Weekly Team Meeting:
5:00-6:00 PM Our github repo: acmucsd-projects/wi23-ai-team-1 (github.com)

Team Meeting Time: 2/18/23 at 5pm in CSE Basement
Attendees: **Aniket Gupta** **Jeffrey Lee** **Vincent Tu**

What have we done so far

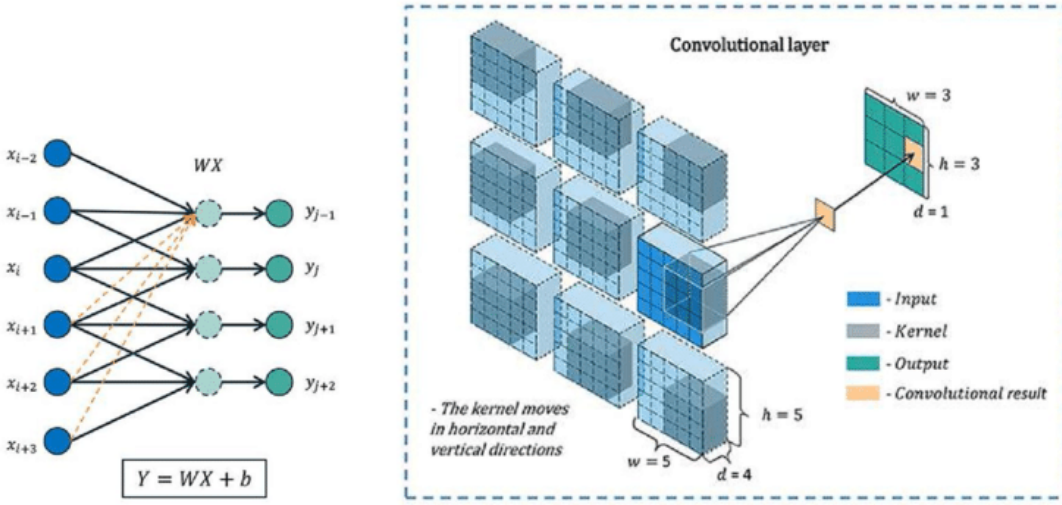
- Added a function to filter out small sentences
- Added a function to filter out all non-ascii characters

What is the point of this meeting, what are we going to discuss

- Filter out numbers and punctuation
- Position embedding to compensate for filtering out punctuation?
- NLTK (natural language toolkit) - library with a bunch of functions for pre-processing
- Code own functions for filtering so that they are as customizable
- Removing other languages: sufficient to just remove non-ascii characters
- Stopwords: filler words in english are used as stop words ("like", "kinda"), used for humans to better read or understand something → these words are wasted text and should be removed
    - Use NLTK.stopwords
    - https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
- Explore the reason for
    - Why there are -1 values in the test label
        - Value of -1 indicates it was not used for scoring?
    - Why there is so much test data
- Must account for class imbalance by checking the amount of data is provided in each class and balance the data between each class
- Filtering Data
    - Removing other languages
    - Tests that are too short
    - Remove special characters (punctuation and new line characters)
    - Remove numbers
    - Remove stopwords

What will we do going forward

- For most collaboration: just write locally → pull and then use and preferred environment (Kaggle)
- Force everything into lowercase first
- Clean dataset
    - Removing other languages
    - Tests that are too short

    - Remove special characters (punctuation and new line characters)
    - Remove numbers
    - Remove stopwords
- After preprocessing and cleaning data → analyse data
- Use Tensorflow for easier way to get into deep learning
    - Tensorflow has all the computation and functions needed
        - Keras is the library that allows you to define layers and networks for deep learning models
        - Layers: collection of nodes
        - 

        - (a) (b)
- Exploration (find patterns and findings to better the models performance)
    - Graph based on
        - Length of the text
        - Which words are most commonly seen
        - Which words are most commonly associated with which kind of toxicity
    - Finding weaknesses and problems
    - Find anything you can exploit
- Preprocessing:
    -
- Modeling (Dense neural network)
    - CNN
        - Recommended to follow this guide using the dataset: https://www.tensorflow.org/text/tutorials/text_classification_rnn
    - BERT:
        - Paper in NLP AI - a novel way to train a transformer (composed of complex layers)
        - A model that can do very well on English text
        - May be used later on along the line
        - Only when moving to a model along the lines of BERT may require using Kaggle
- Experimentation after modeling