# Misinformation Concierge: A Proof-of-Concept with Curated Twitter Dataset on COVID-19 Vaccination

Shakshi Sharma
shakshi.sharma268@gmail.com
Institute of Computer Science, University of Tartu, Tartu, Estonia

Anwitaman Datta
anwitaman@ntu.edu.sg
School of Computer Science and Engineering, Nanyang Technological University, Singapore

Vigneshwaran Shankaran
vigneshwaranpersonal@gmail.com
Department of Computer Science, University of Surrey, Guildford, UK

Rajesh Sharma
rajesh.sharma@ut.ee
Institute of Computer Science, University of Tartu, Tartu, Estonia

## ABSTRACT

We demonstrate the Misinformation Concierge, a proof-of-concept that provides actionable intelligence on misinformation prevalent in social media. Specifically, it uses language processing and machine learning tools to identify subtopics of discourse and discerns non/misleading posts; presents statistical reports for policy-makers to understand the big picture of prevalent misinformation in a timely manner; and recommends rebuttal messages for specific pieces of misinformation, identified from within the corpus of data - providing means to intervene and counter misinformation promptly. The Misinformation Concierge proof-of-concept using a curated dataset is accessible at: https://demo-frontend-uy34.onrender.com/

## CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

## KEYWORDS

misinformation, social media, visual exploratory tool, decision support system, COVID-19 vaccines, Twitter.

**ACM Reference Format:**
Shakshi Sharma, Anwitaman Datta, Vigneshwaran Shankaran, and Rajesh Sharma. 2023. Misinformation Concierge: A Proof-of-Concept with Curated Twitter Dataset on COVID-19 Vaccination. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3614746

## 1 INTRODUCTION

Misinformation is a major societal menace. Many vested interests are invested in manufacturing and propagating misinformation, including by engaging paid personnel to do so, besides co-opting 'useful idiots', wreaking havoc on societies.

Being able to monitor the kinds of misinformation gaining traction, and countering them using a dedicated team of personnel or (semi-)automated tools immediately rebutting rumors, particularly those with the propensity to persist, needs to be a part of the response and policing strategy against misinformation. The envisioned tool, **Misinformation Concierge**, is a step in that direction. Foremost, it captures the overall discourse pertaining to a broad subject of interest (for example, COVID-19 vaccines) from social media posts, e.g., Tweets. Employing language processing tools, it then identifies distinct (sub)topics and also discerns non/misleading posts, and visualizes the descriptive analysis results for users to better understand the big picture. Users could then drill down within specific topics (e.g., Pfizer vaccines) to investigate further aspects such as temporal behavior (e.g., which kind of topics are being discussed more or less over time?) as well as browse actual instances of misleading information pertaining to said topic. Finally, it allows one to zoom into individual misleading posts and explore post-specific information, e.g., which topic it belongs to, and identify other very similar misleading posts but also non-misleading posts. The latter could be used to refute the original offending post. Human users (or automated bots) could then repurpose these to promptly counter misleading posts. Such a platform would allow policymakers, as well as 'first responders' in social media space, to both (i) identify and understand the prevalent misinformation in a timely and concise manner and also (ii) exploit ready to use responses derived automatically from the social media corpus itself.

We envision Misinformation Concierge as a general purpose platform to accommodate a variety of media and social media sources spanning a wider range of subjects to provide actionable intelligence. Still, in order to keep things tractable, the initial proof-of-concept is restricted to a curated Twitter dataset on COVID-19 vaccination [16]. The proof-of-concept using the curated dataset can be accessed at https://demo-frontend-uy34.onrender.com/.

## 2 RELATED WORK

Previous studies, e.g., [2, 14] identify many factors that contribute to the rapid digital dissemination of false news, including low factual understanding and an inability to recognize fake news. Thus, recommending correct information to the users is the first step to combat misinformation. The detection of misinformation and interpretation of black-box models has been a major objective of existing misinformation studies [4, 10, 15], with data analysis receiving little


*CIKM '23, October 21–25, 2023, Birmingham, United Kingdom*

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614746

attention [3, 7]. Few works explore recommendation approaches to fake news [18, 19]. However, in these cases, the recommendation is either in the form of the whole network [8, 11] or based on fact-checking URLs exclusive to a small number of fact-checkers [12, 19]. The works that are specific to COVID-19 misinformation include releasing the new COVID-19 datasets [6] or evaluating the deep learning models on publicly available datasets [1, 9]. Few studies also work on the impact of COVID-19 misinformation on mental health [17] and how to tackle effectively [5].

The work closest to ours is the personalized recommendation [18] which is based on user-level recommendations, as opposed to our tweet-centric recommendation approach. The limitation of a user-level approach is that if new events occur, it is difficult to collect the user's relevant reading history to personalize recommendations. A user might also want to explore new genres of news. Furthermore, user-level approaches risk creating a bubble or tunnel vision of their own. In these regards, a tweet-centric recommendations mechanism complements and can help. Specifically, suppose a user is interested in a certain post that happens to carry misinformation. In that case, our approach can extract the topics, sub-topics (in the form of entities), and the associated sentiment, and then it can recommend similar but factual non-misleading posts to the user, irrespective of the user's own biases or consumption history. Thus, in this work, keeping in mind the issue of confirmation bias of social media, we focus on recommending users with correct information targeted on the same topic as a more practical and reasonable approach to combat spreading misinformation on social media.

# 3 SYSTEM ARCHITECTURE

Misinformation Concierge has a modular architecture (Figure 1). The individual tasks, e.g., data staging, classification, topic extraction, etc., can be accomplished using a different mechanism than what we currently use so as to organically upgrade the system with better tools for individual tasks. e.g., with better quality of results. The emphasis of this demo is as such on the overall framework, rather than maturity and quality of the solutions for the individual tasks. Next, we explain how these tasks are accomplished currently.

## 3.1 Data acquisition, staging & classification

We use a publicly accessible dataset [16] for this work, which contains 114,635 COVID-19 vaccination related tweets for the seven months from September 2020 to March 2021 labeled as misleading or non-misleading using ML techniques trained with originally 1,500 manually annotated data.

In general, to deploy Misinformation Concierge on a new subject, a user would (1) need to provide relevant keywords to be used for acquiring Tweets through Twitter API, which is then (2) cleaned mostly through automated tools, e.g., removing non-English tweets; but also identifying tweets which match a single keyword from step-1, which might need a manual inspection to determine relevance to the subject, e.g., for the chosen case study, some vaccine related tweets were not relevant to COVID-19 and had to be eliminated. Finally, the collected dataset has to be (3) classified as non/misleading using a semi-supervised classifier [16], which in turn needs a small subset of manually labeled data.

Subsequently, as the system is used, end-users could provide feedback on the quality of recommendation in a low-effort manner, to be used to transparently expand the manually annotated data and regularly rerun and refine the ML models in the background.

## 3.2 Recommendation of Non-Misleading Tweets

To recommend related non-misleading tweets to counter misleading tweets, we use three Natural Language Processing (NLP) techniques: LDA topic modeling [16] and Named Entity Recognition[1] (NER) for extracting general and specific topics, accompanied by matching sentiments[16] of the tweet.

*3.2.1 General (Broad) Topic extraction.* It is crucial first to understand the broad topic of a given tweet. In this work, similar to [16], we utilize LDA topic modeling to assign topics to each tweet. The names of the topics are manually ascribed following automatic detection. Specifically, if a particular topic is present in the tweet, we label it as one of the topics. This approach gives us 12 topics in total, and 4,063 tweets are left without labels (Unknown label). To further improve our process for the Unknown label, we used synonyms[2] for each topic and checked if the synonyms of each topic were present in the tweets that contained an Unknown label. 817 tweets could thus be labeled as one of the 12 topics, leaving 3,426 tweets with an Unknown label, as shown in Table 1.

**Table 1: Absolute number (and %) of Tweets per topic.**

| Topics | Tweets: number (%) |
|---|---|
| Choices | 33,150 (28.9%) |
| Politics | 25,276 (22%) |
| Vaccine Efficacy | 21,936 (19.1%) |
| Shots | 9,568 (8.34%) |
| Trump | 8,432 (7.35%) |
| Data & Facts | 3,601 (3.14%) |
| Unknown | 3,426 (2.98%) |
| Trials | 3,217 (2.8%) |
| Myths | 2,376 (2%) |
| Operation Warp Speed | 1,369 (1.19%) |
| Real Side-Effects | 1,216 (1.06%) |
| Approval | 883 (0.77%) |
| Availability | 185 (0.16%) |

*3.2.2 Specific Topic Extraction.* Next, we dig deeper into the tweets to find named entities used in each tweet. We explored various NER (Named Entity Recognition) models, for example, the latest pre-trained roberta-base NER model, en_core_web_trf (which identified entities in 57% of our data). The NER model en_core_web_sm gave us the best performance containing 88% data with at least one entity. We observed two problems with the use of such pre-existing models. First, there are words specific to COVID-19, such as Pfizer, Shots, and Johnson & Johnson, that were being wrongly labeled as PERSON, GPE (Geopolitical entity). Second, there were still 12% of rows that contained zero (or null) entities.

Consequently, we fine-tuned the NER model on a random subset of manually labeled 100 instances from the dataset that contained non-null entities. We added a new entity type called VAC_TYPE for the words that contain vaccine names using a list of manually

[1] https://spacy.io/api/entityrecognizer
[2] https://www.nltk.org/howto/wordnet.html

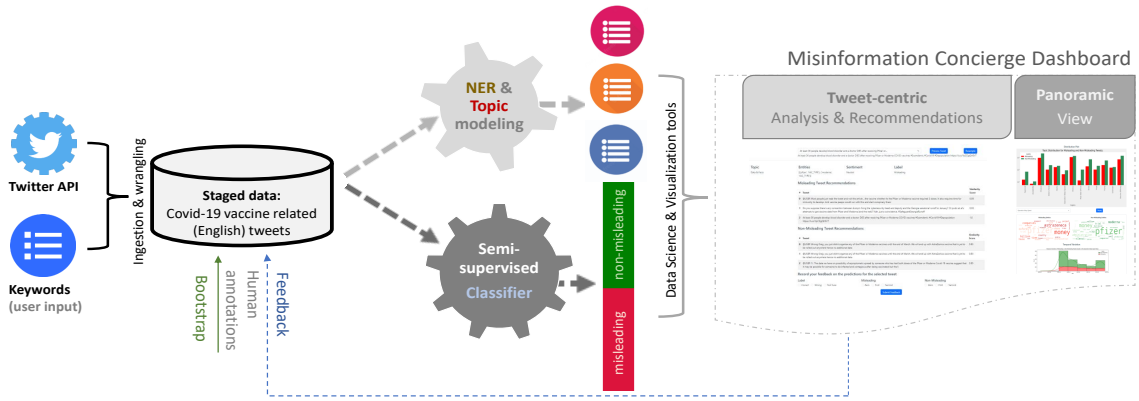


**Figure 1: A high-level view of the Misinformation Concierge architecture.**

labeled VAC_TYPE entities: [pfizer, astrazeneca, mrna, astrazenca, jnj, oxford, sputnik, modern, variants, #pfizer, booster, #astrazeneca, biontech, Covidshield]. To make the model robust and based on our previous observations, we also included spelling errors in the entities, such as 'modern' instead of moderna, in the list. Additionally, we did not remove the hashtags from the dataset as the hashtags play a major role in identifying misinformation these days. We trained the augmented model for 30 epochs with a dropout rate of 0.3 utilizing five-fold cross-validation. Next, we compare the performance of the augmented (fine-tuned) model with the unaugmented model on the same subset of manually labeled 100 examples shown in Table 2. We observe that the augmented model performs better in all the metrics with a good margin. Finally, we labeled all the tweets in the dataset with the augmented model. We obtained the list of VAC_TYPE entities that are labeled by the augmented model, which included new entries beside the ones we had manually provided during training: [phizer, myrna, zenca, novavax, johnsonandjohnson, johnson, mirna].

**Table 2: Comparison of the Spacy NER model with and without fine-tuning the model on the manually labeled data.**

| Spacy model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| w/o training | 0.27 | 0.25 | 0.26 | 0.25 |
| w/ training | 0.89 | 0.90 | 0.88 | 0.87 |

We notice that 112,994 (98.5%) rows have non-identical entities with the augmented model (up from 88% with the unaugmented one), indicating a profound improvement in the detection of entities and entity type. Among the 36,829 previously mislabeled entities, 35,480 (96.3%) and 1,349 (3.7%) entities represent VAC_TYPE and non-VAC_TYPE entities. Table 3 represents the top entities divided into VAC_TYPE and non VAC_TYPE entities that are mislabeled by the unaugmented model. For example, as shown in Table 3, the entity Pfizer has been incorrectly labeled (in the given context) as ORG by the unaugmented model and correctly labeled as VAC_TYPE by the augmented model. VAC_TYPE entities detected by the model are in such a big number since most of the dataset contains vaccine names in their tweets.

In very few instances, some entities were labeled correctly by the unaugmented model but incorrectly by the augmented model. Particularly, a few geographic places were mislabeled as VAC_TYPE by the augmented model but correctly as GPE by the unaugmented model. Additionally, there have been cases where both models fail to identify the entity accurately, e.g., the un/augmented models have incorrectly labeled Ohio as MONEY and VAC_TYPE, respectively.

**Table 3: Top entities mislabeled by the unaugmented NER model. The left and right sides represent vaccine names and non-vaccine names entities.**

| VAC_TYPE | | | Others | | |
|---|---|---|---|---|---|
| Entity | Mislabeled | Correct | Entity | Mislabeled | Correct |
| pfizer | ORG | VAC_TYPE | millions | CARDINAL | MONEY |
| moderna | GPE | VAC_TYPE | billions | CARDINAL | MONEY |
| astrazeneca | ORG | VAC_TYPE | trump | ORG | PERSON |
| johnson and johnson | PERSON | VAC_TYPE | biden | ORG | PERSON |
| novavax | ORG | VAC_TYPE | lock down | NORP | EVENT |

*3.2.3 Sentiments Extraction.* Users suffering from confirmation bias tend to prefer to read posts on a certain topic that reflect their sentiments. We employ VADER API[16] to detect the sentiment of the tweet as positive, negative, or neutral.

*3.2.4 Recommend Similar (Non-)Misleading Tweets.* To determine the similarity between a given misleading tweet and other (non-)misleading tweets, we tested two approaches: direct string matching (using methods like Hamming distance) and vector dimension matching (such as Word2Vec) to match the misleading tweets with equivalent (non-)misleading tweets. The optimal similarity matching for our data was obtained with vector dimension matching. After experimenting with several combinations, we utilized GloVe embedding[13] with cosine similarity in particular.

For a given Misleading tweet, we use three criteria, i.e., the general topic, entities extracted by the augmented NER model, and the sentiment. Next, we match these three criteria over the corpus of the (non-)Misleading tweets. Once such (non-)Misleading tweets have been retrieved, the top-K similar (non-)Misleading tweets are identified using the GloVe embedding and cosine similarity.

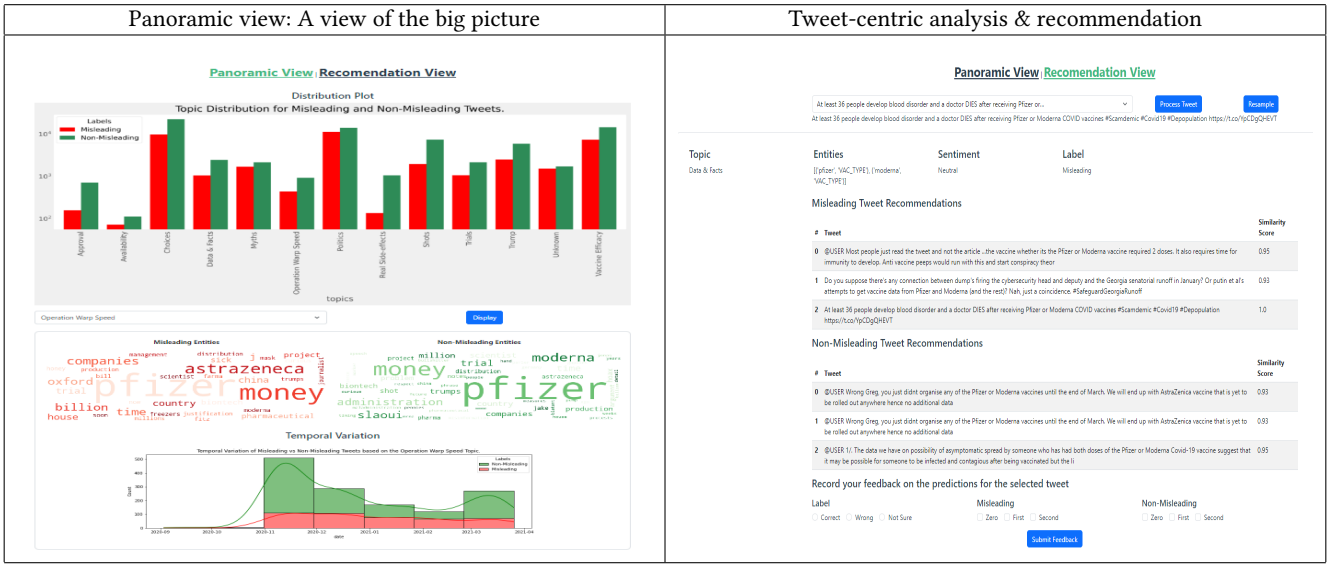


**Figure 2: Recommendation Dashboard. The left and right sides represent panoramic and tweet-centric views. The NM and M denote the Non-Misleading and Misleading tweets, respectively.**

## 4 DEMONSTRATION

For the demonstration, we use a dataset preprocessed as discussed above, which includes columns for sentiments, topics, entities, labels, and tweets. During the live demo, we will explain these steps while showcasing the Misinformation Concierge functionalities comprising broadly two parts: A Panoramic View capturing the big picture and a Tweet-centric Analysis & Recommendation View.

**The Panoramic View** provides an overview of the dataset. In particular, displaying the distribution of the topics for (non-)misleading labels in the dataset serves as the first step. Based on the topic selected by the user, word clouds of entities across both labels are displayed, along with temporal characteristics of how and when the topic has gained traction.

**The Tweet-centric Analysis & Recommendation View** analyses individual tweets, identifying its topic, involved entities, sentiments, identifies other similar misleading tweets, and recommends non-misleading tweets that can be used to counter the chosen tweet. The number of such recommendations is a user chosen parameter (with a default set to three). Finding other similar misleading tweets could be used to identify homophilic users and echo chambers, while the non-misleading tweets can be reused for rebuttal.

The underlying algorithms are inherently imperfect, furthermore, they have been trained with very small samples of human annotated data to bootstrap the system. To ameliorate the latter issue, the dashboard captures feedback from the end-user on the correctness of labels assigned to tweets. This is used as feedback to generate more human annotations organically over time without the onerous burden of labeling data explicitly. The underlying models can thus be retrained and refined over time as the system is used (this is implemented but not tested, since our system currently does not have a user base), and augmented with extrinsic resources such as large language models to improve the quality of recommendations and incorporate live data in the future.

**Challenges:** The deployment of the classification model might run into a notable challenge known as the "cold start" problem, which could hamper its capacity to effectively categorize non-COVID-19 topics. However, by employing a fine-tuning approach to the classification model using even a small dataset particularly labeled for non-COVID-19 topics, this issue can be efficiently overcome. This approach makes it possible to greatly improve the model's performance across diverse datasets. In addition, user feedback about misclassified examples would help in model refinement. The second challenge is that the effectiveness of the demo's underlying components, such as the LDA, NER, and semi-supervised learning modules, is closely related to the utility of the demo. In fact, the demo's current version makes use of conventional methods like VADER and a model that was trained on annotated examples. We purposefully chose to demonstrate functionality above seeking state-of-the-art (SOTA) performance because our goal was to offer a working model that could be utilized as a starting point. With this strategy, we are able to establish a foundation for future development while also providing useful insights into the possibilities of more advanced methods in subsequent iterations. The third challenge involves the recent changes in Twitter API, we are committed to actively monitoring and adapting the tool to align with the most recent developments and ensure its continued relevance and utility.

**Acknowledgment:** S. Sharma and R.Sharma's work has received funding from the EU H2020 program under the SoBigData++ project (grant agreement No. 871042), by the CHIST-ERA grant No. CHIST-ERA-19-XAI-010, (ETAg grant No. SLTAT21096), and partially funded by HAMISON project.

# REFERENCES


[1] Hosam Alhakami, Wajdi Alhakami, Abdullah Baz, Mohd Faizan, Mohd Waris Khan, and Alka Agrawal. 2022. Evaluating Intelligent Methods for Detecting COVID-19 Fake News on Social Media Platforms. *Electronics* 11, 15 (2022), 2417.

[2] V. Balakrishnan, N.w. Zhen, S.M. Chong, G.J. Han, and T.J. Lee. 2022. Infodemic and fake news–A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction* (2022), 103144.

[3] Sabur Butt, Shakshi Sharma, Rajesh Sharma, Grigori Sidorov, and Alexander Gelbukh. 2022. What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses. *Computers in Human Behavior* 135 (2022), 107345.

[4] Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. 2022. Game-on: Graph attention network based multimodal fusion for fake news detection. *arXiv preprint arXiv:2202.12478* (2022).

[5] Federico Germani, Andrew B Pattison, and Monta Reinfelde. 2022. WHO and digital agencies: how to effectively tackle COVID-19 misinformation online. *BMJ Global Health* 7, 8 (2022), e009483.

[6] Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

[7] Raj Jagtap, Abhinav Kumar, Rahul Goel, Shakshi Sharma, Rajesh Sharma, and Clint P George. 2021. Misinformation detection on YouTube using video captions. *arXiv preprint arXiv:2107.00941* (2021).

[8] D. Kempe, J. Kleinberg, and É. Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 137–146.

[9] Ziyi Kou, Lanyu Shang, Yang Zhang, and Dong Wang. 2022. Hc-covid: A hierarchical crowdsource knowledge graph approach to explainable covid-19 misinformation detection. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–25.

[10] Mohit Mayank, Shakshi Sharma, and Rajesh Sharma. 2022. DEAP-FAKED: Knowledge graph based approach for fake news detection. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 47–51.

[11] N.P. Nguyen, G. Yan, M.T. Thai, and S. Eidenbenz. 2012. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*. 213–222.

[12] V. Nguyen and K. Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 275–284.

[13] J. Pennington, R. Socher, and C. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[14] S. Sharma, E. Agrawal, R. Sharma, and A. Datta. 2022. FaCov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1312–1321.

[15] Shakshi Sharma and Rajesh Sharma. 2021. Identifying possible rumor spreaders on twitter: A weak supervised learning approach. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[16] S. Sharma, R. Sharma, and A. Datta. 2022. (Mis)leading the Covid-19 vaccination discourse on Twitter: An exploratory study of infodemic around the pandemic. *IEEE Transactions on Computational Social Systems* (2022).

[17] Gaurav Verma, Ankur Bhardwaj, Talayeh Aledavood, Munmun De Choudhury, and Srijan Kumar. 2022. Examining the impact of sharing COVID-19 misinformation online on mental health. *Scientific Reports* 12, 1 (2022), 8045.

[18] S. Wang, X. Xu, X. Zhang, Y. Wang, and W. Song. 2022. Veracity-aware and Event-driven Personalized News Recommendation for Fake News Mitigation. In *Proceedings of the ACM Web Conference 2022*. 3673–3684.

[19] D. You, V. Nguyen, K. Lee, and Q. Liu. 2019. Attributed multi-relational attention network for fact-checking url recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1471–1480.