Data Analysis of the Web News Headlines based on Natural Language Processing

Hrvoje Karna, Maja Braović, Linda Vicković, and Damir Krstinić

***Abstract*—This paper explores the problem of media content data analysis with the focus on the phenomenon of vaccination, closely related to the COVID-19 pandemic. The presented research is an extension of the previous work, but it differs in two main areas. Firstly, the text corpus submitted to the analysis has been considerably increased. Secondly, the previous data analysis was performed on the body part of the posts, while now it is focused on the most prominent part of the news posts, their headlines. This change from body to headline analysis was provoked by significant differences in their characteristics and the fact that most people read only headlines. Described data acquisition uses an advanced content collection approach followed by the modeling process, during which a set of natural language processing algorithms were applied. To enable the comparison, the model uses the same set of algorithms in the modeling phase like in previous work. The main contributions of the work are manifested in: i) approaching the problem from a new perspective, ii) applying more efficient method of data collection, and crucially iii) enabling the comparison of analysis results for individual parts of the content, which ensured a comprehensive insight into the characteristics of news posts.**

***Index terms*—data mining, information extraction, natural language processing, news portals, text analysis.**

I. Introduction

In modern society, digital news have taken a key role in informing the public. The Internet has long been the most popular source of obtaining information in a number of fields, ranging from education [1], health [2] and various others but it is certainly the most prominent choice for accessing the news [3]. Following the decline of printed media, digital news platforms or web portals practically took over the role of information source for wide population [4, 5]. Online news consumption is constantly growing [6] especially among younger audience [7]. Nowadays, consumers of online content are not faced with the lack of information but with the information overload [8]. Readers often struggle to find the relevant content in the flood of material they are exposed to [9] where headings are expected to provide the relevant and most important information of the news [10].

Manuscript received April 21, 2023; revised June 9, 2023. Date of publication June 29, 2023. Date of current version June 29, 2023.

H. Karna is with the Ministry of Defence of the Republic of Croatia and University of Split, Croatia (e-mail: hrvoje.karna@ gmail.com).

M. Braović, L. Vicković and D. Krstinić are with the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, Department of Electronics and Computing, University of Split, Croatia (e-mails: {maja.braovic, linda.vickovic, damir.krstinic}@fesb.hr).

Digital Object Identifier (DOI): 10.24138/jcomss-2023-0047

Headlines are the most distinctive news feature [11]. The word 'headline' itself is defined as the words printed in large letters at the top of a news post, which should summarize the story [12]. They are the most visible part of the webpage and serve to encourage the audience to further read the post content [13]. These headline characteristics motivated us to explore their features in more detail. Web headlines, to be effective, must comply with the requirements set forth by search engines, they have their own grammar and typically reflect freshness, not news value [14] contrary to the typical expectations of the reader.

However, studies report that the majority of the news portal consumers do not move beyond the homepage and the headlines [15]. This is the result of a number of factors, including the following: increased number of people accessing news via online platforms [16], rapidly rising presence of smart devices [17], passive mindset in approaching news [18], etc. In this constellation, the task of the digital content creator is to attract the attention of a reader and how to do it depends on the type of audience being addressed [19].

This paper builds on top of our previous work [20], but it considers the problem from another perspective. The presented study intends to determine the characteristics of the information that readers were exposed to through the most prominent and impressionable part of the posts [21], their headlines. News headlines are extremely important because they attract attention and their wording influences the perception of readers [22]. Therefore, it is critical to determine what information they convey to the readers. In contast, the base study explored the body i.e. the content of the post.

The content collection in initial study was carried by using portal search engines and semi-automated procedure of text retrieval. This study approaches the problem differently, relying on a proprietary system [23] that continuously scans the news portals in Croatia. This mechanism proved to be significantly more efficient, enabling us to form a representative data set of news headlines that entered the modeling process. To enable the comparison of analysis results, the data mining model itself remains intact. The results of the analysis indicated the differences in the headline content characteristics providing better understanding of how the online media operates.

The foregoing clearly defines the main contributions of this work, reflected in: a) a new approach to the problem of web news content analysis, b) advanced method of data collection, and c) the results that provided an advanced insight into the analyzed data. Given that the study was conducted in the aftermath of the COVID-19 pandemic, it is a useful addition to the existing body of knowledge that considers this issue.

The rest of the paper is organized as follows. Section II provides an overview of the studies related to this research. Section III delinates the modeling process, from data identification and collection, application of algorithms, up to the generation of outputs from the model. Section IV provides an interpretation of the results of the analysis accompanied by conclusions. Finally, section V provides guidelines for a future research in the field of application of natural language processing models in analysis of the news portal contents.

II. Related Research

Data analysis has been extensively used in processing of the digital content of various types [24]. Special attention is directed towards the processing of written content in order to gain various insights [25]. Due to the increase in the number of online news portals, the research has been conducted with the aim of efficiently extracting information from the published corpus of texts [26]. The focus is on the preprocessing techniques applied to the content [27] enabling efficient use of different natural language processing (NLP) algorithms [28], for instance Keyword Extraction, Topic Modelling, Text Summarization, Sentiment Analysis or other.

The print media consumption has been in decline for several years while the trend of obtaining the news through the various Internet based services is rising because of the advantages provided by digital alternatives [29]. Report [30] gives key figures how COVID-19 impacted the news industry.

Using digital newspaper corpuses for media analysis has certain challenges and caution has to be taken when this kind of research is performed [31]. The analysis of the digital media content has a number of challenges and approaches used on analog formats cannot be simply applied [32]. The researches have confirmed that data mining methods can be efficiently applied to the analysis of web news data [33] and various techniques have been used for the extraction of useful information from the web news [34].

The phenomenon of vaccination has been studied from different perspectives, especially the one of public health [35], but there is also a considerable number of studies that investigate it from a media perspective [36]. The analysis of news coverage of the vaccination issue during COVID-19 pandemic aroused the interest of the scientific community [37], which also approached it from the position of data analysis in order to gain insight and allow more efficient interpretation [38].

III. Study

In this section the details of the modelling process applied in the study are presented. The implemented stages of data analysis follow the sequence that begins with data source selection, succeeded by the content acquisition and structuring, and finally the text corpus was processed with the selected NLP algorithms. The algorithms used in the analysis are respectively: Word Cloud, Extract Keywords, Topic Modelling using Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI), and Concordance. The results of processing are provided for each algorithm and accompanied by the interpretation and comments.

A. Data Source Selection

In order to perform the analysis of post headlines, initially it was necessary to determine the sources from which the data collection would be carried out. Given that analysis is focused on news portals in Croatia, the Reuters Institute Digital News Reports [38] of online brands was used as a criteria for their selection. These reports are published by Reuters Institute for the Study of Journalism (RISJ) on an annual basis and provide key indicators for various news media. The scores for online news brands operating in Croatia are available for the last five years (2017-2022). Table I. contains the brand ratings and usage scores of top news portals used for their selection.

TABLE I
News Portal (Index, 24sata, Jutarnji, Dnevnik, Net) Rating and Usage Scores for Croatia

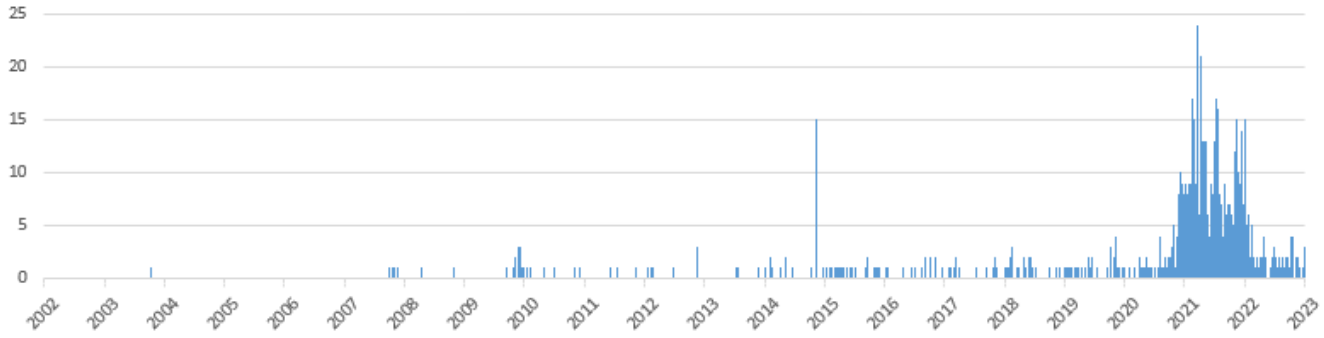
2017	2018	2019	2020	2021	2022
Index (55)	Index (57)	24sata (57)	Index (61)	Index (64)	Index (56)
24sata (54)	24sata (55)	Index (56)	24sata (56)	24sata (57)	24sata (49)
Jutarnji (46)	Jutarnji (45)	Jutarnji (46)	Jutarnji (49)	Jutarnji (48)	Jutarnji (39)
Dnevnik (39)	Net (42)	Net (43)	Dnevnik (41)	Net (39)	Dnevnik (38)
Net (38)	Dnevnik (38)	Dnevnik (36)	Net (41)	Dnevnik (37)	Net (37)

From the data provided in Table I. one can conclude that for the observed period, during which the rating scores are available, the same five most popular news portals (Index.hr, 24sata online, Jutarnji online, Dnevnik online and Net.hr) are constantly present with certain changes in the ranking list. According to [39] these news portals are also among the most trusted online brands. Taking into account the previous indicators and the goal of the research, it is clear why they have been chosen as a source of data.

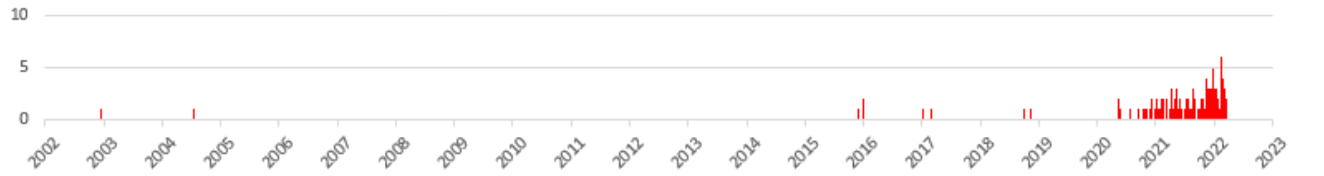
B. Content Identification and Extraction

The data source selection was followed by the process of content identification and collection which was carried out through the TakeLab Retriever [23]. TakeLab Retriever is a platform that scans articles and their metadata from the news portals in Croatia and performs text mining in real-time. For the selected news portals it is possible, using the explorer, to execute queries according to a given phrase (in our case “cijepljenje”, eng. vaccination) and collect the search results. TakeLab Retriever therefore carries out web crawling activities - discovering links to targeted content on selected portals and scraping - extracting data (postId, newsPortal, publishDate and postTitle) from the selected website(s).

By using this data collection approach, a total of 36201 instances of news posts were identified. Considering that the retriever searches the entire content of posts, within the identified instances there were also those containing the keyphrase “cijepljenje” in other parts of the post. Given that the goal of this research is the analysis of headlines characteristics, these instances were filtered out after which the data set was formed from a total of 2743 instances.



(a) Distribution for JCOMSS'23



(b) Distribution for SoftCOM'22

Fig. 1. Post distribution in time for the analysed data sets

Figure 1 shows a comparison of the distribution of the posts between the initial [20] (SoftCOM'22) and hereby presented (JCMOSS'23) research over time. From the charts, it is obvious that the previous research was based on a significantly smaller number of entire posts, while this one uses a much larger set of post headlines that form the corpus for later analysis. The first posts appeared in 2003 and the last one dates the end of 2022. Evidently, the method of finding content applied in the second case is significantly more effective.

Having used the TakeLab Retriever [23] instead of custom search engines and Data Miner [40] for scraping, not only more posts are identified, but also dynamics of publishing both by time (posts by period) and intensity (posts at a selected time) are revealed. Further data inspection determined the results containing a certain number of duplicates - posts that appear multiple times having the same headline but different URL. These were identified only for Index.hr portal and removed. The final dataset is thus formed from 2264 instances, and their distribution per news portal is provided in Table II.

TABLE II
NUMBER OF INSTANCES PER NEWS PORTAL

NEWS PORTAL	NO. OF INSTANCES
INDEX	684
DNEVNIK	408
24SATA	251
JUTARNJI	615
NET	306

Regarding the distribution of posts, it is also worth noting that the trend of post appearance is upward, i.e. more and more posts containing the keyphrase "cijepljenje" appeared over time. In the period leading up to 2020s, we can say that it appeared sporadically. From then until the end of the observed time range, the usage of the keyphrase was increasingly intense, with peak being observed in early 2021. The news portals publishing the greatest amount of texts with the headlines containing the phrase "cijepljenje" are, respectively: Index.hr, Jutarnji.hr, Dnevnik.hr, Net.hr and 24sata.hr.

C. Modeling using Natural Language Processing

The modeling phase of the analysis applies the standard text mining procedure which begins with the data preparation that involves structuring of the corpus into a format suitable for loading into Orange Data Mining [41]. This is followed by the data preprocessing after which the process of modeling with the selected algorithms is performed. Mining process ends with the generation of results at the output of the model, which are subsequently subjected to the analyst interpretation.

C.1 Corpus formation

The elements of the retrieved corpus instances i.e. news posts and their description are provided in Table III:

TABLE III
RETRIEVED ELEMENTS OF CORPUS INSTANCES

ELEMENT	DESCRIPTION
ID	UNIQUE IDENTIFIER OF THE NEWS POST
DATE	THE DATE NEWS POST WAS PUBLISHED
PORTAL	NAME OF THE NEWS PORTAL FROM WHICH THE POST WAS RETRIEVED
URL	PATH TO THE SPECIFIC NEWS POST
HEADLINE	THE HEADLINE OF THE NEWS POST

This means that after the content identification, it was exported into a data file format (.xls) suitable for import into a data mining environment, and it was structured in the form shown in Table III. A comparison of our earlier work (SoftCOM'22 paper [20]) corpus features with those used in this study are shown, as follows:

	SoftCOM'22	*JCOMSS'23*
Text feature:	Body	Headline
Corpus: - Instances:	250*	2264**
- Features:	categorical	categorical
- Types:	9512	3799
- Tokens:	72994	19777

* 50 per portal; ** per portal see Table II.

From the data presented, it is clear that this study (JCOMSS'23 paper) deals with a smaller content in terms of a word count but a more diverse set of textual data. The textual data characteristics will be determined by the following analysis.

C.2 Analytical tool and data loading

Data analysis was performed by using Orange Data Mining environment [41] an open source machine learning and visualization tool based on Python. The text analysis model implementation is depicted in Figure 2. In the initial processing phase, the corpus of texts was loaded into the model using the *Corpus* widget, which begins the formation of a data stream after which the data undergoes processing and analysis procedures. Using the widget *Unique* duplicated data instances were removed, solving the problem described above. From this point begins the modeling part in which the texts are subjected to various processing procedures following the rules of natural language processing.

C.3 Text preprocessing

One section of the model bifurcates into the *Corpus Viewer* where it is possible to perform inspections of post headlines and their respective elements.

After this, the data enters the Preprocess Text element where following transformations are performed:

- *Transformation*: converts the text to lowercase letters
- *Tokenization*: performed via the regular expression: \w + which separates the text into words
- *Filtering*: applies a) a custom built-in script for filtering of redundant words (Croatian stop words), and b) using regular expressions for excluding punctuation marks and digits, see (RegEx) bellow
- *Normalization*: uses the UDPipe Lemmatizer for Croatian language [42].

Regular expressions (RegEx) used during filtering phase:
\d|\.|,|:|;|!|\?|\(|\)|\||\+|\'|\"|‘|’|“|”|\'|\’|…|\-|–|—|\$|&|*|>|<|\/|\[|\]

C.4 Data analysis using NLP algorithms

The content processing performed by using selected NLP algorithms is presented next, accompanied by their outputs.

Word Cloud: displays the tokens in the corpus where their size denotes the "weight" or the word frequency in the corpus i.e. the tokens that appear more often are displayed in larger format, as demonstrated in Figure 3. The order of top ranked words by their weight: "cijepljenje" (eng. vaccination, 2281); "protiv" (eng. against, 342); "cjepivo" (eng. vaccine, 201); "čovjek" (eng. human, 162); "obavezno" (eng. obligatory, 153); "covid" (eng. covid, 149); "početi" (eng. start, 139); "nov" (eng. new, 128); "dijete" (eng. child, 127); "obvezno" (eng. mandatory, 123). Expectedly the most common word in the headlines and the first one on the list was "vaccination". The following words indicate related subjects: "human" and "child". As posts coincided with the onset of COVID-19, the appearance of the keyword "covid" rates high denoting the type of infection.

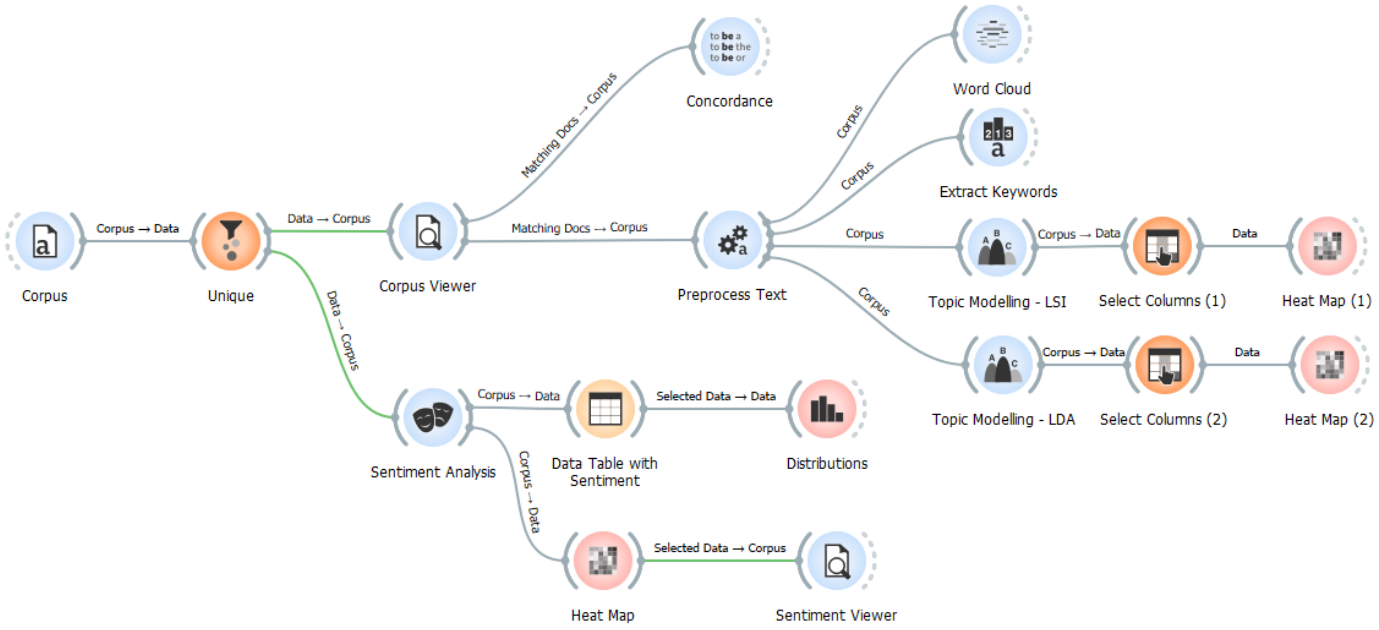


Fig. 2. Text analysis model implementation in Orange Data Mining

Fig. 3. Word Cloud

Extract Keywords: ranks the words by using Term Frequency – Inverse Document Frequency (TF-IDF) factor, that is term frequency weighted by inverse document frequency. A higher value of TF-IDF factor indicates that the words assigned to the factor could be a keyword within a text. The highest-ranking words using this method are: "protiv" (eng. against, 0.023); "cjepivo" (eng. vaccine, 0.016); "obavezno" (eng. obligatory, 0.015); "početi" (eng. start, 0.014); "čovjek" (eng. human, 0.013); "gripa" (eng. flu, 0.013); "covid" (eng. covid, 0.012); "dijete" (eng. child, 0.012); "astrazeneca" (eng. astrazeneca, 0.012); "nov" (eng. new, 0.012). It is interesting to observe the difference in keyword ranking compared to the previously applied method. The word "cijepljenje" is not at the top of the list generated by the TF-IDF factor, while the word "protiv" dominates, indicating the campaign goal "fight *against* the pandemic". Table IV provides a comparison of the results of the two methods applied.

TABLE IV
COMPARISON OF WEIGHT AND TF-IDF WORD FACTORS

WORD_EN	WORD_HR	WEIGHT	TF-IDF
VACCINATION	CIJEPLJENJE	2281	0.000
AGAINST	PROTIV	342	0.023
VACCINE	CJEPIVO	201	0.016
HUMAN	ČOVJEK	162	0.013
OBLIGATORY	OBAVEZNO	153	0.015
COVID	COVID	149	0.012
START	POČETI	139	0.014
NEW	NOV	128	0.012
CHILD	DIJETE	127	0.012
MANDATORY	OBVEZNO	123	0.011

The TF-IDF score for "vaccination" (0.000), the lowest possible, would indicate that the keyword, around which the research was formed, is unimportant. However, after a closer inspection on how it is calculated, it becomes clear why this value was obtained. TF-IDF works by increasing in proportion to the number of occurrences of the word in the document, which is compensated by the number of documents that contain the word. Consequently, the words that are common in every document rank low even though they appear many times. As the criterion for inclusion of the headline into the analysis was that it contains this keyword it is always present. Actually, it is the most present word in the corpus (see WordCloud weights) and consequently TF-IDF classifies it as unimportant, even though this is not the case for this study.

The conclusion is that TF-IDF is not applicable for cases when the corpus is formed in the way given in this study i.e. it functions well for all other words except the keyword based on which the sources for corpus were selected. That is why we will assess the importance of all other words in the corpus by comparing the values of the two indicators used in all cases except for the keyword "vaccination". While we evaluate its importance only on the basis of the weight value generated by WordCloud, which positions it six times higher than any other word in the ranking.

Topic Modelling: two algorithms were used, Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA). LSI and LDA are both techniques used in natural language processing and information retrieval to extract meaning from text data. However, they are based on different mathematical models and have different applications. LSI is based on singular value decomposition, with the goal to find a low-dimensional representation of the text data that captures the main topics or concepts. LSI is tipically used to improve the analysis of the relationships between terms and documents. LDA, on the other hand, is based on a probabilistic generative model that assumes that text data is generated by a mixture of latent topics. The goal of LDA is to discover the main topics and the distribution of words within those topics. Both techniques are extremely suitable for the headline analysis being carried out in this study. The results are provided next, with five topics for each model.

LSI algorithm:

Topic 1: hr: cijepljenje, protiv, cjepivo, čovjek, covid, obavezno, početi, nov, dijete, obavezno
en: vaccination, against, vaccine, human, covid, obligatory, start, new, child, mandatory
Topic 2: hr: protiv, gripa, korona, covid, cijepljenje, počinjati, čovjek, astrazeneca, koronavirus, nov
en: against, flu, corona, covid, vaccination, begin, human, astrazeneca, coronavirus, new
Topic 3: hr: cjepivo, obavezno, obvezno, covid, početi, nov, počinjati, masovno, uvoditi, uvesti
en: vaccine, obligatory, mandatory, covid, start, new, begin, massively, introduce, induct
Topic 4: hr: čovjek, cjepivo, obavezno, masovno, red, dijete, uvesti, zagreba, covid, obvezno
en: human, vaccine, obligatory, massively, order, child, induct, zagreba, covid, mandatory
Topic 5: hr: covid, obavezno, čovjek, dijete, cjepivo, potvrditi, obvezno, gripa, korona, trebati
en: covid, mandatory, human, child, vaccine, confirm, mandatory, flu, corona, need

The algorithm generates a list containing five topics, the keywords within the topic were presented in green or red. Green indicates words with positive connotations, those that appear more often within the topic. Red indicates words with negative connotations, those that must not be found within the text to which the topic is assigned. The first topic has all the

positive words and most often it contains words like: “vaccination”, “against”, “vaccine” “human”, “covid”, “obligatory” and others. The second topic has similar words, but for the text to belong to another topic, it must not contain words “against”, “covid” and similar.

When the results of the LSI algorithm are streamed into the Heat Map widget, a visual representation of the representation of topics in the articles is generated, as shown in Figure 4a. The representation of a topic within an article is assigned a numeric value, if the number is small it means that the topic is not represented in a particular instance. For example, if we take the instance at the upper portion of the Heat Map where Topic 1 is assigned white, Topics 2&3 green and Topics 4&5 light blue color. It is expected that Topic 1 is highly represented and has a high numerical value, Topics 2-5 marginally and the last two similarly but negatively inclined.

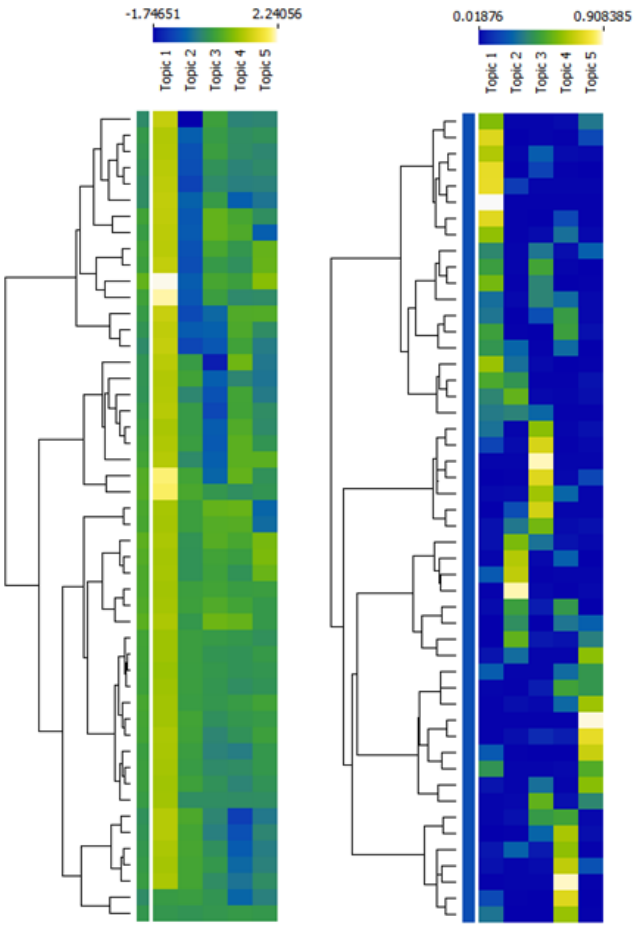
By looking into the details of the numerical values, we can see that Topic 1 really has a high positive value (2.035), Topic 2&3 are practically neutral (0.564; 0.208) while Topics 4&5 have low negative values (-0.076; -0.015). Through thematic modeling, hidden themes within the corpus can be discovered and texts with similar themes identified.

The LDA algorithm has the same capabilities as LSI. Below are shown the results of the application of the LDA algorithm. The topics found and their keywords are very similar to those discovered by the LSI algorithm, but still not exactly the same. This algorithm has no positive and negative words, only keywords that appear in the topic. The first topic the algorithm discovered has the words “vaccination”, “human” and “covid”, similar to LSI. It also provided a few other words that were expected like “dose” and “measure”.

LDA algorithm:

Topic 1: hr: cijepljenje, čovjek, nov, početi, doza, dijete, imati, mnogo, mjera, protiv
en: vaccination, human, new, start, dose, child, have, many, measure, against
Topic 2: hr: cijepljenje, obavezno, covid, protiv, punkt, nov, beroš, video, trudnoći, cjepivo
en: vaccination, mandatory, covid, against, point, new, beroš, video, pregnancy, vaccine
Topic 3: hr: cijepljenje, protiv, počinjati, gripa, cjepivo, booster, doza, dijete, obavezno, korona
en: vaccination, against, start, flu, vaccine, booster, dose, child, mandatory, corona
Topic 4: hr: cijepljenje, obavezno, protiv, kazna, austrija, obvezno, dijete, uvesti, covid, ovaj
en: vaccination, mandatory, against, penalty, austria, obligatory, child, induct, covid, this
Topic 5: hr: cijepljenje, početi, astrazeneca, obavezno, čovjek, protiv, kretati, europski, zemlja, dan
en: vaccination, start, astrazeneca, mandatory, human, against, move, european, country, day

Figure 4b presents results obtained by the LDA algorithm. It can be noticed that the scale of values in the LDA algorithm is different compared to LSI. There are no negative values but the percentage of how much a particular topic is represented in that instance is calculated. For a randomly selected instance in the upper portion of Heat Map it showed the dominant presence of Topic 1 (0.885) and small of other topics (0.028).



a) LSI algorithm b) LDA algorithm

Fig. 4. Heat Map display of results

Concordance: are representative examples of the words that appear around the keyword “vaccination”. By analyzing these instances, one can compare different contexts in which a word was used. Reading the concordances linearly, according to the order of appearance, even from this sample the development of events as well as factors and subjects became obvious.

Query: cijepljenje/vaccination
#1
: hr: U Zagrebu od ponedjeljka **cijepljenje** protiv gripe
: en: **Vaccination** against flu in Zagreb from Monday
#12
: hr: Struka: Jedni za **cijepljenje**, drugi tvrde da je prekasno
: en: Experts: Some for **vaccination**, others claim that it is too late
#23
: hr: **Cijepljenje** je jedina prava zaštita
: en: **Vaccination** is the only real protection
#45
: hr: Izaziva li **cijepljenje** autizam
: en: Does **vaccination** cause autism
#46
: hr: Ljudi odbijaju **cijepljenje** samo da bi se nečemu protivili
: en: People refuse **vaccination** just to be against something
#56
: hr: **Cijepljenje** kombiniranim cjepivima
: en: **Vaccination** with combined vaccines

#75
: hr: Ekskluzivno Roditelji koji odbijaju **cijepljenje** zapravo svojom glupošću ubijaju vlastitu djecu
: en: Exclusive Parents who refuse **vaccination** of own children are actually killing them with their stupidity
#159
: hr: Koalicija za **cijepljenje** održala prvi sastanak u Bruxellesu
: en: The **vaccination** coalition held its first meeting in Brussels
#296
: hr: Cjepivo blizu a pripreme za **cijepljenje** na respiratoru
: en: Vaccine close to preparations for **vaccination** on the respirator
#366
: hr: Počelo **cijepljenje** Broš i Mrkić o nuspojavama
: en: **Vaccination** started Broš and Mrkić about side effects
#486
: hr: Popunjenost bolničkih kapaciteta 67% **cijepljenje** kasni
: en: Hospital capacities are filled 67%, **vaccinations** are late
#1026
: hr: Nema teoretske šanse da **cijepljenje** bude obavezno
: en: There is no theoretical chance that vaccination will be mandatory
#1598
: hr: Ginekolog EU je za **cijepljenje** trudnica
: en: Gynecologist EU is for **vaccinations** of pregnant women
#1886
: hr: Covid potvrde su ustavne jer **cijepljenje** nije obavezno
: en: Covid certificates are constitutional because **vaccination** is not mandatory
#2009
: hr: Studija **Cijepljenje** je prepolovilo broj umrlih od korone
: en: Study **Vaccination** halved the number of deaths from corona
#2062
: hr: Počinje **cijepljenje** protiv gripe
: en: Flu **vaccination** begins
#2075
: hr: današnji dan prije dvije godine počelo **cijepljenje** protiv korona virusa u Hrvatskoj
: en: today two years ago **vaccination** against the corona virus began in Croatia

Sentiment Analysis: allows to determine whether a text is written in a positive or negative context. As headlines are written in Croatian the *Multilingual sentiment* option for the Croatian language was used. To explore the implied emotions, this model uses the Heat Map widget and data table with exact sentiments from the corpus and categorize them as positive or negative ones. The Heat Map for the corpus is shown in Figure 5, the more negative the emotion means that more negative words appear, and the more positive the emotion means that the article has more positive words. The overall result for all articles is that emotions range from -25 to 25.

TABLE V
RESULTS OF THE SENTIMENT ANALYSIS PER PORTAL

NEWS PORTAL	SENTIMENT		
	RANGE	MEAN	VARIANCE
INDEX	-25 - 25	0.121232	5.45162
DNEVNIK	-25 – 21.4286	0.637977	5.44342
24SATA	-25 – 17.6471	-0.678477	6.11971
JUTARNJI	-22.2222 - 20	-0.015229	5.11922
NET	-20 – 16.6667	0.693062	5.47009

From the results of the sentiment analysis it is obvious that the values acquired in this study are much more extreme in comparison to the results from our previous study [20]. This is somewhat expected, as headlines tend to use a lot more emotionally charged words to grab the reader's attention. Detailed results per individual news portal are provided in Table V and visualized in Figure 6. In this illustaration the colored stacks depicted under the curves represent the frequencies based on which sentiment distributions are formed for each news portal. It can be observed that the sentiments generally follow a normal distribution, but there are still certain differences between portals.

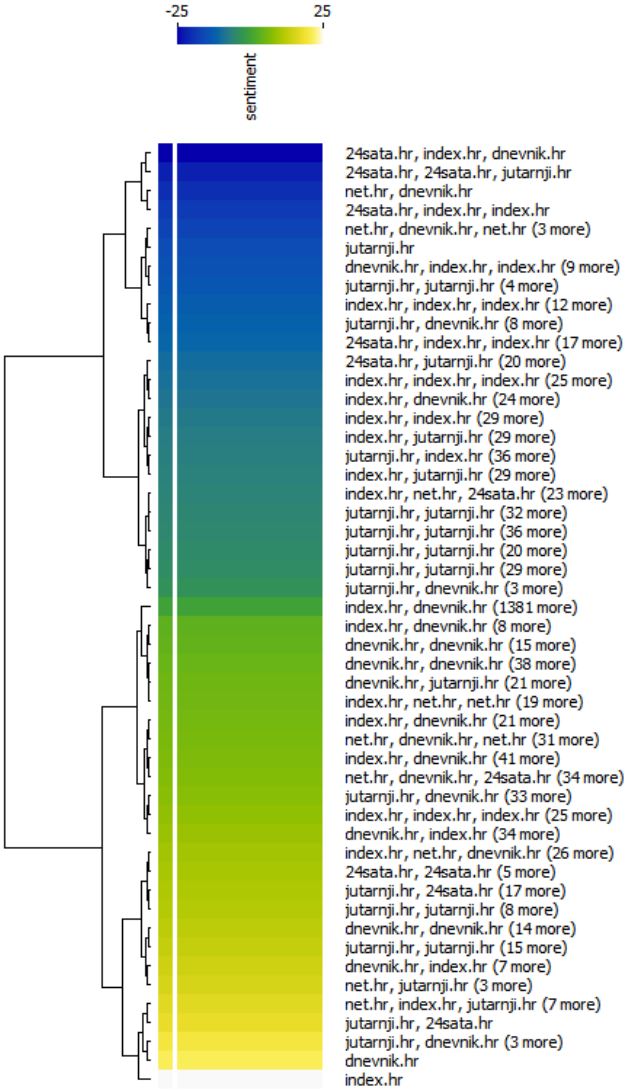


Fig. 5. Results of Sentiment Analysis via Heat Map

Index.hr covers the whole spectrum of sentiment distribution given that its end values range from -25 to 25. Dnevnik online and 24sata online also score the maximum values on the negative side, but on the positive side they achieve a lower score -21.4286 and 17.6471 respectively. Jutarnji online and Net.hr have smaller marginal values, which can be interpreted as a more moderate tone of the headlines compared to the previously mentioned portals.

Overall, Index.hr, Dnevnik online and 24sata online had the most negative ratings, while 24sata and Net.hr scored the lowest positive ratings. It is also worth noting that from the total of 2264 instances i.e. headlines even 1382 of them achieved a score of 0, indicating neutral sentiment. This indicates that in as many as 61% of cases, a neutral message prevailed.

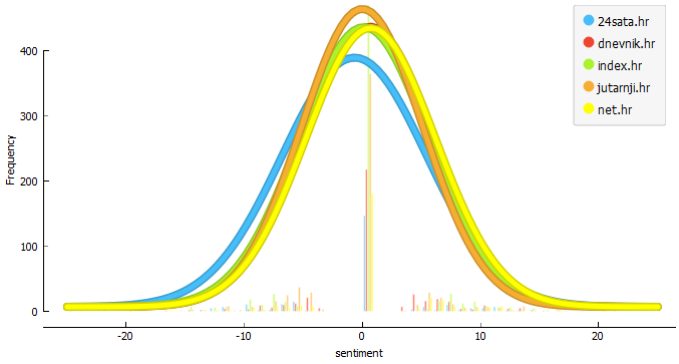


Fig. 6. Sentiment Distribution per News Portal

IV. Result Interpretation and Concluding Remarks

Following the research argument, this study performs the text analysis of post headlines trying to resolve their characteristics instead of deep-diving into the post content. This is an alternative approach and a suitable deviation from the previous research to the analysis of the contents provided on the most popular news portals in Croatia. The headlines are considered as the content from which the corpus has been formed instead of the body of the posts itself. The argument is based on the known fact that readers of digital content in most cases, overwhelmed by the amount of the contents and limited in time, do not even go beyond the headlines of news posts.

This means that a significant part of the messages, and thus the attitudes formed, are actually based on what is said in short highlighted texts that often try to attract the attention of the audience. The previous and this work with the obtained results enable an objective view of the analyzed problem, which is important because it reduces misunderstanding in conditions of extraordinary phenomena such as a pandemic.

The analysis performed on the corpus of news post headlines has identified the various characteristics of this textual data. The modeling itself followed a standard process that is characteristic of the data mining methodology. The collected data were structured in an appropriate form, what was followed by their preprocessing and later analysis using a selected natural language processing algorithms.

By looking at the distribution graph of posts that contain the keyword “vaccination” in the headline, it is evident that they have been present for a longer period of time. This is not surprising because the issue of vaccination has been constantly present, however what was obvious is that their accumulation coincided with the onset of the COVID-19 pandemic. At that point, vaccination became the dominant theme, and due to the specificity of the infection and the availability of the Internet and platforms, we witnessed a significant increase in the number of such news portal posts.

The research uses a significantly more efficient way of collecting the content compared to the earlier study. In this way, it was possible to retrieve a significantly larger number of instances in a wide time interval, which resulted in a representative data set that was subjected to the subsequent analysis. The dynamics of the content appearance shows that most articles were created in the period that correlates with the COVID-19 pandemic. The set of algorithms follows the one of the previous research, thus enabling a comparison of the results. The results indicate important differences between analysed corpora, bellow we refer to the most significant ones.

The keywords used in the headlines are similar to those appearing in the content of the posts, but they have a different order in terms of frequency. Words such as "against", "mandatory", "massive" appear more often here, and they often indicate some necessity and emphasize the essential message of the inscription. TF-IDF does not work well for this type of content. In case when the keyword appears in every headline, it is better to rely on the simple weight factor.

The concordances on such a set, which covers the entire timeframe of events related to the analyzed term, give an excellent insight into the course of events and are actually a summary of key developments. The analysis of sentiments, on the other hand, showed that they are much more emphasized in this part of the post, both in a positive and negative direction. This was somewhat expected as headlines are parts of the posts that are highlighted and have the function of sending an important message but also influencing the readers' attitudes. Insight into the data also allowed us to determine which portals are on one or the other side of the sentiment range and the ratio of nominally neutral posts.

The concept of this research, which uses an advanced method of data collection supported by a specialized system that continuously scans the news portals and a self-built text mining model based on a set of NLP algorithms, can serve as a framework for fastening the web mining process and increasing its efficiency in case of future similar research.

In conclusion, this research has contributed to obtaining a deeper insight into the phenomenon of “vaccination” in the textual digital media in Croatia. The analysis of headlines, which play an important role in the functioning of news portals, can provide various useful information on the basis of which the work of the media world can be viewed and better comprehended. In this sense, the application of NLP methods can be of great benefit in creating an objective picture about events, but also the ways in which the analyzed phenomenon was approached by the media.

V. Future Research

The direction of future research in which NLP techniques are applied in the analysis of digital textual content can take a number of paths. In order to cover the publication channels more completely, the authors of this paper plan to continue the analysis of digitalized textual corpora primarily from two aspects. First, towards the analysis of news posts provided on social networks and secondly those provided by print media that publish their contents online in different digital file formats. Additionally, the authors intend to apply additional algorithms for text processing in their future analyses.

REFERENCES

[1] B. Rangaswamy, G. Manjunatha and B. T. S. Kumar, *Internet as a Source of Information: Usage among the Faculty Members and Students*, Library Waves, Vol. 3, No. 1, pp. 36-42, Dec. 2017. [Online]. Available at: https://www.librarywaves.com/index.php/lw/article/view/48, Accessed: March 1, 2023.

[2] J. Hewitt-Taylor, *Using the Internet as a source of information and support: a discussion paper on the risks and benefits for children and young people with long-term conditions*, Journal of Innovation in Health Informatics, Vol. 22, No. 1, pp. 222–226, 2015. DOI: 10.14236/jhi.v22i1.74

[3] *Internet most popular information source: poll*, Reuters, 2009. [Online]. Available at: https://www.reuters.com/article/us-media-internet-life-idUSTRE55G4XA20090617, Accessed: March 1, 2023.

[4] L. Hermans, M. Vergeer and L. d'Haenens, *Internet in the Daily Life of Journalists: Explaining the use of the Internet by Work-Related Characteristics and Professional Opinions*, Journal of Computer-Mediated Communication, Vol. 15, Issue 1, pp. 138–157, Oct. 2009. DOI: 10.1111/j.1083-6101.2009.01497.x

[5] S. Vermeer, D. Trilling, S. Kruikemeier and C. de Vreese, *Online News User Journeys: The Role of Social Media, News Websites, and Topics*, Digital Journalism, Vol. 8, Issue 9, pp. 1114-1141, 2020. DOI: 10.1080/21670811.2020.1767509

[6] *Consumption of online news rises in popularity - Eurostat*, 2022. [Online]. Available at: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20220824-1, Accessed: March 1, 2023.

[7] S. Boulianne and A. Shehata, *Age Differences in OnlineNews Consumption and Online Political Expressionin the United States, United Kingdom, and France*, The International Journal of Press/Politics, Vol. 27, No. 3, pp. 763–783, 2022.

[8] B. A. Al-Youzbaky and R. D. Hanna, *The Effect of Information Overload, and Social Media Fatigue on Online Consumers Purchasing Decisions: The Mediating Role of Technostress and Information Anxiety*, Journal of System and Management Sciences, Vol. 12, No. 2, pp. 195-220, 2022. DOI: 10.33168/JSMS.2022.0209

[9] C. Kiili et al., *Reading to Learn From Online Information: Modeling the Factor Structure*, Journal of Literacy Research, 2018, Vol. 50, No. 3, pp. 304 –334, 2018. DOI: 10.1177/1086296X18784640

[10] T. A. van Dijk, *Discourse and Communication - Structures of News in the Press*, 1985. The Centre of Discourse Studies. [Online]. Available at: https://discourses.org/wp-content/uploads/2022/07/Teun-A.-van-Dijk-1985-Structures-of-news-in-the-press.pdf, Accessed: March 30, 2023.

[11] S. Isani, *Of headlines & headlinese: Towards distinctive linguistic and pragmatic genericity*, Open Edition Journals, Vol. 60, pp. 81-102, Nov. 2011. DOI: 10.4000/asp.2523

[12] *Glossary of Newspaper Terms*, Colorado NIE. [Online]. Available at: https://nieonline.com/coloradonie/downloads/journalism/GlossaryOfNewspaperTerms.pdf, Accessed: March 22, 2023.

[13] Y. Chen, N. J. Conroy and V. L. Rubin, *Misleading Online Content: Recognizing Clickbait as "False News"*, Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (WMDD'15), pp. 15-19, Nov. 2015. DOI: 10.1145/2823465.2823467.

[14] S. Saxena, *What are the keycharacteristics of webheadlines*, [Online]. Available at: https://www.easymedia.in/what-are-the-key-characteristics-of-web-headlines/, Accessed: Feb. 15, 2023.

[15] K. Olmstead, A. Mitchell and T. Rosenstiel, *Navigating News Online: Where People Go, How They Get There and What Lures Them Away*, Pew Research Center's - Project for Exelence in Journalism, [Online]. Available at: https://www.pewresearch.org/journalism/2011/05/09/navigating-news-online/, Accessed: March 22, 2023.

[16] L. Schwaiger, D. Vogler, and M. Eisenegger, *Change in News Access, Change in Expectations? How Young Social Media Users in Switzerland Evaluate the Functions and Quality of News*, The International Journal of Press/Politics, Vol. 27, No. 3, pp. 609–628, 2022. DOI: 10.1177/19401612211072787

[17] *Scrolling news: The changing face of online news consumption - A report for Ofcom*, Revealing Reality [Online]. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/115915/Scrolling-News.pdf, Accessed: Feb. 17, 2023.

[18] *How Young People Consume News and The Implications For Mainstream Media*, Reuters Institute for the Study of Journalism, Oxford University. [Online]. Available at: https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-02/FlamingoxREUTERS-Report-Full-KG-V28.pdf, Accessed: Feb. 17, 2023.

[19] K. C. Schrøder, *What Do News Readers Really Want to Read About? How Relevance Works for News Audiences*, Reuters Institute for Study of Journalism, Digital News Publications, 2019. [Online]. Available at: https://www.digitalnewsreport.org/publications/2019/news-readers-really-want-read-relevance-works-news-audiences/, Accessed: March 23, 2023.

[20] P. Lovrić, L. Vicković and H. Karna, Hrvoje, *Analysis of the Textual Information Extracted from News Portals*, Proc. of the 30th Conference on Software, Telecommunications and Computer Networks (SoftCOM 2022), pp. 1-6, 2022. DOI: 10.23919/SoftCOM55329.2022.9911444

[21] J. Kuiken, A. Schuth, M. Spitters and M. Marx, *Effective Headlines of Newspaper Articles in a Digital Environment*, Digital Journalism, Vol. 5, Issue 10, pp. 1300-1314, 2017. DOI: 10.1080/21670811.2017.1279978

[22] K. Lamot, T. Kreutz and & M. Opgenhaffen, *"We Rewrote This Title": How News Headlines Are Remediated on Facebook and How This Affects Engagement*, Social Media + Society, Vol. 8, Issue 3, July-Sep. 2022. DOI: 10.1177/20563051221114827

[23] S. Ćurković, D. Dukić, M. Petričević, and J. Šnajder. *TakeLab Retriever*. [Online]. Available at: https://retriever.takelab.fer.hr/explorer, Accessed: Feb. 9, 2023.

[24] E. Hyvönen, *Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery*, pp. 187–193, Jan. 2020.

[25] I. Kim and J. Kuljis, *Applying Content Analysis to Web based Content*, Journal of Computing and Information Technology (CIT), Vol. 4, pp. 369–375, 2010. DOI: 10.2498/cit.1001924

[26] N. B. C. Eembi et al., *A Systematic Review on the Profiling of Digital News Portal for Big Data Veracity*, Procedia Computer Science Vol. 72, pp. 390–397, 2015. DOI: 10.1016/j.procs.2015.12.15

[27] S. Vijayarani, J. Ilamathi and Nithya, *Preprocessing Techniques for Text Mining - An Overview*, International Journal of Computer Science & Communication Networks, Vol. 5, No. 1, pp. 7-16, 2016.

[28] R. Talib et al., *Text mining: techniques, applications and issues*, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 11, 2016. DOI: 10.14569/IJACSA.2016.071153

[29] H. Nossek, H. Adoni and G. Nimrod, *Is Print Really Dying? The State of Print Media Use in Europe*, International Journal of Communication, Vol. 9, No. 1, pp. 365–385, 2015.

[30] N. Newman, *Executive summary and key findings of the 2021 report*, Reuters Institute for the Study of Journalism, Oxford University. [Online]. Available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/dnr-executive-summary, Accessed: June 15, 2023.

[31] D. Deacon, *Yesterday's Papers and Today's Technology: Digital Newspaper Archives and 'push Button' Content Analysis*, European Journal of Communication Vol. 22, No. 1, March 2007. DOI: 10.1177/0267323107073743

[32] M. Karlsson and H. Sjøvaag, *Content Analysis and Online News*, Epistemiologies of analysing the ephemeral Web, pp. 177-192, Oct. 2015. DOI: 10.1080/21670811.2015.1096619

[33] K. Verma and A. K. Malviya, *An Efficient Way of Handling Large Scale Web News Data using Data Mining Techniques*, Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), March 2019. DOI: 10.2139/ssrn.3350999

[34] L.-F. Hsu, *Mining on Terms Extraction from Web News*, Proceedings of the Second international conference on Computational collective intelligence: technologies and applications, Nov. 2010. DOI: 10.1007/978-3-642-16693-8_21

[35] G. Troiano and A. Nardi, *Vaccine hesitancy in the era of COVID-19*, Public Health, Vol. 194, pp. 245-251, May 2021. DOI: 10.1016/j.puhe.2021.02.025

[36] J. E. Marco-Franco et al., COVID-19, *Fake News, and Vaccines: Should Regulation Be Implemented?*, Int. J. Environ. Res. Public Health, Vol. 18. No. 2, 2021. DOI: 10.3390/ijerph18020744

[37] K. K. W. Ho; J. Y. Chan and D. K. W. Chiu, *Fake News and Misinformation During the Pandemic: What We Know and What We Do Not Know*, IT Professional, Vol. 24, No. 2, pp. 19-24, March-April 2022. DOI: 10.1109/MITP.2022.3142814

[38] A. E. Varol et al., *Understanding COVID-19 News Coverage using Medical NLP*, Proceedings of the Text2Story'22 Workshop, Stavanger (Norway), April 2022. DOI: 10.48550/arXiv.2203.10338

[39] Reuters Institute of the Study of Journalism - Digital News Report, University of Oxford, [Online]. Available at: https://www.digitalnewsreport.org/, Accessed: March 7, 2023.

[40] *Data Miner*, Google Chrome extension. [Online]. Available at: https://dataminer.io/, Accessed: March 10, 2023.
[41] *Orange Data Mining*, [Online]. Available at: https://orangedatamining.com/, Accessed: January 10, 2023.
[42] M. Straka and J. Strakova, *UDPipe 1*, [Online]. Available at: https://ufal.mff.cuni.cz/udpipe/1, Accessed: March 10, 2023.

Hrvoje Karna received his PhD in the field of technical sciences at the University of Split. From 2007 to 2018, he worked in the IT sector for Siemens and Atos in various positions. In 2018, he joined the Ministry of Defense of the Republic of Croatia and since then he has been working at the Croatian Defense Academy. Currently, he holds an Assistant Professor position and teaches several courses in the field of information and communication technology. His interest include analytics, artificial intelligence, data mining, natural language processing, machine learning, software engineering.

Maja Braović is an Assistant Professor at the Department of Electronics and Computer Science at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Croatia. She is a member of the Department for Modelling and Intelligent Systems. She received her BSc and MSc degrees in Computer Science in 2008 and 2010, respectively, at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Croatia. She received her PhD in Artificial Intelligence at the same university in 2015. Her research interests include artificial intelligence, machine learning, image processing, natural language processing, computational linguistics and cryptography.

Linda Vicković received her PhD degree in 2007 in the field of technical sciences at the University of Split in 2007. Since then, she has been working as a professor for computing science at the Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture (FESB), University of Split, Croatia. She has been a member of ALICE collaboration at Center for Nuclear Research (CERN) since 2004. In 2023 she obtained a Full Professor degree. Her research interests include software engineering, data mining, deep learning, convolutional neural networks, image processing and discrete event simulation.

Damir Krstinić received the Ph.D degree from the University of Split, Split, Croatia in 2008. He is Full Professor on Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture of the University of Split, Department for Modelling and Intelligent Systems. His main area of research interest are computer vision, image understanding and machine learning. He has been involved in several research and technological projects related to forest fire research and is project leader of Wildfire Early Detection and Monitoring System for Croatian Forests.