The paper that we choose to implement as your baseline model is *NULI at SemEval-2019 Task 6: Transfer Learning for Offensive Language Detection using Bidirectional Transformers*. This paper uses a dataset from SemEval-2019 Task 6, which collected tweets through Twitter API and labeled them as offensive tweets or not.

They first clean the dataset by substituting Emoji, segmenting hashtags, and converting all the text into lower case. This solves the challenge of the characteristics of language on social media such as out-of-vocabulary words.

Then, they create three features. They first use cross-validate hyper-parameters of different vectorizers to build a bag of the word representation. Then, they utilize the word2vec and aggregate the maximum and average value in each dimension. Finally, they use the dictionary Hatebase API to aggregate the hate words in each category. With these three features, they train the logistic regression model and obtain a 0.4004 Macro-F1 score for ALL NOT, a 0.2494 Macro-F1 score for ALL OFF.

We choose this paper because this paper utilized a popular twitter dataset. We want to use this dataset to build our model and compare our result with the model in this paper. What’s more, this paper utilized logistic regression as their model. We think it will be a suitable baseline model since it is not a deep learning model. In Milestone 4, we could implement some deep learning models such as BERT and compare the result.

Citation:

Liu, Ping, Wen Li, and Liang Zou. "NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers." Proceedings of the 13th international workshop on semantic evaluation. 2019.