Toxicity Classification

1. **Problem Identification**

   The toxic comments are a severe issue on social media, because it leads to people expressing their opinion irrationally and trying to harm each other's feelings. The goal of this project is to use a deep learning model to identify toxicity in text and try to suggest the solution for classifying the toxic comments in various categories using natural language processing concepts.

2. **Data Source**

   https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

- We have one single csv file for training and one cvs file to test.
- Columns in train data:
  - Comment_text: This is the data in string format which we have to use to find the toxicity.
  - target: Target values which are to be predicted (has values between 0 and 1)
  - Data also has additional toxicity subtype attributes: (Model does not have to predict these)
    - severe_toxicity
    - obscene
    - threat
    - insult
    - identity_attack
    - sexual_explicit
  - Comment_text data also has identity attributes carved out from it, some of which are:
    - male
    - female
    - homosexual_gay_or_lesbian
    - christian
    - jewish
    - muslim
    - black
    - white
    - asian

        - latino
        - psychiatric_or_mental_illness
    - Apart from above features the train data also provides meta-data from jigsaw like:
        - toxicity_annotator_count
        - identity_anotator_count
        - article_id
        - funny
        - sad
        - wow
        - likes
        - disagree
        - publication_id
        - parent_id
        - article_id
        - created_date

Reference:

https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1184/reports/6837517.pdf
https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview
https://machinelearningknowledge.ai/natural-language-processing-github-projects-to-inspire-you/