Sentiment Analysis
Due November 30, 2021

For this assignment, you will work with the Sentiment140 dataset. The description of the dataset is in the given link. You will be focusing on the first column containing the sentiment of the text (0 = negative, 2 = neutral, 4 = positive) and the last column which contains the text itself.

There are two primary tasks you need to complete for this assignment.

1. Pre-process and clean the data. Remove HTTP links and usernames.
2. Train two models for classification
    a. Fine-tune a pre-trained BERT model for sentiment analysis. You will use the BERT model provided by the HuggingFace library and add additional layers for classification.
    b. Train a Naive Bayes Classifier using CountVectorizer

You **can not** use a model that is already trained on Twitter data or has inbuilt classification layers. The goal of the assignment is to familiarize yourself with the process of extending pre-trained models for your downstream task.

You are **encouraged** to work in groups. If you choose to do so, mention your teammates in the report.

You will need to submit the following:

1. Details about the pre-processing (cleaning) step you performed.
2. A short description of the layers you added and your reasoning in addition to the training hyperparameters. (Optimizer, Learning rate, batch size, Loss Function,...)
3. The performance of your models (BERT and Bayes Classifier) on the Test-CSV provided with the dataset. For this, you can include a classification report. You are encouraged to use library functions to do this.

https://scikit-learn.org/stable/auto_examples/text/plot_document_classification_20newsgroups.html#sphx-glr-auto-examples-text-plot-document-classification-20newsgroups-py