# Berehulia Mykyta – NLP

To solve this problem I've decided to fine tune all ready existing model. I've selected xlm-roberta-large as it's multi lingual and robust. I've added Stratified K-Fold with 5 splits to ensure that the percentage of samples for each class is preserved in every fold.

Dataset has a significant class imbalance so I decided to use oversampling artificially increasing the presence of rare classes up to 250.

I've first started with a standard Cross Entropy Loss and next moved to the Focal Loss to down-weights "easy" examples "hard" examples. Also, I calculated class_weights and passed them into the loss function. This should assign a higher penalty for misclassifying minority classes.

I've also tried to find best_weights based on Out-Of-Fold (OOF) probabilities, as standard argmax (taking the class with the highest probability) is suboptimal for the Macro F1 metric with imbalanced data.

The submission attached is for 0.67210 score. I hasn't preserved the submission for 0.71251.