Topic Modelling of Croatian News During COVID-19 Pandemic

Petar Kristijan Bogović[1], Ana Meštrović[1,2], Slobodan Beliga[1,2], Sanda Martinčić-Ipšić[1,2]
[1]Department of Informatics, [2]Center for Artificial Intelligence and Cybersecurity,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: petar.kristijab@student.uniri.hr, amestrovic@uniri.hr, sbeliga@uniri.hr, smarti@uniri.hr

***Abstract*—This paper addresses topic modelling in Croatian news articles related to COVID-19 pandemics and corresponding comments. We identify and analyze Croatian online news media's main topics for the first nine months of pandemics shedding some light on the leading themes covered in news articles and corresponding comments. Topics are derived automatically by training the model and calculating topics' coherence values. We report the results by listing the top 15 detected words in top 10 detected topics from the content of articles and corresponding comments. Our findings include the analysis of intersected topics and discussion of dissents. Obtained results are the first step toward better information monitoring and hopefully mitigating the infodemics effect in Croatia.**

Keywords - topic modelling; Latent Dirichelt Allocation; coherence; Croatian news; COVID-19 infodemic

I. Introduction

We have recently witnessed a dramatic increase in textual information, which is hard to organize into meaningful chunks. The outbreak of COVID-19 pandemics amplified the information tsunami into new phenomena in digital society - infodemics [8], [12], [24], [32]. World Health Organization (WHO) defines an infodemic as an overabundance of information, accurate or not, which impedes the finding of trustworthy sources and reliable facts [24]. Besides pandemics research community lead by the WHO initiative is actively fighting the infodemic response. For example, authors in [29] propose a framework with six criteria for policymaking to manage infodemics during a health crisis. Similarly, the author in [11] proposes four pillars of infodemic management: (1) information monitoring (infoveillance); (2) building eHealth Literacy and science literacy capacity; (3) encouraging knowledge refinement and quality improvement processes such as fact-checking and peer-review; and (4) accurate and timely knowledge translation, minimizing distorting factors such as political or commercial influences. Our research is an attempt to contribute to information monitoring (infoveillance) in Croatian news spaces. Specifically, we propose topic modelling of central themes in online media for semantically organizing and better monitoring of the information deluge.

Topic modelling is a natural language processing (NLP) technique tasked with discovery of semantic structures in documents. The goal of topic modelling is to detect the latent semantic structure or topics from texts [4], [13], [6]. Topics can contribute to better organization of documents in the collection, enabling semantically driven clustering and classification of documents [4], [20]. Extracted topics can be considered as the condensed descriptions of the content; hence they contribute to the informed summarization of the documents [22]. Additionally, topics can contribute to novelty detection, and longitudinal tracking of interest in written material [5], as well as the basis for assessing the similarity of documents and relevance of the context [4].

This paper addresses topic modelling in Croatian news articles that mention coronavirus and corresponding comments published during the first nine months of pandemics. We identify and analyze the main topics and shedding light on the leading themes related to COVID-19 covered in news articles and comments.

This paper is structured as follows: Section II overviews related work and methods; Section III presents experimental setup and results; Section IV discusses obtained results while concluding remarks and directions for the future research are elaborated in Section V.

II. Methods and Related work

Topic modelling is tasked with automatically detecting latent semantic structure or topics from the co-occurrence of words in texts [4], [13], [6]. The topic is a probability distribution over words, while a document is a mixture of topics. The topic is the collection of words that are likely to appear in the same context, forming the coherent semantic structure [2]. The discovered topics are usually represented with the top N highest-ranked terms [6].

There are several branches of methods developed for topic modelling. Latent Semantic Analysis (LSA) is the seed topic modelling principle proposed in [9]. LSA contributes the approach to automatic indexing and information retrieval that maps documents and terms to a representation in the latent semantic space. A prominent group of methods followed LSA, Latent Semantic Indexing (LSI) and Probabilistic Latent Semantic Indexing (PLSI) [14]. PLSI is an automated document indexing model based on a statistical latent class model for factor analysis of count data and overcomes the initial LSI method's limitations by defining a proper generative model of the textual data. Definition of a proper generative model motivated the next generation of topic modelling methods with Latent Dirichlet Allocation (LDA) proposed in [4], as

a pominent topic modelling method. Our research uses the LDA method for topic modelling of the text and comments, so the LDA method is elaborated in the continuation.

Besides topic modelling of larger documents and document collections topic modelling of short texts like messages, tweets, microblogs, and comments has attracted the research community's attention as well [15], [31], [7], [30]. Since short texts do not carry enough information for statistical modelling, some aggregation strategies have been shown beneficial [15]. The occurrence of terms in the short text is not discriminating enough to derive a coherent topic of the message. For instance, one tweet can belong to multiple topics or have no topics at all [31]. Only topics with sufficient precision are of interest in a single short message setup, as typical for multi-label classification tasks. However, we are often interested in the general monitoring process, as the level of anxiety in the population or polarity of the sentiment [16]. Then, the generalized topic can be of interest. In the general analysis setup, strategies for topic modelling of aggregated (compound) texts should be considered. In this case, we opt to monitor the content of a larger scale of comments that are not necessarily related to corresponding news articles' content. Monitoring a larger scale of comments is especially important when we opt to analyze human reactions and opinions during COVID-19 pandemics as an essential part of infodemic monitoring, as shown in [30], [16].

1) Latent Dirichlet Allocation: Latent Dirichlet Allocation (LDA) is a generative probabilistic model [6]. The basic idea behind LDA is to represent documents as random mixtures over latent topics, where a topic is characterized by a distribution of words in a document. Latent Dirichlet allocation, proposed by Blei, Ng, and Jordan in [4], is one of the most prominent methods for topic modelling. Given a corpus D consisting of M documents, and document d containing N_d words $(d \in 1, ..., M)$.

LDA assumes that document d is generated by first sampling a topic z from the document-topic distribution ϕ, and word w is derived according to the corresponding topic-word distribution θ. Authors in [4] proposed to use the Dirichlet prior on ϕ distribution with associated parameter α, and parameter β on θ distribution. Collapsed Gibbs sampling is used to estimate ϕ and θ distributions by iteratively estimating the probability of assigning each word w_i to the topic z_i [13]. The probability assignment is conditioned on the current topic assignment of all other words in topic-word count matrices C^{WT} and document-topic matrix C^{DT}:

$$P(z_i = j | z_{-i}, w_i, d_i, .) \propto \frac{C^{WT}_{w_{i,j}} + \beta}{\sum_{w=1}^{W} C^{WT}_{w,j} + W\beta} \frac{C^{DT}_{d_i,j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i,t} + T\alpha} \quad (1)$$

The distribution (θ^j) for sampling a word i from topic j, and distribution (ϕ^d) for sampling topic j for document d are:

$$\theta^j = \frac{C^{WT}_{ij} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta}, \quad (2)$$

$$\phi^d = \frac{C^{DT}_{dj} + \alpha}{\sum_{t=1}^{T} C^{DT}_{dt} + T\alpha}. \quad (3)$$

2) Topic Coherence: Topic coherence reflects the semantic interpretability of extracted terms that describe a topic and can serve to evaluate the topic modelling task [6]. Coherence evaluates the quality of a given topic regarding the human perception of semantic understandability [23]. Authors in [23] calculated the correlation between human judgments and a set of proposed measures, and confirmed that a Pointwise Mutual Information (PMI) measure outperforms all other evaluated measures (i.e. MIW-Milne Witten, RACO-related article concept overlap and document similarity). Recently, authors in [26] proposed a unifying framework that represented coherence measures as a composition of parts, aiming to achieve a higher correlation with human judgments. They systematically search the space of all proposed coherence measures and show that the combination of measures outperforms single solutions. Namely, the cosine measure for the segmentation of word set into smaller units combined with NPMI (Normalized Pointwise Mutual Information) for scoring the agreement and Boolean sliding window over words for calculating topic probabilities outperforms 20 other combined measures.

Practically, coherence measures are calculated over the different number of topics in data collection. The highest coherence value is the indicator of the best selection of the number of topics. Choosing the number of topics is of high importance since LDA is an unsupervised method and accompanied by automatic coherence evaluation contributes to the extraction of semantically interpretable and coherent topics from data.

Topic coherence pipeline used in the research consists of four stages, as presented in [26]. The four-stage pipeline is comprised of (1) segmentation of the data into word pairs, (2) probability estimation of words or word pairs, (3) confirmation measure calculation to see how strongly a word set supports another word set, and (4) aggregation of all the individual confirmation scores into an overall coherence score.

The first step is to segment the word set W into two subsets $W^{'}$ and W^{*}. Let W be the set of a topic's top-N most probable words $W = \{w_1, \cdots, w_N\}$, then let S_i be a segmented pair of each word $W' \in W$ paired with all other words $W^* \in W$, and S the set of all pairs [10]. Approach used for this part of the pipeline was presented in [1]. The approach compares words to the total

$$S = \{(W^{'}, W^{*}) | W^{'} = \{w_i\}; w_i \in W; W^{*} = W\} \quad (4)$$

The second step is the estimation of word probabilities P. Boolean sliding window (P_{sw}) is used for probability

estimation from data. This method determines word counts using a sliding window which slides over the document one word token per step. Each step defines a new virtual document by copying the window content. After defining the virtual documents, Boolean document (P_{bd}) is used to calculate the word probabilities. The boolean document estimates the probability of a single word as the number of documents in which the word occurs divided by the total number of documents, but does not consider the frequencies and distances of words. By including the Boolen sliding window, the method captures the word token proximity. The final results of [26] have shown that the sliding window of 110 gives the best results in comparison to human-reviewed topics - $P_{sw(110)}$.

In the third step, the confirmation measure takes a single pair $S_i = (W^{'}, W^{*})$ of words with corresponding probabilities P_i to compute how strong the conditioning word set W^{*} supports $W^{'}$. The confirmation measure is used to calculate the agreements η of these pairs. Computation of agreement η can be done directly or indirectly, although, the indirect method, proposed by [1] has shown better results. Indirect confirmation computes the similarity of words in $W^{'}$ with respect to direct confirmations of all words. This means that, for example, words such as *Ford* and *Ferrari* that might rarely be mentioned together in some document have low similarity using direct confirmation, will be recognized as similar with the indirect confirmation. Both words strongly correlate with words like wheels and road and are mutually semantically supported, reflected by the indirect measure.

Authors in [26] show that NPMI (Normalized Pointwise Mutual Information) is the superior confirmation measure. Hence, the formula for computing the value of vector $\vec{v}(W^{'})$ is defined as:

$$\vec{v}(W^{'}) = \sum_{w_i \in W} NPMI(w_i, w_j)^{\gamma} \tag{5}$$

Where NPMI is defined as:

$$NPMI(w_i, w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i, w_j)+\varepsilon}{P(w_i) \cdot P(w_j)}}{-\log(P(w_i, w_j)+\varepsilon)} \right)^{\gamma} \tag{6}$$

Then, the confirmation measure η of a pair S_i is obtained by calculating the cosine vector similarity of all context vectors $\eta_{S_i}(\vec{u}, \vec{w})$ within S_i, with $\vec{v}(W^{'}) \in \vec{u}$ and $\vec{v}(W^{*}) \in \vec{w}$ and is defined as:

$$\eta_{S_i}(\vec{u}, \vec{w}) = \frac{\sum_{i=1}^{|W|} u_i \cdot w_i}{\|\vec{u}\|_2 \cdot \|\vec{w}\|_2} \tag{7}$$

The final coherence score is the arithmetic mean of all confirmation measures η. Additionally, ε is used to account for the logarithm of zero and γ to place more weight on higher NPMI values [28].

In this paper, we use Latent Dirichlet Allocation to detect latent topics in Croatian news texts and corresponding comments. We set the number of extracted topics according to calculated coherence scores of extracted topics.

III. Experiments and Results

A. Dataset

The dataset used for this research contains texts from two major Croatian daily and tabloid news. Analyzed data contain only the textual part of the articles and disregards visual material like photos or video. Specifically, we retain only textual parts of articles (title, body text, tags, and readers' comments if they exist in the particular article). Comments are written by a large number of anonymous online readers. Therefore, it is impossible to obtain demographic or authorship information, so we concatenate comments into one large text for further analysis.

Texts from the daily news portal are dated between January 1st and September 22nd of 2020 and are composed of 7,016 news articles and 17,355 corresponding comments. Texts from the tabloid news are dated between January 1st and September 29th 2020 and are composed of 6,902 news articles and 156,150 corresponding comments. The total count of news articles and comments before data cleaning is 13,918 articles and 173,505 corresponding comments. In the dataset, we have included only those articles that include words' corona', 'covid-19', 'coronavirus' or 'sars-cov-2' in their title or text. Before proceeding with topic modelling, as the standard step in natural language processing, we performed data cleaning including the removal of duplicates, stop words, and emojis, which resulted in a final count of 13,744 news articles and 171,695 comments. The stop word list used in the research contains 2,482 different Croatian stop words.

B. Experimental Setup

Here we report the implementation details and used tools. The central tool is Gensim - a free, open-source Python library developed to automatically extract semantic topics from a set of documents by representing documents as semantic vectors and processing them using unsupervised machine learning algorithms [25]. Gensim includes multiple algorithms such as Word2Vec, FastText, Latent Semantic Indexing (LSI, LSA, LsiModel), Latent Dirichlet Allocation (LDA, LdaModel). We use Gensim in multiple steps: first to automatically detect collocations in a stream of sentences, second to create a dictionary by mapping words, third to convert the documents into corresponding bag-of-words (BoW) representation. The BoW is a document representation model that uses an unordered set of word or term frequency [19].

MALLET is a Java-based natural language processing (NLP) toolkit used for the analysis of the unlabeled text in document classification, clustering, topic modelling, information extraction, and other text mining tasks [18]. MALLET contains several algorithms for topic extraction: Pachinko Allocation Model (PAM), LDA, and hierarchical LDA. MALLET incorporates the quick and scalable implementation of the Gibbs Sampling. For this experiments, we wrapped MALLET by Python Gensim library to train the LDA model.

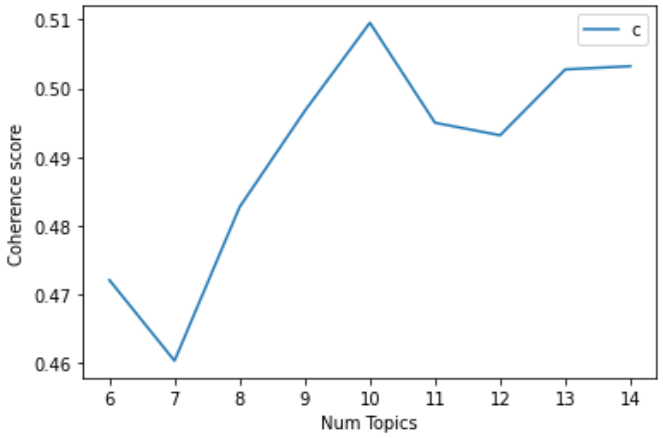
For the reduction of the high number of morphological word forms in the Croatian language, we use stemmer [27]. The

stemmer performs a series of transformations that take care of morphological changes and a series of rules that remove the suffixes and on all parts of speech performs with a precision of 0.98 and recall of 0.92 [17].

C. Results

Here we report results on topic modelling and coherence in texts of news articles. First, we detect a suitable number of topics with a high coherence score by calculating a coherence score for each of models in the range of 6 to 14 topics as reported in Figure 1. Specifically, we train an LDA model for each selected number of topics and calculate a coherence score. According to the obtained results in Figure 1 for the news articles' content, the number of ten topics exhibit the highest coherence value of 0.509.

Fig. 1. Coherence scores for different number of topics obtained with LDA models trained on content of Croatian news articles



Next, we present the top ten topics obtained according to elaborated methods with the top 15 words. Although the top 30 words results were coherent and easily interpretable into the meaningful topics, we report only the top 15 keywords per topic, due to the available space limitation. The ordering of topics is reported as the model derives them, and we provide the general label of the topics reach by the agreement between authors of this research. Additionally, along with lemmatized word forms in the Croatian language, we list an English translation. In Table I are results for the first two detected topics, which we named "Earthquake during pandemics" (i.e. the earthquake in Zagreb in March 2020) and "Elections in Croatia" (i.e. parliamentary elections in July 2020, or domestic politics). Table II lists the top words in topics of "Crime" and "Online education". Table III shows the top words in "Anti-pandemics measures protests" (i.e. protests against governmental measures for the prevention of pandemics spreading in Croatia) and "Pandemics worldwide" (i.e. the pandemics related cases, events and deaths world-wide). Table IV contains words for the topic of "Travel and EU borders crossing" and the topic of "Pandemics in Croatia" covering the local development of the pandemics. The last two topics of "Economy" and "General" (i.e. topic covering different aspects of daily living, events and problems) derived from the content of news articles are in Table V.

TABLE I
TOP 15 KEYWORDS FOR TOPICS: EARTHQUAKE DURING PANDEMICS, ELECTIONS IN CROATIA / POLITICS

	Earthquake during pandemics	**Elections in Croatia / Politics**
1	Bolnica / Hospital	Hrvatski / Croatian
2	Zagreb	Reći / To say
3	Grad / City	Predsjednik / President
4	Zaštita / Protection	Vlada / Government
5	Dom / Home	Izbori / Elections
6	Rad / Work	Zakon / Legislation
7	Sustav / System	Plenković
8	Gradjani / Citizens	Stranka / Party
9	Ministarstvo / Ministry	Politički / Political
10	Zdravlje / Health	Poručiti / Say
11	Potres / Earthquake	Gradjani / Citizens
12	Hitno / Urgent	Premijer / Prime Minister
13	Dobiti / To get	Pitanje / Question
14	Dan / Day	Odluka / Decision
15	Ravnatelj / Director	HDZ / HDZ

TABLE II
TOP 15 KEYWORDS FOR TOPICS: CRIME, ONLINE EDUCATION

	Crime	**Online education**
1	Godina / Year	škola / School
2	Policija / Police	Djeca / Children
3	Imati / To have	Rad / Work
4	Kuća / House	Godina / Year
5	Dan / Day	Hrvatski / Croatian
6	Žena / Wife	Nastava / Teaching
7	Obitelj / Family	Roditelj / Parent
8	Znati / To know	Novi / New
9	Suprug / Husband	Proizvod / Product
10	Sat / Hour	Učenik / Pupil
11	Zatvor / Jail	Poseban / Special
12	Automobil / Car	Prostor / Space
13	Reći / To say	Obrazovanje / Education
14	Majka / Mother	Način / Method
15	Kasno / Late	Potreban / Necessary

D. Topics Derived from Comments

In this section, we report results derived from comments of the news articles. Comments are short texts usually without semantic context or with a limited one; therefore, we aggregate all available comments into one large text and proceed with the topic modelling methods as defined above. First, we analyze the coherence scores of the number of topics. From Figure 2 it is apparent that LDA models trained on comments have higher coherence scores as the number of topics rise, as opposed to the models trained on the content of the news articles

TABLE III
TOP 15 KEYWORDS FOR TOPICS: ANTI PANDEMIC MEASURES PROTEST, PANDEMICS WORLDWIDE

	Anti pandemic measures protest	Pandemics worldwide
1	Policija / Police	Koronavirus / Coronavirus
2	Grad / City	Virus / Virus
3	Reći / To say	Pandemija / Pandemic
4	Ljudi / People	Zemlja / Country
5	Prosvjed / Protest	Svijet / World
6	Objaviti / To publish	Kina / China
7	Velik / Huge	Godina / Year
8	Crkva / Church	Ljudi / People
9	Dan / Day	Američki / American
10	Sat / Hour	Bolest / Disease
11	Mediji / Media	Cjepivo / Vaccine
12	Policijski / Police	Novo / New
13	Napad / Attack	Istraživanje / Research
14	Vlast / Government	Zdravlje / Health
15	Subota / Saturday	Svjetski / Worldwide

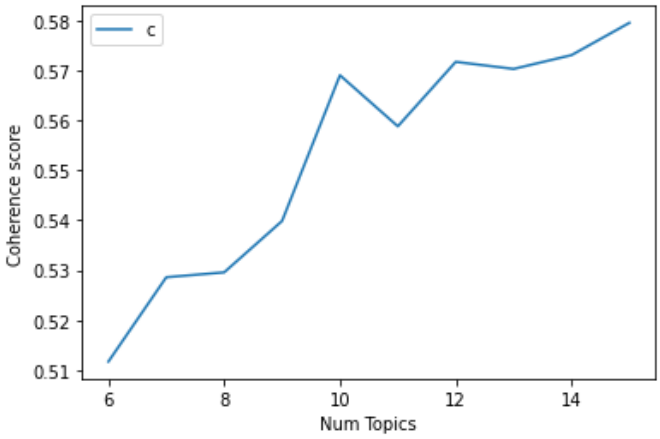
TABLE IV
TOP 15 KEYWORDS FOR TOPICS: EU BORDERS/TRAVEL, PANDEMICS IN CROATIA

	EU borders/Travel	Pandemics in Croatia
1	Hrvatski / Croatian	Osoba / Person
2	Zemlja / Country	Koronavirus / Coronavirus
3	Njemačka / Germany	Mjera / Measure
4	Granica / Border	Slučaj / Case
5	Vlada / Government	Zaražen / Infected
6	Europski / European	Zaraza / Contagion
7	Reći / To say	Broj / Number
8	Turist / Tourist	Dan / Day
9	Putovanje / Traveling	Sat / Hour
10	Država / State	Samoizolacija / Self-isolation
11	Europska Unija / European Union	Ukupno / Total
12	Europa / Europe	Epidemija / Epidemic
13	španjolski / Spanish	Oboljeli / Diseased
14	Državljanin / Citizen	Novi / New
15	Brod / Ship	Kontakt / Contact

TABLE V
TOP 15 KEYWORDS FOR TOPICS: ECONOMY, GENERAL

	Economy	General
1	Godina / Year	Ljudi / People
2	Kuna / Croatian currency	Kazati / To tell
3	Hrvatski / Croatian	Znati / To know
4	Mjera / Measure	Dobar / Good
5	Javni / Public	Pun / Full
6	Tvrtka / Company	Dan / Day
7	Gospodarstvo / Economy	Raditi / To work
8	Kriza / Crisis	Situacija / Situation
9	Vlada / Government	Problem / Problem
10	Država / State	Veliki / Large
11	Plaća / Salary	Reći / To say
12	Rad / Work	Jak / Strong
13	Iznos / Amount	Život / Life
14	Euro	Ne htjeti / To not want
15	Sektor / Sector	Moći / To be able

Fig. 2. Coherence score depending on the number of topics for the LDA model trained on comments



which have an isolated peak of coherence score value with the number of topics equal to ten. Coherence scores of the comments indicate that many more topics can be derived from comments. Nevertheless, we have chosen the number of ten topics to stay consistent with the articles' reported topics, also motivated by limited space for reporting.

Next, we report the top 15 words for each of the ten derived topics from comments. The table VI reports the first two topics that we refer to as "Zagreb politics and crime" (i.e. local politics and affairs in Zagreb) and "Elections in Croatia" (i.e. domestic politics). Table VII lists words defining topics of "Crime" and "Online education and work". Table VIII presents results for topics of "Law and measures" (including protests) and "Pandemics". Table IX reports the results for topics derived from commenting "General information in the news" and "Pandemics in Croatia". Table X shows the words in topics of "Finances" and "General topic".

IV. DISCUSSION

In this section, we interpret and discuss the obtained results. First, we reflect upon topics derived from the content of articles. For this part of the data, we derived ten highly coherent topics. For at least seven topics we have easily interpreted the meaning (i.e. we straightforwardly named the topic), additional two topics were interpretable after discussion. The last one is a general topic where we did not reach a consensus of meaning; hence it remained general. Overall, the top extracted topics summarized the main themes covered in news articles that mention COVID-19 in the first nine months of 2020.

For the results reported for comments, the essential issue is identified while setting the number of topics to report the

TABLE VI
TOP 15 KEYWORDS FOR TOPICS: ZAGREB POLITICS AND CRIME, ELECTIONS IN CROATIA / POLITICS

	Zagreb politics and crime	Elections in Croatia / Politics
1	Država / State	HDZ / HDZ
2	Bog / God	Izbor / Election
3	Grad / City	Vlada / Government
4	Ostati / To remain	Vlast / Government
5	Voditi / To lead	Plenkovic
6	Crkva / Church	Najbolji / The best
7	Zagreb	Stranka / Party
8	Novac / Money	Predsjednik / President
9	Pomoć / Aid	Glasati / To vote
10	Lova / Cash	Sabor / Parliament
11	Sramota / Emberassment	Milanovic
12	Lopov / Crook	Politički / Political
13	Jadan / Miserable	Plenki / Plenkovic nickname
14	Red / Order	SDP / SDP
15	Bandić	Glas / Vote

TABLE VII
TOP 15 KEYWORDS FOR TOPICS: CRIME, ONLINE EDUCATION AND WORK

	Crime	Online education and work
1	Pravi / Right	Ljudi / People
2	čovjek / Man	Djeca / Children
3	Imati / To have	Rad / Work
4	Život / Life	Raditi / To work
5	Znati / To know	Maska / Mask
6	Kriv / Guilty	Ići / To go
7	Žena / Woman	Normalan / Normal
8	Policija / Police	škola / School
9	Auto / Car	Dom / Home
10	Kazna / Fine	Nemati / To not have
11	Osoba / Person	Kuća / House
12	Reći / To say	Obitelj / Family
13	Mlad / Young	Vrijeme / Time
14	Hitni / Emergent	Roditelj / Parent
15	Teški / Difficult	Cijeli / Whole

TABLE VIII
TOP 15 KEYWORDS FOR TOPICS: LAW AND MEASURES, PANDEMICS

	Law and measures	Pandemics
1	Policija / Police	Koronavirus / Coronavirus
2	Grad / City	Virus / Virus
3	Reći / To say	Pandemija / Pandemic
4	Ljudi / People	Zemlja / Country
5	Prosvjed / Protest	Svijet / World
6	Objaviti / To publish	Kina / China
7	Velik / Huge	Godina / Year
8	Crkva / Church	Ljudi / People
9	Dan / Day	Američki / American
10	Sat / Hour	Bolest / Disease
11	Mediji / Media	Cjepivo / Vaccine
12	Policijski / Police	Novo / New
13	Napad / Attack	Istraživanje / Research
14	Vlast / Government	Zdravlje / Health
15	Subota / Saturday	Svjetski / Worldwide

TABLE IX
TOP 15 KEYWORDS FOR TOPICS: GENERAL INFORMATION IN NEWS, PANDEMICS IN CROATIA

	General information in news	Pandemics in Croatia
1	Znati / To know	Ljudi / People
2	Kazati / To tell	Korona / Corona
3	Vidjeti / To see	Virus / Virus
4	Laž / A lie	Zaražen / Infected
5	Priča / Story	Stožer / Headquarters
6	Vjerovati / To trust	Bolnica / Hospital
7	Stvaran / Veritable	Mjera / Measure
8	Komentar / Comment	Bolest / Disease
9	Istina / Truth	Svijet / World
10	Misliti / To think	Broj / Number
11	Govor / Speech	Corona / Corona
12	Pitati / To ask	Cijeli / Whole
13	Misliti / To think	Cjepivo / Vaccine
14	Pisati / To write	Doktor / Doctor
15	Novinar / Journalist	Zaraza / Contagion

results. We decided on ten topics to be consistent, but here we have several additional remarks. As short texts generated by thousands of readers, comments are dispersed according to the reader's interests, standpoints, beliefs, and values. So, they reflect the heterogeneity of society. Additionally, when checking topics beyond the selected ten, we notice several topics that are usually not in the articles' content but always present when expressing an opinion (e.g. Croatian War of Independence, relations with neighbouring countries, hate speech). Surprisingly, in the top ten, we derived "General" topics with the positive polarity of content. Nevertheless, the number of topics in comments should be set much higher, (even above 14 - please see Figure 2), and this remains an open research question, which should be addressed in the future.

The intersection contains six coherent topics: "Crime", "Elections in Croatia/Politics", "Online education and work", "Law and measures", "Politics", "Pandemics in Croatia", while "Economy" - "Finances" are related. Topics of "Anti-pandemic measures protest" and "Earthquake in Zagreb" are mainly covered in articles. Still, when deriving more topics from comments, we speculate that the earthquake will be identified as a topic as well. Deriving more topics from comments also remains an open research question.

Finally, the obtained results can be of use for monitoring infodemics caused by COVID-19 pandemic. Within the proposed solution, topics can be automatically detected for each day, week and month of the pandemics contrasting the content in news articles and interest of the broader readers community

TABLE X
TOP 15 KEYWORDS FOR TOPICS: FINANCES, GENERAL

	Finances	General
1	Godina / Year	Dobar / Good
2	Dan / Day	Bolji / Better
3	Dobiti / To get	Pun / Full
4	Plaća / Salary	Ne htjeti / To not want
5	Par / Couple	Jak / Strong
6	Mjesec / Month	Bravo / Good job
7	Kuna / Croatian currency	Hrvatska / Croatia
8	Cijena / Price	Dalji / Farther
9	Zadnji / Last	Lijep / Pretty
10	Pola / Half	Živ / Alive
11	Sat / Hour	Star / Old
12	Veći / Larger	Hvaliti / Praise
13	Kupiti / To buy	Drag / Nice
14	Euro	čekati / To wait
15	Prošli / Last	čast / Honor

expressed in comments. Hence, our results can contribute to COVID-19 information monitoring (infoveillance) in Croatian news spaces. Specifically, topics can be modelled and analyzed longitudinally, providing the tool for tracking the main themes related to coronavirus covered in online media and the readers' interest. Hopefully, this can lead to systematical monitoring of the information deluge in Croatian online news space.

V. CONCLUSION

In this paper, we perform topic modelling Croatian online news media during the first nine months of COVID-19 pandemics. We identify main topics in the dataset of articles related to COVID-19 and corresponding comments. Topics are derived automatically by training the Latent Dirichlet Allocation model and calculating reported topics' coherence values. We list the top 15 detected words in top 10 detected topics from the content of articles and corresponding comments, respectively. Our findings include the analysis of intersected content and discussion of dissents. Three topics extracted from articles and comments are directly related to pandemics ("Anti-pandemics measures protests", "Pandemics worldwide", "Pandemics in Croatia"), additional three are related to the aspect of change in daily living caused by pandemics ("Online education", "Travel and EU borders crossing", and "Economy-Finances"), and the rest of the topics are not related to pandemics. These results indicate that many articles discuss topics that are not directly related to the pandemic but mention coronavirus in some context.

Reported results are the first step toward a better understanding of infodemic caused by COVID-19 and better information monitoring in Croatian news space in general. Our future work plans employ temporal topic modelling and deriving methods for longitudinal analysis. Additionally, we will opt to cover broader sets of data (i.e. expanding the study with comments in social media), extend our previous work on the semantic context extraction [21] and test derived topics against automatically extracted keywords with Selectivity Based Keyword Extraction Method [3].

VI. ACKNOWLEDGEMENT

This work has been supported in part by the Croatian Science Foundation under the project IP-CORONA-04-2061, "Multilayer Framework for the Information Spreading Characterization in Social Media during the COVID-19 Crisis" (InfoCoV) and by University of Rijeka project number uniri-drustv-18-20.

REFERENCES

[1] N. Aletras, and M. Stevenson, (2013, March). Evaluating topic coherence using distributional semantics. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers (pp. 13-22).

[2] B. V. Barde and A. M. Bainwad, "An overview of topic modelling methods and tools," in 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), 2017.

[3] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "Selectivity-Based Keyword Extraction method," Int. J. Semant. Web Inf. Syst., vol. 12, no. 3, 2016, pp. 1–26.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. the Journal of machine Learning research, 3, 993-1022. (2003)

[5] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in Proceedings of the 23rd international conference on Machine learning - ICML '06, 2006.

[6] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modelling," Expert Syst. Appl., vol. 42, no. 13, 2015, pp. 5645–5657.

[7] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic modelling over short texts," IEEE Trans. Knowl. Data Eng., vol. 26, no. 12, 2014, pp. 2928–2941.

[8] M. Cinelli, W. Quattrociocchi, A. Galeazzi et al. The COVID-19 social media infodemic. Sci Rep 10, 16598 (2020). https://doi.org/10.1038/s41598-020-73510-5

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," J. Am. Soc. Inf. Sci., vol. 41, no. 6, 1990, pp. 391–407.

[10] I. Douven and W. Meijs. Measuring coherence. Synthese, 156(3):405-425, 2007.

[11] G. Eysenbach, "How to fight an infodemic: The four pillars of infodemic management," J. Med. Internet Res., vol. 22, no. 6, p. e21820, 2020.

[12] R. Gallotti, F. Valle, N. Castaldo, P. Sacco, and M. De Domenico . Assessing the risks of 'infodemics' in response to COVID-19 epidemics. Nature Human Behaviour, 4(12), 2020, 1285-1293.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," Proc. Natl. Acad. Sci. U. S. A., vol. 101 Suppl 1, no. Supplement 1, 2004, pp. 5228–5235.

[14] T. Hofmann, "Probabilistic latent semantic indexing," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99, 1999.

[15] L. Hong and B. D. Davison, "Empirical study of topic modelling in Twitter," in Proceedings of the First Workshop on Social Media Analytics - SOMA '10, 2010.

[16] B. Kleinberg, I. van der Vegt, and M. Mozes, "Measuring emotions in the COVID-19 Real World Worry Dataset," arXiv [cs.CL], 2020.

[17] N. Ljubešić, D. Boras, and O. Kubelka, Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. 2007.

[18] "MALLET homepage," Edu. [Online]. Available: http://mallet.cs.umass.edu. [Accessed: 10-Jan-2021].

[19] H. Schütze, C. D. Manning, and P. Raghavan, Introduction to information retrieval. 2008. (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

[20] S. Martinčić-Ipšić, T. Miličić, and A. Todorovski, "The influence of feature representation of text on the performance of document classification," Appl. Sci. (Basel), vol. 9, no. 4, 2019, p. 743.

[21] N. Matas, S. Martinčić-Ipšić, and A. Meštrović, "Comparing network centrality measures as tools for identifying key concepts in complex networks: A case of Wikipedia," J. Digit. Inf. Manag., vol. 15, no. 4, p. 203, 2017.
[22] A. Nenkova and K. McKeown, "A survey of text summarization techniques," in Mining Text Data, Boston, MA: Springer US, 2012, pp. 43–76.
[23] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, Automatic evaluation of topic coherence. In: Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics. 2010. p. 100-108.
[24] Understanding the Infodemic and Misinformation in the fight against COVID-19, Pan American Health Organization, 2020. https://iris.paho.org/handle/10665.2/52052
[25] R. Rehurek, and P. Sojka, (2010). Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks.
[26] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15, 2015.
[27] "Stemmer for Croatian," Ffzg.hr. [Online]. Available: http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian/. [Accessed: 10-Jan-2021].
[28] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation," in 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017.
[29] V. Tangcharoensathien et al., "Framework for managing the COVID-19 infodemic: Methods and results of an online, crowdsourced WHO technical consultation," J. Med. Internet Res., vol. 22, no. 6, p. e19659, 2020.
[30] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu, "Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modelling on Twitter," PLoS One, vol. 15, no. 9, p. e0239441, 2020.
[31] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta, "Large-scale high-precision topic modelling on twitter," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014.
[32] J. Zarocostas, "How to fight an infodemic," Lancet, vol. 395, no. 10225, p. 676, 2020.