Journal of Database Management
Volume 37 • Issue 1 • January-December 2026

A FAIR Principles-Driven Quality Assessment of Social Media Datasets for Natural Language Processing-Based Pandemic Surveillance

Yang Liu
https://orcid.org/0000-0002-4307-9738
North Carolina Central University, USA

May Almousa
Princess Nourah Bint Abdulrahman University, Saudi Arabia

Mohd Anwar
https://orcid.org/0000-0002-2653-7987
North Carolina A&T State University, USA
Received: March 16th, 2025 | **Accepted:** January 15th, 2026

ABSTRACT

Social media has become integral to daily interactions and a key data source for researchers. Using COVID-19 as a case study, this work compares 24 social media datasets to address three research questions: 1) Is the dataset in compliance with the FAIR principles of being Findable, Accessible, Interoperable, and Reusable? 2) To what extent have people utilized social media to voice and exchange their apprehensions during the COVID-19 pandemic? 3) To what extent can social media datasets be utilized for natural language processing (NLP)-based COVID-19 pandemic surveillance? Leveraging the evaluation questions derived from the FAIR principles, the authors assess 24 social media datasets related to the COVID-19 pandemic. Additionally, they comprehensively analyze each dataset, including their composition, and the specific instances and features they encompass. They have initiated an attempt hoping that more researchers will join to create a data community where information can be repurposed and reused.

KEYWORDS

Social Media Dataset, Pandemic Surveillance, Natural Language Processing, FAIR Principles, Findable, Accessible, Interoperable, Reusable

INTRODUCTION

According to Statista (2025), 5.66 billion people used social media globally as of October 2025, reaching 68.7% of the world's population. People use social media platforms to voice and exchange their opinions and concerns regarding issues including public health (Conway et al., 2019; Zhang et al., 2024), politics (Kruse et al., 2018), culture (Sheldon et al., 2020), economics (Karami et al., 2018), and education (Antelmi et al., 2023; Liu & Anwar, 2022). Social media platforms play a crucial role in producing huge amounts of data over the Internet every day. They provide rich resources for information retrieval, information extraction, and large language model training.

DOI: 10.4018/JDM.399759

More and more researchers are using social media as an alternative data source for understanding public health crises such as Middle East respiratory syndrome coronavirus (Fung et al., 2013, 2018), avian influenza (Schaeffer et al., 2022; Soltani et al., 2025), coronavirus (Whitfield et al., 2021; Zang et al., 2023; Zhang et al., 2024), and mpox (Leslie et al., 2023; Liu et al., 2022). One study used Weibo to track public reactions to two different outbreaks: the 2012 Middle East respiratory syndrome coronavirus (MERS-CoV) outbreak and the 2013 outbreak of human infection with avian influenza A (H7N9) in China. The results demonstrated that social media serves as a valuable tool for measuring public awareness of disease outbreak information released by health authorities, as well as the public's responses to this information. Another study explored the use of social media data (from daily Naver blog posts and Twitter) and search engine query data (from the Korean website Daum) to estimate influenza outbreaks (Woo et al., 2016). The results demonstrated the feasibility of using search queries to enhance influenza surveillance in South Korea. Additionally, query selection based on social media data proved to be effective.

Nevertheless, challenges remain. The quality, volume, interoperability, and reusability of data are difficult to assess when dealing with high-volume, complex datasets from disparate sources that include raw, unstructured, and noisy data. Moreover, the discoverability and usability of existing datasets are major concerns for potential re-users.

To address these challenges, a set of guiding principles—findable, accessible, interoperable, and reusable (FAIR)—offers a valuable framework for responsible research data stewardship (Boeckhout et al., 2018; Wilkinson et al., 2016). To better understand the application of these principles to pandemic-related datasets, specifically using COVID-19 as a case study, this study compares 24 datasets containing COVID-related posts from social media platforms including Twitter/X, Reddit, and Weibo. More specifically, we compare social media datasets to answer these research questions:

- Is the dataset in compliance with the FAIR principles of being findable, accessible, interoperable, and reusable?
- To what extent have people utilized social media to voice and exchange their apprehensions during the COVID-19 pandemic?
- To what extent an social media datasets be utilized for natural language processing (NLP)-based COVID-19 pandemic surveillance?

The main contributions of this study are the following:

- We identify 24 datasets about COVID-related posts from multiple social media platforms and evaluate them based on FAIR principles.
- We analyze the advantages and disadvantages of each dataset.
- We discuss the reusability of COVID-19 social media datasets for NLP-based pandemic surveillance.

To our knowledge, this is the first comparative study of COVID-19 social media datasets for NLP-based pandemic surveillance.

The rest of the paper is organized as follows. In the next section, we briefly introduce the background of social media platforms (i.e., Twitter/X, Reddit, and Weibo) and the COVID-19 pandemic. Then, we present the methodology for data search and FAIR principles. Afterward, we demonstrate the results of the analysis and discuss its limitations. In the final section, we provide the conclusions of this study.