Towards Computational Content Analysis of Crises-Related News in Electronic Media

Vedran Orešković
Faculty of Informatics and Digital Technologies
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
vedran.oreskovic@student.uniri.hr

Ana Meštrović, Slobodan Beliga
Faculty of Informatics and Digital Technologies
Center for Artificial Intelligence and Cybersecurity
University of Rijeka
Radmile Matejčić 2, 51000 Rijeka, Croatia
{amestrovic, sbeliga}@inf.uniri.hr

Abstract. *In this study, a computational content analysis of crisis-related articles published in electronic media for two crises, namely COVID-19 and the Russia-Ukraine war, is conducted. A set of methods for content analysis is proposed, which involves the combination of exploratory analysis, main topic filtering, and performing named entity recognition enhanced with network analysis. The main finding of this study is a list of the most frequently mentioned persons, locations, and organizations in the news articles of both crises and how they are connected. The results obtained affirm the suitability of the proposed approach for crisis-related content analysis.*

Keywords. natural language processing, NER, computational content analysis, electronic media news, COVID-19, Russia-Ukraine war, eVaražd in.hr

1 Introduction

Analysis of electronic media content is a powerful tool for gaining insights into public opinion. This type of analysis is particularly important in domains related to global crises such as wars, pandemics, climate change, migration crises, etc. as it can help in crisis management. Crisis-related communication in electronic media leads to a massive amount of news articles that can cause an infodemic (Eysenbach, 2002, Eysenbach, 2009). Recently, we have experienced an infodemic caused by the COVID-19 pandemic (Beliga et al. 2021; Beliga et al. 2022; Eysenbach, 2020). The main characteristic of an infodemic is an overabundance of information, misinformation, and disinformation (Eysenbach, 2002). This is a serious problem because, in the event of a crisis, the need for information increases beyond the usual level and people try to get informed as quickly as possible. Thus, automated computational analysis can be of great help in providing fast and relevant information during a crisis.

Many studies use various natural language processing (NLP) techniques to uncover certain aspects of crisis-related content published in electronic media. To identify public perceptions, opinions, and attitudes about specific crisis-related topics, researchers typically combine topic modelling and sentiment analysis (Boon-Itt & Skunkan, 2020) and/or perform named entity recognition (Imran et al., 2015).

An extensive NLP-based analysis of COVID–19-related content in the Croatian language published in electronic media have already been performed (Beliga et al., 2021; Beliga et al., 2022; Bogović et al., 2021, Ilić & Beliga, 2021; Pandur et al., 2021) as well as in social media (Babić et al., 2021a; Babić et al., 2021b; Meštrović et al., 2022). In (Beliga et al. 2021) the authors performed a longitudinal analysis of the COVID-19–related content on electronic media based on natural language processing, extracted main keywords, terminology, and entities, and analysed trends and changes. In (Beliga et al., 2022) the authors combined NLP and statistics and showed that there was an infodemic present in the Croatian news. Next, in (Ilić & Beliga, 2021) the authors performed sentiment analysis of COVID-19 news articles published in electronic media, while in (Pandur et al., 2021) they analysed COVID-19 news articles by combining topic modelling and sentiment analysis.

The results of the authors' previous studies provide detailed insights into the COVID-19–related content published in electronic media. One possible direction for expanding previous research is to compare the content of different crises. Therefore, in this study, the authors extend their previous research to the new domain of news about the Russia-Ukraine war (RUW). The main objective of this study is to analyse and compare the content of news published in electronic media related to the two crises. The authors perform content analysis in the sense of exploratory analysis and apply methods from the field of NLP, specifically topic filtering and named entity recognition (NER).

First, news articles from the *eVarazdin.hr* electronic media were collected and the *eVarazdin-19-23* dataset was prepared. In the next step, articles related to COVID-19 and articles related to the RUW were filtered, afterwards an exploratory analysis was per-

formed. The authors identify different topics in the dataset and compare the number of articles related to different topics and categories. In the last step, NER was applied to extract important entities from the news articles. Based on the NER results, the most frequent entities were organised into the network connecting entities which appear together in the same article. This way it is possible to further explore relations between entities.

The proposed approach provides detailed insights into the content of both crises and allows us to better understand crises. Results of this study indicate that COVID-19–related topics were much more present in news articles than RUW-related topics. Furthermore, NLP-based methods in combination with other techniques such as network analysis can reveal the specifics of the content of both crises. The proposed methodology is applicable to different crises, and the same approach can be easily ported to another dataset.

The rest of the paper is organised as follows. In the next section, an overview of the related work is given. The third section explains the dataset and methods. In the fourth section, the details of the proposed approach and the main results are described. The last section concludes this study with a discussion and recommendations based on the results.

2 Related Work

Fu et al. (2022) suggest an approach based on natural language processing techniques and machine learning to mitigate **flood hazards**. In order to avoid floods, it is very important to identify flood-prone areas. In the conducted study, they emphasize the importance of collecting textual data from the media in assessing the sensitivity of urban areas to floods. In a case study of the city of Dalian, China, the locations of floods were identified from news media data using a NER model. Then, a method based on frequency or distance was used to improve the quality of the extracted flood locations. Finally, flood conditioning factors that included information about historical flood locations were entered into a Support Vector Machine (SVM) model to assess flood susceptibility. In the obtained flood susceptibility map, the high flood susceptibility areas got a recall of 90% compared with the high flood hazard areas in the planning report (Fu et al., 2022).

Liu et al. (2021) have explored how useful the Reddit social media platform is, to monitor the **COVID-19** pandemic. In addition, they tried to answer the question of how human behavior changes over the course of the COVID-19 pandemic in North Carolina. They collected data for the research from the COVID-19 pandemic-related posts from North Carolina subreddit communities. In the task of named entities recognition, they developed a customized system in which they focused on 5 specific categories: Distancing, Disinfection, Personal Protective Equipment, Symptoms and Testing. In addition to NER, the monitoring of the COVID-19 pandemic was carried out through other NLP tasks: feature engineering (using CBOW, Skip-Gram, Glove, and BERT-based sentence clustering), topic entities discovery (Cosine Similarity and LDA topic modeling) and frequency statistics. They concluded that the comprehensiveness of the mentioned NLP methods shows effective monitoring and discovering of how the public's concerns changed over the course of the pandemic, all based on data collected from Reddit. However, the presented results show that representative social media data can be utilized to surveil the epidemic situation in a specific community (Liu et al., 2021).

Truong et al. (2021) presented a dataset of approximately 10k informative sentences about **COVID-19** patients filtered from news articles published on reputable Vietnamese online news sites. When filtering sentences, care was taken to ensure that they reported confirmed, suspected, recovered, or dead cases, as well as the travel history or location of the cases. The authors defined 10 types of entities associated with COVID-19 patients that may be particularly useful for downstream applications in any type of future epidemic. The created dataset represents the largest Vietnamese dataset with corresponding NER annotations (35k entities). Experiments on new a NER dataset compare strong baselines and find that the input representations and the pre-trained language models all have influences on this COVID-19 related NER task (Truong et al., 2021). Datasets like this certainly have an important role in future COVID-19 research and NLP applications related to the Vietnamese NLP community.

Alam et al. (2020) conducted case studies of **hurricanes** Harvey, Irma, and Maria and produced descriptive summaries using artificial intelligence techniques – NLP and computer vision. The studies were conducted using text and image data from Twitter posted during the three major disasters in 2017. One of the main questions of this study was whether and how computer analysis techniques of natural language processing, such as topic modelling, sentiment analysis, NER, etc. can be used to process text and image data in social media to improve situational awareness. The Stanford NER toolkit, which is based on CRFs, was used in the study. The authors investigated entities based on humanitarian categories, especially focused on donation and volunteering category. This focuses on the donation needs that people are writing about, the things that are needed, and in what places. The authors provide visualizations using word clouds to point out which organizations and celebrity people have the most frequent entity mentions (i.e., donate the most). Ultimately, the study showed numerous visualizations made on the basis of data obtained by NLP techniques (including NER). In addition, the successful combination of textual and image data was proven for a comprehensive analysis of crisis situations. Various types

of useful information can inform crisis managers and responders and facilitate the development of future automated systems for disaster management (Alam et al., 2020).

Eligüzel et al. (2022) propose a deep learning approach (RRN-based) for named entity recognition on tweets during **earthquake disaster** in Nepal. They used two different tools to recognize named entities, first the Natural Language Toolkit (NLTK) and then the General Architecture for Text Engineering (Gate). The obtained entities were then used as input data for RNN models. In addition, on the initially labeled data, they trained different variants of models with the original labeled data, from GloVe word embeddings to different variants of RNN models such as LSTM and BLSTM. They concluded that RNN-based approaches perform well in finding named entities. In emergencies, the results of this work can help reduce the effort required to detect event locations and enable better disaster management (Eligüzel, Çetinkaya & Dereli, 2022).

Garcia & Ladeira (2021) conducted a study that aims to propose an approach for the application of NER in the identification of possible adverse events related to the application of COVID-19 vaccines. NER is used for information extraction from social media (websites related to health topics and similar sources). In addition, the authors also conducted an extensive survey on the application of NER for information extraction, in the domain of **postmarket surveillance** for the supplementation of traditional post-market surveillance. The contribution of this research is in the systematization of the positive and negative sides of NER in the information extraction task to supplement traditional postmarket surveillance systems. More practical use of information extraction and NER was shown in work conducted by Nemes and Kiss (2021), where sentiment analysis on Twitter messages is augmented with information extraction and NER to get an even more comprehensive picture of the sentiments of people and to determine how people have related to **COVID-19** over a given period of the pandemic. By extracting additional information and analyzing named entities on the sentiment results categorized and labeled by RNN, they get a better insight into the sentiment. In other words, one gets a detailed insight into how people write on Twitter, what are the characteristics of their expression, which words are used in positive and which in negative attitudes/tweets, which people, locations and other entities are mentioned, but also the events that can affect tweets. With such an analysis they provide a whole new picture to traditional sentiment analysis (Nemes & Kiss, 2021).

Phopli et al. (2021) investigate the events of **natural disasters** by using information from a social network to report on crisis situations. The detection of natural disaster locations is based on the recognition of named entities (NE) in messages from Twitter. In doing so, they use Complex, or Social Network Analysis (CNA or SNA) techniques. Apropos, they analyze the network of named entities, and compute values for centrality measures for certain named entities at the local level of the network, such as: degree, betweeness, and closeness centrality measure. They conducted the experiment only on tweets written in Thai language in the topic of Super Typhon Hagibis blowing in Japan. Based on the graph theory, mentioned centrality measures and majority voting, they determine first 5 words that are related to the event for particular method, and finally only one most important word which can be used to indicate a location that had a natural disaster.

The authors found motivation for their investigation in this approach. They used NER to detect not only locations that are mentioned in electronic media news related to a particular crisis, but also named entities from other categories. However, they measured the importance of a single entity in the corpus based on graph theory using the centrality of a node in the network, which will be discussed in more detail in the following sections.

Related work on the crisis situation related to the **Russia-Ukraine war** mostly analyzes data collected from Twitter with the aim of sentiment analysis (Wadhwani et al., 2023) or topic detection and tracking (De Santis et al., 2023). To a slightly lesser extent, data collected from other social networks have also been analyzed, such as Facebook (Ngo et al., 2022), where sentiment analysis was performed, then data from the Chinese microblogging website Weibo, where emotion recognition and opinion clustering and mining were performed (Chen et al., 2022), and Reddit (Guerra & Karakuş, 2023), where sentiment analysis was also performed. In addition to the above works, there are many other works with sentiment analysis or topic modeling of data related to the Russia-Ukraine war. However, to the best of the authors' knowledge, they have not found any work that uses NER. With this work, they would like to fill the gap in the analysis of news in Croatian electronic media about the Russia-Ukraine war and use NER to analyse important entities.

As far as we know, this is the first study that provided a dataset of newspaper articles written in the Croatian language, collected from electronic media, for the domain that is related to the Russia-Ukraine war. In addition, it is also the first report on NER and the computational content analysis of the crisis-related to the RUW on Croatian electronic news texts.

3 Methodology

3.1 Dataset

News articles were collected from the *eVarazdin.hr*[1] web portal. *eVarazdin* is an independent Internet portal for the area of Varaždin County, which is read in the

[1] https://evarazdin.hr/

wider area of northern Croatia. According to available statistics from July 2023, about 64,000 users followed this portal on the social network Facebook[2].

eVarazdin was launched on March 4, 2010, and since 2013 it has been operating as a media company within the Varaždin consulting financial investment holding *Fine's Grupa*. With the transfer to the ownership of Fine's, the portal was recapitalized in 2015 in the amount of HRK 4,16 million, which makes it one of the most capitalized local media in the Republic of Croatia.

The data collection process included web scraping techniques to extract relevant data from electronic news portal *eVarazdin*. Python libraries such as *BeautifulSoup* (Richardson, 2023) were used to extract textual content from web pages.

The news on the portal are classified into nine newspaper categories: *regional news, sports, focus, politics, black chronicle, eTV, entertainment, lifestyle*, and *interviews*. The data crawler retrieved textual content from all nine categories, encompassing all website URLs in a time span of 4 years, i.e. all news published in the period from April 5, 2019 to April 12, 2023. For the mentioned period, 21,562 newspaper records were collected.

The collected data was stored in a comma-separated values file format, which provides a simple and widely supported structure for tabular data storage. The dataset consist of 12 columns, each representing a specific attribute of the posts. These columns include: *title, subtitle, text, author name, photographer name, URL, publication date, publication time, news category, tags,* and *number of images in news.* Dataset *eVarazdin-19-23* [3] is publicly available for use. It has been cleaned up of duplicate posts and inconsistencies in text formatting.

3.2 Methods

A short overview of methods that were applied in the crisis-related analysis was given in this section. The authors propose an approach based on three main steps: (i) exploratory analysis, (ii) topic filtering and (iii) NER enriched with network analysis.

In the first step, an exploratory analysis of the dataset in terms of giving a timeline with the number of news articles published during the observed period was performed. The number of crisis-related news articles was compared with the number of other news articles. As well, this analysis enables a comparison between the number of articles related to COVID-19 and RUW. In order to provide a more detailed analysis, the distribution of news articles classified into news categories was additionally performed.

Next, a set of the main topics related to COVID-19 and another set of the main topics related to RUW were identified. Then, news articles were filtered according to selected topics. The goal of this step is to determine the number of articles on every topic and identify the most popular topics within the crisis.

In the last step, the named entity recognition task was performed in combination with network analysis. NER is a natural language processing task that extracts information from text to identify and classify named entities into predefined categories, such as personal names, locations, organizations, dates, numeric values, monetary values, and sometimes very specific entities, depending on the domain or application in which they are used. For example, Liu et al. (2021) in their study on the monitoring of the COVID-19 pandemic on the Reddit network introduce their own categories by which they monitor specific entities related to the pandemic: distancing, disinfection, personal protective equipment, symptoms and testing.

In this research, the implementation of NER was performed using the programming language Python and spaCy (v3.6)[4], a well-known open-source software library for advanced natural language processing. A pretrained pipeline and predefined NER model for the Croatian language is used. In this experiment, a small amount of data is used for training. A small Croatian model, *hr_core_news_sm* was trained for NER on written web text (news, media), that includes vocabulary, syntax and entities. The model uses training corpus hr500k 1.0 (Ljubešić et al., 2016) which contains about 500,000 manually annotated tokens on the levels of tokenisation, sentence segmentation, morphosyntactic tagging, lemmatisation and named entities. Accuracy evaluation in terms of the F1-score for the Croatian NER pipeline is 76.1%.

Entities were extracted and divided into four categories: *persons* (PER), *locations* (LOC), *organisations* (ORG) and *miscellaneous* (MISC). Then, the top 15 most frequent entities in each category were selected and four entity networks for both crises were constructed. All networks are weighted and undirected. Nodes represent entities, and two nodes are linked if these entities appear together in at least one news article. The weight on the link represents the number of different news articles in which both entities appear.

The details of every method applied within the proposed approach are described in the next section.

4 Content Analysis and Results

In this section, details of the content analysis were provided, along with explanations of results obtained by the proposed approach.

[2]https://www.facebook.com/evarazdin.hr

[3]https://github.com/sbeliga/InfoCoV/tree/main/CECIIS2023/eVarazdin-19-23.csv

[4]https://spacy.io/models/hr

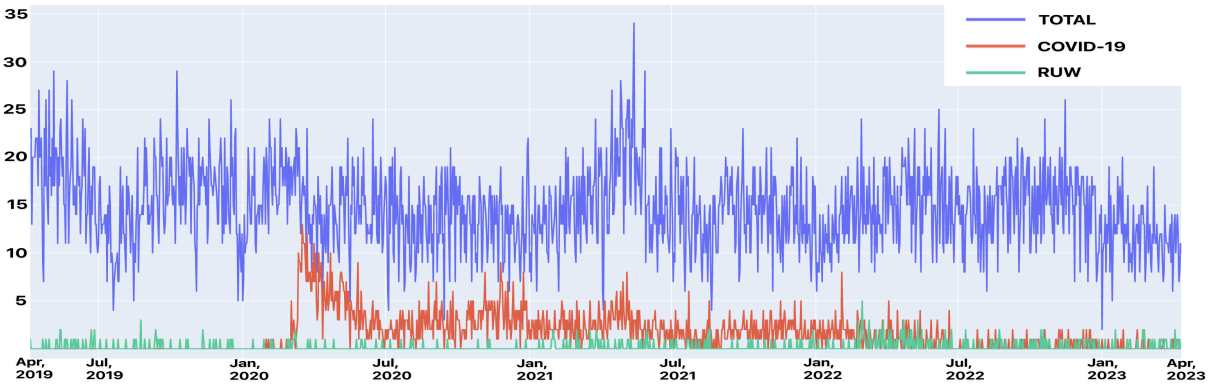

Figure 1: Dynamics of publication of newspaper articles on the *eVarazdin* news portal (blue line) in the time span from April 5, 2019 to April 12, 2023, and announcements related to the COVID-19 (red line) and RUW domain (green line).

4.1 Exploratory Analysis and Topic Filtering

In addition to the observed time period from April 5, 2019 to April 12, 2023, consideration of experiment is also limited to articles on two specific domains: COVID-19 or RUW. Before a deeper analysis of the created dataset *eVarazdin-19-23*, filtering based on keywords has been implemented. Keywords are a set of terms that best describe the subject of a document (Beliga et al., 2015). Filtering was performed using two different thesauruses [5] which were created manually and contain words that succinctly describe a particular topic discussed in the text. One with keywords from the domain of RUW (such as Ukrajina, Rusija, Zelenski, etc.), and the other from the domain of COVID-19 pandemic (such as coronavirus, COVID-19, covid, etc.). Articles that did not meet certain filtering criteria (e.g., news that is not primarily related to the domain of COVID-19 or RUW, but only marginally or incidentally mentions COVID-19 or the war) were excluded from the analysis.

It is important to emphasize that when filtering articles the morphological variations of words were taken into account. Given that the Croatian language is morphologically very rich, news were filtered by considering a truncated form of the (key)word in order to include into extraction criteria all the morphological variations that can arise from the suffixal formation of an individual (key)word.

The articles satisfying the previous condition were extracted along with their associated metadata.

Fig. 1 shows the dynamics of newspaper article publication on the *eVarzdin* portal on the timeline (from April 5, 2019 to April 12, 2023). The red line represents the newspaper publications related to the domain of COVID-19, the green line those related to the RUW domain, while the blue line represents the total number of all publications on the portal. The appearance of the SARS-CoV-2 virus in the world occurred at the end of 2019, and the first newspaper publications on the *eVarazdin* portal mentioned the coronavirus at the very end of January 2020. It was a newspaper article about the suspicion that a patient has COVID-19 and denial of the director of Varaždin Hospital because the patient had the ordinary flu. It is interesting that in the title of this publication, there is a statement "*Coronavirus is disinformation – there is no reason to worry*". The first appearance of the coronavirus in Croatia was recorded on February 25, 2020. The trend of news publications related to this topic on the graph also begins to increase immediately after that, i.e. at the beginning of 2020. After that, the red line begins to grow steeper and reaches its highest peak (March 17, 2020). This is where the real infodemic begins. Press discussions on measures for entrepreneurs, passes for passing through neighboring counties, and the Civil Protection Headquarters begins to prescribe the work regime in the time of the coronavirus crisis. The news reports that public gatherings and manifestations are being canceled due to the bad epidemiological situation, that the local hospital in Varaždin is expanding its capacities for those infected with the coronavirus, and that emergency meetings of the Civil Protection Headquarters are being held to determine measures and guidelines for action. This is followed by a flood of inscriptions about the way catering establishments and service activities work, about the closure of cafes, gyms, hair salons, shopping centers, and banks, but also about people who do not respect the prescribed self-isolation. The announcements are also related to sports topics, primarily about athletes who must report going to countries where COVID-19 is present, the postponement of competitions by UEFA, etc. A large peak on the graph is also visible in mid-March 2020, immediately after the Minister of Health declared the beginning of the COVID-19 epidemic in the Republic of Croatia, and the World Health Organization declared the previous epidemic a pandemic. Electronic media

[5] COVID-19 and Russia-Ukraine thesaurus used for main article filtering: https://github.com/sbeliga/InfoCoV/tree/main/CECIIS2023

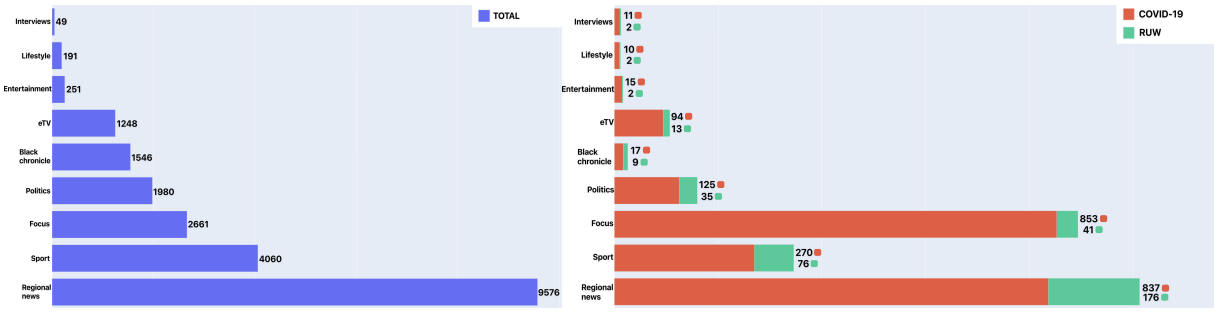

Figure 2: Total number of collected news articles in the time span from April 5, 2019 to April 12, 2023 on *eVarazdin* news portal, classified to corresponding news category (left barplot), and comparison of filtered articles (right barplot) from the COVID-19 (red bars) or RUW domain (green bars).

eVarazdin then intensively reported on the guidelines of the Varaždin General Hospital, Varaždinske Toplice also gave recommendations for the situation with the coronavirus, news were written about the mass closure of sports facilities, announcements were written about measures asking the elderly to avoid going out unnecessarily from their homes, etc.

Later, the steep fall and regrowth of the red line to a new peak occur on several occasions, which follows waves of new coronavirus infections. These results are in accordance with findings in previous research (Beliga et al., 2021). The last big peak was detected at the beginning of February 2022. There is still mass writing about the number of PCR tests and the number of infected, announcements are written about politicians and their responsibility during the pandemic, but also about the lack of blood supplies because people stopped donating their blood during the pandemic. As the virus weakens and the number of infected people decreases, the publication frequency drops slightly. The end of the epidemic in the Republic of Croatia was declared on May 11, 2023. The trend of a small number of publications on the graph is already visible from the very beginning of 2023.

The beginning of news publications about the unrest in Ukraine goes back far before the mass invasion that took place on February 24, 2022. The unrest in Ukraine has been recorded since 2014, and a fairly constant number of posts on the graph is manifested from the very beginning of our defined timeline, which is April 2019. The dynamics of publishing news related to RUW are not as steep and sudden as those related to the coronavirus pandemic. However, some minor peaks can be found. For example, the one at the beginning of January 2022, when the portal reports on the humanitarian organization Red Cross, how the city of Varaždin accepts refugees from the war-affected Ukraine, how the first refugees arrive in Varaždin spa, about accommodation centres, about sports halls where people will sleep and live, about private accommodation offered by local residents. It is also written about the need for Ukrainian language translation.

The scale of the crisis in the domain of COVID-19 and RUW, which manifests itself in the form of the publication of newspaper articles related to that topic, is stronger when it comes to COVID-19 than RUW. The explanation one can find is that Croatia, like the rest of the world, was directly involved in the pandemic. In the case of the Russian-Ukrainian war, it is not directly involved, although it is suffering from an economic crisis and strong inflation, which is partly a consequence of that conflict. Another reason for the stronger intensity of announcements related to the COVID-19 crisis is certainly the two strong earthquakes that occurred on the territory of the Republic of Croatia during the COVID-19 epidemic, which made the fight against the virus more difficult in the areas affected by the earthquake.

The statistics of the collected articles in the observed period are presented in Fig. 2. Overall, (10.35%) of the news items refer to the COVID-19 domain, while only (1.65%) refer to the RUW domain. As expected, among all categories of newspaper articles, the same three categories (*regional news, sports* and *focus*) have the largest number of published articles related to COVID-19, but also to the RUW domain. This is to be expected, as these are the most popular newspaper categories on the portal. However, if we observe them in ratios, the first three places for the COVID-19 domain are occupied by the categories *focus* (32.06%), *interviews* (22.45%) and *regional news* (8.74%). In the RUW domain, the first three places are occupied by the categories *interviews* (4.08%), *regional news* (1.84%) and *politics* (1.87%). It is interesting to note that the categories *politics* and *black chronicles* do not rank higher, as one might expect, considering that we are analyzing crisis situations. The reason could be that some news from COVID-19 or RUW fell into the category *focus* or *regional news*, considering that we are dealing with a local web portal where the type of news can be of a very local nature and therefore be placed in the mentioned categories.

In order to define more specific topics within individual domains, additional topic filtering was done.

Based on additional keywords[6] in the domains about COVID-19 and RUW, the news was filtered into four more sophisticated subtopics of the domain, which gives a more detailed insight into the specificity and importance of each topic. Topic filtering in the domain of COVID-19 was carried out for the topics of *vaccines, measures, infected* and *famous personalities*, and in the domain of RUW for *victims, locations, humanitarian* and *economy*.

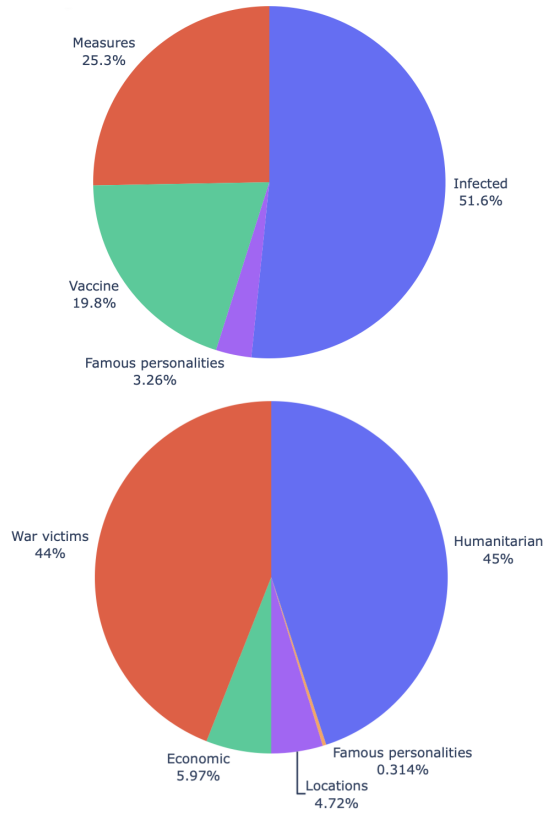

Figure 3: Comparison of the ratios of published news items with regard to the topic filtering for the COVID-19 (top) and RUW domain (bottom).

Additional topic filtering of news articles was performed automatically in such a way that only those news records that met the following criteria were taken into consideration. First, the article should meet the criteria of basic filtering with the previously mentioned filtering dictionary to become a member of the COVID-19 or RUW domain. Second, topic filtering ensures the checking of subtopics within the domain. True positive cases are those articles that meet the filtering criteria of the basic COVID-19 thesaurus, but also of individual thesauruses for a specific COVID-19 topic (e.g. for vaccines or infected subtopic). True negative cases are represented by articles that generally satisfy filtering based on the basic COVID-19 filter, but not the thesaurus of secondary filtering topics for a specific subtopic. Such records are not considered in topic filtering statistics. The same applies to the RUW domain.

The total number of news records is 2229 and 344, of which 2083 and 308 are true positives and 146 and 36 are false positives, in the COVID-19 and RUW domains, respectively. Only true positive news records are considered in further topic filtering analysis (the results of which are shown in Fig. 3).

By analyzing and filtering topics in the domain of COVID-19 (see Fig. 3 - top pie chart), it was found that the most dominant topic in the newspaper articles is related to *infected*, after which equally dominant topics are related to *measures* and *vaccines*. Articles related to the topic of *famous personalities* were published least frequently. Articles related to the topic of *famous personalities* mostly included statements or interviews from famous people such as politicians, people from a specific profession such as epidemiologists, etc. When analyzing the RUW domain (see Fig. 3 - bottom pie chart), the most dominant topic is *humanitarian* and only slightly less dominant is *victims*. About the same number of contributions refer to the topics of *economy* and *locations*, and the topic of *famous personalities* is the least frequent, as is also the case in the domain of COVID-19. It can be expected that during the pandemic, the most relevant newspaper articles are to expect statements or interviews of specialized persons who can speak professionally about the epidemic caused by the coronavirus. However, the analysis shows that the number of publications on such topics was the rarest. The source of the infodemic can be detected and confirmed precisely with the largest number of articles related to the topic *infected*, which was the most talked about and written about during the pandemic. Related to the RUW domain, perhaps one could have expected a greater number of articles written about the economic consequences of the war, but it was still not the most dominant topic. The reason for this may be that a local newspaper portal is analyzed, and not some mainstream portal. It is interesting that most of the coverage was on humanitarian topics such as who and how much donations (money, food, medicines and even weapons) were sent to the victims and the vulnerable.

4.2 NER Enriched with Network Approach

Named entities extracted from newspaper articles are *persons*, *locations*, *organizations* and *miscellaneous*. The approach used to visualize the results is enriched with concepts from the field of complex network analysis (i.e. graph theory). Nodes in a constructed graph represent entities and connections between two nodes in the graph (named entities) exist if they are mentioned in the same newspaper article. If they appear in more than one article, the connection is stronger. In other words, the number of newspaper articles in which the entities appear together is proportional to the weight of

[6]Keyword lists for additional topic filtering: https://github.com/sbeliga/InfoCoV/tree/main/CECIIS2023

the link.

In the visualisations of these NER-based networks, the links with small weights are shown with a thinner red line, while links with larger values of weights are shown with a thicker blue line. The size of a node is expressed with the value of the node's degree centrality. In other words, larger nodes are more important because they have a higher degree centrality.

Due to limited space, only two visualizations for some selected categories of named entities are presented below. The left graph in Fig. 4 illustrates the 15 most significant entities for the *organisations'* category extracted from articles that are in the COVID-19 domain. The entity "City-Municipal Council" stood out as the most significant. In addition, the "Croatian Institute for Public Health", "headquarters" and "general hospital" also stood out as very significant entities in this domain, as well as the larger cities in this region – "Varaždin" and "Ivanec". Here they are recognized as separate entities even though they do not represent organizations. However, they have strong connections with the hospital, the headquarters and the city-municipal council, as they are often mentioned right next to their names, and sometimes as an integral part of the name, for example, "General Hospital Varaždin". In this case, due to the lack of NER tools, they are recognized as separate entities. Two more groups of entities are recognized here as important. One is related to political organizations, parties such as HDZ, SDP and HNS. The second group is related to sports organizations such as HNL (Croatian football league), NK (football club) and football club Dinamo. At the very end, COVID-19 is not an organization, but it is singled out as a very important entity. This is also one of the disadvantages of the used NER tool, but it is reasonable that this entity is important for the domain we are analyzing. The advantage of the combined NER and network approach is that we can also analyze the links between entities. In this way, we can explain the strength of the relationship between individual entities. For example, we see that there are connections between health organizations and headquarters, sports organizations and cities, and, for example, strong mutual connections between political parties.

The right part of Fig. 4 illustrates the extracted entities for the *locations* category from articles that are in the RUW domain. Russia, Ukraine and Croatia are the three central entities with very strong ties, which implies that they are mentioned in the same press releases and that they are also the three most important entities in this domain. Russia is the most frequent entity. According to the visualization, in addition to the capital of the county – the city of Varaždin, other smaller towns and larger cities also appear in the announcements. These are Ludbreg, Ivanec, Novi Marof, Čakovec and Zagreb. It is often written about Europe and Australia, as well as other countries, except for the aforementioned Russia, Croatia and Ukraine. The entity "Stričak" does not represent the location. Due to the imperfection of the NER tool, it is recognized as a location, but it actually refers to a politician and the mayor of Varaždin County. Interestingly, "Municipality" entity is also mentioned, which by itself represents an incomplete location because it does not specify an exact municipality. Nevertheless, the "Municipality" node has many connections to other nodes and still belongs to one of the 15 most significant entities in this analysis.

5 Conclusion

In this study, an approach for the content analysis of crisis-related news articles in the Croatian language published in electronic media is proposed. The authors analyse and compare two domains: COVID-19 and RUW. Exploratory analysis, topic filtering and NER were performed. The proposed approach extends NER with the network analysis by constructing NER-based networks of entities that co-occur in news articles. In this way, analysis was enriched by gaining better insight into the relationships of entities involved in the crisis. Moreover, this overcomes the lack of the NER technique by noticing strong connections between entities that can form a whole, which can not be recognized by NER. This is a particularly important extension in the case of crisis-related content analysis, as the vocabulary of the language expands with new terminology and previously trained NLP-based algorithms do not perform well with the new terms.

From the results presented in the previous section, several conclusions about the content can be drawn. Firstly, the number of news articles about COVID-19 is higher than the number of news articles about RUW in all categories. This can be explained by the fact that the pandemic was present in Croatia while the war was taking place in other countries. Secondly, the most represented category is *regional news* which is to be expected in this type of electronic media. Next, the frequencies of news articles over the observed period reveal peaks that occur during crises. Finally, the network analysis of the main entities helps in the detection of most present persons/locations/organisations in the news articles about the crisis and how they are connected.

Overall, the presented results confirm that the proposed approach provides a good insight into crisis-related content for two different crises. This indicates that the same approach could be applied to the analysis of any other crisis-related content.

This is only preliminary research which the authors intend to expand into a comprehensive methodology for analysing crisis-related communication in general. The next step in their future work will be to extend the proposed approach with new methods and techniques suited for crisis-related communication analysis.

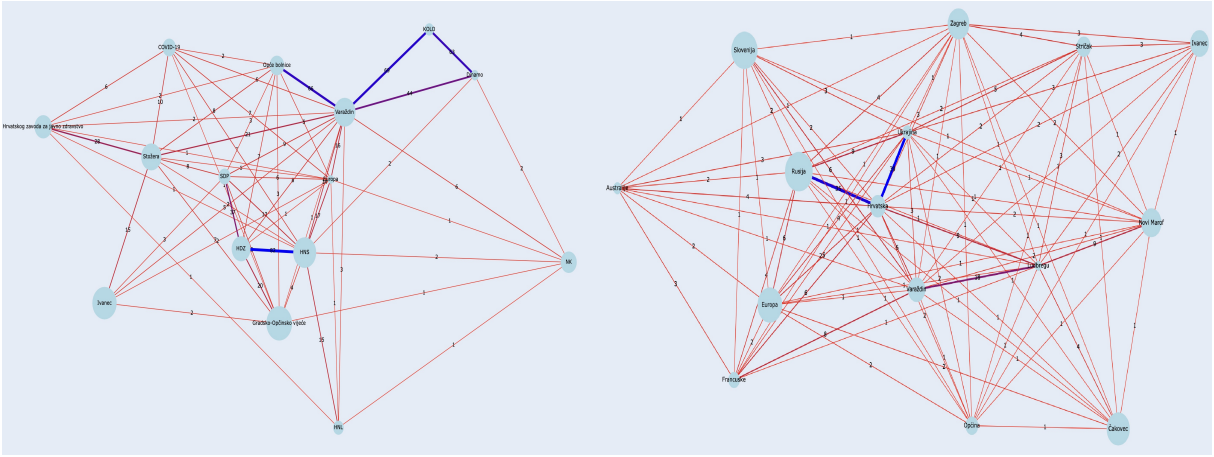

Figure 4: Networks of named entities: the left graph represents entities of class ORGANIZATIONS for press releases from the COVID-19 domain, and the right graph represents entities of class PERSONS for press releases from the RUW domain.

Acknowledgments

This work has been fully supported by the University of Rijeka projects: uniri-mladi-drustv-22-39 and uniri-drustv-18-38.

References

Alam, F., Ofli, F., & Imran, M. (2020). Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria. *Behaviour & Information Technology*, 39(3), 288-318. doi:10.1080/0144929X.2019.1610908

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2021a). Characterisation of COVID-19–related tweets in the Croatian language: framework based on the Cro-CoV-cseBERT model. *Applied Sciences*, 11(21), 10442. doi:10.3390/app112110442

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Jarynowski, A., & Meštrović, A. (2021b). Covid-19–related Communication on Twitter: Analysis of the Croatian and Polish Attitudes. In *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 3* (pp. 379-390). Singapore: Springer. doi:10.1007/978-981-16-1781-2_35

Beliga, S., Meštrović, A., & Martinčić-Ipšić, S., (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of information and organizational sciences*, 39(1), 1-20.

Beliga, S., Martinčić-Ipšić, S., Matešić, M., Petrijevčanin Vuksanović, I., & Meštrović, A. (2021). Infoveillance of the Croatian Online Media During the COVID-19 Pandemic: One-Year Longitudinal Study Using Natural Language Processing. *JMIR public health and surveillance*, 7(12), e31540. doi:10.2196/31540

Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2022). Natural language processing and statistic: The first six months of the COVID-19 infodemic in Croatia. *The Covid-19 Pandemic as a Challenge for Media and Communication Studies*, (pp. 78-92) Routledge: London, Delhi. doi:10.4324/9781003232049-9

Bogović, P. K., Meštrović, A., Beliga, S., & Martinčić-Ipšić, S. (2021). Topic modelling of Croatian news during COVID-19 pandemic. In *Proceedings of 44th International Convention on Information, Communication and Electronic Technology (MIPRO)* (pp. 1044-1051). IEEE. doi:10.23919/MIPRO52101.2021.9597125

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. textitJMIR Public Health and Surveillance, 6(4), e21978.

Chen, B., Wang, X., Zhang, W., Chen, T., Sun, C., Wang, Z., Wang, F.-Y. (2022). Public Opinion Dynamics in Cyberspace on Russia-Ukraine War: A Case Analysis With Chinese Weibo, In *IEEE Transactions on Computational Social Systems*, 9(3), pp. 948-958, doi: 10.1109/TCSS.2022.3169332

De Santis, E., Martino, A., Ronci, F., & Rizzi, A. (2023). An Unsupervised Graph-Based Approach for Detecting Relevant Topics: A Case Study on the Italian Twitter Cohort during the Russia-Ukraine Conflict. *Information*, 14(6), 330. doi:10.3390/info14060330

Eligüzel, N., Çetinkaya, C., & Dereli, T. (2022). Application of Named Entity Recognition on Tweets During Earthquake Disaster: a Deep Learning-Based Approach. *Soft Computing* 26(1), 395-421. https://doi.org/10.1007/s00500-021-06370-4

Eysenbach, G. (2020). How to fight an infodemic: the four pillars of infodemic management. *Journal of medical Internet research*, 22(6), e21820.

Eysenbach, G. (2002). Infodemiology: The epidemiology of (mis) information. *The American journal of medicine*, 113(9), 763-765.

Eysenbach, G. (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), e1157.

Fu, S., Lyu, H., Wang, Z., Hao, X., & Zhang, C. (2022). Extracting historical flood locations from news media data by the named entity recognition (NER) model to assess urban flood susceptibility. *Journal of Hydrology*, 612(C), 128312, https://doi.org/10.1016/j.jhydrol.2022.128312

Garcia, G., & Ladeira, M. (2021). A proposal for the supplementation of traditional Post-market Surveillance systems based on Named Entity Recognition, In *Proc. of 16th Iberian Conference on Information Systems and Technologies (CISTI)*, Chaves, Portugal, (pp. 1-6), doi:10.23919/CISTI52073.2021.9476560.

Guerra, A. & Karakuş, O. (2023). Sentiment analysis for measuring hope and fear from Reddit posts during the 2022 Russo-Ukrainian conflict. *Front Artif Intell.* 5;6:1163577. doi:10.3389/frai.2023.1163577

Ilić, A., & Beliga, S. (2021). The Polarity of Croatian Online News Related to COVID-19: A First Insight. In *Proceedings of 32nd Central European Conference on Information and Intelligent Systems-CECIIS 2021* (pp. 237-246). Faculty of Organization and Informatics, University of Zagreb.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4), 1-38.

Liu, Y., Whitfield, C., Zhang, T., Hauser, A., Reynolds, T., & Anwar, M. (2021) Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Inf Sci Syst*, 9(25). https://doi.org/10.1007/s13755-021-00158-4

Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec, I.-P. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian, In *Proceedings of the Tenth International Conference on LLREC 2016*, ELRA, (pp. 4264-4270).

Meštrović, A., Petrović, M., & Beliga, S. (2022). Retweet Prediction Based on Heterogeneous Data Sources: The Combination of Text and Multilayer Network Features. *Applied Sciences*, 12(21), 11216. https://doi.org/10.3390/app122111216

Ngo, V.M., Huynh, T.L.D., Nguyen, P.V. & Nguyen, H.H. (2022). Public sentiment towards economic sanctions in the Russia-Ukraine war. *Scottish Journal of Political Econom*, 69, 564-573. https://doi.org/10.1111/sjpe.12331

Nemes, L., & Kiss, A. (2021). Information Extraction and Named Entity Recognition Supported Social Media Sentiment Analysis during the COVID-19 Pandemic. *Appl. Sci.*, 11(11017). https://doi.org/10.3390/app112211017

Pandur, M. B., Dobša, J., Beliga, S., & Meštrović, A. (2021). Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal. In *Proceeding of Conference on Data Mining and Data Warehouses 2021*, Slovenian KDD Conference on Data Mining and Data Warehouses (SiKDD 2022), Ljubljana, Slovenia.

Phopli, W., Maliyaem, M., Haruechaiyasak, C. & Ketmaneechairat, H. (2021). Microblog Entity Detection for Natural Disaster Management. *Journal of Advances in Information Technology*, 12(4), 351-356. doi:10.12720/jait.12.4.351-356

Richardson, L. (2023). Beautiful soup documentation. Retrieved from https://www.crummy.com/software/BeautifulSoup/ (1 June 2023)

Truong, T. H., Dao, M.H., & Nguyen, D. Q. (2021). COVID-19 Named Entity Recognition for Vietnamese. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp.2146-2153). ACM. https://aclanthology.org/2021.naacl-main.173

Wadhwani, G.K., Varshney, P.K., Gupta, A., & Kumar, S. (2023). Sentiment Analysis and Comprehensive Evaluation of Supervised Machine Learning Models Using Twitter Data on Russia-Ukraine War. *SN COMPUT. SCI.* 4, 346. https://doi.org/10.1007/s42979-023-01790-5