CoVID-19 Fake News Infodemic Research (CoVID19-FNIR) Dataset — Documentation

JULIO A. SAENZ, University of Wyoming, USA
SINDHU REDDY KALATHUR GOPAL, University of Wyoming, USA
DIKSHA SHUKLA, University of Wyoming, USA

*This document provides a detailed description of CoVID19-FNIR (**CoVID19- Fake News Infodemic Research**) Dataset.*

DATASET SUMMARY

CoVID19-FNIR is a CoVID-19-specific dataset consisting of fact-checked fake news scraped from Poynter and true news from the verified Twitter handles of news publishers. The data samples were collected from India, The United States of America, and European regions and consist of online posts from social media platforms between February 2020 to June 2020. The dataset went through prepossessing steps that includes removing special characters and non-vital information.

DATA FORMAT AND FILE STRUCTURE

The CoVID19-FNIR.zip folder contains the whole dataset. The folder has two files; (1) *fakeNews.csv*, and (2)*trueNews.csv*.

fakeNews.csv The file *fakeNews.csv* is organized as follows. It contains the columns and the corresponding information as listed below. The last column, **label**, shows the classification label for the corresponding news item. Each row is one news item.

- `Date`: The date that the article was published
- `Link`: The Poynter link of the article
- `Text`: The text found in the article
- `Region`: The region the article is from
- `Country`: The country the article is from
- `Explanation`: The explanation as to why the article was false
- `Origin`: The website origin of the article
- `Origin_URL`: The URL for the website origin of the article
- `Fact_checked_by`: Name given of who fact-checked the article
- `Poynter_Label`: The multi-class classification label given by Poynter
- `Label`: The binary classification label we provided of 0 for false

trueNews.csv The file *trueNews.csv* contains the following columns with last column **label** being the classification label for the corresponding news item. In this file all news items come from the twitter handles of trusted news sources and were assigned a classification label as *'True'*.

- `Date`: The date the tweet was posted
- `Link`: The Twitter link of the tweet

Authors' addresses: Julio A. Saenz, jsaenz5@uwyo.edu, University of Wyoming, Department of Computer Science, Laramie, WY, 82072, USA; Sindhu Reddy Kalathur Gopal, skalathu@uwyo.edu, University of Wyoming, Department of Computer Science, Laramie, WY, 82072, USA; Diksha Shukla, dshukla@uwyo.edu, University of Wyoming, Department of Computer Science, Laramie, WY, 82072, USA.

- `Text`: The text found in the tweet
- `Region`: The region the tweet is from
- `Username`: The Twitter handle username of the news publication
- `Publisher`: The official name of the news publication organization
- `Label`: The classification label as True

Twitter Handles. The verified Twitter handles of news publishers from where the news samples for the *trueNews.csv* were collected.

- `India`: NDTV, The Hindu, India Today
- `The United States of America`: CDC, The New York Times, The Washington Post
- `Europe`: Guardian News, BBC News UK, Reuters UK