Results discussion

Due to time and resources constraints only a limited number of choices has been tested (like pre-trained BERT model or usage of only dropout and a linear layer in the model). The dataset could have been explored more carefully to identify using simple machine learning models (random forest, XGBoost, …) to detect categories based on predefined list of words. In addition to that, 3 categories (severe_toxic, threat and identity_hate) contained very small number of samples, therefore, either more samples could have been collected for these categories or some augmentation techniques could be used.

From the confusion matrix results we can see that the model detects most of the samples correctly, however, the F1 score is low (macro being 56.76), as for rare categories model performed poorly, especially for 'threat' category with only 2 samples as true positive, even though the dataset has been split proportionally and training/validation/testing datasets had the same proportion of rare categories. The total number of samples where this category is marked as 1 have been 475, which is much smaller than for any other dataset. Some sort of augmentation technique can be used, or a separate type of machine learning technique can be used to get prediction for this category. Other categories performed better with "toxic", "obscene" and "insult" having the highest relative number of true positive and true negative. "Identity_hate" and "severe_toxic" had smaller number of true positive than either false positive or false negative, even though their true positive value still being high.

As a future work, a different pre-trained embedding can be used, which focuses more on emotions or anger. Additionally, more complex model should be created, as having only dropout with linear network makes it hard for the model to learn anything. The dataset can be extended, or a separate machine learning/deep learning model can be applied for specific categories.