Bridging the Gap: Commonality and Differences between Online and Offline COVID-19 Data

Nayoung Kim[1], Ahmadreza Mosallanezhad[1], Lu Cheng[1,2], Baoxin Li[1], and Huan Liu[1]

[1] Arizona State University, Tempe, AZ, USA
[2] University of Illinois Chicago, Chicago, Illinois, USA
{nkim48,amosalla,lcheng35,baoxin.li,huanliu}@asu.edu

Abstract. With the onset of the COVID-19 pandemic, news outlets and social media have become central tools for disseminating and consuming information. Because of their ease of access, users seek COVID-19-related information from online social media (i.e., online news) and news outlets (i.e., offline news). Online and offline news are often connected, sharing common topics while each has unique, different topics. A gap between these two news sources can lead to misinformation propagation. For instance, according to the Guardian, most COVID-19 misinformation comes from users on social media. Without fact-checking social media news, misinformation can lead to health threats. In this paper, we focus on the novel problem of bridging the gap between online and offline data by monitoring their common and distinct topics generated over time. We employ Twitter (online) and local news (offline) data for a time span of two years. Using online matrix factorization, we analyze and study online and offline COVID-19-related data differences and commonalities. We design experiments to show how online and offline data are linked together and what trends they follow.

Keywords: Matrix Factorization · COVID-19 · Social Media Analysis

1 Introduction

The COVID-19 pandemic has been accompanied by a massive infodemic involving a variety of topics, such as vaccination, inflation, and so on. Communication of important information during emergency situations is critical [2, 17] for taking actions to contain disease. With its convenience, easy access, and large volume, news media becomes an expected means to a pandemic response. People seek and receive information from numerous sources such as newspapers, television, and increasingly, social media. For example, during the early stages of the COVID-19 pandemic, Twitter and news outlets such as Washington post reported about US states' social distancing mandates to help prevent the growth of the virus.

Due to the different nature of online social media data and offline news articles, information gaps may exist among online social media users and offline news media readers. These users and readers consist of scientists, medical and

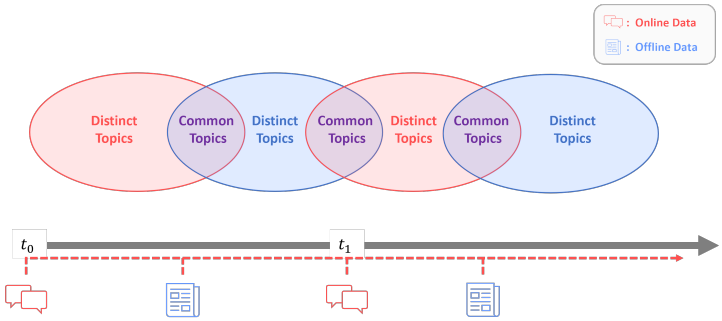


Fig. 1: Illustration of the common and distinct topics identified within two types of datasets: *offline* and *online*.

public health professionals, and the general public. A large information gap could contribute to the spread of misinformation in news media. Therefore, to better understand the relations between online social media data and offline news articles, we aim to monitor their common and distinct topics generated over time. In particular, connecting offline with online data can be beneficial in two main ways: (1) Researchers can measure public interest over time and identify new themes of COVID-19-related discussion emerging quickly through data streams. (2) By providing users with useful information about COVID-19, it reduces the chance of exposing them to misinformation, i.e., a gap between offline and online topics can alert users about possible misinformation propagation.

Figure 1 shows the problem setting. Our goal is to discover and link topics from the traditional and social media news sources, to understand their association. We study this problem on a COVID-19-related dataset [7] collected through Twitter and Google News. This problem presents the following challenges: (1) The collected dataset contains a significant amount of noisy and irrelevant data. In the online data, not all the tweets follow a specific topic or provide useful information about an actual event. Consequently, the integration approach should discard noise and irrelevant data. (2) Identifying topics across two different data sources in a separate and joint setting. The model should be able to find the association between them by leveraging the dataset properties. Typically, Twitter data are short texts, including hashtags, user mentions, and covering any topics. In contrast, news articles published on media websites are long and structured.

To address the challenges above, we use a variant of the online Non-negative matrix factorization (NMF) method, joint ONMF. NMF methods are capable of dimensionality reduction and data clustering simultaneously. Because of our data's nature and its streaming property, we cannot use vanilla NMF. Further, the online NMF method cannot model discovered topics over time. As a result, we adopt joint ONMF to find common and distinct topics over time and link online and offline data. Specifically, we overcome the challenges by making the following contributions: (1) We study the novel problem of linking data related to COVID-19 from online social media and offline news sources, (2) We use joint

ONMF to identify common and distinct topics across time, and (3) We conduct extensive experiments on real-world datasets to draw conclusions.

2 Related Work

There have been a growing number of research on social media discourse associated with disasters or crises. For example, Martin et al. [14] leveraged the spatiotemporal features of tweets to assess the responses on Hurricane Matthew, and Yeo et al. [22] analyzed Twitter communications on 2016 Southern Louisiana flood recovery process. Cheng et al. [2] studied the problem of tracking disaster events through streaming data from Twitter. They proposed an online matrix factorization-based method to identify topics across time periods.

As people share their emotions [11] and opinions [3] related to the COVID-19 pandemic on social media (e.g. Twitter, Facebook), a number of studies have also been conducted to explore the use of social media during COVID-19 epidemic. Although numerous researches have focused on online discourses [6, 19, 21], a few recent studies have tried to disclose the impact of various offline data delivered by traditional news sources. Crupi et al. [4] showed the change of topics regarding COVID-19 vaccination in Italian user communities and the polarization of vaccine stances with common and distinct topics. Another work by Poddar et al. [16] identified distinct topics in online data and investigated how vaccine-related discourse has changed between pre and during COVID times. A few studies have utilized geo-tagged data on social media to map online data to real-world events and statistics [5, 18]. Others tried to uncover the association between vaccine hesitancy and online user behavior [13, 20] or online misinformation [12, 15].

Our work is different than previous studies as we study the novel problem of linking COVID-19-related offline and online data. The aim of this problem is to examine how common and distinct topics around offline and online data change over time, and how offline events affect online data.

3 Proposed Method

In this section, we first review the basic NMF model and its online variant which is a modified matrix factorization for temporal data. Then, we discuss the main components of the proposed method in detail. Inspired by Cheng et al. [2], we employ the joint ONMF algorithm to extract the latent factors of streaming data using matrix factorization. Considering a time span of 24 months, in every one-month time period, historical data is combined with newly arriving data to identify common and distinct topics. We perform the joint ONMF on COVID-19 tweets and its related news data to extract the topics.

3.1 Non-negative Matrix Factorization (NMF)

NMF [10] is a popular non-negative matrix decomposition algorithm which is widely used for analysis of multivariate data such as latent representation of

text. Given a non-negative data matrix $\mathbf{V}$, NMF factorizes $\mathbf{V}$ into two lower rank non-negative matrices $\mathbf{W}$ and $\mathbf{H}$ as $\mathbf{V} \approx \mathbf{WH}$. Matrix $\mathbf{V} \in \mathbb{R}^{m \times n}$ indicates m data, each having n attributes. Matrix $\mathbf{H}$ is a feature matrix with each row represents a latent factor. $\mathbf{W} \in \mathbb{R}^{m \times k}$ is a coefficient matrix, reflecting the association weights between the data samples and the extracted latent factors. NMF is solved by minimizing the distance between the original matrix $\mathbf{V}$ and the reconstructed matrix $\mathbf{V}' = \mathbf{WH}$.

The NMF method assumes both latent factors and input data are static and do not change over time. Thus, we cannot directly apply this method for our problem. One simple solution to apply NMF on streaming data is to feed NMF with updated data matrix as new subset of the data joins at each time step t. However, this approach underutilizes the previous NMF results in the upcoming factorization results that is calculated for the newly collected data [1]. Cao et al. [1] proposes Online-NMF (ONMF) to decompose matrix $\mathbf{V}_{t+1}$ based on a new batch of data $\mathbf{U}$ at time step $t+1$ and the previous matrix $\mathbf{V}_t$. Formally, the ONMF problem is defined as $\mathbf{V}_{t+1} = \begin{pmatrix} \mathbf{V}_t \\ \mathbf{U} \end{pmatrix} \approx \mathbf{W}_t\mathbf{H}_t$, which is solved by minimizing the mean squared error between $\mathbf{V}$ and $\mathbf{WH}$ [1].

3.2 Linking Offline and Online Data

ONMF can efficiently update the latent factors (i.e., topics) in streaming data. However, it is hard to see the global semantic changes in the generated topics. To solve this problem, inspired by [2], we use a modified ONMF-based approach, joint ONMF, to discover commonness and distinctiveness in topics. Assuming there are k topics in the given documents, joint ONMF identifies k_c and $k_d = k - k_c$ which are defined as the number of *common topics* and *distinct topics*, respectively. To alleviate the massive computational cost proportional to the growing data matrix $\mathbf{V}$, this method uses one of the decomposed low rank matrices $\mathbf{H}_t$ to derive topics. Since $\mathbf{H}_t$ is fixed at time step t, we perform the linear transformation $\mathbf{H}^* \approx \mathbf{L}^*\mathbf{H}_t$ to dynamically tune the dependency between $\mathbf{H}_t$ and newly arrived data $\mathbf{U}$. In this transformation, ($\mathbf{L}^* \in \mathbb{R}^{k \times k}$) is a low rank matrix that is used to adjust the dependency between $\mathbf{H}_t$ and $\mathbf{U}$.

Joint ONMF aims to minimize the distance between common topics in $\mathbf{H}^*$ and $\mathbf{H}_U$ and maximize the distance between distinct topics in $\mathbf{H}^*$ and $\mathbf{H}_U$. Finally, we use the following objective function to find common and distinctive topics:

$$\min_{\mathbf{W}_U, \mathbf{H}_U, \mathbf{H}^*, \mathbf{L}^*} \frac{1}{2}\|\mathbf{H}^* - \mathbf{L}^*\mathbf{H}_t\|_F^2 + \frac{1}{2}\|\mathbf{U} - \mathbf{W}_{U_c}\mathbf{H}_{U_c} - \mathbf{W}_{U_d}\mathbf{H}_{U_d}\|_F^2 + \alpha f_c(\mathbf{H}_c^*, \mathbf{H}_{U_c}) + \beta f_d(\mathbf{H}_d^*, \mathbf{H}_{U_d}), \quad (1)$$

where the second term denotes the reconstruction error on matrix $\mathbf{U}$, which is factorized into $\mathbf{W}_U$ and $\mathbf{H}_U$ ($\mathbf{U} \approx \mathbf{W}_U\mathbf{H}_U$). f_c and f_d are the commonness score and distinctiveness score described in section 4. We leverage parameters α and β to find the optimal balance between three different objectives: the accuracy of matrix reconstruction, common, and distinct topics.

This method identifies the common and distinct topics over time within one dataset. To examine the relationship between the datasets with different sources, we apply the method on both datasets simultaneously. Specifically, we split each dataset by months as delimiters and concatenate the subsets of tweets and news pieces one after another along the timeline. (e.g., Tweets of January 2020 || News of January 2020 || Tweets of February 2020 || News of February 2020 $\cdots$). Starting with the Tweets of January 2020, the model outputs the common topics between each month (i.e., topics that appear both in t and $t+1$, considering t as month) and distinct topics (i.e., two unique topics at time before t and time $t+1$) whenever new data arrives at every time stamp t. The result distinct topics could give an insight on the main theme and statements raised within a group of people on Twitter, along with the important events recorded by the press happened in the specific time period. On the contrary, the output of common topics implies the tweets' topics are closely related to the news pieces.

4 Experiments

We conduct qualitative and quantitative evaluation on the performance of the joint ONMF model for extracting topics and gaining insights from both online and offline data. We aim to answer two research questions: (**Q1**) How well joint ONMF extracts common and distinct topics across online and offline data? (**Q2**) How does the offline data impact the online data? To answer **Q1**, we compare the performance of joint ONMF with baselines based on various evaluation metrics. To answer **Q2**, we design experiments to measure the distinctiveness of online and offline data.

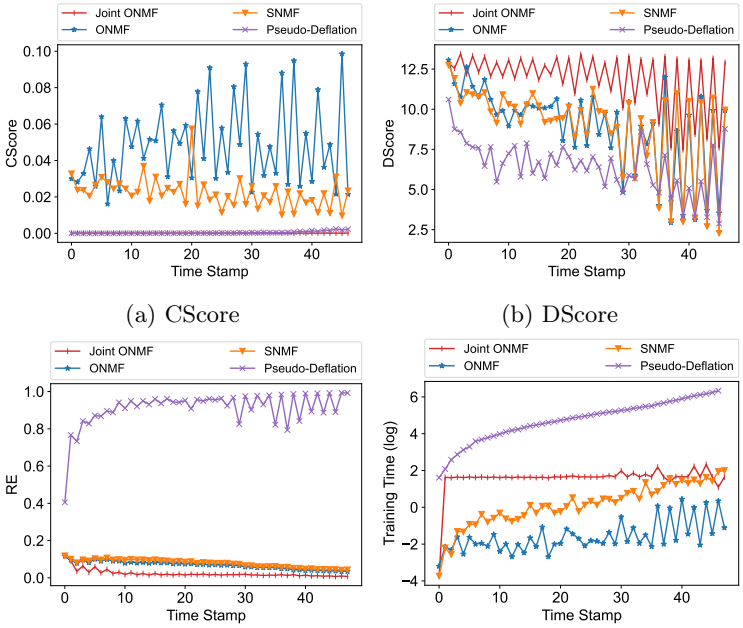
4.1 Experimental Design

We use the CoVaxNet dataset[3] that includes $4,790$ offline news articles and $1,831,220$ online tweets within a time span of two years from January 2020 till December 2021 [7]. We further use Newspaper3k[4] to extract *title*, *body*, and *summary* from news pieces. The online data is first filtered to select $20K$ representative random tweets. Finally, we extract the initial textual features using TF-IDF scores from both datasets. A large TF-IDF value indicates that the term can better distinguishes documents from each other.

To design experiments, we start with online data from January 2020 followed by offline data from the same time period. Assuming the topics of tweets vary in accordance with newly uploaded news pieces within certain period, we assign tweets and news by turns monthly. We choose $k_c = 2$ and $k_d = 3$. The optimization parameters α and β of Equation 1 are set to $1,000$ and 0.1, respectively. Finally, the step size of the joint ONMF method is set to 100.

To evaluate the effectiveness of joint ONMF, we consider the three baselines. **Standard NMF (SNMF)** [9] approach repeatably calculates the latent factors

[3] `https://github.com/jiangbohan/CoVaxNet`
[4] `https://github.com/codelucas/newspaper`



(a) CScore (b) DScore

(c) RE (d) Training Time

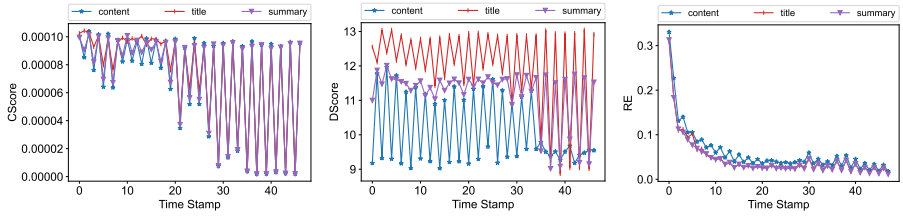
Fig. 2: Performance comparison of different methods

using the entire dataset whenever a new batch of data is received. **ONMF** [1] method is a modified variant of NMF that focuses on handling large-scale and streaming data. Finally, **Pseudo-Deflation** [8] approach uses a variant of ONMF, to identify common and distinct topics between two documents. For fair comparison between joint ONMF and the baselines, we evaluate the retrieved topics within the same time period and hyperparameters.

4.2 Experimental Results

Following Kim et al. [8], we evaluate the performance of joint ONMF using commonness score, distinctiveness score, and reconstruction error:
Commonness Score (CScore) stands for the similarity between the k common topics at time t and $t+1$ for every time stamp, calculated as $\text{CScore} = \frac{1}{k_c}\|\mathbf{H}_c^* - \mathbf{H}_{\mathbf{U}_c}\|_F^2$. A smaller CScore indicates better efficiency of the method on extracting common topics. **Difference Score (DScore)** uses average symmetric Kullback-Leibler (KL) divergence between all the distinct topic pairs. A large DScore indicates the distribution of the obtained distinct topics are discrete.

$$\text{DScore} = \frac{1}{2k_d^2}\sum_{i}^{k_d}\sum_{j}^{k_d}[h_d^{*i}\log(h_d^i)^T + h_{\mathbf{U}_d}^i\log(h_{\mathbf{U}_d}^i)^T - \\ - h_d^{*i}\log(h_d^j)^T - h_{\mathbf{U}_d}^j\log(h_{\mathbf{U}_d}^{*i})^T] \quad (2)$$



(a) CScore (b) DScore (c) RE

Fig. 3: Performance comparison with different data metrics of offline data

Reconstruction Error (RE) estimates the loss of the NMF on the newly arriving data $\mathbf{U}$ at each time stamp t based on $\min_{\mathbf{W},\mathbf{H}} \|\mathbf{V} - \mathbf{WH}\|$. Smaller RE indicates the initial data matrix $\mathbf{U}$ can be reconstructed accurately.

We answer the two questions **Q1** and **Q2** using quantitative and qualitative experiments to evaluate the quality of common and distinct topics in the integrated dataset, and study the effect of offline data on the online data.

Quantitative Results:

Performance (Q1). To calculate CScore and DScore, we directly apply them on JOINT ONMF and PSEUDO-DEFLATION as they explicitly state which topics are the common and distinct. For other baseline models we assign topics with the smallest CScore as common topics k_c, while the others are treated as distinct ones k_d. Figure 2 illustrates the evaluation results including computation time of each method. From these results, we conclude that the JOINT ONMF method outperforms ONMF and SNMF with a large gap in CScore and DScore. PSEUDO-DEFLATION has comparable CScore, however, the DScore is significantly lower than JOINT ONMF. DScores of all four methods fluctuates severely in later time periods due to the noisy and unbalanced nature of the dataset. Among the baselines, PSEUDO-DEFLATION presents remarkably smaller RE. Besides PSEUDO-DEFLATION, SNMF can often obtain the best RE because the goal of SNMF focuses on minimizing RE while other online topic models seek to balance between RE and computational efficiency. The JOINT ONMF method presents competitive RE compared to SNMF and ONMF as it separately conducts matrix factorization on data matrix $\mathbf{U}$.

Considering all three evaluation metrics and the computational cost, the JOINT ONMF approach attains the most optimal balance among the baselines. For instance, it can have the smallest CScore and the largest DScore, meanwhile achieving negligible reconstruction error similar to SNMF. Figure 2d shows that JOINT ONMF is slightly slower than two baselines yet maintains consistent training time, making it comparatively computationally efficient.

Impact (Q2). According to Figure 3, one interesting observation is that the amount of offline information inconsistently influence the performance of JOINT ONMF over each measurements. As mentioned earlier, we measure the performance on {'title', 'summary', 'body'} of each news piece, assuming that 'title'

Table 1: Examples of topics from Jan 1st to Mar 31st, 2020. D_T and D_N show distinct topics of tweet and news title, respectively. C shows the common topics of between offline and online data.

Jan 2020			**Feb 2020**			**Mar 2020**		
D_T	C	D_N	D_T	C	D_N	D_T	C	D_N
misinformation	oil	antigen	deadly	flu	Remdesivir	misinformation	patent	immunity
children	Japan	China	fund	antivaccine movement	influenza season	worry	mask wear	Donald Trump
thank	religious exemption	HPV cancer	happy	pediatric	South Korea	child killer	candidates	CDC
kill	Ebola	Donald Trump	thank	conspiracy	afford insurance	propaganda	INOVIO (Pharmaceuticals)	adjuvants
amazing	Johnson & Johnson	tuberculosis	autism	Pneumonia	peptid	prevent disease	egg-based vaccine	Summit

text has the smallest, and the 'body' text has the largest quantity of information. Results show that given more information, the model can better reconstruct original data but worse in distinguishing distinct topics.

Qualitative Study:

Performance (Q1). We perform an extensive qualitative study on the validity of the identified topics. Table 1 shows top five ranked topics discovered by the JOINT ONMF method on the integrated online and offline data during the first three time periods of the pandemic. The extracted common topics are closely related to temporal and/or newly rising events (e.g., Ebola, Remdesivir, egg-based vaccine) which are compelling and provocative as well (i.e., topics discussed online change over time and selective real-world events). Meanwhile, distinct topics consist of specific keywords relevant to COVID-19 or medical events. Note that according to Table 1, exceptionally large portion of emotional words account for the common topics of the online data.

Impact (Q2). We take a deeper analysis on the common topics to observe how the two datasets are linked. For example, in Table 2, we sample tweets and news headlines containing the topic keyword *oil*. It is worth mentioning that the users not only talk about the facts and opinions, but also they expose their strong emotions (e.g., `i will laugh, lol`) and groundless claims (e.g., `oil companies just hid their own research data`) that could falsely lead to misinformation. We could find similar characteristics in multiple other tweet samples. This implies that the offline data (i.e., local events and news) play a substantial role in the shift of online behavior and potential biases of users.

5 Conclusion and Future Work

In this paper, we study the novel problem of integrating online and offline COVID-19-related data to seek common and distinct topics across them. Finding the commonness and distinctiveness between online and offline data is an important

Table 2: Examples of different sources from common topic

Topic	Source	Examples
oil	News Articles	- Russia halts oil to Belarus, but transit to Europe still flowing (Jan 3, 2020)
	Tweets	- Lol oh yes definitely max single handedly created the need smh he is just one of the ... antivaxxers of the oil world (Jan 4, 2020) - The oil companies just hid their own research data now all the sceptics have left like flat earthers ... (Jan 5, 2020) - I will laugh at anyone who sneers at China for this given that America is a rapidly decaying shithole filled with ... climate denialism antivaxxers and oil companies that own multiple politicians (Jan 31, 2020)
trial	News Articles	- Coronavirus Vaccine: NIH Goes Straight to Human Trials (Mar 16, 2020)
	Tweets	- Let's go science NIH clinical trial of investigational vaccine for covid 19 begins ... (Mar 16, 2020) - Why i am volunteering to get the fastest startup of a vaccine trial ever ... (Mar 31, 2020)
delta	News Articles	- Pfizer says COVID vaccine is highly effective against Delta variant (Jun 24, 2021) - Moderna Says Studies Show Its Vaccine Is Effective Against The Delta Variant (Jun 30, 2021)
	Tweets	- This is what I've been saying the delta variant is growing and now there is a new lamda variant while we can't live in fear over every variant that pops up we will still need to be cautious until we learn what they do and whom they affect (July 8, 2021) - Get vaccinated or get sick but do not punish those who are vaccinated I am fed up with restrictions on us because of and citing rise of delta variant Los Angeles reports 165 percent increase in covid cases (July 9, 2021)

research problem because it provides an effective way for researchers to study how users' opinions change over time and how offline news sources affect online social media discussions. Furthermore, studying this problem paves the way for internet users to gather information on COVID-19-related topics and avoid possible misinformation effectively. To this end, we use the joint ONMF method to compute common and distinct topics efficiently. We design qualitative and quantitative experiments to measure the performance of joint ONMF and provide valuable insights into how offline data affects online data. For future work, we will study methods that can automate the process of finding interesting points of view and user attributes contributed to online discussions about offline events.

Acknowledgement

This work was supported by the Office of Naval Research under Award No. N00014-21-1-4002. Opinions, interpretations, conclusions, and recommendations are those of the authors.

References

1. Cao, B., et al.: Detect and track latent factors with online nonnegative matrix factorization. In: IJCAI (2007)
2. Cheng, L., et al.: Tracking disaster footprints with social streaming data. In: IAAI (2020)
3. Cotfas, L.A., et al.: The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. IEEE Access (2021)
4. Crupi, G., et al.: Echoes through time: Evolution of the italian covid-19 vaccination debate. In: ICWSM (2022)
5. Feng, S., Kirkley, A.: Integrating online and offline data for crisis management: Online geolocalized emotion, policy response, and local mobility during the covid crisis. Scientific Reports (2021)
6. Glandt, K., et al.: Stance detection in covid-19 tweets. In: IJCNLP (2021)
7. Jiang, B., Sheth, P., Li, B., Liu, H.: Covaxnet: An online-offline data repository for covid-19 vaccine hesitancy research. arXiv preprint arXiv:2207.01505 (2022)
8. Kim, H., et al.: Simultaneous discovery of common and discriminative topics via joint nonnegative matrix factorization. In: KDD (2015)
9. Lee, D., Seung, H.S.: Algorithms for non-negative matrix factorization. NeurIPS (2000)
10. Lee, D.D., et al.: Learning the parts of objects by non-negative matrix factorization. Nature (1999)
11. Lwin, M.O., et al.: Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. JMIR public health and surveillance (2020)
12. Lyu, H., et al.: Misinformation versus facts: Understanding the influence of news regarding covid-19 vaccines on vaccine uptake. Health Data Science (2022)
13. MacDonald, N.E., et al.: Vaccine hesitancy: Definition, scope and determinants. Vaccine (2015)
14. Martín, Y., et al.: Leveraging twitter to gauge evacuation compliance: Spatiotemporal analysis of hurricane matthew. PLoS one (2017)
15. Pierri, F., et al.: Online misinformation is linked to early covid-19 vaccination hesitancy and refusal. Scientific reports (2022)
16. Poddar, S., et al.: Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users. In: ICWSM (2022)
17. Reynolds, B., et al.: Crisis and emergency risk communication (2002)
18. Shen, C., et al.: Using reports of symptoms and diagnoses on social media to predict covid-19 case counts in mainland china: Observational infoveillance study. JMIR (2020)
19. Wicke, P., Bolognesi, M.M.: Covid-19 discourse on twitter: How the topics, sentiments, subjectivity, and figurative frames changed over time. Frontiers in Communication (2021)
20. Wilson, S.L., Wiysonge, C.: Social media and vaccine hesitancy. BMJ global health (2020)
21. Xiong, Z., et al.: Social media opinions on working from home in the united states during the covid-19 pandemic: Observational study. JMIR medical informatics (2021)
22. Yeo, J., et al.: Disaster recovery communication in the digital era: Social media and the 2016 southern louisiana flood. Risk analysis (2020)