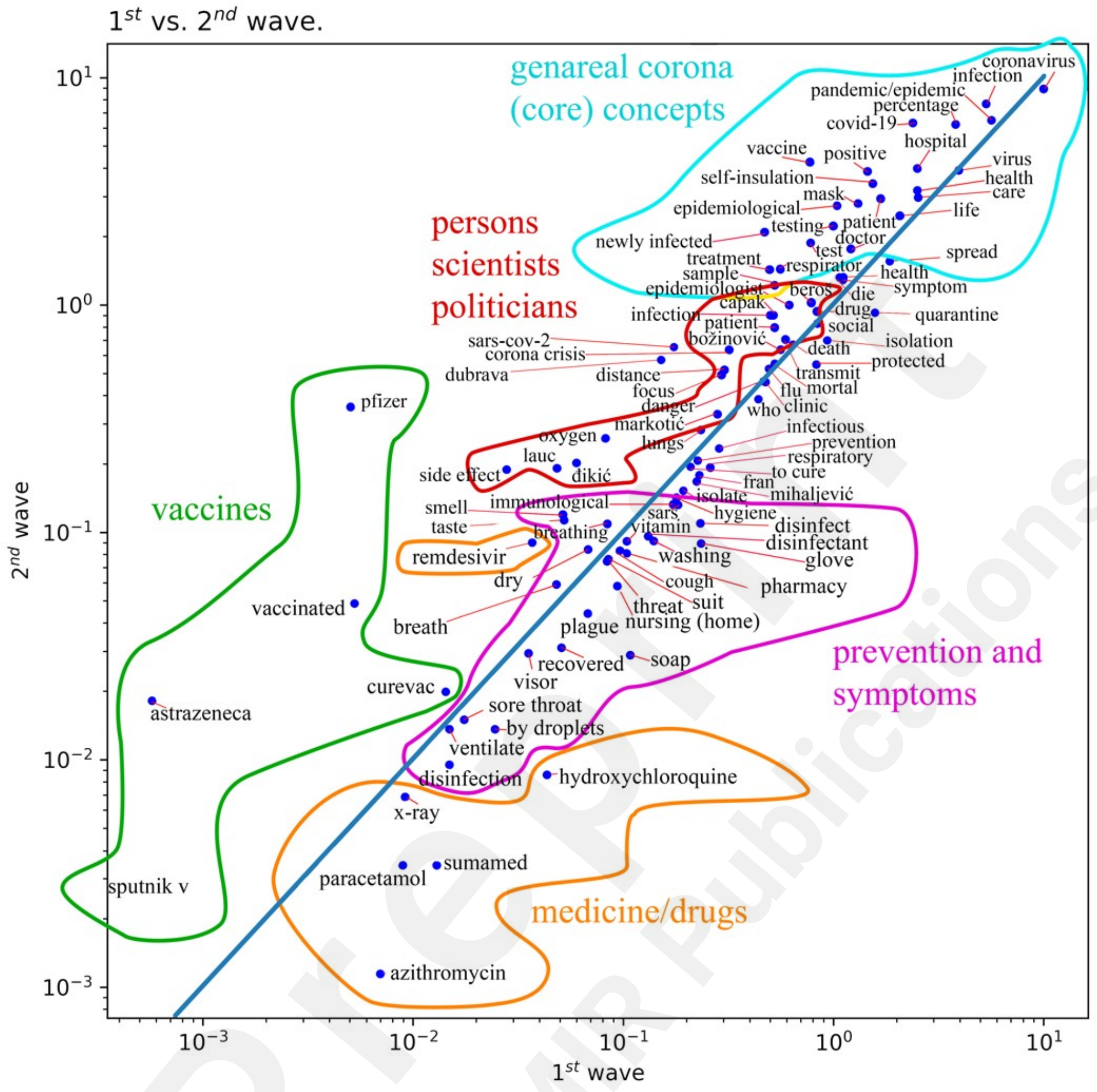


Figure 5. The relationship of epidemiological vocabulary between the first (upper left) and second (lower right) pandemic waves.

The Main Subjects in the Pandemic

An analysis of the ratio of unique entities and the total number of entities in the pandemic articles is obtained by NER. The results indicate that the proposed longitudinal tracking of focal entities can serve as one aspect of infoveillance providing insights into the trends of public interest. Figure 6 reveals that the number of persons and organizations was significantly higher than the number of locations and general (miscellaneous) entities. Combined with the insights from Figure 6 (right) where the total number of detected entities in the categories persons, organizations and locations is fairly equal and the general category (MISC) is marginal, it is possible to enable the consistent tracking of public interest during the pandemic. The left part of Figure 6 provides a numerical insight into the representation of individual subjects in media coverage during the pandemic. In the pandemic year, among the four studied groups of entities, persons (the blue graph area) are the most frequent subjects in coronavirus-related news. In addition to personal names, nationalities also belong to a group of entities called persons.

The personal names were mostly referring to leading figures of the political scene, the presidents of the state and government, the heads of the civil protection headquarters, ministers, scientists, hospital directors, infectious diseases specialists, etc. The second most frequent group were the organizations (green). During the pandemic year, most journalists wrote about hospitals, public health schools, testing centers, civil protection headquarters, WHO, EMA, vaccine companies, and, surprisingly, about the most popular social networks such as Facebook and Twitter, occasionally about football or sports clubs, and often about political organizations and parties.

Locations are the third group of entities (red): states, cities, counties. The captured location entities are the foci of the epidemic or areas where important pandemic related events are happening, e.g. the first vaccines appear, anti-maskers protesting, running out of oxygen for clinical treatment, infection entering nursing homes, state borders closing, borders opening for the tourist season, schools closing, presidential elections, massive earthquake, etc. Basically, during the pandemic the news articles mentioned only a limited and consistent set of locations since not much traveling and migration was allowed. Hence, the number of locations is constantly below the number of persons and organizations.

The last place is occupied by the group of general or miscellaneous entities (violet). The MISC category includes the titles of events, commercial products and brands, documents, TV channels, viruses and diseases, etc. Their occurrence is highly dependent on the time of year or month in which an event, competition, concert or promotion takes place.

Finally, both graphs in Figure 6 reach the maximal values during the peaks of the first and the second waves of the pandemic.

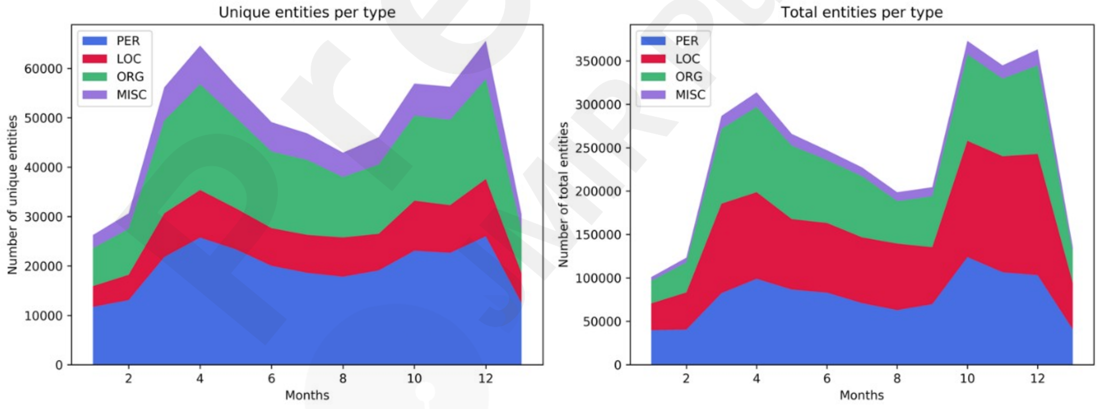


Figure 6. The ratio of the representation of unique entities (left graph) and the total number of recognized entities (right graph) in COVID-19 related media releases in summary for all observed online news media.

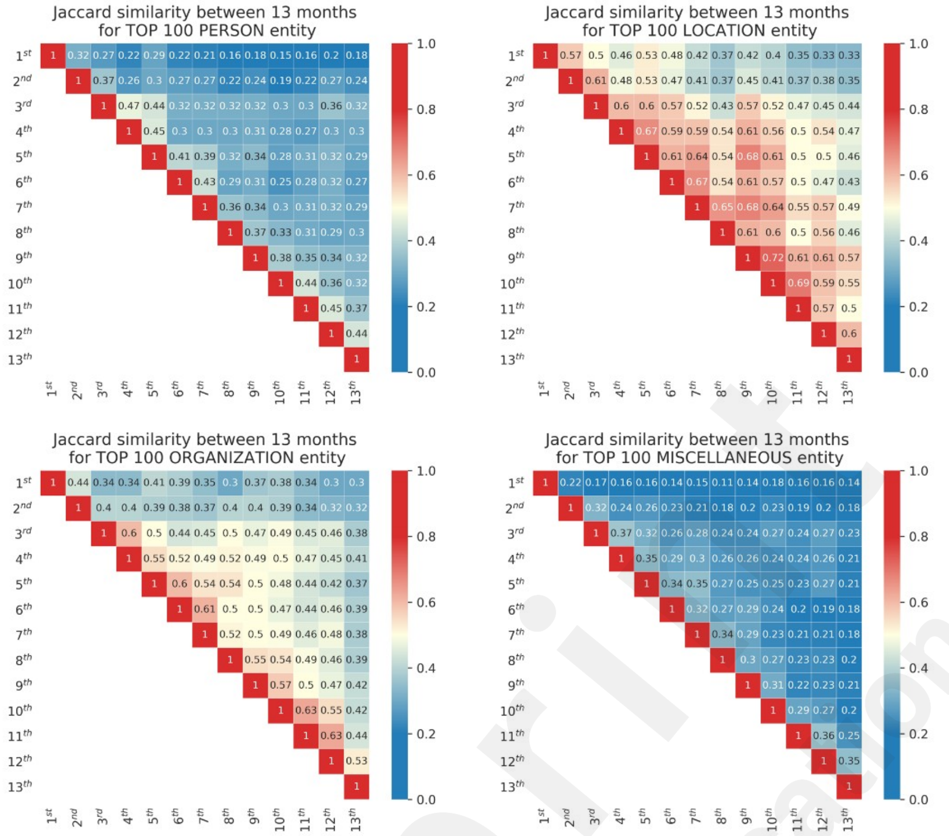


Figure 7. The Jaccard similarity between 13 months for the 100 most frequent entities per four traditional categories: persons (a), locations (b), organizations (c), and other general entities (miscellaneous) (d).

Quantifying the similarity of the top 100 entities by months during the observed pandemic period, we report that the heat map reveals higher similarity Jaccard values between sets of persons (a) and locations (b) than between sets of organizations (c) and miscellaneous (d) entities. The bright red color indicates a stronger overlap, while the dark blue indicates disjunction, i.e. no overlap between the observed entities. The yellow color indicates only a mediocre overlap.

According to the results, locations have the highest overlap, while persons and miscellaneous entities show the lowest overlap across the months. This indicates news is dispersed across many persons appearing in daily events. In contrast, locations are fairly constant during the pandemic, due to the 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 like Wuhan, Lombardy, etc. This reflects the fact that countries closed their borders and the majority of events took place inside the country. That is why Croatian cities and regions are the predominant location entities during the whole period. Similar observations are made for organizations as well. The principal organizations in focus are WHO, local infectious disease clinics, hospitals, etc. Besides medical institutions, the focus is on government entities like the national headquarters, ministries, the Croatian parliament, and political parties. In the second view focused upon are entities related to vaccination. The names of the most popular social networks, Facebook and Twitter, are also always present because news articles were reporting COVID-19 related discussions on Facebook and Twitter. A difference can be noticed (see Figure 7 part c) for January and February 2020 (when the epidemic had not yet been declared in Croatia). The fields are in shades of blue, which indicates that online news media wrote about different organizations until the epidemic broke out. After that (from March 2020 onwards), the color changes to yellow and slightly red. The online news media then predominantly wrote about the same set of organizations for the whole duration of the pandemic even in the month of June in which there was a break between the two epidemic waves.

Furthermore, we perform an entity analysis between the two pandemic waves. Here, we focused on the analysis of the 250 most frequent entities per entity type (PER, LOC, ORG and MISC) and observed their overlap between the two epidemic waves. In this case, the Jaccard similarity shows the largest overlaps for the location entity type (0.5337). It is slightly less for organizations (0.4793) than for persons (0.4045) and the least for the miscellaneous (0.333) category. The interpretation of the results is identical to that in the analysis by months.

Discussion

Principal Results

In this work, we characterize the online media response to the COVID-19 pandemic in Croatia by examining the amount and the content of news articles related to COVID-19. Since most of the studies dealing with the media response to the previous world epidemics were conducted without using NLP for the task of infoveillance, such as [12]–[16], our study is not fully comparable with them. In response to the other infoveillance studies related to the COVID-19 media coverage [20]–[23], [27], [26], [28], this study brings methodological extension. Speciffically, it proposes an integrative infoveillance approach based on NLP methods combined with the Jaccard similarity coefficient for the longitudinal tracking the dynamics of changes across the first 13 months of the pandemic.

Our results show that the number of COVID-19 articles is relatively high, on average, around 40% of the total news articles. This property remains the same during both waves of the pandemic. These results differ from those described in [2] which shows that COVID-19 media coverage has decreased after the initial intense attention at the beginning of the crisis. It seems that the online news media in Croatia tend to write a lot about the pandemic during both waves, as well as during the period after the first wave (which, in fact, in the following three weeks turned out to be a break before the second wave).

The high amount of pandemic-related articles is one of the three indicators of the dramatized media coverage [15], which may indicate an infodemic. However, this alone is not a sufficient condition to confirm an infodemic. Clearly, during the first wave it was necessary to inform the public about the COVID-19 pandemic. The online media plays an important role in informing the public and perhaps this is the main reason why the number of COVID-19 articles is high although the number of COVID-19 cases was lower during the first wave. Consequently, our findings show that there is no high correlation between the number of news articles related to COVID-19 and the number of new cases of COVID-19. These findings are in line with results reported in [26] where the authors show that Zika-related tweeting dynamics were not significantly correlated with the underlying Zika epidemic. Additionally, we show that the number of articles and number of COVID-19 new cases are repeating in cycles within the time window of one week. There is a constant pattern: the number of articles is smaller during the weekends, and fewer new cases of COVID-19 are reported on Sundays and Mondays.

Capturing the dynamics of changes in the most frequent terms across the 13 months shows the highest similarities from May to September 2020. This was the period with the lower number of COVID-19 cases and it is probable that the news articles were less informative and featured similar topics. Additional examination of the similarities between pandemic-related terms indicates that all the general terms (such as coronavirus, infection, pandemic, and hospital) are equally present in both waves. The pandemic-related terminology shifted from some possible remedies and medicines that could be used to prevent or cure COVID-19 (e.g. disinfectant, paracetamol, Sumamed, azithromycin, hydroxychloroquine, etc.) in the first wave to the vaccination process (Pfizer, AstraZeneca, Sputnik V, vaccination) in the second wave. This can be interpreted as a sign of adequate online media coverage – in the sense that the online media provided the available information.

The results of NER show that the online news media concentrates mostly on the people from the state administration – even the scientists featured are often in fact involved as members of the various state bodies. There is a similar pattern as in [19] where it is shown that politicians appear in media coverage more frequently than scientists. The online news media shows low dynamics of changes regarding the locations, while persons, organizations and other entities were frequently changing over the monitored months.

The inclusion of NER as a method for infoveillance makes the longitudinal tracking of the dynamics of changes richer by introducing the insights of focal entities. This approach, however, is not a replacement for the topic modeling that is also used as a part of infoveillance methodology [38], [39]. In fact, thanks to its certain advantages, NER can be a complementary approach to the characterization of the content of the information sources. While topic modeling, relying on the annotator's viewpoint, raises potential ambiguities in detecting and naming the topics and meets the challenges regarding the inter-annotator agreement or consistency [31], NER enables unambiguous monitoring since there is no need for an additional interpretation of annotation, which clearly speaks in the favor of NER as the complement method of the topic modeling.

To the best of our knowledge, this longitudinal study is the first of its kind to use NLP techniques in combination with the Jaccard similarity coefficient for tracking the changes in the most frequent subjects. In addition, since this study is oriented to the Croatian online news media response during the first year of pandemic, it can provide useful data for further comparisons with data collected from other countries.

Limitations

This research has several limitations. First, we characterized media content related to the COVID-19 pandemic by considering only Croatian online news media. However, a large amount of information is present in social media, especially the social networks that were not included in this study. Additionally, individuals are also exposed to COVID-19 related information through traditional sources. Therefore, to obtain a more realistic picture of media content related to the pandemic it would be advisable to extend the analysis of all sources. Hence, in future work we plan to extend this study by integrating heterogeneous data sources such as online social networks and similar social media platforms, online forums and all other sources of textual data in social media such as user comments on online news media. Second, this study is focused only on the Croatian language, however, the same longitudinal approach can be applied to any other language and/or country while the whole methodology is portable and only dependent on the available data sources and the maturity of the NLP methods per selected language.

Furthermore, there are many possible extensions of the reported research. For example, in the part of the research with inferential statistics we use only one variable (the number of new COVID-19 cases), but there are also some other variables (e.g. the number of deaths, number of hospitalizations, number of patients in an ICU or on a respirator) that can be studied as potentially related to the amount of published articles. Moreover, a number of NLP methods can be applied to the infoveillance (i.e. topic modeling combined with polarity of the sentiment, attitudes in comments, etc.). Another important direction of our future research is to develop a full stack of NLP-based methods focused on the longitudinal monitoring of the infodemia, infoveillance, health-crisis communication and infodemic management.

Conclusion

The presented approach enables the infoveillance of online media in response to the COVID-19 pandemic through the quantification of the share of COVID-19 articles. Specifically, in this study, we address three open research questions and our main findings are as follows.

The low correlation between the number of COVID-19 related articles and new cases indicates that the amount of media content is not driven solely by the number of new COVID-19 cases, but rather by external processes. In the first wave, the large amount of news articles was necessary to inform the public about the new disease and the pandemic outbreak. In the second wave, the large amount of news articles was important to communicate findings such as vaccines and other epidemiological measures.

The deeper insights are provided by analyzing the media’s content. The quantification of the dynamics of the changes captured by the Jaccard similarity reveals that there are slow changes in key terminology, locations and institutions. The similarity between the most frequent terms is higher than 50% across all the observed months (except for January 2020) and higher than 70% from May to September 2020. This may indicate the narrow focus of reporting by online media during certain periods. However, the additional analysis of the frequencies of the pandemic-related terms between the two waves indicates that there was a shift from the initial medical terminology known in the first wave to the novel medicine approaches and vaccines in the second wave.

To conclude, the online media had a prompt response to the pandemic in the sense of quantity (the number of articles) in both waves that occurred during the first 13 months of the pandemic. Despite the high amount of COVID-19 related articles, the key terms and entities encountered slow changes. However, the results based on tracking the dynamics of the changes of pandemic-related terminology suggest that the media covered the important changes during the pandemic (e.g. the number of infected people, prevention measures, vaccine production etc.).

Overall, the proposed infoveillance approach based on NLP for the longitudinal tracking of the dynamics of the changes enables a better insight into the online news media’s response. This research contributes to a better understanding of the published content related to COVID-19 in the Croatian online news media and can be further exploited for improving crisis communication.

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 under the number uniri-drustv-sp-20-58.

Conflicts of Interest

None declared.

Abbreviations

ACF: Autocorrelation function
CCF: cross-correlation function
MI: mutual information
NER: name entity recognition
NLP: natural language processing
NMI: normalized mutual information
PMI: pointwise mutual information

References

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

[2] O. Pearman *et al.*, “COVID-19 media coverage decreasing despite deepening crisis,” *The Lancet Planetary Health*, vol. 5, no. 1, pp. e6–e7, 2021.

[3] Y. Liu and J.-A. Cui, “The impact of media coverage on the dynamics of infectious disease,” *International Journal of Biomathematics*, vol. 1, no. 01, pp. 65–74, 2008.

[4] Y. Wang, J. Cao, Z. Jin, H. Zhang, and G.-Q. Sun, “Impact of media coverage on epidemic spreading in complex networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 23, pp. 5824–5835, 2013.

[5] C. Xia *et al.*, “A new coupled disease-awareness spreading model with mass media on multiplex networks,” *Information Sciences*, vol. 471, pp. 185–200, 2019.

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

[7] G. Eysenbach, “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*, vol. 11, no. 1, p. e11, 2009.

[8] G. Eysenbach, “How to fight an infodemic: the four pillars of infodemic management,” *Journal of medical Internet research*, vol. 22, no. 6, p. e21820, 2020.

[9] V. Tangcharoensathien *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.

[10] World Health Organisation, WHO, “An ad hoc WHO technical consultation managing the COVID-19 infodemic: call for action, 7-8 April 2020,” 2020.

[11] K. R. Chowdhary, “Natural language processing,” *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[12] A. D. Dudo, M. F. Dahlstrom, and D. Brossard, “Reporting a potential pandemic: A risk-related assessment of avian influenza coverage in US newspapers,” *Science communication*, vol.

28, no. 4, pp. 429–454, 2007.

[13] T. R. Berry, J. Wharf-Higgins, and P. J. Naylor, “SARS wars: an examination of the quantity and construction of health information in the news media,” *Health communication*, vol. 21, no. 1, pp. 35–44, 2007.

[14] D.-H. Choi, W. Yoo, G.-Y. Noh, and K. Park, “The impact of social media on risk perceptions during the MERS outbreak in South Korea,” *Computers in Human Behavior*, vol. 72, pp. 422–431, 2017.

[15] C. Klemm, E. Das, and T. Hartmann, “Swine flu and hype: a systematic review of media dramatization of the H1N1 influenza pandemic,” *Journal of Risk Research*, vol. 19, no. 1, pp. 1–20, 2016.

[16] L. Safarnejad, Q. Xu, Y. Ge, A. Bagavathi, S. Krishnan, and S. Chen, “Identifying Influential Factors in the Discussion Dynamics of Emerging Health Issues on Social Media: Computational Study,” *JMIR public health and surveillance*, vol. 6, no. 3, p. e17175, 2020.

[17] E. Almazán-Ruiz and A. Orrequia-Barea, “The British Press’ Coverage of Coronavirus Threat: A Comparative Analysis Based on Corpus Linguistics,” *Çankaya University Journal of Humanities and Social Sciences*, vol. 14, pp. 1–22, 2020.

[18] J. N. Ogbodo *et al.*, “Communicating health crisis: a content analysis of global media framing of COVID-19,” *Health promotion perspectives*, vol. 10, no. 3, p. 257, 2020.

[19] P. S. Har, S. Chinn, and P. S. Hart, “Politicization and polarization in COVID-19 news coverage.(Special Issue: Communicating risk and uncertainty in the face of COVID-19.),” *Science Communication*, pp. 679–697, 2020.

[20] H. Jang, E. Rempel, D. Roth, G. Carenini, and N. Z. Janjua, “Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis,” *Journal of medical Internet research*, vol. 23, no. 2, p. e25431, 2021.

[21] M. S. Satu *et al.*, “TClustVID: a novel machine learning classification model to investigate topics and sentiment in COVID-19 Tweets,” *Knowledge-Based Systems*, p. 107126, 2021.

[22] M. O. Lwin *et al.*, “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.

[23] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, and G. S. Choi, “A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis,” *Plos one*, vol. 16, no. 2, p. e0245909, 2021.

[24] W. S. Paka, B. Rachit, K. Abhay, S. Shubhashis, and T. Chakraborty, “Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection,” *Applied Soft Computing*, no. 107393, 2021.

[25] S. Gundapu and R. Mamid, “Transformer based Automatic COVID-19 Fake News Detection System,” *arXiv preprint arXiv:2101.00180*, 2021.

[26] H. Jelodar, Y. Wang, R. Orji, and S. Huang, “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*, vol. 24, no. 10, pp. 2733–2742, 2020.

[27] R. Chandrasekaran, V. Mehta, T. Valkunde, and E. Moustakas, “Topics, trends, and sentiments of tweets about the COVID-19 pandemic: Temporal infoveillance study,” *Journal of medical Internet research*, vol. 22, no. 10, p. e22624, 2020, doi: 10.2196/22624.

[28] T. de Melo and C. M. Figueiredo, “Comparing News articles and tweets about COVID-19 in Brazil: sentiment analysis and topic modeling approach,” *JMIR Public Health and Surveillance*, vol. 7, no. 2, p. e24585, 2021, doi: 10.2196/24585.

[29] S. Eurobarometer, “Media use in the European Union,” *TNS Opinion and Social, Survey coordinated by the European Commission, Directorate-General Communication, Tech. Rep*, 2012.

[30] “Minitab tutorial: Methods and formulas for Autocorrelation.” [Online]. Available:

https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/time-series/how-to/autocorrelation/methods-and-formulas/methods-and-formulas/

[31] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "An overview of graph-based keyword extraction methods and approaches," *Journal of information and organizational sciences*, vol. 39, no. 1, pp. 1–20, 2015.

[32] D. Fišer, N. Ljubešić, and T. Erjavec, "The Janes project: language resources and tools for Slovene user generated content," *Lang Resources & Evaluation*, vol. 54, no. 1, pp. 223–246, Mar. 2020, doi: 10.1007/s10579-018-9425-z.

[33] N. Ljubešić, M. Stupar, T. Jurić, and Ž. Agić, "Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene," *Slovenščina 2.0*.

[34] K. Babić, M. Petrović, S. Beliga, S. Martinčić-Ipšić, M. Pranjić, and A. Meštrović, "Prediction of COVID-19 related information spreading on Twitter," presented at the MIPRO 2021, Opatija, Croatia, 2021.

[35] K. Babić, M. Petrović, S. Beliga, S. Martinčić-Ipšić, and A. Meštrović, "COVID-19 related communication on Twitter: analysis of the Croatian and Polish attitudes," presented at the International Congress on Information and Communication Technology, London, 27.2 2021.

[36] P. K. Bogović, S. Beliga, S. Martinčić-Ipšić, and A. Meštrović, "Topic Modelling of Croatian News During COVID-19 Pandemic," presented at the (MIPRO 2021, Opatija, Croatia, 2021.

[37] S. Beliga, S. Martinčić-Ipšić, M. Matešić, and A. Meštrović, "Natural Language Processing and Statistic: The First Six Months of the COVID-19 Infodemic in Croatia,"

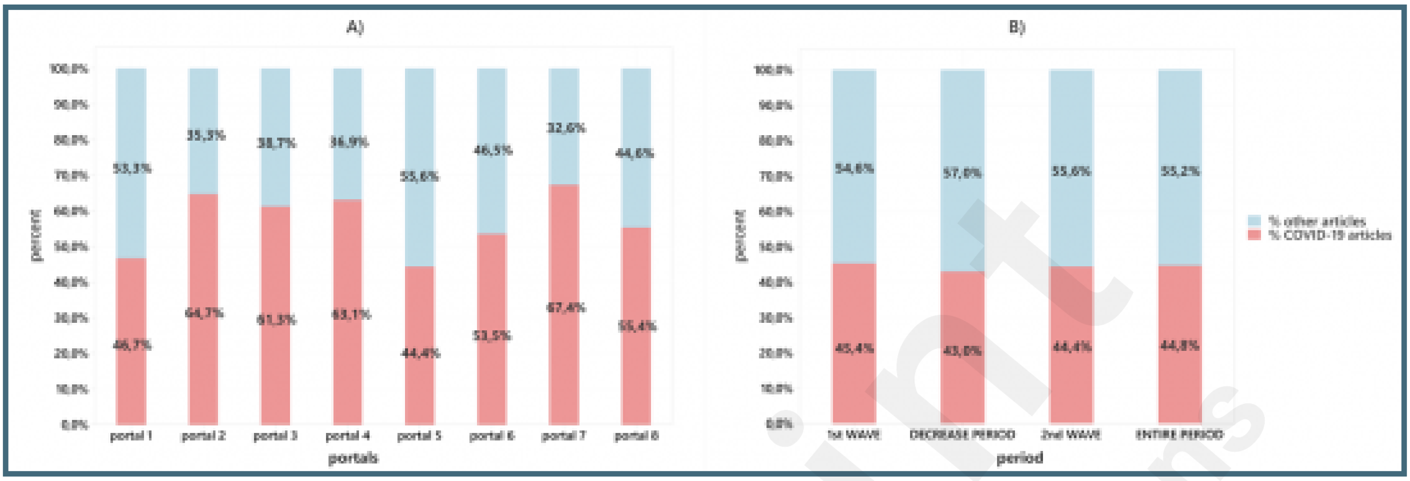
[38] T. K. Mackey *et al.*, "Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales: Infoveillance Study on Twitter and Instagram," *JMIR public health and surveillance*, vol. 6, no. 3, p. e20794, 2020.

[39] N. Gozzi *et al.*, "Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis," *Journal of medical Internet research*, vol. 22, no. 10, p. e21597, 2020.

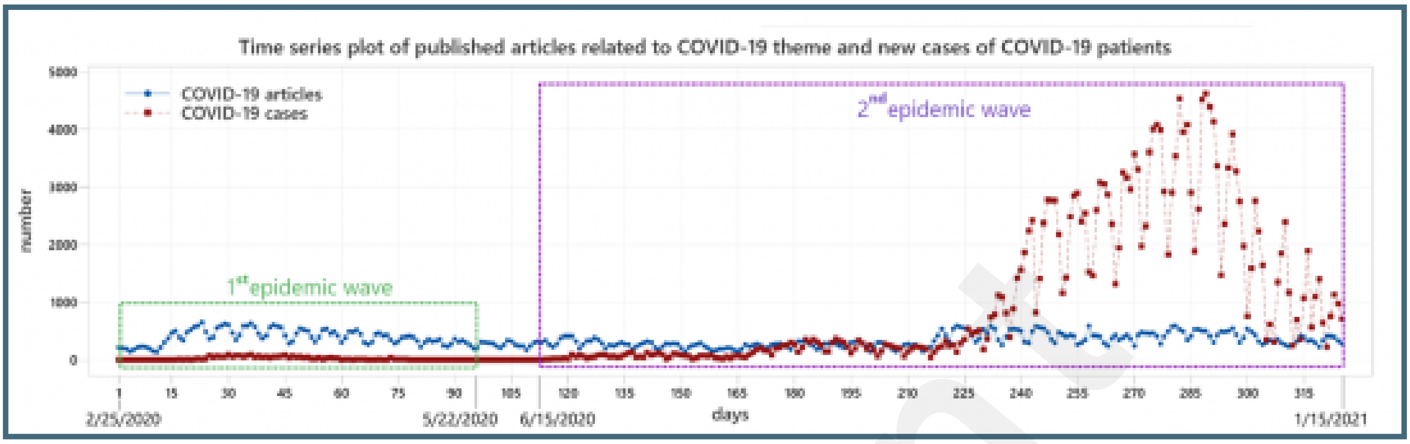
Supplementary Files

Figures

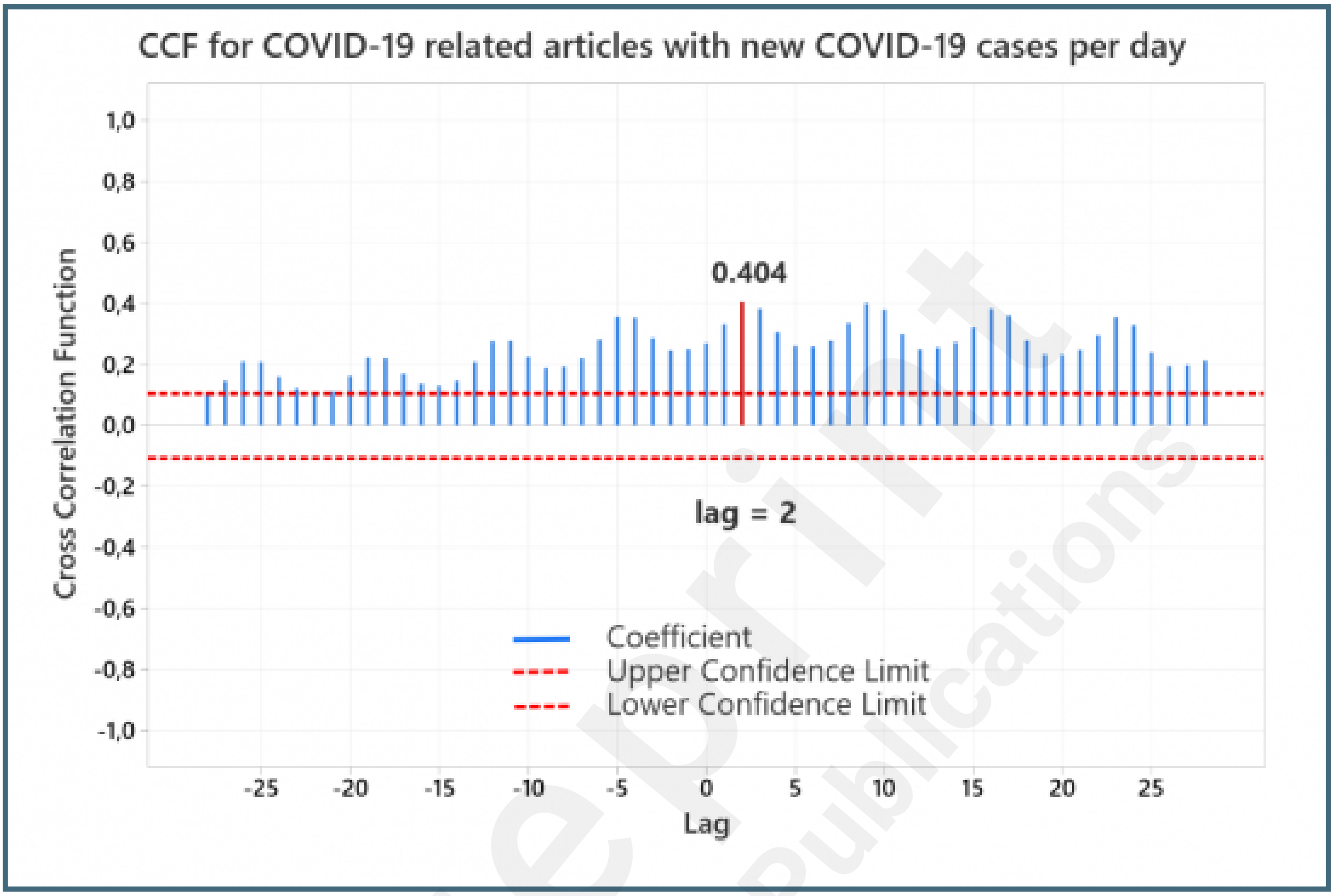
The percentage of COVID-19 related articles summarized per each of the eight online news media during the pandemic in Croatia (February 25, 2020 – January 15, 2021) (A), and the percentage of COVID-19 related articles in the total number of articles summarized across eight online news media for different periods during the pandemic (B).



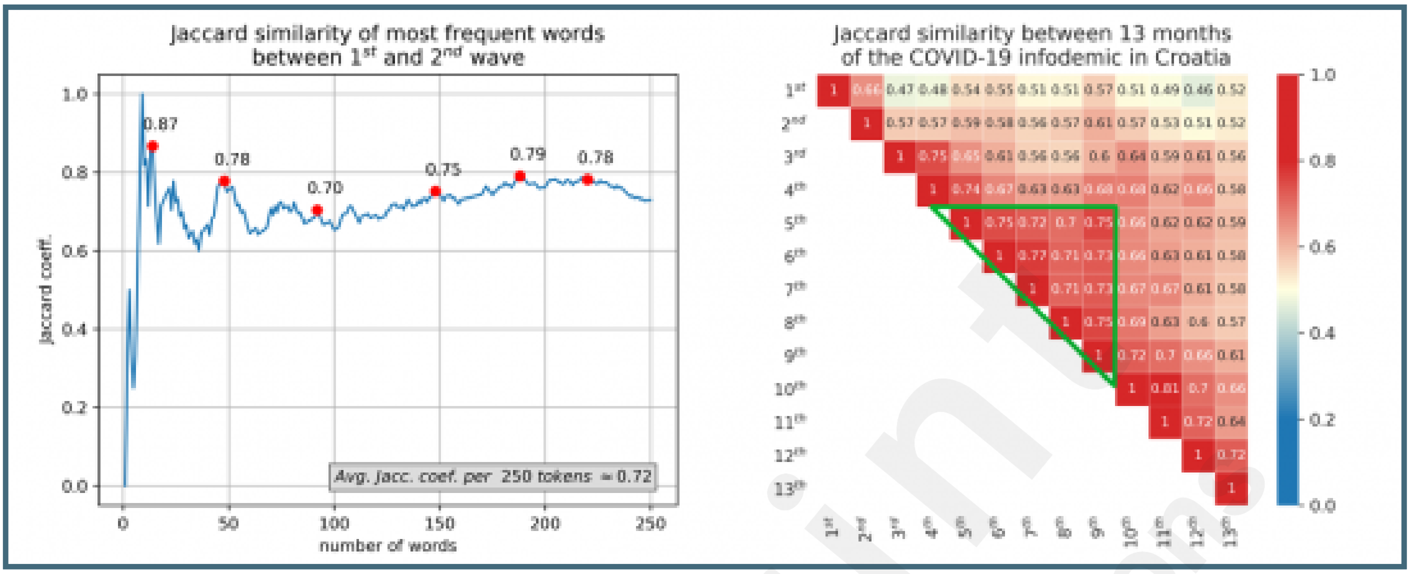
A time series plot comparing the number of published COVID-19 articles per day (blue) and the number of new COVID-19 cases (red) from February 25, 2020 to January 15, 2021.



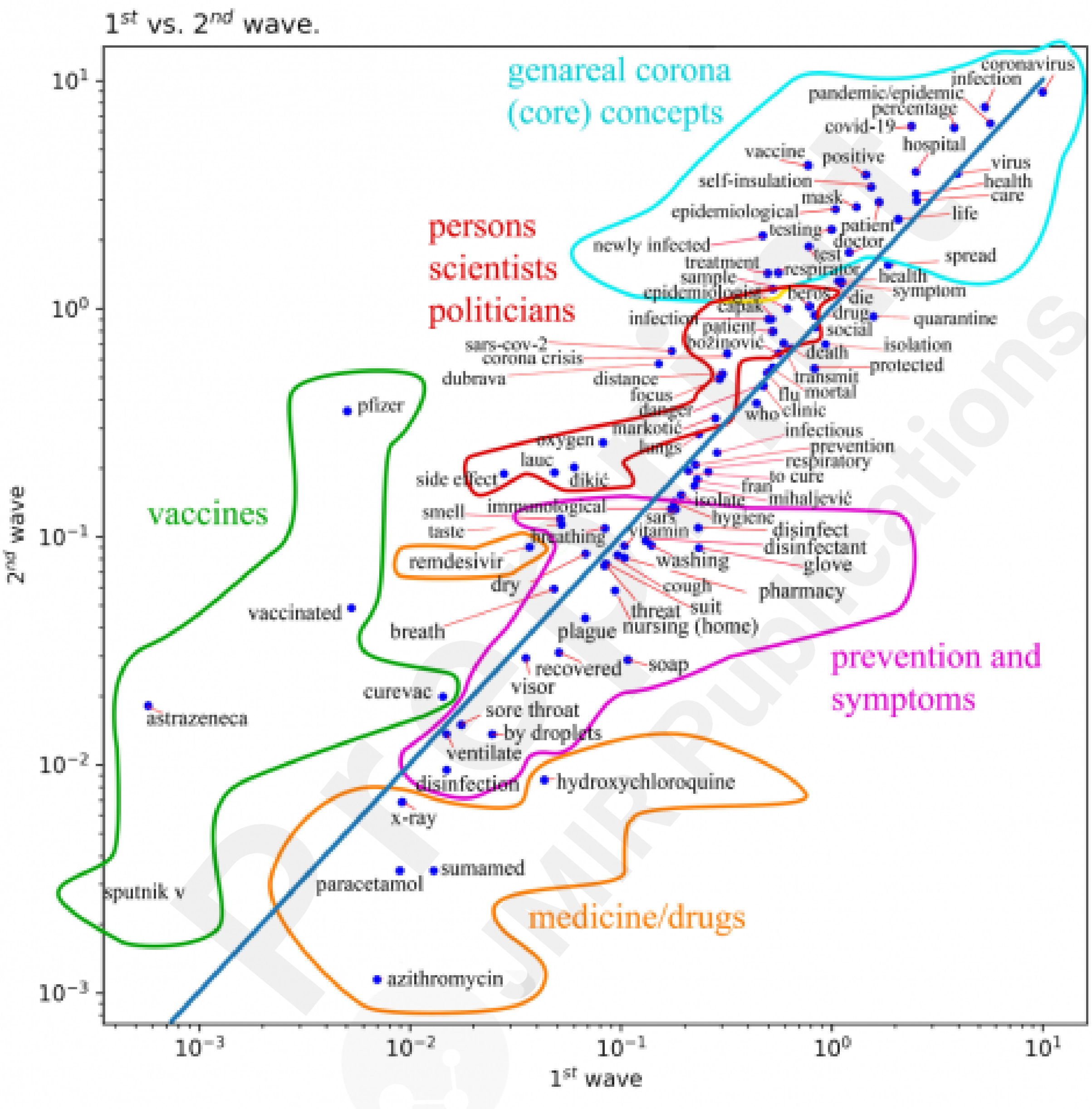
Cross-correlation function (CCF) between the published COVID-19 related articles counts per day and the number of confirmed COVID-19 cases per day.



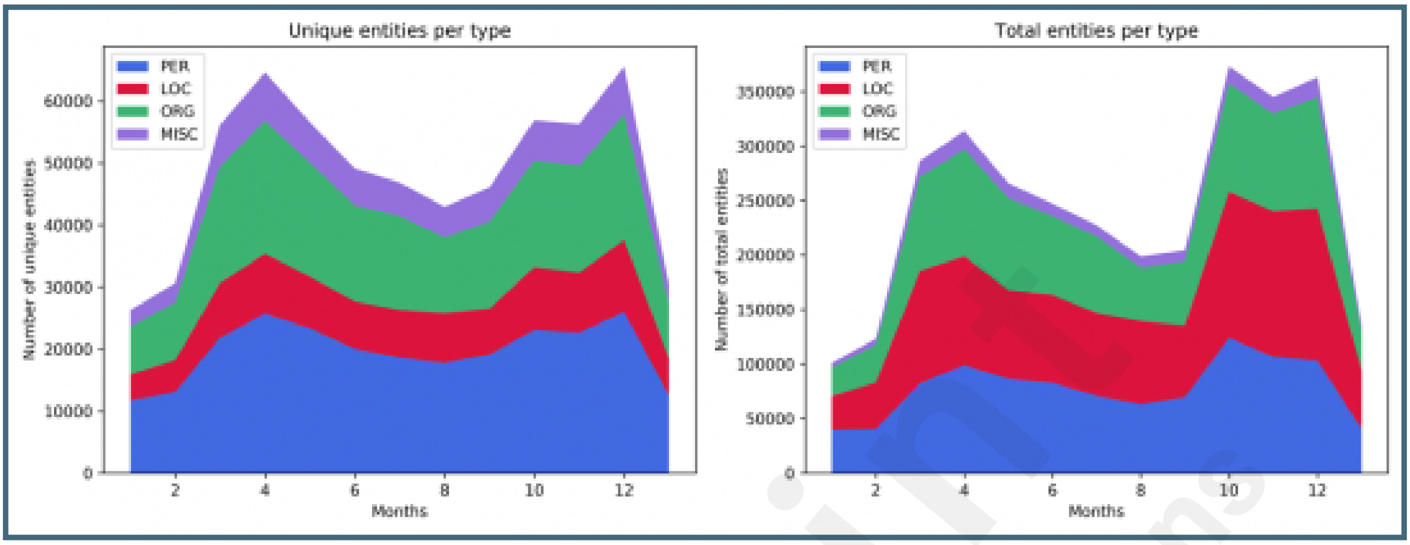
The Jaccard similarity of most frequent words (terms) between the first and second waves (left) and the Jaccard similarity between 13 months of the COVID-19 pandemic in Croatia (right).



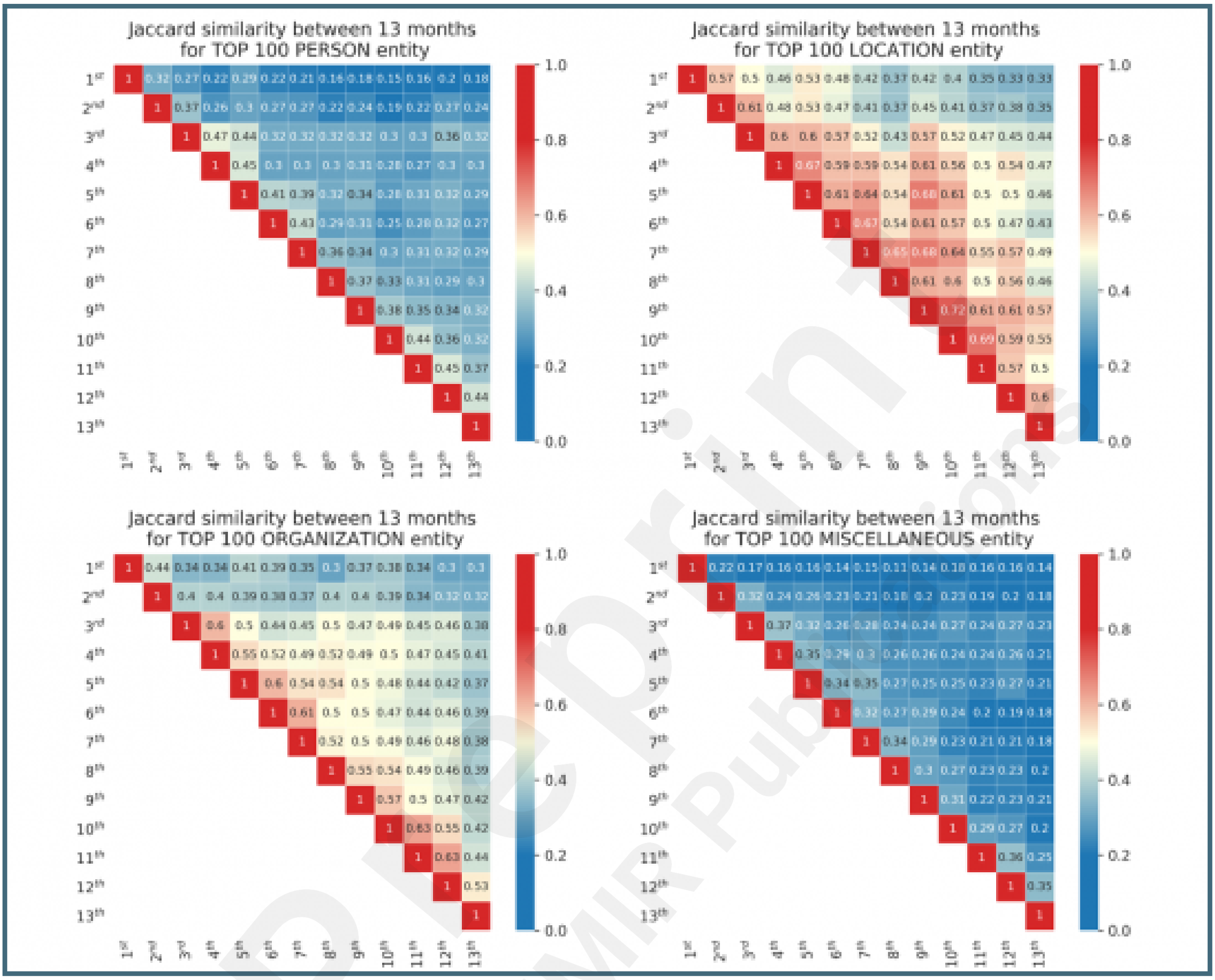
The relationship of epidemiological vocabulary between the first (upper left) and second (lower right) pandemic waves.



The ratio of the representation of unique entities (left graph) and the total number of recognized entities (right graph) in COVID-19 related media releases in summary for all observed online news media.



The Jaccard similarity between 13 months for the 100 most frequent entities per four traditional categories: persons (a), locations (b), organizations (c), and other general entities (miscellaneous) (d).



Multimedia Appendixes

Original Multimedia Appendix.
URL: http://asset.jmir.pub/assets/42550a0b73a1e7e73fff3840ea1c00ea.docx