Title and Group:

Group members: Chow Jun Wei, Esraa Sultan.
Internship Batch: NLP 02.
Report Date: November 20, 2021

Problem Description

There are 20 different categories of articles, and our job is to classify, to which category does each article belongs to.

Data Cleaning:

Note: This data cleaning pdf contains all data cleaning done by all members rather than splitted. For splitted, check each individual ipynb notebooks.

- Choose sentences that are larger than a certain threshold (tune-able). (This gives the most significant improvement in model training, anywhere between threshold of 5 to 10 [perhaps larger but not tried] works quite well in increasing prediction accuracy by about 5-9%).
- Removal of strings without words (either empty or just space(s)).
- Cleaning using regex: removing unnecessary signs and separators like ****** and ----------, unnecessary tabs (\t), other symbols (<, >) for example. Also, replace all emails with empty strings. And if required, replace away newlines (\n) with single space (" "). Also replace all \' with '. For example: don\'t → don't. (This had not prove itself to give non-negligible improvements, and sometimes if not handled properly, worse results).
- Removal of data that are mostly composed of xxunk (a.k.a. OOV) with a threshold (originally we set 0.3, but after EDA it is planned to reduce to 0.1 threshold) to remove those texts that contains more than a certain percentage xxunk. (We didn't test this separately to see if there're any improvements or not, for training).
- There are also some preprocessing done by fastai's internals such as: lowering all capital cases into lower case (and put a tagging token xxmaj in front of it signifying it is capital letter), separating contractions, removing unnecessary spaces, tokenization with any tokenizer of your choice (we use Spacy Tokenizer), html are cleaned away with fastai predefined fixed rules, etc.