DFND : Dravidian Fake News Detection Dataset

Eduri Raja, National Institute of Technology Silchar, India
Badal Soni, National Institute of Technology Silchar, India
Samir Kumar Borgohain, National Institute of Technology Silchar, India

This document provides a detailed description of DFND (Dravidian Fake News Detection) Dataset.

DATASET SUMMARY

DFND is a Dravidian fake news dataset for detecting fake news in Dravidian languages, namely Telugu, Kannada, Tamil, and Malayalam. We collected the data from different sources: for real news articles, we scrapped the data from various news websites like Eenadu, Dinamalar, Kannadaprabha, Malayala manorama, etc.; for fake news articles, we scrapped the data from various fact-checking websites like factly, factcrescendo, etc. We collected the data from January 2021 to December 2022. After collecting the data, we did the data annotation through corresponding language experts for Dravidian languages. Our dataset is preprocessed and cleaned. This dataset contains around 27,000 news articles which are both fake and real news articles for the four Dravidian languages.

DATA FORMAT AND FILE STRUCTURE

The DFND.zip folder contains the whole Dravidian languages dataset. The folder has four files: (1) Telugu, (2) Tamil, (3) Kannada, and (4) Malayalam. Each folder has two files: (1) fake.csv and (2) true.csv.

The Dataset has two columns: text and label.
text: A claim published in the media by a person or an organization.
label: The class for each sample.

Authors' addresses: Eduri Raja, eduri_rs@cse.nits.ac.in, Department of Computer Science, NIT Silchar, India; Badal Sonil, badal@cse.nits.ac.in, Department of Computer Science, NIT Silchar, India; Samir Kumar Borgohain, samir@cse.nits.ac.in, Department of Computer Science, NIT Silchar, India.