# Independent Study Report

Zubair Abid (20171076)

## Objectives

The objective of the Independent Study was to work on Fine and Coarse-grained hate speech detection, and submit results to HASOC at FIRE 2020. I have a background in NLP and some prior knowledge of Deep Learning-based classification projects from earlier courses, and this project provided me with the experience of trying it out on more real-world tasks to get near-state of the art results. I worked in a team of four, along with Sayar Ghosh Roy, Ujwal Narayan, and Tathagata Raha.

## Results

The results we obtained on the provided dataset are given in table 1. Across multiple experiments (with several early-stage ones not mentioned in the final submission), we found that fine tuning the pre-trained XLM-RoBERTa transformer weights worked best for our task, significantly outperforming baselines with frozen transformer weights.

Task 1 (Subtask A on the website) is on coarse-grained classification distinguishing between hate speech and otherwise, whereas Task 2 (Subtask B) is fine-grained classification, where hate speech can be offensive, profane, or just hateful.

Part of the final scoring was done with an unrevealed test dataset. Over the six leaderboards, we did reasonably well in English – as the only team to have top 5 in both subtasks – and also in Hindi, subtask B. Our performance relative to other teams is given in table 2.

## Work done

### Initial work

We started off with a literature review of work submitted at the previous iteration of the workshop to get an idea of what could be done. The best performers had almost done some form of fine-tuning on pretrained large neural models such as BERT and XLM-RoBERTa. A more general literature review confirmed the same.

| **Model** | **English** | | **German** | | **Hindi** | |
|---|---|---|---|---|---|---|
| | Task 1 | Task 2 | Task 1 | Task 2 | Task 1 | Task 2 |
| XLMR-freeze-mono | 83.92 | 52.38 | 66.85 | 41.52 | 68.25 | 40.45 |
| XLMR-freeze-multi | 82.02 | 51.02 | 68.34 | 48.60 | 66.27 | 41.59 |
| XLMR-adaptive | **90.29** | 59.03 | 81.04 | 52.99 | **75.40** | 45.87 |
| XLMR-tuned | 90.05 | **60.70** | **81.87** | **53.28** | 74.29 | **49.74** |

Table 1: Performance of the Transformer-based Models (Best results highlighted in bold)

| Task | F1 Macro Average | Position/Participants |
|---|---|---|
| English Task 1 | 50.67 | 4/36 |
| English Task 2 | 25.28 | 4/26 |
| German Task 1 | 50.36 | 12/25 |
| German Task 2 | 25.42 | 12/19 |
| Hindi Task 1 | 49.43 | 17/24 |
| Hindi Task 2 | 26.12 | 4/17 |

Table 2: Final performance

To set baselines, we initially ran a number of experiments with pretrained embeddings, without any fine-tuning. I took care of the work with German, where I tried using various classifiers over pretrained weights from UKPLab's sentencetransformers. I used BERT-based models – explicitly trained for German, and a general multilingual one, with both cased and uncased variants. Various ablations were run with differing features, provided by Sayar as he was in charge of data preprocessing. A grid-search was run as well to optimise for hyperparameters. The features used were hashtags and emojis, in all their permutations along with the tweet text.

## Submission work

The final submissions were based on Ujwal and Sayar's work that setup the training loop over the transformer models. I ran the experiments on multiple models – Multilingual Distilbert and XLM-RoBERTa Base – with various hyperparameters and adaptive learning turned on and off, to get improved results.

As the evaluation was to be done on private data as well, we needed to submit the code in easily runnable scripts. Tathagata and I worked on making the final experiment code more portable, and according to the submission format.