On this page

Training Data

In Machine Learning (ML) and Natural Language Processing (NLP), data is sorted in the following groups:

- Training dataset (text + labels)
- Validation dataset (text + labels)
- Testing dataset (text + label)

Kindly is currently compiling a training dataset to both tranfer learning from the current model (see cardiffnlp/twitter-roberta-base-offensive) using data submitted by children, and to build a new model from scratch.

Once data is started to be compiled, it will be split into the three groups mentioned above to create the model. You can refer to the notebook `KindlyModel.ipynb` as a reference on how this split is currently being done with the initial dataset.

Initial Dataset

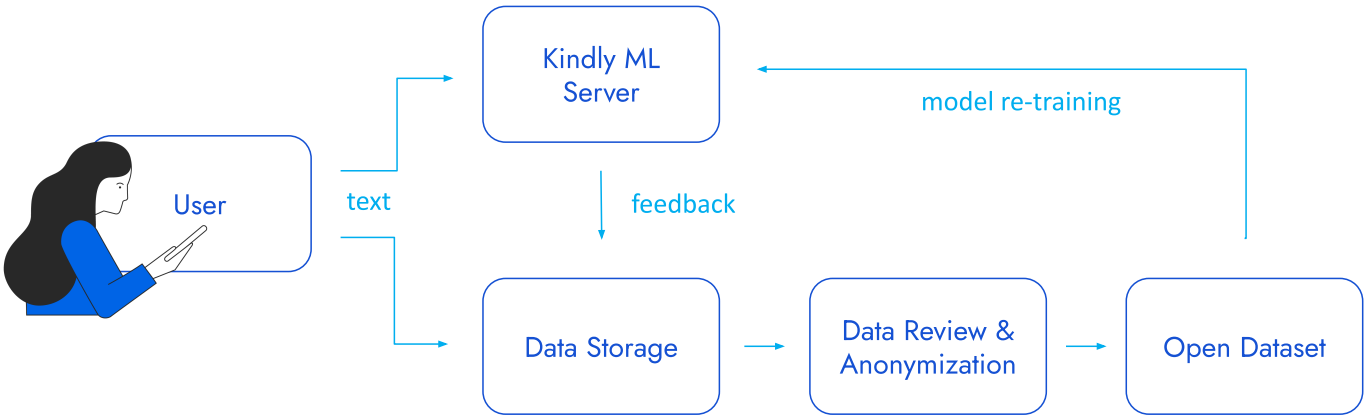
The initial dataset comes from the data provided by Gitanjali Rao (see Acknowledgements), wh was exported from the proof-of-concept NLP model developed using Microsoft Azure's Language Understanding (LUIS) cloud-based engine.

This initial data was extracted from it's original JSON format and put in to two text files, `modeling/dataset/offensive_train_text.txt` and `modeling/dataset/offensive_train_labels.txt`. These two files are the initial format which will be used to build the final model for kindly.

The data is tagged in the file `modeling/dataset/offensive_train_labels.txt` to know which submissions are offensive or not to help the model learn. The numbers `1` and `0` in the labels file correspond to `offensive` and `non-offensive` respectively. The number on each line in the file directly corresponds to the sentence or phrase in each line of the training text file `offensive_train_text.txt`. These numerical tags will be validated as more data comes in through the *Data Colection* process mentioned below.

Data Collection

As part of Kindly's website, a contribute page has been developed to request training data from children. The form simulates a hypothetical chat interaction a child my have with a friend, prompting them to react to a predetermined set of prompts chosen at random.



Every time the child submits text through the form, data is both stored and submitted for analysis by the current model. The model then returns its prediction on whether it detects cyberbulling intent, and the child can validate whether the prediction is acurate or not. The validation from the user is apended to the initial data entry.

No data is collected from the user other than the text that that the child enters in the form and their validation on whether the preditcion is accurate or not. The data submission is completely anonymous.

Data submitted by children is stored pending review by UNICEF staff. The human review validates that the data is relevant and anonymizes it by removing any personal identifiable information (PII) that the user may have submited inadvertently or by mistake (e.g. replacing proper nouns with a generic reference such as "*@user*"). The data cleansing includes the removal of individual names, physical or email addresses, school names, phone numbers or any other personal or device identifiers, as well as location data. At the discretion of UNICEF staff, if we can think of any other data (or combinations thereof) that may lead back to a child, then that will also be removed. After the new data has been reviewed, it is then added to Kindly's open dataset. This open dataset is used to periodically re-train and improve the existing machine learning model.

Edit this page