ANALYSIS OF ROMANIAN ONLINE NEWS ARTICLES REGARDING COVID-19 VACCINES USING NLP TECHNIQUES

Mihnea Andrei CALIN[1], Denis IORGA[2], Irina TOMA[3], Diana OLAR[4], Mihai DASCALU[5], Cezar SCARLAT[6], Gheorghe MILITARU[7]

Understanding the relationship between online media and vaccine-related information is essential for public inoculation strategies. Despite the advent of automated methods for this purpose, there is a gap in terms of applying Natural Language Processing techniques (NLP) to understand information regarding COVID-19 vaccines in Romanian online news. In this sense, this pilot study aims to close the gap by using NLP techniques to analyze information related to vaccines in online news articles. A corpus of 5,670 vaccine-related online news articles published between January and December 2021 was analyzed using sentiment and word cloud analyses to understand the valence and content of COVID-19 vaccine-related information. The results indicate the utility of the proposed method for public and private actors, as well as further required efforts for using NLP techniques to understand and monitor information regarding vaccines present in Romanian online news articles.

Keywords: Natural Language Processing, Vaccines, COVID-19, Sentiment analysis, Romanian online news articles

[1] PhD student, Doctoral School of Entrepreneurship, Business, Engineering and Management, University POLITEHNICA of Bucharest, Romania, e-mail: mihnea.calin@gmail.com

[2] PhD student, Research Technology, 19D Soseaua Virtutii, 060782 Bucharest, Romania; Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania, e-mail: denis.iorga@drd.unibuc.ro

[3] PhD student, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: irina.toma@upb.ro

[4] PhD student, Interdisciplinary School of Doctoral Studies, University of Bucharest, Romania, e-mail: diana.a.olar@gmail.com

[5] Professor, Research Technology, 19D Soseaua Virtutii, 060782 Bucharest, Romania; Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: mihai.dascalu@upb.ro

[6] Professor, Doctoral School of Entrepreneurship, Business Engineering and Management, University POLITEHNICA of Bucharest, Romania, e-mail: cezar.scarlat@upb.ro

[7] Professor, Doctoral School of Entrepreneurship, Business Engineering and Management, University POLITEHNICA of Bucharest, Romania, e-mail: gheorghe.militaru@upb.ro

1. Introduction

Different types of sources of information that individuals use to manifest information-seeking behaviors have an impact on perceptions regarding vaccines. For example, social media fuels negative attitudes towards vaccines through the rapid spread of misinformation [1]. As such, multiple studies were conducted to understand the relationship between social media and vaccine perceptions [2-4]. In contrast, institutional and governmental websites have the opposite effect on attitudes towards vaccines, by encouraging vaccination rather than promoting vaccination delaying or refusal [5].

Of interest to this paper is the information related to vaccines presented in online news articles. Placed at the border between social media and institutional or governmental websites, online news articles can have both positive and negative effects on attitudes related to vaccines. For example, it is argued that certain journalistic practices promote negative attitudes towards vaccines [6, 7], and that simply reading about vaccines in news articles is correlated with vaccine hesitancy attitudes [8]. On the other hand, news articles can be used to promote vaccination through the dissemination of vaccine promotion information [9].

In line with previous information, the necessity to understand existing opinions towards vaccines portrayed by online news articles is becoming more stringent. Such an understanding would provide valuable information for stakeholders in the vaccination process. In this sense, the development of automated Natural Language Processing (NLP) techniques emerges as a possibility. Thus, the following lines focus on a series of studies relevant to this approach.

Bai, Jia, and Chen [10] used a topic modeling approach on an open-source corpus of 3534 COVID-19-related new articles published by the Canadian Broadcasting Corporation (CBC) to understand how the public discourse around the pandemic changed between January and March 2020. First, the authors determined the number of topics based on a dynamic topic modeling approach. Then, the authors defined their topics based on a keyword analysis for each topic. Their analysis managed to illustrate the evolution of various topics, such as the downfall of the “Wuhan” topic in the analyzed corpus.

A similar study [11] focused on articles published by major Chinese news media between January and February 2020. First, the authors applied a Dirichlet topic modeling procedure to a dataset of 7791 relevant articles related to COVID-19. Then, the authors generated keywords and names for each topic. Finally, the resulting topics made them conclude that “prevention and control procedures, medical treatment and research, and global or local social and economic influences” [11] were the top three most popular topics in the analyzed corpus.

Other studies focused on applying NLP methods on social media to understand attitudes toward the COVID-19 pandemic. In one study, Barkur and Vibha [12] analyzed 24,000 Indian posts published on Twitter using keyword and sentiment analysis. Interestingly, their sentiment analysis results revealed that the most prominent valence of the posts was a positive one, arguing that individuals were optimistic regarding their efforts to "flatten the curve". Yet another study on social media data shows the complementary use of topic modeling and sentiment analysis to understand public attitudes towards the COVID-19 pandemic [13].

While NLP techniques were widely used to understand the public discourse around the COVID-19 pandemic both in online newspapers and social media platforms, that is not the case for information related to vaccines specifically. Most of the research conducted in this sense rather focused on analyzing attitudes towards vaccines in social media. One example is a study in which NLP methods were used to study vaccine hesitancy [14]. The study involved conducting a sentiment analysis procedure on a corpus of Twitter posts. First, the authors extracted a corpus of Twitter posts published from June 2011 to April 2019. Then, they developed a supervised sentiment classification algorithm capable of discerning between positive, neutral, and negative posts. Following this, the authors managed to illustrate time and space variations of the valence of vaccine-related posts. For example, their results show that positive tweets regarding vaccines peaked every April within the above interval and were more prevalent in Switzerland.

Two research gaps can be identified based on the information presented above. First, there is a need for studies that seek to understand information regarding vaccines in online newspapers rather than in social media posts. Second, there seems to be a lack of research efforts to apply NLP methods to Romanian information sources, despite the impact of such sources of information on information-seeking behaviors regarding vaccines. In this sense, the current paper aims to close the two gaps by analyzing information related to COVID-19 vaccines in a Romanian online news outlet using sentiment analysis and keyword analysis. The next section introduces the methodology and the collected corpus, followed by results and discussions.

2. Method

2.1 Corpus

A collection of 5,670 online news articles was extracted from the Agerpres (https://www.agerpres.ro/) news website to understand the valence of attitudes towards various COVID-19 vaccines. Given that the analysis aimed to illustrate the applicability of the method in analyzing vaccine-related information, the choice of the data source was made based on criteria such as the high prestige of the agency and its perceived neutrality by the research team. The extraction procedure involved the use of web crawling and web scraping methods. An initial HTML format analysis of the web pages was performed to identify relevant data to be extracted. The title, the main content, the publish date, and the keywords associated with the articles were identified as information of interest for our experiment. This information was obtained through a web crawler implemented using BeautifulSoup [15].

A total of 5670 online news articles published in the Romanian language between the 1st of January 2021 and 31 December 2021 were collected. The aim was to analyze vaccine-related information published over the course of a year by the news outlet. Therefore, we selected the year 2021 as COVID-19 vaccines were already widely available. Table 1 presents the distribution of the extracted articles per month. As can be observed, we selected only articles that contained the "vaccin" keyword in their content, assuming that such articles are providing information related to COVID-19 vaccine.

Table 1

Corpus Statistics

Month (2021)	Nr. articles that contain the "vaccin" keyword	% of total articles extracted
January	656	12%
February	491	9%
March	630	11%
April	513	9%
May	766	14%
June	316	6%
July	495	9%
August	273	5%
September	378	7%
October	523	9%
November	340	6%
December	289	5%
Total	5,670	100%

2.2 Article Processing

The first step of the analysis involved conducting sentiment analysis [16] to identify positive, neutral, and negative articles from the previously introduced corpus. A deep learning approach based on a BERT-based [17] architecture was considered. As such, RoBERT [18] was fine-tuned for sentiment analysis using a corpus of 160,000 product reviews from a Romanian e-commerce website. The model achieved an accuracy of 76.14% in assigning the valence of reviews.

The fine-tuned model receives a text as input and provides a continuous score from 0 (maximum negative sentiment) to 1 (maximum positive sentiment).

For this analysis, the results were recoded into negative (0-0.45), neutral (0.451-0.699), and positive (0.7-1.0) valences. The cutting points were selected while accounting for the initial distribution of reviews; inherently, the model was inclined to provide larger values and we wanted to better grasp negative impressions.

Moreover, we decided to select three-time segments to compare differences in the valence and content of vaccine-related articles. In this sense, we used the number of COVID-19 infections in Romanian territory during 2021. As such, we compared the valence and content of articles using the following temporal categories, using as a criterion the total number of new infections reported by Romanian national authorities: *wave 3* of infections (1st January– 31st May 2021), *in-between waves* (1st July-31st July 2021), and *wave 4* of infections (1st September – 31st December 2021). In this sense, the worldometer website[8] was considered to identify COVID-19 infection spikes in the Romanian territory. It was assumed that a wave period starts when there are more than 1400 new daily COVID-19 infections and is characterized by an ascending trend in the number of infections the following week. Reversely, it was assumed that the end of a wave period starts when there are less than 1400 new daily COVID-19 infections and is characterized by a descending trend in the number of infections the following week.

The second step of the analysis aimed to explore the content of positive, neutral, and negative online news articles using word cloud visualizations. To achieve this, the keywords associated with the news articles by their authors were analyzed via word cloud visualization based on the frequency of lemmas corresponding to content words. A world cloud analysis on the top 30 keywords (except for the words “Coronavirus”, “Covid-19”, and “vaccin”) from *wave 3* and *wave 4* articles was conducted. The following section introduces the results of the sentiment and word cloud analyses. Furthermore, the frequency of specific vaccine keywords was analyzed for articles in *wave 3* and *wave 4* to understand their presence in positive, neutral, and negative articles. The latter analysis is illustrated in the discussion section as it represents an interpretation of the results of the word cloud analysis.

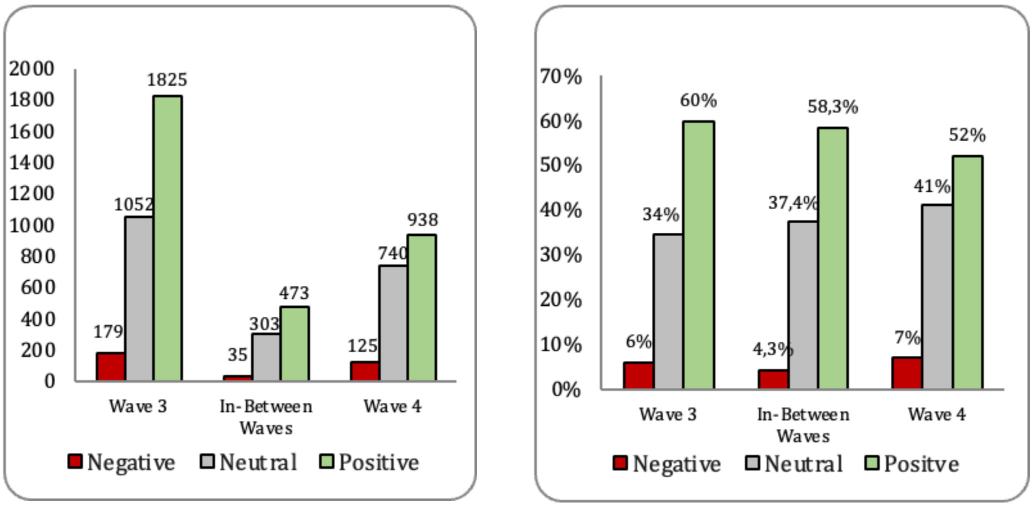
3. Results

The first step of the analysis was to identify the number of negative, neutral, and positive articles related to vaccines by the wave of infection (see Fig 1). Overall, there is a considerable variation between waves 3 and 4 with regard to the number of articles published on vaccine-related matters. It can also be observed that the number of positive articles registered the most considerable

[8] Available at: https://www.worldometers.info/coronavirus/country/romania/

decline during wave 4. In this sense, there were 887 fewer positive articles in wave 4 when compared to wave 3 (-48.6%). Both the number of negative and neutral articles also registered a decline, with 54 fewer negative articles (-30.1%) and 314 fewer neutral articles (-29.6%) in wave 4, when compared to wave 3.

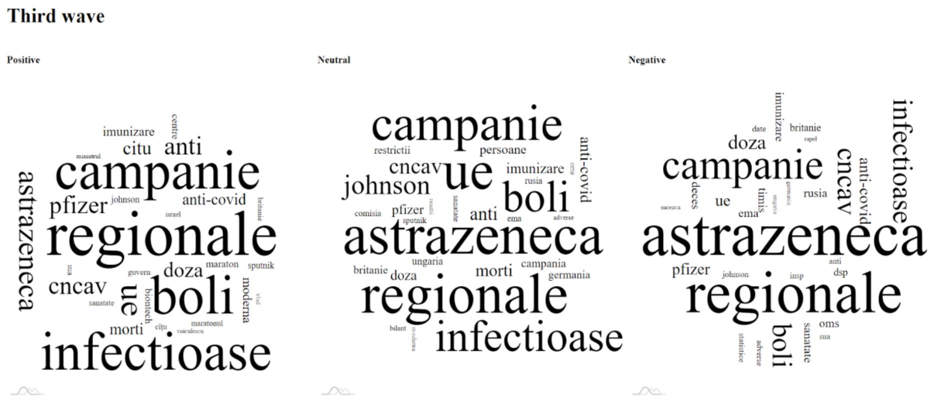
While all three categories registered a decline, it can be observed that the proportion of positive, neutral, and negative articles remained relatively constant both during and between waves (see Fig 1.a). However, an additional analysis was conducted to avoid misleading insights caused by the use of absolute values. Thus, we analyzed the distribution of positive, neutral, and negative articles relative to the total number of articles during each temporal category (see Fig 1.b). An actual increase in the proportion of neutral (+6%) and negative (+1%) articles in the 4th wave, when compared to articles in the 3rd wave, can be observed. Additionally, the decrease in the proportion of positive articles is confirmed (-8%).



(a) (b)

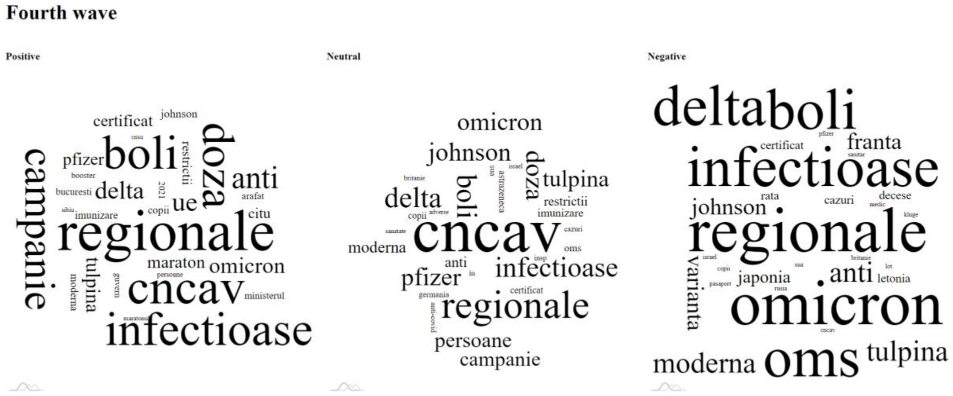
Fig. 1. Valence of vaccine-related articles by infection wave in the analyzed corpus
a) absolute values; b) relative to the number of articles in each of the three segments.

The second step of the analysis involved comparing the content of positive, neutral, and negative articles from *wave 3* and *wave 4* of infections. The results of the word cloud analysis on the top 30 keywords (except for the words “Coronavirus”, “Covid-19”, and “vaccine”) from wave 3 (see Fig. 3) and wave 4 (see Fig. 4) articles are presented below.


Third wave

Positive Neutral Negative

Fig. 3. Top 30 journalist keywords in positive, neutral, and negative articles from Wave 3


Fourth wave

Positive Neutral Negative

Fig. 4. Top 30 journalist keywords in positive, neutral, and negative articles from Wave 4.

As can be observed in the figures above, the word cloud analysis illustrates the frequency of keywords used by journalists for articles classified as being positive, neutral, or negative. Given that the name of specific vaccines appears amongst the keywords, an estimation regarding the positive, neutral or negative attitude towards specific vaccines can be obtained. However, such an analysis faces a series of limitations and is therefore presented in the following section.

4. Discussion and Limitations

The proposed method shows whether a news media outlet uses positive, negative, or neutral language in articles related to vaccines. The results of our analysis highlight that the analyzed media outlet rather used positive and neutral language toward vaccines. In line with existing literature that used sentiment analysis on social media posts [12], there were more positive articles than negative or neutral articles posted on the matter of vaccines.

Additionally, the proposed method can be used to monitor the consistency with which a news media outlet communicates articles related to vaccines. The results of the analysis show that there were more articles posted during wave 3 (1st January 2021 – 31st May) than during wave 4 (1st September – 31st December). In line with existing literature [14], there were more articles related to vaccines posted during the first half of the year compared with the second half of the year.

Moreover, the proposed method can monitor the consistency with which a news media outlet communicates positive, neutral, and negative articles related to vaccines. As the results reveal, the analyzed news media outlet was consistent in terms of the positive/neutral/negative articles ratio.

A qualitative analysis on a random sample of 100 articles was also conducted to evaluate the efficiency of the sentiment analysis model in classifying the valence of the online news articles. A human annotator was requested to label the positive, neutral, or negative valences corresponding to the 100 articles in the random sample. A Cohen's kappa of $k = .289$ ($p < .001$) and an accuracy of 55% between the labels of the annotator and those of the model suggest a fair agreement [19]. At the same time, this evaluation points toward the main limitation of the current study, namely that our sentiment analysis model was trained on a different dataset with a different focus. In this sense, further efforts are required to train a sentiment analysis model for this particular task.

The same qualitative evaluation procedure also revealed the already mentioned tendency of the model to provide more positive results in terms of sentiment polarity due to its training on product reviews. On the contrary, the annotator presented a tendency toward labeling the articles as being neutral. The confusion matrix presented in Table 2 illustrates the comparison between the labels associated given by the annotator versus the ones generated by the model.

Table 2

Table of confusion between labels of the model and labels of the annotator

Annotator / Model	Positive	Neutral	Negative
Positive	5	5	1
Neutral	2	18	3
Negative	7	27	32

The use of word cloud analysis on journalist keywords in conjunction with the sentiment analysis method enables an estimation regarding public attitudes toward certain vaccines. The word clouds from Fig. 3 and Fig. 4 depict keyword references to various vaccines. Thus, the frequency of such keywords associated with positive, neutral, and negative articles may provide a proxy regarding the preference/hesitancy to use certain vaccines by the audience of the respective news outlet. However, this interpretation must consider the second limitation of our current study, namely that the sentiment analysis was conducted at the document-level. In other words, the reference to the vaccine may be positive, while the rest of the article may be negative.

Considering this latter limitation, the frequency of keywords related to specific vaccines is summarized in Table 3. The relative frequency is also represented in Fig. 5. The results reveal, for example, that AstraZeneca was the most negatively perceived vaccine, which was afterward removed from the administration scheme. However, this interpretation must consider that the negative count of negative news articles is considerably lower than neutral or positive articles in the corpus.

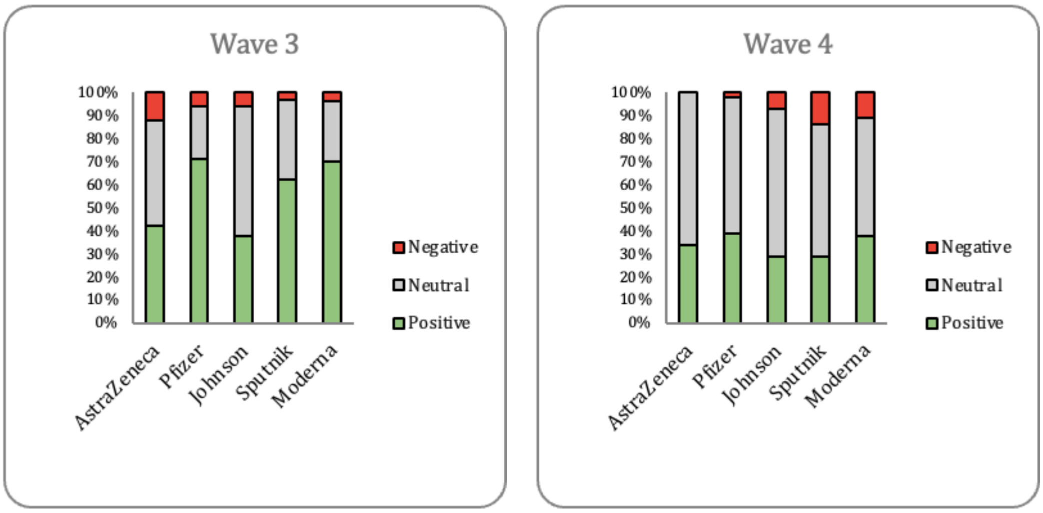
Table 3

Frequency of specific vaccine keywords

Vaccine	Wave 3			Wave 4		
	Positive	Neutral	Negative	Positive	Neutral	Negative
AstraZeneca	115 (42%)	127 (46%)	**32** (12%)	10 (34%)	19 (66%)	0 (0%)
Pfizer	99 (71%)	32 (23%)	9 (6%)	29 (39%)	44 (59%)	2 (2%)
Johnson	29 (38%)	43 (56%)	5 (6%)	8 (29%)	18 (64%)	2 (7%)
Sputnik	43 (62%)	24 (35%)	2 (3%)	2 (29%)	4 (57%)	1 (14%)
Moderna	57 (70%)	21 (26%)	3 (4%)	23 (38%)	31 (51%)	7 (11%)

* Percentages are calculated from the row total of a specific vaccine in Wave 3 and Wave 4. For example, 12% of the total AstraZeneca keyword references from Wave 3 appear in negative articles

Table 3 and Fig. 5 also show that in the 3rd wave references to the Pfizer, Sputnik, and Moderna vaccines were mostly used as keywords in positive articles. On the other hand, all keywords related to vaccines in the 4th wave were mostly used in neutral articles. This provides helpful information to stakeholders on the vaccination process. However, further effort is required to validate the association between the keyword and the valence of the article. In this sense, a Named Entity Recognition (NER) approach and a paragraph-based sentiment analysis may be more appropriate for the matter.



(a) (b)

Fig. 5. Relative frequency of specific vaccine keywords by sentiment and wave of infections.

5. Conclusions and further research paths

This study addressed the research gap identified in the introductory section, namely the lack of studies that seek to use NLP techniques to understand information regarding COVID-19 vaccines in Romanian online news articles. Following a literature survey on the topic, the current work applied a mix of sentiment analysis and word cloud visualization techniques to explore information related to COVID-19 vaccines in Romanian online news articles. As such, the proposed method provides a tool for health officials and other private and public authorities to better understand the communication of news outlets, as well as associated sentiments and attitudes related to vaccines – in particular COVID-19 vaccines.

The proposed method allows the extraction of various indicators, for instance, the quantity, valence, consistency, and content of articles related to vaccines. In turn, these indicators may provide proxy variables to plan and monitor the impact of public health and communicational interventions on vaccine perceptions.

As a pilot study, method-centered primarily, the mix of sentiment analysis techniques and tools has proved its capability for further exploration and finer investigation of sentiments and attitudes by types of sources – *i.e.*, a larger set of online news outlets. In turn, this empowers follow-up analyses on the dissemination mechanisms and the understanding of relationships between healthcare institutions and the population (trust in particular), while striving for a better healthcare system.

Future work considers extending the data sources and using an improved and more granular sentiment analysis model; thus, we plan to compare opinions from different news media outlets using an alternative sentiment analysis model. Furthermore, the use of a topic modeling approach might also provide valuable information in terms of emergent latent topics and groupings of co-occurring words.

Acknowledgment

This work was funded by the “Innovative Solution for Optimizing User Productivity through Multi-Modal Monitoring of Activity and Profiles – OPTIMIZE” / “Soluție Inovativă de Optimizare a Productivității Utilizatorilor prin Monitorizarea Multi-Modala a Activității și a Profilelor – OPTIMIZE” project, Contract number 366/390042/27.09.2021, MySMIS code: 121491

REFERENCES

[1] H. J. Larson, "The biggest pandemic risk? Viral misinformation," Nature, vol. 562, pp. 309-310, 2018.

[2] W. Jennings, G. Stoker, H. Bunting, V. O. Valgarðsson, J. Gaskell, D. Devine, L. McKay, and M. C. Mills, "Lack of trust, conspiracy beliefs, and social media use predict COVID-19 vaccine hesitancy," Vaccines, vol. 9, p. 593, 2021.

[3] N. Puri, E. A. Coomes, H. Haghbayan, and K. Gunaratne, "Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases," Human vaccines & immunotherapeutics, vol. 16, pp. 2586-2593, 2020.

[4] S. L. Wilson and C. Wiysonge, "Social media and vaccine hesitancy," BMJ Global Health, vol. 5, p. e004206, 2020.

[5] C. Reno, E. Maietti, Z. Di Valerio, M. Montalti, M. P. Fantini, and D. Gori, "Vaccine hesitancy towards COVID-19 vaccination: investigating the role of information sources through a mediation analysis," Infectious disease reports, vol. 13, pp. 712-723, 2021.

[6] S. Ahmad Kamboh, M. Ittefaq, and A. A. Sahi, "Journalistic routines as factors promoting COVID-19 vaccine hesitancy in Pakistan," Third World Quarterly, pp. 1-10, 2021.

[7] D. Catalan-Matamoros and C. Elías, "Vaccine hesitancy in the age of coronavirus and fake news: analysis of journalistic sources in the Spanish quality press," International Journal of Environmental Research and Public Health, vol. 17, p. 8136, 2020.

[8] M. Vrdelja, A. Kraigher, D. Verčič, and S. Kropivnik, "The growing vaccine hesitancy: exploring the influence of the internet," European journal of public health, vol. 28, pp. 934-939, 2018.

[9] C. Calloway, C. M. Jorgensen, M. Saraiya, and J. Tsui, "A content analysis of news coverage of the HPV vaccine by US newspapers, January 2002–June 2005," Journal of women's health, vol. 15, pp. 803-809, 2006.

[10] Y. Bai, S. Jia, and L. Chen, "Topic evolution analysis of COVID-19 news articles," in Journal of Physics: Conference Series, 2020, p. 052009.

[11] Q. Liu, Z. Zheng, J. Zheng, Q. Chen, G. Liu, S. Chen, B. Chu, H. Zhu, B. Akinwunmi, and J. Huang, "Health communication through news media during the early stage of the COVID-19 outbreak in China: digital topic modeling approach," Journal of medical Internet research, vol. 22, p. e19118, 2020.

[12] G. Barkur and G. B. K. Vibha, "Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India," Asian journal of psychiatry, vol. 51, p. 102089, 2020.

[13] S. Boon-Itt and Y. Skunkan, "Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study," JMIR Public Health and Surveillance, vol. 6, p. e21978, 2020.

[14] H. Piedrahita-Valdés, D. Piedrahita-Castillo, J. Bermejo-Higuera, P. Guillem-Saiz, J. R. Bermejo-Higuera, J. Guillem-Saiz, J. A. Sicilia-Montalvo, and F. Machío-Regidor, "Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019," Vaccines, vol. 9, p. 28, 2021.

[15] L. Richardson, "Beautiful soup documentation". Retrieved 1st April 2016, ed, 2007.

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," arXiv preprint cs/0205070, 2002.

[17] J. Devlin and M.-W. Chang, "Open sourcing BERT: State-of-the-art pre-training for natural language processing," Google AI Blog, vol. 2, 2018.

[18] M. Masala, S. Ruseti, and M. Dascalu, "Robert–a romanian bert model," in Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6626-6637.

[19] J.R. Landis and G.G.K Koch, "The measurement of Observer Agreement for Categorical Data." Biometrics, 33(1), 159