1 SenWave Dataset

The **SenWave Dataset** is a collection of over 104 million tweets related to COVID-19, collected between March 1 and May 15, 2020, across six languages: English, Spanish, French, Arabic, Italian, and Chinese. The dataset includes 10,000 labeled English tweets and 10,000 labeled Arabic tweets, each categorized into ten sentiment categories: optimistic, thankful, empathetic, pessimistic, anxious, sad, annoyed, denial, official report, and joking. This categorization enables in-depth analysis of the emotional responses during the pandemic. The dataset was collected using (Twint)[1] and includes unlabeled tweet IDs from five languages. To maximize the use of labeled data, English tweets were translated into Spanish, French, and Italian using Google Translate. The dataset also contains tweet IDs organized by language and date, and statistics are provided in separate text files. The labeled tweets are stored in CSV files, with English and Arabic tweets manually annotated, while the other languages were translated from English.

For our research, we used this dataset after fine-tuning our HP-BERT model with the Hinduphobia dataset. The SenWave dataset was then employed for multi-stage fine-tuning to enhance the model's ability to perform sentiment analysis, especially in the context of Hinduphobic content during the COVID-19 pandemic.

The dataset is publicly available and can be accessed from the GitHub repository **SenWave Dataset**[2].

[1] `https://github.com/twintproject/twint`
[2] `https://github.com/gitdevqiang/SenWave`