Reddit moderators spend significant time removing the comments that are offensive, provocative or disrupts any quality discussions. In this assignment, you will train a model that takes the comment text as an input and predicts if it should be kept or removed. Dataset can be found at following link.

https://www.kaggle.com/areeves87/rscience-popular-comment-removal

For this assignment, only use the reddit_200k training and test set, and only use the “body” and “removed” columns.

Be careful about the encoding when loading the data. Pick an appropriate evaluation metric for imbalanced binary classification.

1. Create a baseline model using a bag-of-words approach and a linear model. [20 Marks]

2. Try using n-grams, characters, tf-idf rescaling and possibly other ways to tune the BoW model. Be aware that you might need to adjust the (regularization of the) linear model for different feature sets. [30 Marks]

3. Explore other features you can derive from the text, such as html, length, punctuation, capitalization or other features you deem important from exploring the dataset. [30 Marks]

4. Use a pretrained word-embedding (word2vec, glove or fasttext) instead of the bag-of-words model. Does this improve classification? [20 Marks]