

Irmasari Hafidz¹, Alfado Rafly Hermawan¹, Rahmah Yasinta Rangkuti (MD, Pediatrician)², Tri Hediando (MD, ENT/ Otolaryngologist)², Nur Aini Rakhmawati¹

¹Department of Information Systems, Institut Teknologi Sepuluh Nopember

²Department of Department of Medical Doctor Profession Education, Institut Teknologi Sepuluh Nopember

Abstract

Purpose

This study analyzes social media data on Long COVID to better understand symptoms, daily life impacts, and coping mechanisms. The dataset was transformed into machine-readable formats such as RDF and JSON-LD, using standardized vocabularies and ontologies to enhance interoperability, reusability, and adherence to FAIR principles (Findable, Accessible, Interoperable, Reusable), ensuring its utility for future research.

Methods

The dataset included 28,000 clean tweets and 3,000 labeled social media posts sourced from Hafidz (2024) and annotated by A. R. Hermawan et al. (2024). Methods include:

- Data Preprocessing: Removal of duplicates, handling missing values, and standardizing text formats.
- Entity Extraction and Categorization: Extracting named entities and grouping them (e.g., individuals vs. institutions).
- Data Transformation: Converting the dataset into RDF and JSON-LD formats using standardized vocabularies like Schema.org and Dublin Core.
- Validation and Querying: Validating the transformed data with SPARQL queries to extract insights on symptom prevalence and user experiences.

Results

- Symptom Diversity: Common symptoms include fatigue, loss of taste and smell, brain fog, respiratory issues, and cognitive dysfunction. Sentiment analysis reveals negative emotions such as anxiety, frustration, and despair.
- Coping Mechanisms: Online communities provided support, shared tips, and offered encouragement.

- Frequent Entities: Individuals like Brixx, @MountvSxxxx, and @annalxxxn, and organizations like LongCovid, N95, and @statnews were often mentioned.
- Co-Occurrences: LongCovid co-occurred with terms like breathlessness, fatigue, and brain fog. Comparisons were made with conditions like Lyme disease and fibromyalgia.
- Machine-Readable Formats: RDF and JSON-LD enabled advanced querying, enhancing dataset integration and analysis.

Value

This study highlights Long COVID's significant impact on both physical and mental health. By analyzing social media data, it provides complementary insights to clinical studies. Transforming the dataset into machine-readable formats enhances its reusability and interoperability, aligning with FAIR principles. The dataset and scripts are publicly available for future research:

https://github.com/irhafidz/2025_WOOC_Bologna/ Future work could refine NLP techniques and expand ontologies for broader applications.

Reference

Hafidz, I. (2024). Data_collection_longcovid_2022. Zenodo.

<https://doi.org/10.5281/zenodo.14227098>

Hermawan, A. R., Hafidz, I., Rangkuti, R. Y., Latiffianti, E., & Rakhmawati, N. A. (2024, December). Early Detection of Long COVID Symptoms from Social Media Using BERT. In *2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1-5). IEEE.

<https://ieeexplore.ieee.org/abstract/document/10836286>