Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal

Buhin Pandur, Maja; Dobša, Jasminka; Beliga, Slobodan; Meštrović, Ana

Topic modelling and sentiment analysis of COVID-19 related news on Croatian Internet portal

Maja Buhin Pandur
Faculty of Organization and Informatics,
University of Zagreb
Varaždin, Croatia
mbuhin@foi.hr

Jasminka Dobša
Faculty of Organization and Informatics,
University of Zagreb
Varaždin, Croatia
jasminka.dobsa@foi.hr

Slobodan Beliga
University of Rijeka, Department of Informatics &
University of Rijeka, Center for Artificial
Intelligence and Cybersecurity
Rijeka, Croatia
sbeliga@uniri.hr

Ana Meštrović
University of Rijeka, Department of Informatics &
University of Rijeka, Center for Artificial
Intelligence and Cybersecurity
Rijeka, Croatia
amestorovic@uniri.hr

ABSTRACT

The research aims to identify topics and sentiments related to the COVID-19 pandemic in Croatian online news media. For analysis, we used news related to the COVID-19 pandemic from the Croatian portal *Tportal.hr* published from 1st January 2020 to 19th February 2021. Topic modelling was conducted by using the LDA method, while dominant emotions and sentiments related to extracted topics were identified by National Research Council Canada (NRC) word-emotion lexicon created originally for English and translated into Croatian, among other languages. We believe that the results of this research will enable a better understanding of the crisis communication in the Croatian media related to the COVID-19 pandemic.

KEYWORDS

News media, sentiment, emotions, pandemic, lexicon approach, Latent Dirichlet Allocation

1 INTRODUCTION

There are three major approaches to sentiment and emotions analysis in text: lexicon based, machine learning based approach [12] and the most recent deep-learning approach. In this research, we used a hybrid approach by applying the method of Latent Dirichlet Allocation (LDA) for topic modelling [6] and lexicon approach by using NRC word-emotion lexicon [13] for detection of sentiments (positive or negative) and basic emotions, according to Pluchik's model of emotions [15], in extracted topics.

The main goal of this paper is to analyse sentiments and emotions in crises communication in the news related to the COVID-19 pandemic published on the Croatian online portal. Our goal was aggravated in this research because articles belong rather to objective than to subjective type of reporting. Another problem is the lack of lexical resources for sentiment and emotions in the Croatian language. Glavaš and co-workers [10] developed a Croatian sentiment lexicon called CroSentiLex, which consists of positive and negative lists of words ranked with PageRank scores. Nevertheless, there is no available lexicon for the analysis of emotions for the Croatian language. Our analysis uses the NRC word-emotion lexicon, initially developed for English and translated into 104 languages, including Croatian. Such an approach has disadvantages due to cultural differences, but developing emotion lexicons for low-resource languages as Croatian is very demanding. Sentiment analysis of COVID-19 related texts is conducted mainly for texts written in English, such as research by Shofiya and Abidi [17], where the SentiStrength tool was used to detect the polarity of tweets, and support vector machine (SVM) algorithm was employed for sentiment classification. In [14], tweets about COVID-19 in Brazil written in Brazilian Portuguese due to lack of language resources are analysed by translating original text from Portuguese to English and using available resources for English.

Regarding Croatian social media space, Twitter social network communication was analysed through sentiment analysis [2] and COVID-19 information spreading [3]. Crisis communication of Croatian online portals was already explored by topic modelling of COVID-19 related articles [7]. However, in that research, it is not included further sentiment and emotional analysis of topics. In [4], information monitoring and name entity

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

recognition were conducted on news portal texts related to pandemics.

2 METHODS

2.1 Latent Dirichlet Allocation

LDA is a generative, probabilistic hierarchical Bayesian model that induces topics from a document collection [5,6]. The intuition behind topic modelling using LDA is that documents exhibit multiple topics. The topic is formally defined as a distribution over fixed vocabulary. Induction of topics is done in three steps:

- Each document in the collection is distributed over topics that are sampled using Dirichlet distribution.
- Each word in the document is connected with one single topic based on Dirichlet distribution.
- Each topic is defined as a multinomial distribution over words that are assigned to the sampled topics.

Topic modelling by LDA is conducted using *stm* package in R [16].

2.2 Number of topics estimation

Before performing the LDA topic modelling, it has to be estimated the number of topics. In this research we used four metrics from the R package *ldatuning*: Arun2010 [1], CaoJuan2009 [8], Deveaud2014 [9], and Griffiths2004 [11]. Measures Arun2010 and CaoJuan2009 have to be minimised, while measures Deveaud2014 and Griffiths2004 have to be maximised. However, as measures, Arun2010 and CaoJuan2009 generally decrease with the number of topics, and measures Deveaud2014 and Griffiths2004 increase with the number of topics, we will choose the number of topics as the value when observed measures start to stagnate.

2.3 Detection of sentiments and emotions

For the association of sentiments and emotions to extracted topics it was used NRC word-emotion lexicon [13], which consists of 14,182 words with scores of 0 or 1, according to the association to *positive* or *negative* sentiment or one of eight emotions of Pluchick's model (*anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*) [15]. The lexicon was created manually by crowdsourcing on Mechanical Turk.

For every sentiment and emotion, we created a vector with a distribution of zeros and ones over the words of a controlled dictionary created from the collection. Association of topics to sentiments and emotions is calculated as the cosine similarity between vectors of topics and corresponding vector of sentiment or emotion.

3 EXPERIMENT

3.1 Data set and preprocessing

The data set used for research consists of articles from the Internet portal *Tportal.hr* related to the topics of COVID-19 pandemic crises and collected from 1st January 2020 to 19th February 2021. Each article included in the dataset is defined as a COVID-19 article only if it contains at least one keyword related to coronavirus thematic. We use COVID-19 thesaurus for article filtering, which contains about thirty of the most important words describing the SARS-CoV-2 virus epidemic together with their corresponding morphological variations. From the total of 31,177 articles, according to defined filtering, the dataset used in the experiment consists of 12,080 COVID-19 related articles. Articles on the portal are categorised into one of nine main categories: *Biznis* (*Business*), *Sport* (*Sport*), *Kultura* (*Culture*), *Tehno* (*Techno*), *Showtime*, *Lifestyle*, *Autozona* (*Autozone*), *Funbox*, and *Vijesti* (*News*) (see Table 1).

Documents of a collection are created using text from the article's subcategory, introduction, main text, and tags. The collection is preprocessed by ejection of English and Croatian stop words and numbers and performing a lemmatisation. It is created a term-document matrix using *tf-idf* weighting scheme. The collection is indexed by terms contained in at least four documents of the collection, and the final list of index terms contained 31,121 terms.

Table 1: Number of articles from dataset categorised into one of nine main categories

Category	Number COVID-19 articles
Business	2,767
Sport	2,008
Culture	894
Techno	101
Showtime	1,352
Lifestyle	1,442
Autozone	124
Funbox	58
News	3,334

3.2 Results

As a first step, the number of topics had to be estimated. Since articles on the portal are categorised into nine main categories, we examined a number of topics from 5 to 15. We chose nine topics since the metrics started to stagnate for a higher number of topics (see Figure 1).

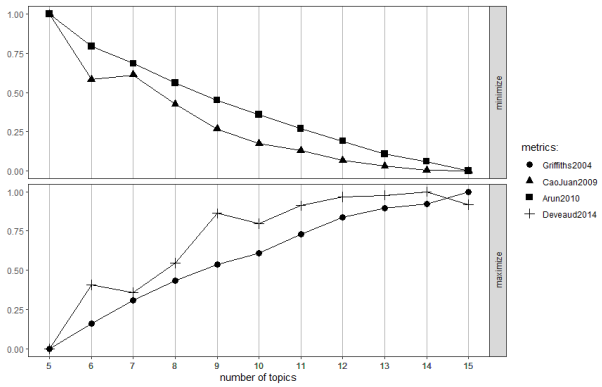

Figure 1: Metrics for estimation of the best fitting number of topics for 5 to 15 topics

Table 2: Top 10 words with the largest probabilities over topics and top 10 words with a *negative* sentiment with the largest probabilities over topics, both sorted in descending order of their probabilities. Topics are sorted by their representation in documents in descending order.

Topic's theme	Top 10 words
Topic 1 – *Sport*	*words by theme:* koronavirus (coronavirus), liga (league), klub (club), nogometni (football), igrač (player), godina (year), utakmica (match), sezona (season), hrvatski (Croatian), nogomet (football)
	words by negative sentiment: igrač (player), velik (big), problem (problem), epidemija (epidemic), odgoditi (to delay), prekinuti (to interrupt), čekati (to wait), borba (fight), napraviti (to make), posljedica (consequence)
Topic 2 – *Vaccination and epidemic measures*	*words by theme:* cijepljenje (vaccination), cjepivo (vaccine), zemlja (country), europski (European), koronavirus (coronavirus), doza (dose), predsjednik (president), vlada (government), mjera (measure), čovjek (man)
	words by negative sentiment: vlada (government), velik (big), epidemija (epidemic), red (order), borba (fight), sud (court), granica (border), problem (problem), potreban (required), upozoriti (to warn)
Topic 3 – *Earthquake and government measures*	*words by theme:* mjera (measure), hrvatska (Croatia), vlada (government), rad (labor), pomoć (help), potpora (support), odluka (decision), potres (earthquake), zaštita (protection), Zagreb
	words by negative sentiment: potres (earthquake), velik (major), pogoditi (to hit), potreban (required), posao (job), šteta (demage), prijava (report), republika (republic), poziv (call), posljedica (consequence)
Topic 4 – *Lifestyle*	*words by theme:* modni (fashion), godina (year), pandemija (pandemic), nov (new), koronavirus (coronavirus), poznat (famous), moda (fashion), obitelj (family), brend (brand), model (model)
	words by negative sentiment: velik (big), nositi (to wear), izolacija (isolation), veza (relationship), majka (mother), dug (debt), djevojka (wench), znak (sign), mali (small), pun (full)
Topic 5 – *Generally stories*	*words by theme:* čovjek (man), vrijeme (time), znati (know), virus (virus), velik (big), život (life), dan (day), dijete (child), koronavirus (coronavirus), dobro (good)
	words by negative sentiment: velik (big), virus (virus), problem (problem), posao (job), napraviti (to make), bolest (disease), mali (small), potreban (required), teško (hard), nositi (to wear)
Topic 6 – *Business 1*	*words by theme:* posto (percentage), godina (year), pad (drop), velik (big), pandemija (pandemic), tržište (market), rast (growth), kuna, gospodarstvo (economy), banka (bank)
	words by negative sentiment: pad (drop), velik (big), kriza (crisis), vlada (government), prihod (income), smanjiti (decrease), mali (small), trošak (expenditure), posljedica (consequence), epidemija (epidemic)
Topic 7 – *Daily reports*	*words by theme:* osoba (person), koronavirus (coronavirus), covid, slučaj (case), mjera (measure), broj (number), županija (county), nov (new), sat (hour), bolnica (hospital)
	words by negative sentiment: bolest (disease), virus (virus), zaraziti (to infect), zaraza (infection), epidemija (epidemic), umrijeti (to die), velik (big), infekcija (infection), zarazan (contagious), simptom (symptom)
Topic 8 – *Culture*	*words by theme:* godina (year), film (film), nov (new), festival (festival), program (program), hrvatski (Croatian), Zagreb, kultura (culture), kazalište (theater), knjiga (book)
	words by negative sentiment: velik (big), mali (small), predstavljati (to present), nastup (appearance), otkazati (to cancel), odgoditi (to delay), smrt (death), rat (war), strana (side), kritika (critique)
Topic 9 – *Business 2*	*words by theme:* nov (new), proizvod (product), automobil (car), velik (big), godina (year), hrvatska (Croatia), proizvodnja (production), tvrtka (company), trgovina (market), kupac (buyer)
	words by negative sentiment: velik (big), nafta (oil), epidemija (epidemic), lanac (chain), smanjiti (decrease), kriza (crisis), mali (small), zaraza (infection), problem (problem), utjecaj (influence)

Topics were labelled based on words with the largest probabilities in topics vectors (keywords) shown in Table 2. Some of the topics are directly connected to main categories on the portal: the first topic is labelled as *Sport*, the fourth topic as *Lifestyle*, and the eighth topic as *Culture*, while the sixth and the ninth topics are connected to the business world and are labelled as *Business 1* and *Business 2*. *Business 1* is associated with the capital market, while *Business 2* is associated with production. Topic 2 is associated with *Vaccination and epidemic measures*, while Topic 3 is associated with *Earthquake and government measures*. Topic 5 seems rather *General on stories* in a pandemic world, while Topics 7 contains *daily reports* on the pandemic state.

We found that all topics are mainly associated with *negative* sentiments. In Table 2 are listed words associated with *negative* sentiment with the largest probabilities across topics, while words associated with *positive* sentiment have coincided with the words from topics theme. This list gives some insight into what "bears" *negative* sentiment in the topics.

Figure 2 shows the association of topics to sentiments and emotions. The ratio of *positive* and *negative* sentiments is the best for categories of *Sport* and *Culture*. These categories and *Lifestyle* are only categories associated with *joy* as one of the dominant emotions. *Surprise* and *anticipation* are dominant emotions across all topics. Categories *Vaccination and epidemic measures*, *Earthquake and government support*, *Generally stories* and *Business 1* are associated with the emotion of *sadness*, while categories *Vaccination and epidemic measures* and *Daily reports* are associated with *fear*.

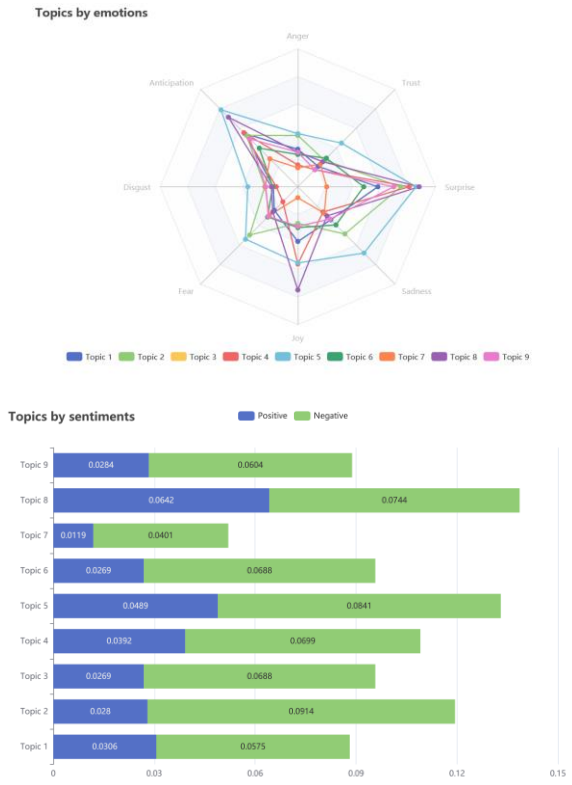

Figure 2: Association of topics to sentiments and emotions

4 CONCLUSIONS AND FURTHER WORK

The main goal of this paper was to analyse sentiments and emotions in crises communication in the news related to the COVID-19 pandemic. For that purpose, we have created our collection of documents from articles on the Internet news portal connected to pandemic crises and analysed it utilising the LDA method for extraction of prevalent topics in the collection and NRC word-emotion lexicon for detection of sentiments and emotions associated with extracted topics.

Application of LDA resulted in relatively intuitive topics. Some of them can be associated with the main categories of the observed portal, and the other are related to the actual situation in a pandemic world in Croatia: *vaccination*, *earthquake* (there were two great earthquakes in Croatia in 2020), *stories*, *daily reports*. It is shown that all extracted topics are associated dominantly with *negative* sentiment, while prevalent emotions are *anticipation*, *surprise*, *sadness* and *fear*.

By this research, we have gained insight into how COVID-19 pandemic crises was communicated to the public. To gain insight into how the public experienced the crises, we could use the same methodology applied to comments of articles or on social networks. This could be a direction for a further work. Also, it would be interesting to investigate how topics and sentiments/emotions are changing and evaluating over time.

ACKNOWLEDGEMENTS

This work has been supported in part by the Croatian Science Foundation under the project IP-CORONA-04-2061, “Multilayer Framework for the Information Spreading Characterization in Social Media during the COVID-19 Crisis” (InfoCoV) and by the University of Rijeka project number uniri-drustv-sp-20-58.

REFERENCES

[1] R. Arun, V. Suresh, C.E. Madhavan and M. Narasima Murty. 2010. On finding the natural number of topics with Latent Dirichlet Allocation: Some observations, In *Proceedings of Advances in Knowledge Discovery and Data Mining, 14th Pacific-Asia Conference* (PAKDD 2010), Hyderabad, India. doi: 10.1007/978-3-642-1357-3_43.

[2] K. Babić, M. Petrović, S. Beliga, S. Martinčić-Ipšić, A. Jarynowski and A. Meštrović. 2022. COVID-19-Related Communication on Twitter: Analysis of the Croatian and Polish Attitudes. In: *Yang XS., Sherratt S., Dey N., Joshi A. (eds) Proceedings of Sixth International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems*, vol 216. Springer, Singapore. Available at https://link.springer.com/chapter/10.1007%2F978-981-16-1781-2_35.

[3] K. Babić, M. Petrović, S. Beliga, S. Martinšić-Ipšić, M. Pranjić and A. Meštrović. 2021. Prediction of COVID-19 related information spreading on Twitter. In *Proceedings of the IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2021)*, accepted for publication.

[4] S. Beliga, S. Martinčić-Ipšić, M. Matešić and A. Meštrović. 2021. Natural Language Processing and Statistic: The First Six Months of the COVID-19 Infodemic in Croatia, In *The Covid-19 Pandemic as a Challenge for Media and Communication Studies*. K. Kopecka-Piech and B. Łódzki, Eds., Routledge, Taylor & Francis Group, accepted for publication.

[5] D. M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84. doi:10.1145/2133806.2133826.

[6] D. M. Blei, A. Y. Ng and M.I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993-1022.

[7] P. K. Bogović, S. Beliga , A. Meštrović and S. Martinčić-Ipšić. 2021. Topic modelling of Croatian news during COVID-19 pandemic. In *Proceedings of the IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2021),* accepted for publication.

[8] J. Chao, L. Tian, Z. Jintao, T. Yongdong and S. Tang. 2009. A density-based method for adaptive LDA model selection, *Neurocomputing*, 72(7-9), 1775-1781. doi: 10.1016/j.neucom.2008.06.0011.

[9] R. Deveaud, E. Sanjuan, P. Bellot. 2014. Accurate and effective latent concept modeling for ad hoc information retrieval, *Document Numérique*, 17(1). doi: 10.3166/dn.17.1.61-84.

[10] G. Glavaš, J. Šnajder and B. Dalbelo Bašić. 2012. Semi-supervised acqusition of Croatian sentiment lexicon. In *Proceedings of 15th International Conference on Text, Speech and Dialogue, TSD 2112,* Brno, 166-173.

[11] T.L. Griffiths, M. Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences 101 Suppl* 1(1), 5228-35, doi: 10.1073/pnas.0307752101.

[12] H. Lane, C. Howard and H. Hapke. 2019. *Natural Language Processing in Action*. Manning Publications, New York, NY.

[13] S. M. Mohammad and P.D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.

[14] T. Melo and C. M. S. Figueiredo. 2021. Comparing news articles and tweets about COVID-19 in Brasil: Sentiment analysis and topic modeling approach. *JMIR Public Health and Surveillance*, 7(2), doi: 10.2196/24585.

[15] R. Plutchik. 1962. *The Emotions*. Random House, New York, NY.

[16] M. Roberts, B.M. Stewart and D. Tingley. 2019. stm: An R package for structural topic models, *Journal of Statistical Software*, 91(2), 1-40. doi: 10.18637/jss.v091.i02.

[17] C. Shofiya and S. Abidi. 2021. Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data. *International Journal of Environmental Research and Public Health*, 18(11), 1-10.