The Open Public Health Journal
DOI: 10.2174/0118749445406703250824134524, 2025, 18, e18749445406703
ISSN: 1874-9445

RESEARCH ARTICLE
OPEN ACCESS

Patterns and behaviors during the Coronavirus Disease 2019 pandemic in Germany: A natural language processing application

Chiazam Izuchukwu[1,2], Hayden Wimmer[1,2,*], Jessica Schwind[2] and Joana Tome[2,3]

[1]Department of Information Technology, Georgia Southern University, Statesboro, GA, USA
[2]Institute for Health Logistics and Analytics, Georgia Southern University, Statesboro, GA, USA
[3]Department of Biostatistics, Epidemiology & Environmental Health Sciences, Georgia Southern University, Statesboro, GA, USA

Abstract:

Introduction: This study aimed to identify the underlying patterns and behaviors during the Coronavirus Disease pandemic for future preparedness and response strategies.

Methods: We applied natural language processing techniques to interview data of qualitative nature collected from 40 German participants across various phases of the study. We then preprocessed the data well, getting rid of stop words, tokenizing, stemming, and lemmatizing the text, all done to ensure that the analysis would be meaningful and accurate.

Results: Significant terms from the term frequency-inverse document frequency analysis included noting the terms people, mask, vaccination, and vaccinated. Latent semantic analysis expressed major topics in phase I including discussions of experiences, vaccination, government, preventive measures, and public sentiment. Phase II consisted of vaccination efforts, government trust, and public coronavirus opinions, whereas phase III encompassed long-term impacts, trust in preventive measures, and changes in vaccination efforts. Sentiment analysis showed that negative sentiments are more (> 60%).

Discussion: The analysis showed that public concerns moved from compliance to skepticism and identified central themes, including vaccination, trust, and emotional burden. TF-IDF and LSA shed light on an evolving discourse in the pandemic, and sentiment analysis showed a pervasive distress. Such insights reinforce the importance of effective communication and mental health interventions during public health emergencies.

Conclusions: These findings help us to know more about the pandemic's impact a decade later that may inform future research, public health strategies, and policymaking.

Keywords: Patterns, Behaviors, Coronavirus Disease of 2019, Term frequency-inverse document frequency, Latent semantic analysis, Sentiment analysis.

*Address correspondence to this author at the Department of Information Technology, Georgia Southern University, Statesboro, GA, USA; E-mail: hwimmer@georgiasouthern.edu

Cite as: Izuchukwu C, Wimmer H, Schwind J, Tome J. Patterns and behaviors during the Coronavirus Disease 2019 pandemic in Germany: A natural language processing application. Open Public Health J, 2025; 18: e18749445406703. http://dx.doi.org/10.2174/0118749445406703250824134524

Received: May 01, 2025
Revised: June 01, 2025
Accepted: June 18, 2025
Published: August 27, 2025

1. INTRODUCTION

The Coronavirus Disease of 2019 (COVID-19) pandemic altered daily life across the globe, prompting extensive research into its social, emotional, and behavioral impacts. Creating public health strategies and policies that work requires understanding how people navigate these unprecedented times. This pandemic occurred at a time when large-scale data (structured and unstructured) was accessible to public health and healthcare institutions. Unstructured data comes in various forms that do not neatly fit into traditional data models. Managing and analyzing unstructured data poses challenges, which have historically hindered analysis and search efforts, making unstructured data less useful for these institutions [1-3]. However, with the rise of innovative technologies such as data-driven artificial intelligence (AI), the landscape has changed, enabling more effective analysis of unstructured data [1, 4].

NLP is an area of AI technology with a lot of promise [5-7], especially considering the vast amounts of free-text data that are already accessible and constantly being produced through various channels. NLP, including SA algorithms and machine learning techniques, has been successfully applied in automobile insurance fraud detection, online retail branding, customer service and satisfaction, job satisfaction, political lean, geoscience, and cybersecurity [8-15]. NLP has also been used in the COVID-19 pandemic for disease forecasting, early detection and prognosis (non-imaging), drug repurposing and early drug development, social media data analysis, genomic, transcriptomic, and proteomic data analysis, as well as medical imaging-based diagnosis and prognosis [16-18]. Nonetheless, most research and review articles on deep learning techniques for COVID-19 concentrate on image classification applications [19-24].

Few studies have examined NLP use in COVID-19 and its social, emotional, and behavioral impacts, and most research were focused on analyzing social media data [25-31]. Hence, there is a need to apply NLP to more robust data. Examining COVID-19 social, emotional, and behavioral impacts offers important insights into the impact of the COVID-19 pandemic on the general population and identifies strategies that can increase future pandemic preparedness and response. The study by [32] utilizes NLP to analyze interviews about the impact of COVID-19 in rural communities and compare it to thematic analyses obtained traditionally from a subset of the interviews. The literature we reviewed highlights research gaps in the application of NLP to datasets in a structured qualitative interview form rather than social media data, especially in the context of the sentiments during the pandemic. By demonstrating the power of modern NLP techniques in analyzing unstructured data, this study can offer a foundation for future research and inform public health strategies and policymaking efforts. Therefore, the aim of this research was to explore the underlying patterns and behaviors during the COVID-19 pandemic in the German population by analyzing interview data over different time intervals.

2. METHODS

In this work, we first collect data from Herbig, *et al*. [33] then advance to data preprocessing. Following preprocessing, we generate TF-IDF analysis and word clouds. We proceed to topic modeling and then sentiment analysis. Finally, we present the interpretation of the results. Our process is detailed in Fig. (**1**).

2.1. Data Source

In this study, we explored the experiences and sentiments of individuals in Germany during the COVID-19 pandemic by using comprehensive text analyses to examine the qualitative data collected in German from a longitudinal interview study conducted as part of the larger “Viral Communication” by Herbig, *et al*. [33] from interviews with 40 participants who were carefully selected from a nationally representative survey based on gender, age, and socioeconomic status. This full set of interview transcripts was treated as this study’s corpus with each participant, except for two who dropped out, was interviewed three times over [phase I (December 2020, 40 interviews), phase II (April 2021, 38 interviews), and phase III (September 2021, 38 interviews)] 10 months between December 2020 and September 2021. These semi-structured interviews were designed to delve deeper into survey responses and provide additional insights into topics and controversies surrounding the pandemic in Germany. Key focus areas included information and misinformation, trust and distrust, compliance, vaccination, and conspiracy beliefs [33]. The database is publicly available [33] and was downloaded in August 2024. This study used publicly available, de-identified data; hence, no Institutional Review Board approval or informed consent was needed. The dataset can be downloaded directly from https://zenodo.org/records/6673833.

2.2. Preprocessing

We translated the interviews from German to English using the Microsoft Word language translation tool, then leveraged the natural language toolkit to implement various data preprocessing techniques in Python (including figure production), ensuring an effective and unbiased analytical process. Translation from German to English was necessary to enhance preprocessing, as certain NLP tools are more reliable in English, and Microsoft Word with the CoPilot translation tool was picked mainly for its accessibility. Furthermore, the investigative team is not fluent in German, so translation was necessary to confirm results. These preprocessing steps included punctuation removal, tokenization, stop word removal, stemming, and lemmatization. We proceeded with several advanced text analysis methods after completing the data preprocessing. These included TF-IDF vectorization using TfidfVectorizer, topic modeling/LSA, word cloud generation, and SA. These techniques allowed us to gain deeper insights and uncover significant patterns within the dataset.

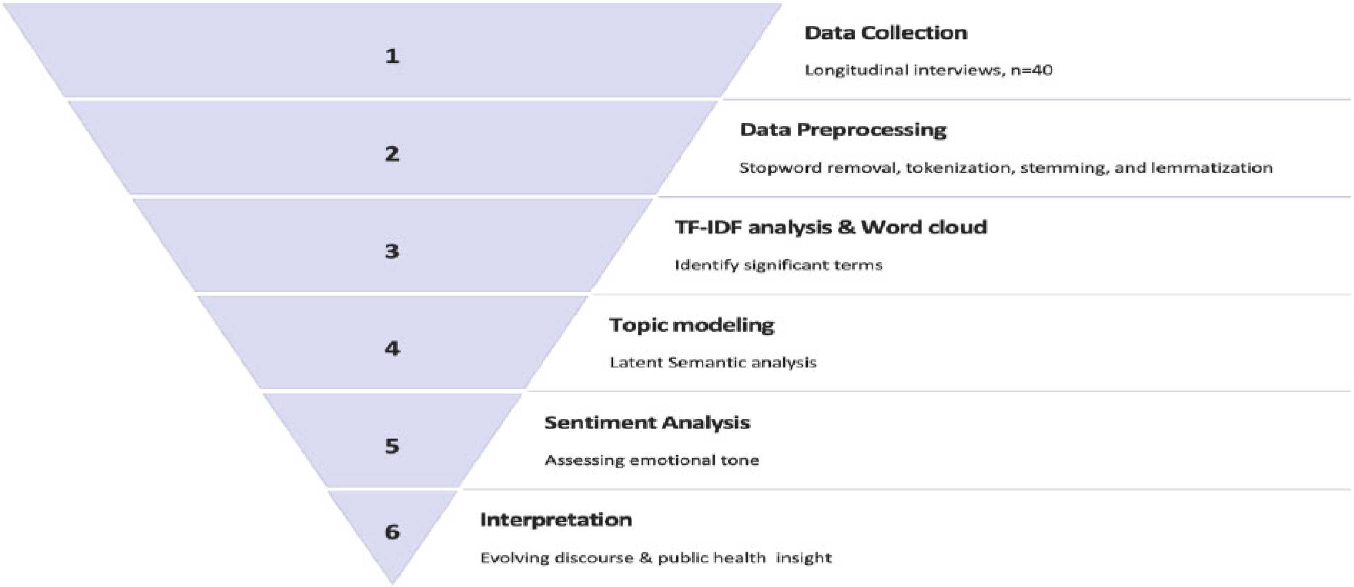
1
Data Collection
Longitudinal interviews, n=40
2
Data Preprocessing
Stopword removal, tokenization, stemming, and lemmatization
3
TF-IDF analysis & Word cloud
Identify significant terms
4
Topic modeling
Latent Semantic analysis
5
Sentiment Analysis
Assessing emotional tone
6
Interpretation
Evolving discourse & public health insight

Fig. (1). Study design flowchart.

2.3. TF-IDF

Machine learning algorithms typically require numerical data to function. Therefore, when working with textual data or any NLP tasks, it is necessary to transform the text into numerical data through a process called vectorization [34]. Hence, after cleaning the dataset, we began our text analysis by performing TF-IDF vectorization. We imported the “scikit-learn” Python library and utilized “sklearn.feature_extraction.text.Tfidf Vectorizer” to transform the collection of raw documents into a matrix of TF-IDF features. TF-IDF, a statistical measure used in text mining and information retrieval, was needed to evaluate the importance of specific words identified. As Schütze, *et al*. [35] mentioned, the central idea behind TF-IDF was to weigh the frequency of a word in a document against its frequency in the entire corpus (a collection of documents), thereby highlighting words significant to a specific document while diminishing the weight of commonly occurring words that were less informative. We identified the TF, the number of times the word appeared in the document, and IDF, the importance of the word across the corpus. The words were ranked based on the TF-IDF score (TF*IDF), which increased with the number of times a word appeared in a document (TF) but was offset by the frequency of the word in the entire corpus (IDF).

2.4. Word Cloud

We used the word cloud, also known as tag cloud, to visually represent textual data produced in Python software, where the importance of each work was indicated by its size or color. This visualization technique was chosen because it can highlight the most significant terms in a body of text, providing a quick and intuitive understanding of the main themes and concepts [36]. The more frequently a word appeared in the text, the larger or more prominent it was displayed in the word cloud.

2.5. Topic Modeling

We used topic modeling, a machine learning technique, to uncover the hidden thematic structure within a large collection of documents. Topic modeling helped us organize, understand, and summarize this large textual dataset, as well as identify the main topics. Additionally, it provided insights into the observed patterns and relationships [37]. The decision regarding how many topics to specify for each interview phase was vital; therefore, we employed domain knowledge and a series of tests (trial and error). This process identified the topics that provided optimal results and those that yielded distinct and interpretable topics.

2.6. Latent Semantic Analysis (LSA)

We applied LSA, a natural language processing technique using Scikit-learn’s TruncatedSVD in conjunction with a TF-IDF matrix. It was used to analyze relationships between a set of documents and the terms they contain, to uncover the underlying structure in our textual data. LSA transformed textual data into a mathematical space where the similarities and differences between words and documents were analyzed more effectively through a process that involves TF-IDF matrix creation, singular value decomposition, and dimensionality reduction [35].

2.7. Sentiment analysis (SA)

To gauge the tone, opinions, and emotions conveyed in the interviews, we conducted SA using Hugging Face transformers, leveraging their pre-trained pipeline model. This process, often referred to as opinion mining, was used to ascertain the emotional tone of the text. Based on SA, the text was categorized into positive, negative, or neutral sentiments [38].

2.8. ChatGPT

With the advancement in large language models, many now have the ability to perform various levels of text analytics. Based on this, this work incorporates ChatGPT as a mechanism to extract keywords from text to augment the results from classic TF-IDF.

Our choice of methods is based on layers of simple but foundational techniques that allow for vectorizing text to identify important terms in a document, uncover underlying structures in the data and reveal shifts in the psychological response of the public. Together, our methods provide a robust framework of both the emotional and informed experiences of participants.

3. RESULTS

3.1. TF-IDF

Based on the highest score of TF-IDF, we selected 10 words as the most significant or relevant captured in the interview. Table 1 highlights the most significant words for the interviews based on 3 phases. “people” (0.417468, 0.374612, 0.365573) achieved the highest score in all the phases. “mask” (0.222576, phase I), “vaccination” (0.201241, phase II), and “vaccinated” (0.281210, phase III) were the second words with the highest scores in their corresponding phases.

In Table 2, ChatGPT selected some words (“people,” “vaccination,” “trust,” and “pandemic”) for our TF-IDF analysis, meaning these words were common concerns across participants. TF-IDF provided information on words or expressions across interview phases like “wear,” “laughs,” and “question,” with term scores to indicate frequency. At a broader level, ChatGPT declared additional themes, such as “government,” “restrictions,” and “health”, without any metric.

3.2. Word Cloud

(Fig. 2, 3, and 4) visually represent the most frequent words from the three interview phases using word clouds. The word “people” appeared in all phases as the largest and most prominent in the word cloud. In phase I, the other most frequent words were “time,” “bit,” “simply,” “laughs,” “yes,” “lot,” “work,” “corona,” and “um.” In phase II, the other most frequent words were “laugh,” “bit,” “time,” “vaccination,” “vaccinated,” “simply,”” work,” “lot,” and “vaccine”/” difficult.” As well, the other most frequent words in phase III were “time,” “vaccinated,” “laugh,” “bit,” “vaccination,” “yes,” “year,” “work,” and “good.”

Table 1. Top 10 words by interview phases selected by TF-IDF.

Phase I		Phase II		Phase III	
Term	TFIDF Score	Term	TFIDF Score	Term	TFIDF Score
people	0.417468	people	0.374612	people	0.365573
mask	0.222576	vaccination	0.201241	vaccinated	0.281210
time	0.137680	vaccinated	0.196536	vaccination	0.213792
wear	0.118117	laughs	0.173009	pandemic	0.208745
trust	0.113687	pandemic	0.158532	time	0.205500
vaccination	0.111842	mask	0.151655	changed	0.153224
simply	0.110365	time	0.135729	laughs	0.145653
vaccinated	0.108520	trust	0.122337	question	0.138803
die	0.104090	masks	0.113651	life	0.112845
corona	0.098923	question	0.111117	trust	0.099145

Table 2. Top 10 words by interview phases selected by ChatGPT.

Phase I	Phase II	Phase III
Term	Term	Term
Pandemic	Vaccination	Vaccination
Challenges	Lockdown	Corona
Health	Government	Masks
Trust	Restrictions	Government
Government	Trust	Restrictions
Vaccination	Virus	Trust
Mask	Masks	Virus
Social	Testing	Pandemic
Restrictions	Corona	Testing
Freedom	Pandemic	Health

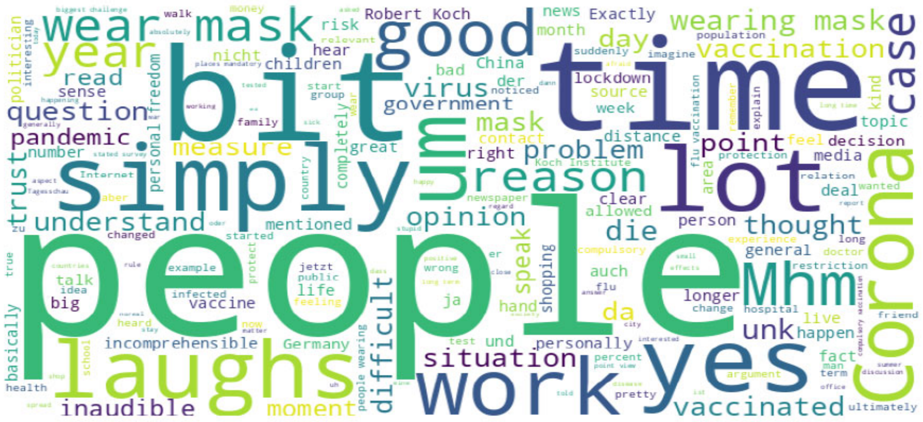

Fig. (2). Interview phase I – most frequent words.

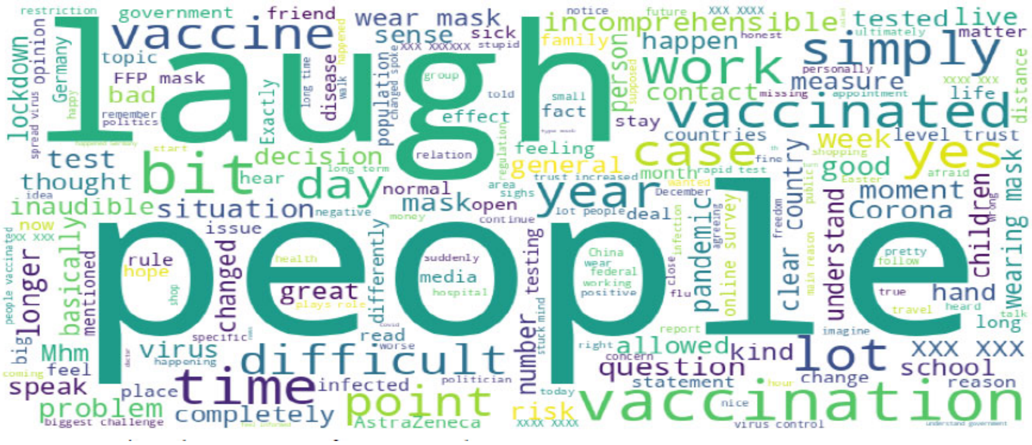

Fig. (3). Interview phase II – most frequent words.

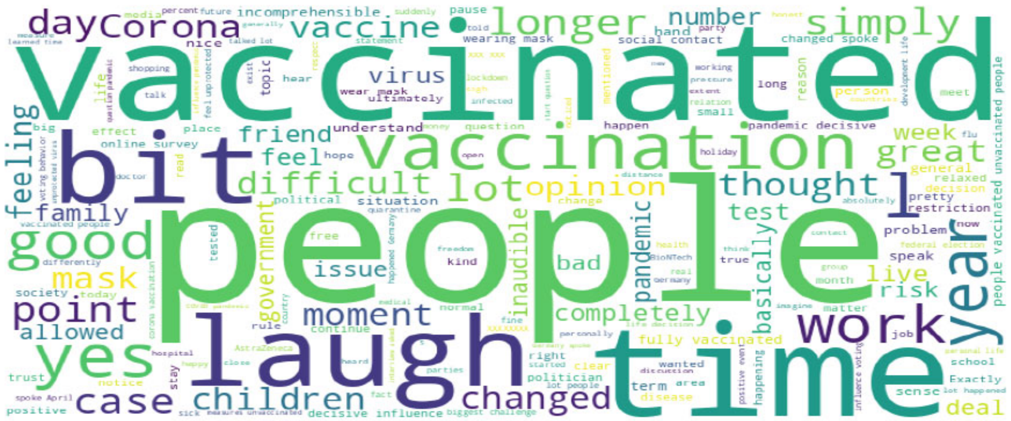

Fig. (4). Interview phase III – most frequent words.

3.3. Latent Semantic Analysis (LSA)

LSA uncovered distinct themes within the interview transcripts, providing insights into the primary subjects of discussion. Key themes included discussions about the personal experience during the COVID-19 pandemic, vaccination, trust in government and preventive measures, and public sentiment. Below are presented the main topics by interview phases (including terms weight and interpretations) to better understand the content of the conversations or text data analyzed.

Phase I

Topic 0

- Key Terms: “people” (0.031), “time” (0.014), “bit” (0.011), “lot” (0.010)
- Interpretation: This topic likely revolved around general discussions involving people and their experiences or activities over time during the COVID-19 pandemic. The terms “bit” and “lot” may be related to conversations about quantities or degrees of personal experiences during the COVID-19 pandemic.

Topic 1

- Key Terms: “mask” (0.078), “wear” (0.038), “people” (0.031), “wearing” (0.029)
- Interpretation: This topic was likely focused on the use of masks, including discussions about people wearing masks or related to health guidelines and safety measures during the COVID-19 pandemic.

Topic 2

- Key Terms: “man” (0.006), “war” (0.004), “okay” (0.003), “halt” (0.002)
- Interpretation: This topic appears to reflect personal or societal struggles during the COVID-19 pandemic. The term “man” could refer to individuals or broader discussions about the human condition during the crisis. “War” suggests a metaphorical reference to the fight against the pandemic, which was often framed as a battle.

Topic 3

- Key Terms: “yes” (0.019), “mhm” (0.012), “trust” (0.012), “question” (0.011)
- Interpretation: This topic may have involved conversational elements with affirmative responses (“yes,” “mhm”) and themes around trust and questioning. This could have been part of a dialogue or interview where trust is substantial.

Topic 4

- Key Terms: “trust” (0.029), “coronavirus” (0.014), “situation” (0.014), “china” (0.012)
- Interpretation: This topic addresses trust issues in the context of the coronavirus situation, potentially discussing the origins of the virus or the situation in China.

Phase II

Topic 0

- Key Terms: “people” (0.034), “vaccinated” (0.014), “work” (0.008), “risk” (0.008)
- Interpretation: The emphasis on “people” suggests that this topic concerns the broader population, while the terms “vaccinated” and “risk” indicate a focus on immunization and associated safety concerns.

Topic 1

- Key Terms: “people” (0.034), “time” (0.011), “lot” (0.010), “vaccinated” (0.009)
- Interpretation: This topic may have involved discussions about people and time, possibly related to vaccination efforts and their impact on the population.

Topic 2

- Key Terms: “laughs” (0.029), “pandemic” (0.019), “question” (0.019), “bit” (0.012)
- Interpretation: This topic may have included conversational elements with laughter and discussions about the pandemic, likely addressing questions and experiences during this period.

Topic 3

- Key Terms: “trust” (0.037), “government” (0.028), “vaccinated” (0.025), “vaccination” (0.023)
- Interpretation: This topic may have focused on trust in the government and vaccination efforts, reflecting public sentiment and opinions on these issues.

Topic 4

- Key Terms: “survey” (0.018), “coronavirus” (0.017), “opinion” (0.017), “situation” (0.014)
- Interpretation: This topic may have involved questions and opinions about the coronavirus situation, likely gathering public views and attitudes.

Phase III

Topic 0

- Key Terms: “vaccinated” (0.026), “yes” (0.025), “people” (0.022), “question” (0.016)
- Interpretation: This topic may have included discussions about vaccination, affirmative responses (“yes”), and general questions about people's experiences or opinions.

Topic 1

- Key Terms: "people" (0.039), "time" (0.022), "lot" (0.010), "longer" (0.009)
- Interpretation: This topic may have revolved around discussions about people and time, possibly related to long-term impacts or experiences.

Topic 2

- Key Terms: "trust" (0.031), "measures" (0.019), "survey" (0.018), "pandemic" (0.012)
- Interpretation: This topic may have addressed trust, pandemic measures, and interview results, reflecting public sentiment on these issues.

Topic 3

- Key Terms: "vaccination" (0.052), "vaccinated" (0.045), "changed" (0.038), "test" (0.012)
- Interpretation: This topic focuses on vaccination, changes brought by vaccination efforts, and testing, highlighting the dynamic aspects of the pandemic response.

Topic 4

- Key Terms: "laughs" (0.024), "situation" (0.019), "survey" (0.015), "pandemic" (0.013)
- Interpretation: This topic may have involved conversational elements with laughter, discussions about the pandemic situation, and survey results, reflecting public sentiment and experiences.

3.4. Sentiment Analysis (SA)

A sentiment score usually ranges between -1 and 1, with 1 indicating a strong positive sentiment in a task and -1 indicating a strong negative sentiment. In certain cases, there can be a score of 0, indicating a neutral sentiment or no emotional tone. (Figs. **5** and **6**) illustrate Phase 1 sentiment distribution and scores. (Figs. **7** and **8**) represent Phase 2, and (Figs. **9** and **10**) illustrate Phase 3 sentiment distribution and scores. The figures represent the proportion of positive and negative sentences in the dataset that help to understand the overall sentiment distribution. Throughout all interview phases, the sentiments were negative (> 60%). The sentiment analysis indicated models were generally confident in their predictions, especially for negative sentiments.

4. DISCUSSION

This analysis of interview data collected during the COVID-19 pandemic (December 2020- September 2021) in Germany provided valuable insights into participants' underlying patterns, behaviors, and sentiments. By employing various text processing and analysis techniques, including TF-IDF, LSA, and SA, we were able to uncover key themes and emotional tones present in the conversations. The TF-IDF analysis revealed terms such as "people," "mask," "vaccination," and "vaccinated" were highly relevant, indicating their central role in their discussions. This finding suggested conversations often revolved around people's experiences and actions during the pandemic as it relates to interpersonal interactions or the lack thereof. These findings were similar to other studies that documented the pandemic's impact on individual well-being, particularly examining isolation and its mental health impact [30, 39, 40]. Our findings also revealed the nuanced insights, concerns, experiences, and general sentiments of interviewees, highlighting the multifaceted nature of the human experience as it evolved during this global crisis, similarly observed in other research studies [28, 41, 42]. Finally, this research showed how AI tools can streamline the review and analysis of complex interview data, particularly on health-related topics such as the COVID-19 pandemic [43].

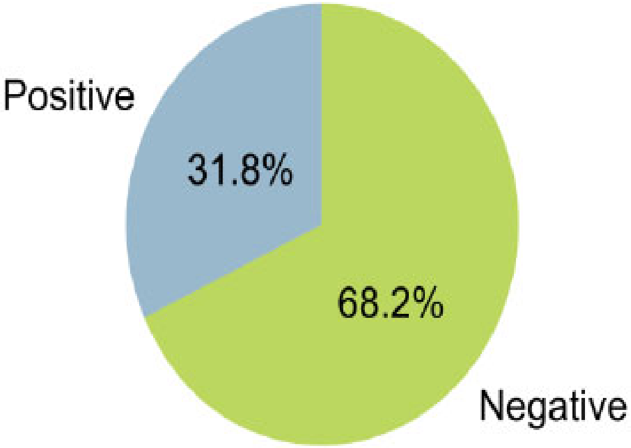

Fig. (5). Phase I – Sentiment distribution.

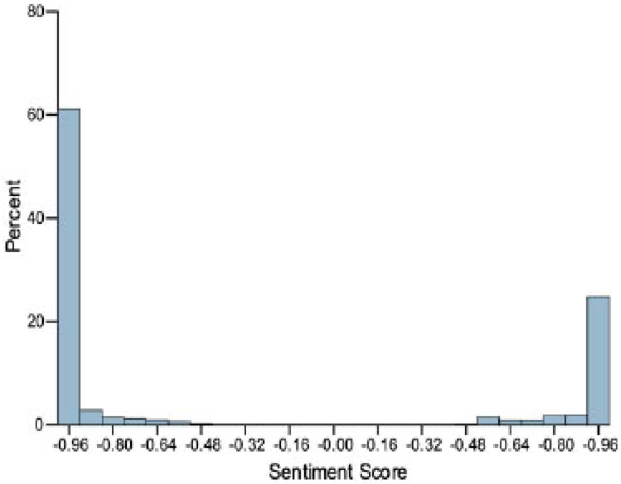

Fig. (6). Phase I – Sentiment score distribution.

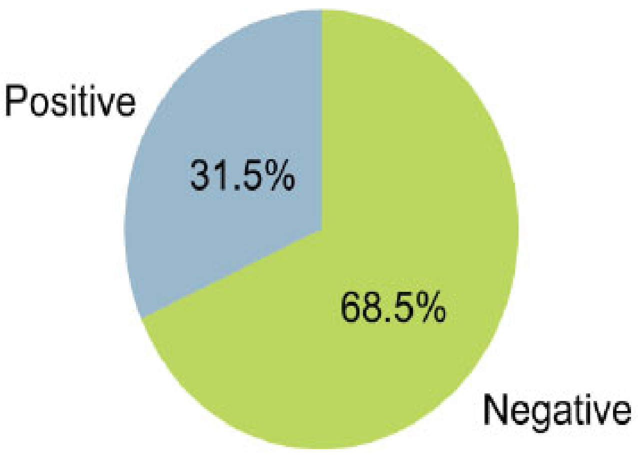

Fig. (7). Phase II – Sentiment distribution.

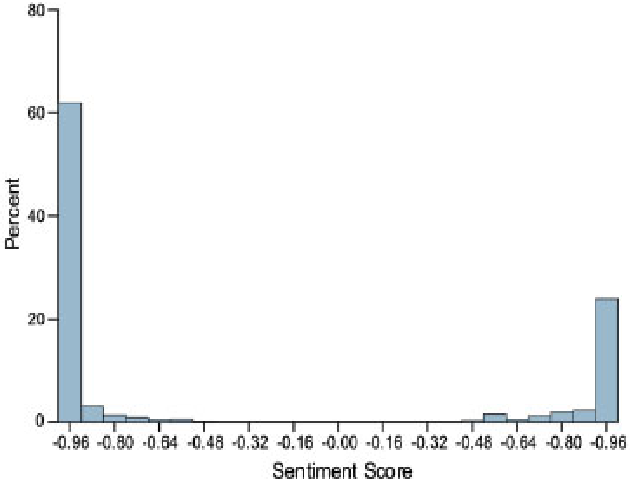

Fig. (8). Phase II – Sentiment score distribution.

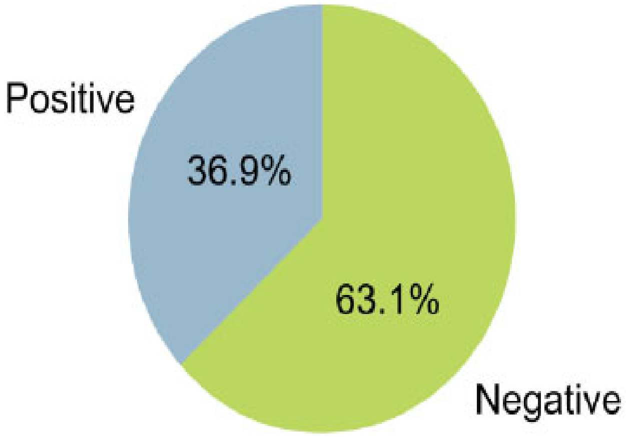

Fig. (9). Phase III – Sentiment distribution.

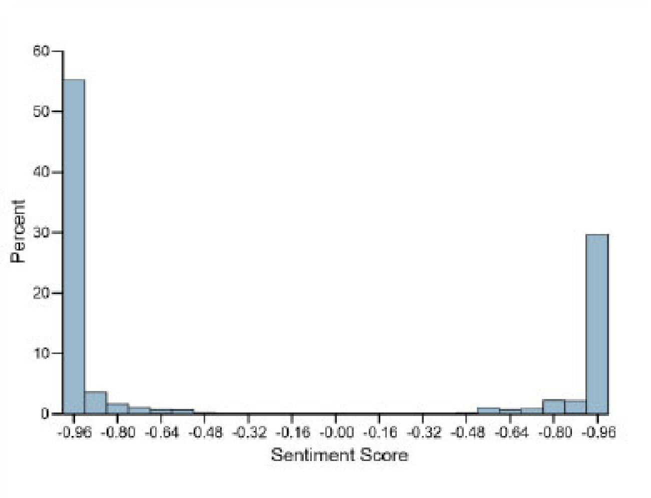

Fig. (10). Phase III – Sentiment score distribution.

LSA illuminated the thematic structure of the interviews by identifying key topics across different sessions. For instance, the phase I interview topics ranged from general discussions about people and their personal experiences over time to more specific conversations about mask-wearing and trust issues related to the coronavirus situation. Phase II interview topics included sensitive discussions with redacted information, as well as themes around vaccination efforts, government trust, and public opinions on the pandemic. Phase III interview topics highlighted ongoing discussions about vaccination status, the long-term impacts of the pandemic, and public trust in health measures and surveys. The issue of trust was a recurring theme with both skepticism toward the government and certain health protocols, particularly those related to vaccination. Some trusted the government, and some doubted the fruitfulness of pandemic containment. Masks, vaccines, and how people might adjust over time were all common topics. Those struggles (often metaphorically described as a "war") still rankled, and people looked to humor for coping strategies. Public attitudes about time and lifestyle changes underscored the continued turmoil wrought by the pandemic [44-47].

A study by Han and Wang [48] also explored semantic evolution during the COVID-19 pandemic among social media users, finding a constant change of topics throughout the pandemic, marked by the ultimate narrowing of a few topics as time progressed. In our research, SA showed a predominance of negative sentiments over positive ones, reflecting the emotional toll and widespread impact of the pandemic on individuals' lives. This finding underscored the challenges and adversities faced by the participants during the COVID-19 period. The evolution of discussion topics and sentiment through the pandemic reflected the changing nature of the health crisis itself, moving from the immediate impacts of an individual's change in their social life (*i.e.*, lockdown, isolation, and quarantine) to broader themes such as public health communication.

By combining these analytical approaches, we have gained a multi-dimensional perspective on the COVID-19 pandemic's impact on selected individuals in Germany. The insights into key terms, thematic relationships, and emotional responses provided a comprehensive understanding of the participants' experiences. These findings can inform future research, public health strategies, and policymaking to better address the needs and concerns of the population in similar crises. For example, if public health practitioners can quickly identify the salient themes arising in large amounts of interview data, health messaging, and social marketing campaigns can be tailored to address the immediate needs of the population. Further, these techniques can be used to highlight and design interventions to decrease the health disparities seen in the hard-to-reach or vulnerable sub-populations observed both during public health emergencies and in non-pandemic times [49]. Additionally, emotional insights gained through sentiment analysis underscore the necessity of holistic approaches to health crises. While the immediate negative sentiment was likely related to the virus itself or the individual’s decision to vaccinate, the continued negative sentiment reflected the mental health toll of the prolonged pandemic response. This information would be particularly useful to public health practitioners and social marketers as they are helping communities navigate the long-term effects of a global pandemic through targeted interventions [50].

5. LIMITATIONS

This research demonstrated the power of modern NLP techniques in analyzing unstructured data, offering a foundation for future research and informing public health strategies and policymaking efforts. While this study offered valuable insights, it is important to note some of the limitations. The data used in this analysis were derived from interview transcripts translated from German to English from a diverse range of participants across several sociodemographic backgrounds. Nevertheless, all participants were in Germany during the interview time periods; thus, care should be taken when determining the generalizability of the results to other populations. Additionally, the interview data came from forty participants, so nuanced findings should be explored further with a higher-powered dataset used to research specific associations. Human emotions and experiences are extremely complex, and NLP techniques are constantly being refined to better understand the interactions and perceptions between individual experiences and global events [51]. As additional NLP techniques are developed and implemented, additional findings may provide further insight into human responses to pandemics. Despite these limitations, this qualitative data provided robust material for analysis to better understand how people coped with the COVID-19 pandemic over time.

CONCLUSIONS

Overall, this study highlighted the importance of understanding public sentiment and thematic concerns during public health emergencies, offering valuable lessons for tracking and managing such events in the future. The detailed examination of the interview data shed light on the critical issues and sentiments that shaped people's experiences during the COVID-19 pandemic. They also suggest that health responders should place emphasis on effective communication and emotional support strategies that reflect the needs of the public. Future research should consider how to integrate quantitative survey data with qualitative interview data to complement the findings derived from each method, whilst enhancing generalizability and interpretation by exploring divergent cases. By leveraging AI-driven methodologies, we can enhance our understanding of health issues and improve response to public health events, especially as it relates to vaccination uptake and health communication.

AUTHORS' CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: study conception and design: H.W., J.S.; Writing the Paper: C.I.; Writing - Reviewing and Editing: J.T. All authors reviewed the results and approved the final version of the manuscript.

LIST OF ABBREVIATIONS

AI	=	Artificial intelligence
COVID-19	=	The Coronavirus Disease of 2019

ETHICAL APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not Applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIAL

All data generated or analyzed during this study are included in this published article.

FUNDING

None.

CONFLICT OF INTEREST

The author(s) declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We thank Deborah Harvey for her assistance with the manuscript figures.

REFERENCES

[1] Mahadevkar SV, Patil S, Kotecha K, Soong LW, Choudhury T. Exploring AI-driven approaches for unstructured document analysis and future horizons. J Big Data 2024; 11(1): 92. http://dx.doi.org/10.1186/s40537-024-00948-z

[2] Schmidt L, Mohamed S, Meader N, Bacardit J, Craig D. Automated data analysis of unstructured grey literature in health research: A mapping review. Res Synth Methods 2024; 15(2):

178-97.
http://dx.doi.org/10.1002/jrsm.1692 PMID: 38115736

[3] Awotunde JB, Adeniyi EA, Kolawole PO, Ogundokun RO. Application of big data in COVID-19 epidemic. Data Science for COVID-19. Elsevier 2022; pp. 141-65.
http://dx.doi.org/10.1016/B978-0-323-90769-9.00023-2

[4] Chang Z, Zhan Z, Zhao Z, *et al.* Application of artificial intelligence in COVID-19 medical area: A systematic review. J Thorac Dis 2021; 13(12): 7034-53.
http://dx.doi.org/10.21037/jtd-21-747 PMID: 35070385

[5] Bilal M, Hamza A, Malik N. NLP for analyzing electronic health records and clinical notes in cancer research: A review. J Pain Symptom Manage 2025; 69(5): e374-94.
http://dx.doi.org/10.1016/j.jpainsymman.2025.01.019 PMID: 39894080

[6] Wang Z, Ma Y, Song Y, Huang Y, Liang G, Zhong X. The utilization of natural language processing for analyzing social media data in nursing research: A scoping review. J Nurs Manag 2024; 2024(1)2857497
http://dx.doi.org/10.1155/jonm/2857497 PMID: 40224767

[7] Eguia H, Sánchez-Bocanegra CL, Vinciarelli F, Alvarez-Lopez F, Saigí-Rubió F. Clinical decision support and natural language processing in medicine: Systematic literature review. J Med Internet Res 2024; 26e55315
http://dx.doi.org/10.2196/55315 PMID: 39348889

[8] Wang Y, Xu W. Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. Decis Support Syst 2018; 105: 87-95.
http://dx.doi.org/10.1016/j.dss.2017.11.001

[9] Ibrahim NF, Wang X. A text analytics approach for online retailing service improvement: Evidence from Twitter. Decis Support Syst 2019; 121: 37-50.
http://dx.doi.org/10.1016/j.dss.2019.03.002

[10] Jung Y, Suh Y. Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. Decis Support Syst 2019; 123: 113074.
http://dx.doi.org/10.1016/j.dss.2019.113074

[11] Jeong B, Yoon J, Lee JM. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. Int J Inf Manage 2019; 48: 280-90.
http://dx.doi.org/10.1016/j.ijinfomgt.2017.09.009

[12] Kowsik V V S, Kishore A, S R, Jose AC, v DM. Sentiment analysis of twitter data to detect and predict political leniency using natural language processing. J Intell Inf Syst 2024; 62(3): 765-85.
http://dx.doi.org/10.1007/s10844-024-00842-3

[13] Kosar M, Lee F. Using topic modeling to identify factors influencing job satisfaction in the it industry. J Inf Syst Appl Res 2024; 17(1): 4-20.
http://dx.doi.org/10.62273/TVTU9122

[14] Narock T, Wimmer H. Linked data scientometrics in semantic e-Science. Comput Geosci 2017; 100: 87-93.
http://dx.doi.org/10.1016/j.cageo.2016.12.008

[15] Omakwu S, Wimmer H, Rebman C. Using textual analytics to process information overload of cyber security subreddits. J Inf Syst Appl Res 2024; 17(1): 64-74.
http://dx.doi.org/10.62273/AJJR5232

[16] Raza S, Schwartz B. Entity and relation extraction from clinical case reports of COVID-19: A natural language processing approach. BMC Med Inform Decis Mak 2023; 23(1): 20.
http://dx.doi.org/10.1186/s12911-023-02117-3 PMID: 36703154

[17] Mermin-Bunnell K, Zhu Y, Hornback A, *et al.* Use of natural language processing of patient-initiated electronic health record messages to identify patients with COVID-19 infection. JAMA Netw Open 2023; 6(7): e2322299-.
http://dx.doi.org/10.1001/jamanetworkopen.2023.22299 PMID: 37418261

[18] Guo Y, Zhang Y, Lyu T, *et al.* The application of artificial intelligence and data integration in COVID-19 studies: a scoping review. J Am Med Inform Assoc 2021; 28(9): 2050-67.
http://dx.doi.org/10.1093/jamia/ocab098 PMID: 34151987

[19] Shorten C, Khoshgoftaar T M, Furht B. Deep learning applications for COVID-19. J Big Data 2021; 8(1): 18.
http://dx.doi.org/10.1186/s40537-020-00392-9

[20] Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals 2020; 139: 110059.
http://dx.doi.org/10.1016/j.chaos.2020.110059 PMID: 32834612

[21] Islam MM, Karray F, Alhajj R, Zeng J. A review on deep learning techniques for the diagnosis of novel coronavirus (COVID-19). IEEE Access 2021; 9: 30551-72.
http://dx.doi.org/10.1109/ACCESS.2021.3058537 PMID: 34976571

[22] De Felice F, Polimeni A. Coronavirus disease (COVID-19): A machine learning bibliometric analysis. In Vivo 2020; 34(3): 1613-7.
http://dx.doi.org/10.21873/invivo.11951

[23] Alzubaidi M, Zubaydi HD, Bin-Salem AA, Abd-Alrazaq AA, Ahmed A, Househ M. Role of deep learning in early detection of COVID-19: Scoping review. Comput Methods Programs Biomed Update 2021; 1100025
http://dx.doi.org/10.1016/j.cmpbup.2021.100025 PMID: 34345877

[24] Riedel P, von Schwerin R, Schaudt D, Hafner A, Späte C. ResNetFed: Federated deep learning architecture for privacy-preserving pneumonia detection from COVID-19 chest radiographs. J Healthc Inform Res 2023; 7(2): 203-24.
http://dx.doi.org/10.1007/s41666-023-00132-7

[25] Kwon J, Grady C, Feliciano J T, Fodeh S J. Defining facets of social distancing during the COVID-19 pandemic: Twitter analysis. J Biomed Inform 2020; 111: 103601.
http://dx.doi.org/10.1016/j.jbi.2020.103601 PMID: 33065264

[26] Sanders AC, White RC, Severson LS, *et al.* Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. AMIA Jt Summits Transl Sci Proc 2021; 2021: 555-64.
PMID: 34457171

[27] He L, He C, Reynolds TL, *et al.* Why do people oppose mask wearing? A comprehensive analysis of U.S. tweets during the COVID-19 pandemic. J Am Med Inform Assoc 2021; 28(7): 1564-73.
http://dx.doi.org/10.1093/jamia/ocab047 PMID: 33690794

[28] Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis. J Med Internet Res 2021; 23(2)e25431
http://dx.doi.org/10.2196/25431 PMID: 33497352

[29] Cotfas LA, Delcea C, Roxin I, Ioanăş C, Gherai DS, Tajariol F. The longest month: Analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. IEEE Access 2021; 9: 33203-23.
http://dx.doi.org/10.1109/ACCESS.2021.3059821 PMID: 34786309

[30] Oyebode O, Ndulue C, Mulchandani D, *et al.* COVID-19 pandemic: Identifying key issues using social media and natural language processing. J Healthc Inform Res 2022; 6(2): 174-207.
http://dx.doi.org/10.1007/s41666-021-00111-w PMID: 35194569

[31] Daluwatte C, Khromava A, Chen Y, *et al.* Application of a language model tool for COVID-19 vaccine adverse event monitoring using web and social media content: Algorithm development and validation study. JMIR Infodemiology 2024; 4e53424
http://dx.doi.org/10.2196/53424 PMID: 39705077

[32] Holmes A, Sachar AS, Chang YP. Perceived impact of COVID -19 in an underserved community: A natural language processing approach. J Adv Nurs 2025; 81(6): 3201-12.
http://dx.doi.org/10.1111/jan.16522 PMID: 39373025

[33] Herbig L, Wagoner B, Watzlawik M, Jensen EA, Lorenz L, Pfleger A. Trajectories of experience through the pandemic: A qualitative longitudinal dataset. Front Polit Sci 2022; 4791494
http://dx.doi.org/10.3389/fpos.2022.791494

[34] Kumar V, Subba B. A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus. National Conference on Communications (NCC). Kharagpur, India, 2020, pp. 1-6 http://dx.doi.org/10.1109/NCC48643.2020.9056085.

[35] Schütze H, Manning CD, Raghavan P. Introduction to information retrieval. Cambridge University Press Cambridge 2008.

[36] Heimerl F, Lohmann S, Lange S, Ertl T. Word cloud explorer: Text analytics based on word clouds. 47th Hawaii International Conference on System Sciences (HICSS). Waikoloa, HI, USA, January 6-9, 2014, pp. 1833-1842 http://dx.doi.org/10.1109/HICSS.2014.231

[37] Blei DM, Ng AY, Jordan MI. Latent dirichl*et al*location. J Mach Learn Res 2003; 3(Jan): 993-1022.

[38] Sharma NA, Ali AS, Kabir MA. A review of sentiment analysis: Tasks, applications, and deep learning techniques. Int J Data Sci Anal 2024; 2: 1-38.

[39] Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. J Med Internet Res 2020; 22(10)e22635 http://dx.doi.org/10.2196/22635 PMID: 32936777

[40] Liu Y, Whitfield C, Zhang T, Hauser A, Reynolds T, Anwar M. Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. Health Inf Sci Syst 2021; 9(1): 25. http://dx.doi.org/10.1007/s13755-021-00158-4 PMID: 34188896

[41] Chang CH, Monselise M, Yang CC. What are people concerned about during the pandemic? Detecting evolving topics about COVID-19 from Twitter. J Healthc Inform Res 2021; 5(1): 70-97. http://dx.doi.org/10.1007/s41666-020-00083-3 PMID: 33490856

[42] Nemes L, Kiss A. Social media sentiment analysis based on COVID-19. J Inform Telecommun 2021; 5(1): 1-15. http://dx.doi.org/10.1080/24751839.2020.1790793

[43] Hitch D. Artificial intelligence augmented qualitative analysis: the way of the future? Qual Health Res 2024; 34(7): 595-606. http://dx.doi.org/10.1177/10497323231217392 PMID: 38064244

[44] Ruiz-Roso MB, de Carvalho Padilha P, Mantilla-Escalante DC, *et al.* Covid-19 confinement and changes of adolescent's dietary trends in Italy, Spain, Chile, Colombia and Brazil. Nutrients 2020; 12(6): 1807. http://dx.doi.org/10.3390/nu12061807 PMID: 32560550

[45] Gao C. General population's psychological perceptions of COVID-19: a systematic review. Psychol Res Behav Manag 2023; 16: 4995-5009. http://dx.doi.org/10.2147/PRBM.S440942 PMID: 38107446

[46] Lee YC, Wu WL, Lee CK. How COVID-19 triggers our herding behavior? Risk perception, state anxiety, and trust. Front Public Health 2021; 9587439 http://dx.doi.org/10.3389/fpubh.2021.587439 PMID: 33659231

[47] Reid JC, Brown SJ, Dmello J. COVID-19, diffuse anxiety, and public (mis) trust in government: Empirical insights and implications for crime and justice. Crim Justice Rev 2024; 49(2): 117-34. http://dx.doi.org/10.1177/07340168231190673

[48] Han X, Wang J. Modelling and analyzing the semantic evolution of social media user behaviors during disaster events: A case study of COVID-19. ISPRS Int J Geoinf 2022; 11(7): 373. http://dx.doi.org/10.3390/ijgi11070373

[49] Shah GH, Shankar P, Schwind JS, Sittaramane V. The detrimental impact of the COVID-19 crisis on health equity and social determinants of health. J Public Health Manag Pract 2020; 26(4): 317-9. http://dx.doi.org/10.1097/PHH.0000000000001200 PMID: 32433385

[50] Evans WD, Bardus M, French J. A vision of the future: Harnessing artificial intelligence for strategic social marketing. Businesses 2024; 4(2): 196-210. http://dx.doi.org/10.3390/businesses4020013

[51] Al-Garadi MA, Yang Y-C, Sarker A. The role of natural language processing during the COVID-19 pandemic: Health applications, opportunities, and challenges. Healthcare 10(11): 2270.2022; http://dx.doi.org/10.3390/healthcare10112270

DISCLAIMER: The above article has been published, as is, ahead-of-print, to provide early visibility but is not the final version. Major publication processes like copyediting, proofing, typesetting and further review are still to be done and may lead to changes in the final published version, if it is eventually published. All legal disclaimers that apply to the final published article also apply to this ahead-of-print version.