A BERT-Based Transfer Learning Approach for Multi-lingual Hate Speech Detection and Text Classification

Deep Natural Language Processing

Politecnico di Torino

A.Y. 2021-22

Hafez Ghaemi s289963
Juan José Márquez Villacís s287313
Ionut Cosmin Nedescu s292495

Professor Luca Cagliero
Dr. Moreno La Quatra

Outline

- Problem statement and introduction
- Methodology
- Experiments
  - Implementation and results
  - Extensions
- Discussion
- Conclusion

Problem statement and introduction

Social networking platforms such as Twitter, Meta, and Reddit allowed people to express their opinion.

↓

Despite the advantages in communication and information propagation, problems such as **hate speech** emerge that need to be addressed.

Problem statement and introduction

- We replicate the fine-tuning strategies and experiments by Mozafari et al. [1]
- The first work to propose a transformer-based fine-tuning strategy for hate speech detection; previous efforts in the pre-transform era focused on traditional NLP, and deep learning techniques
- We propose two extensions:
  1. Apply the proposed fine-tuning strategies to other two text classification datasets:
     - IMDB review dataset for sentiment classification
     - Dataset for determining the eligibility of cancer patients for clinical trials
  2. Adaptation of the fine-tuning models to be compatible with multi-lingual hate speech detection and evaluating the model on tri-lingual hate speech dataset

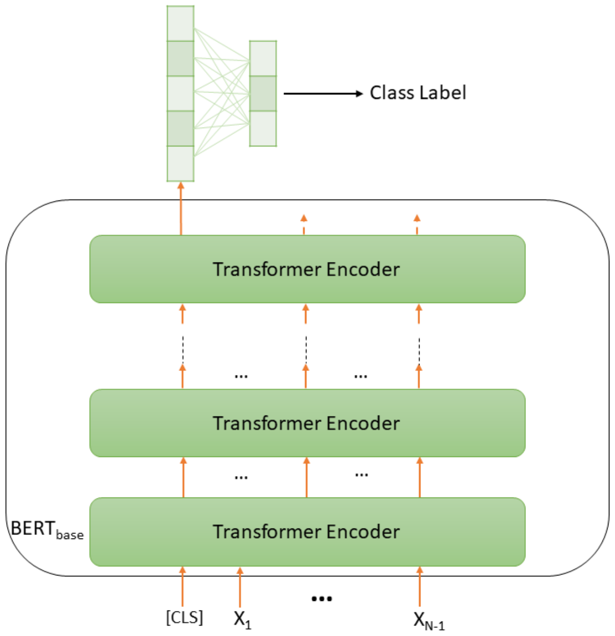
[1] M. Mozafari, R. Farahbakhsh, and N. Crespi, “A bert-based transfer learning approach for hate speech detection in online social media,” in International Conference on Complex Networks and Their Applications. Springer, 2019, pp. 928–940.

Methodology - Architectures Employed

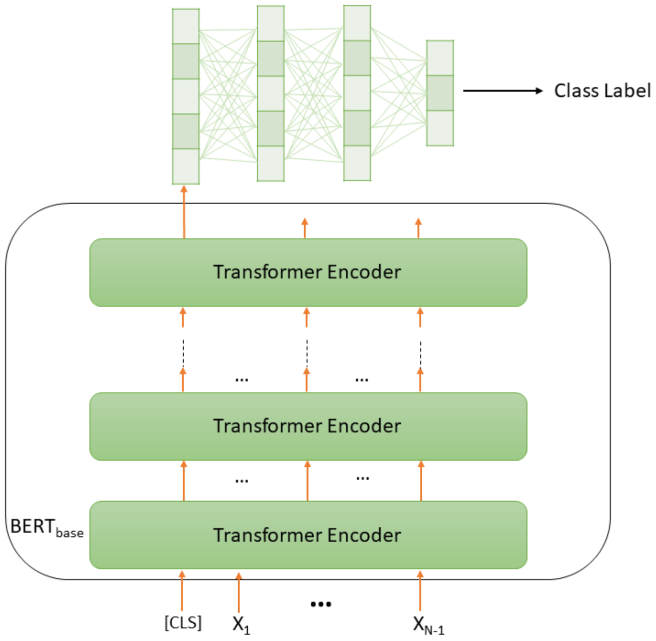
- For all the architectures, we use BERT as the backbone network.
- To replicate of the experiments by Mozafari et al.:
  - We use $\text{BERT}_{\text{base}}$ that is pre-trained on English Wikipedia (2500M tokens) and Book Corpus (800M tokens).
  - BERTbase has 12 layers (transformer blocks), 12 self-attention heads with embeddings of size 768
- For multi-lingual tasks, we use multi-lingual BERT pre-trained on 104 languages with the largest Wikipedia.

Methodology – Architectural Modifications (Fine-tuning Strategies)

A) Single linear layer

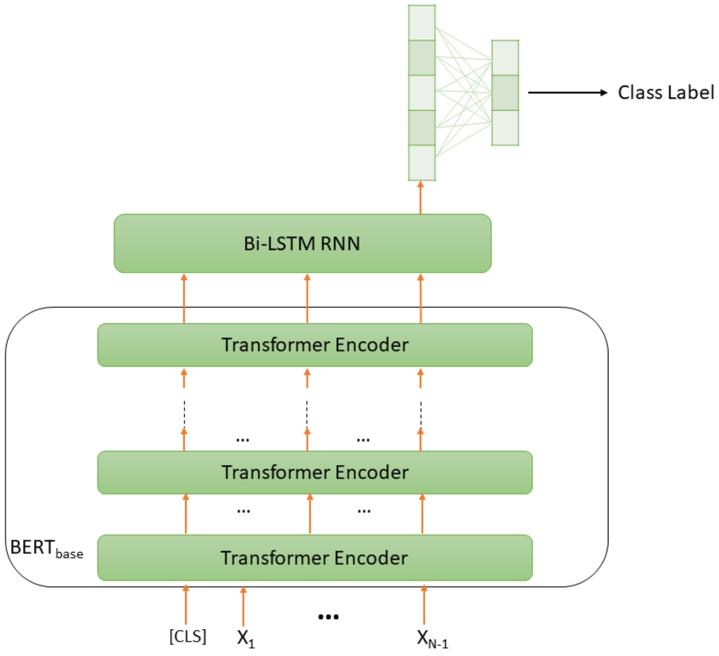


B) Multi-layer ANN

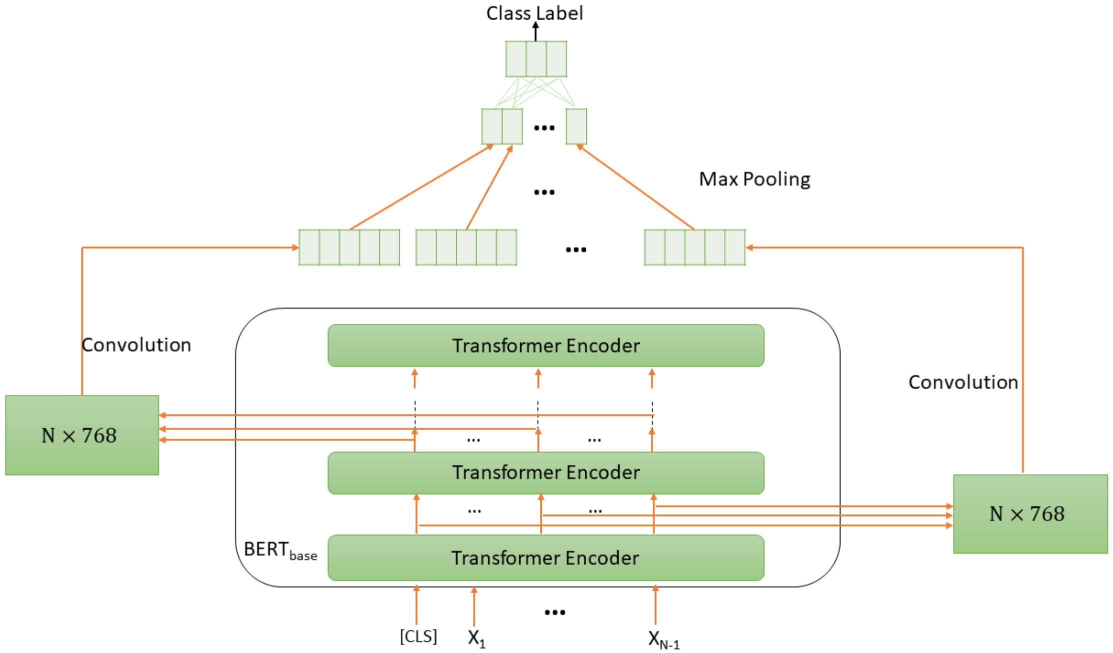


Methodology – Architectural Modifications (Fine-tuning Strategies)

C) Sequential Bi-LSTM

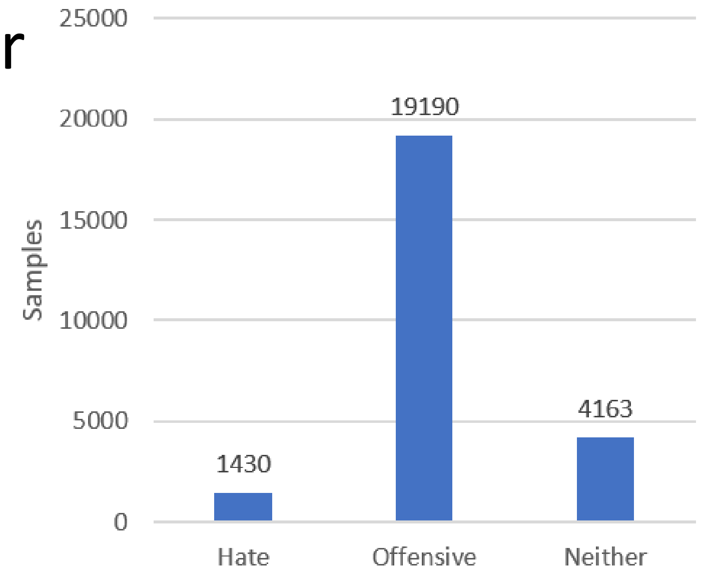


D) CNN Layer



Experiments - Datasets

- For hate speech detection, we use the dataset proposed by Davidson et al.
  - 25k random sampled tweets from 84.4 million containing hate speech words and phrases
  - 3 different classes: hate, offensive, neither



Experiments – Pre-processing

- Replacement of some textual components with their corresponding BERT tokens
    - mentions of users -> <user>
    - numbers -> <number>
    - hashtags -> <hashtag>
    - links -> <url>
    - emoticons -> <emoticon>
- Removal of punctuation marks and invalid characters
- All text samples converted to lower case

Experiments – Implementation

- For each experiment the classifier is trained with the following parameters:

Parameter	Value
# of epochs	3
Batch size	32
Dropout	0.1
Learning rate	2e-5
BERT tokenizer max length	64

- Due to dataset imbalance of the Davidson dataset, we performed stratified sampling to avoid overfitting, and separated the data into 3 sets with proportional class distribution:
  - 80% training data
  - 10% validation data
  - 10% test data

Experiments – Results on Davidson Hate Speech Dataset

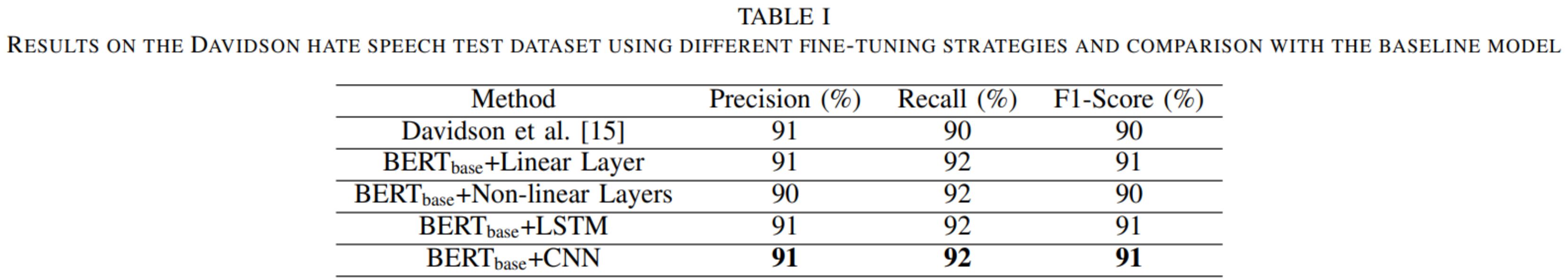
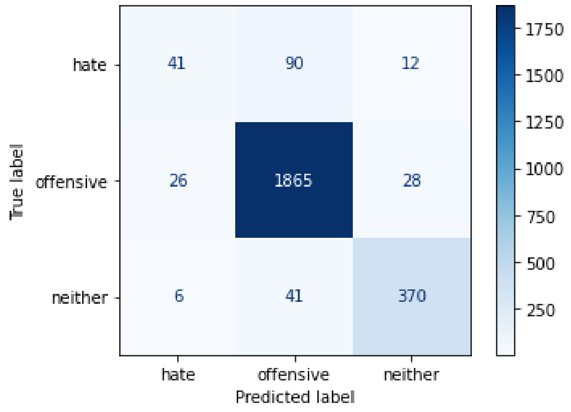


TABLE I
RESULTS ON THE DAVIDSON HATE SPEECH TEST DATASET USING DIFFERENT FINE-TUNING STRATEGIES AND COMPARISON WITH THE BASELINE MODEL

Method	Precision (%)	Recall (%)	F1-Score (%)
Davidson et al. [15]	91	90	90
BERT$_{\text{base}}$+Linear Layer	91	92	91
BERT$_{\text{base}}$+Non-linear Layers	90	92	90
BERT$_{\text{base}}$+LSTM	91	92	91
BERT$_{\text{base}}$+CNN	**91**	**92**	**91**

The combination of BERTbase + CNN performs slightly better than the other combinations.

On the right side, you can see the confusion matrix for this architecture.



Extension 1 – Text Classification for Sentiment Analysis and Clinical Trial Eligibility Detection

- Sentiment Classification Dataset
  - The dataset contains IMDB movie reviews annotated with positive or negative classes [1]; it has a balanced class distribution and 25k train and 25k test samples
  - Use 20k samples for training and 5k for test
  - Hyperparameters are the same as the original experiment, and only the max length of BERT tokenizer is set to 128 (reviews are longer than tweets).

Results in terms of F1-score

Method	F1-score
BERTbase + Linear layer	97%
BERTbase + Non-linear layers	88%
BERTbase + Bi-LSTM	94%
BERTbase + CNN	97%

[1] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics

Extension 1 – Text Classification for Sentiment Analysis and Clinical Trial Eligibility Detection

- 2nd dataset:
  - The dataset is taken from a Kaggle competition, the objective is to predict if a patient diagnosed with some type of cancer is eligible for a treatment given its intervention studies and diagnoses from clinical trials
  - The dataset consists of 500k samples belonging to each class
  - Max length of the BERT tokenizer is set to 50 (due to the large dataset and computational limits).

Results in terms of F1 – score

Method	F1-score
Baseline (CNN) [1]	93.00%
BERTbase + Linear layer	95.31%
BERTbase + Non-linear layers	95.60%
BERTbase + Bi-LSTM	95.45%
BERTbase + CNN	95.29%

[1] A. P. Aurelia Bustos, "Learning eligibility in cancer clinical trials using deep neural networks," 2018

Extension 2 – Multi-lingual Hate Speech Detection

- We use multi-lingual BERT instead of BERT$_{base}$ in all the architectures.
- We evaluate the multi-lingual models on the dataset HASOC 2020 competition.
- The dataset is made up of 10k doubly annotated tweets in 3 Indo-European languages (English, German, Hindi)

- We solve 2 sub-tasks:
  1. Binary classification into hate and offensive (HOF) or non-hate-offensive (NOT) (a binary classification problem).
  2. Fine-grain classification of the HOF class into 3 classes named hate speech (HATE), offensive (OFFN), and profane (PRFN) (a classification problem with four classes).

Extension 2 – Multi-lingual Hate Speech Detection

- The dataset is unbalanced in terms of hate speech classes and almost balanced in terms of languages
- We again use stratified sampling
  - 80% training data
  - 10% validation data
  - 10% test data
- We use the same hyperparameter configurations as the original experiment

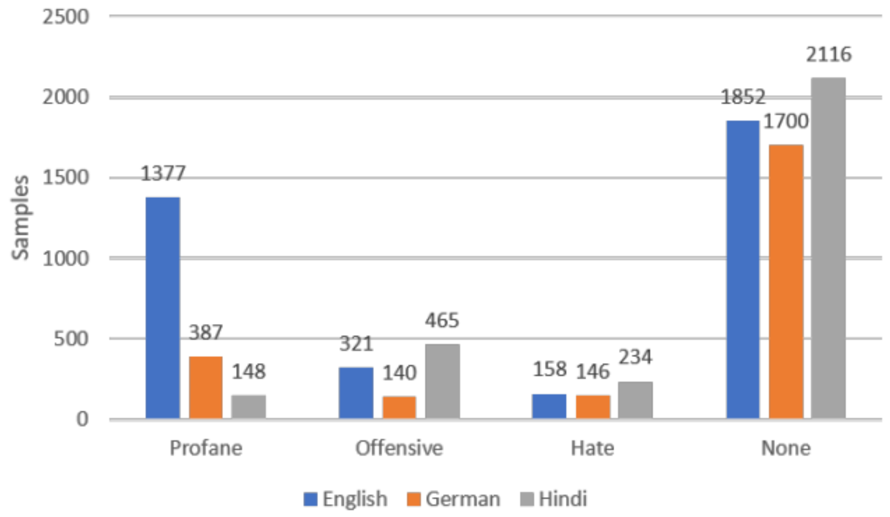


Fig. 3. HASOC multi-lingual hate speech dataset class distribution

Extension 2 – Multi-lingual Hate Speech Detection - Results

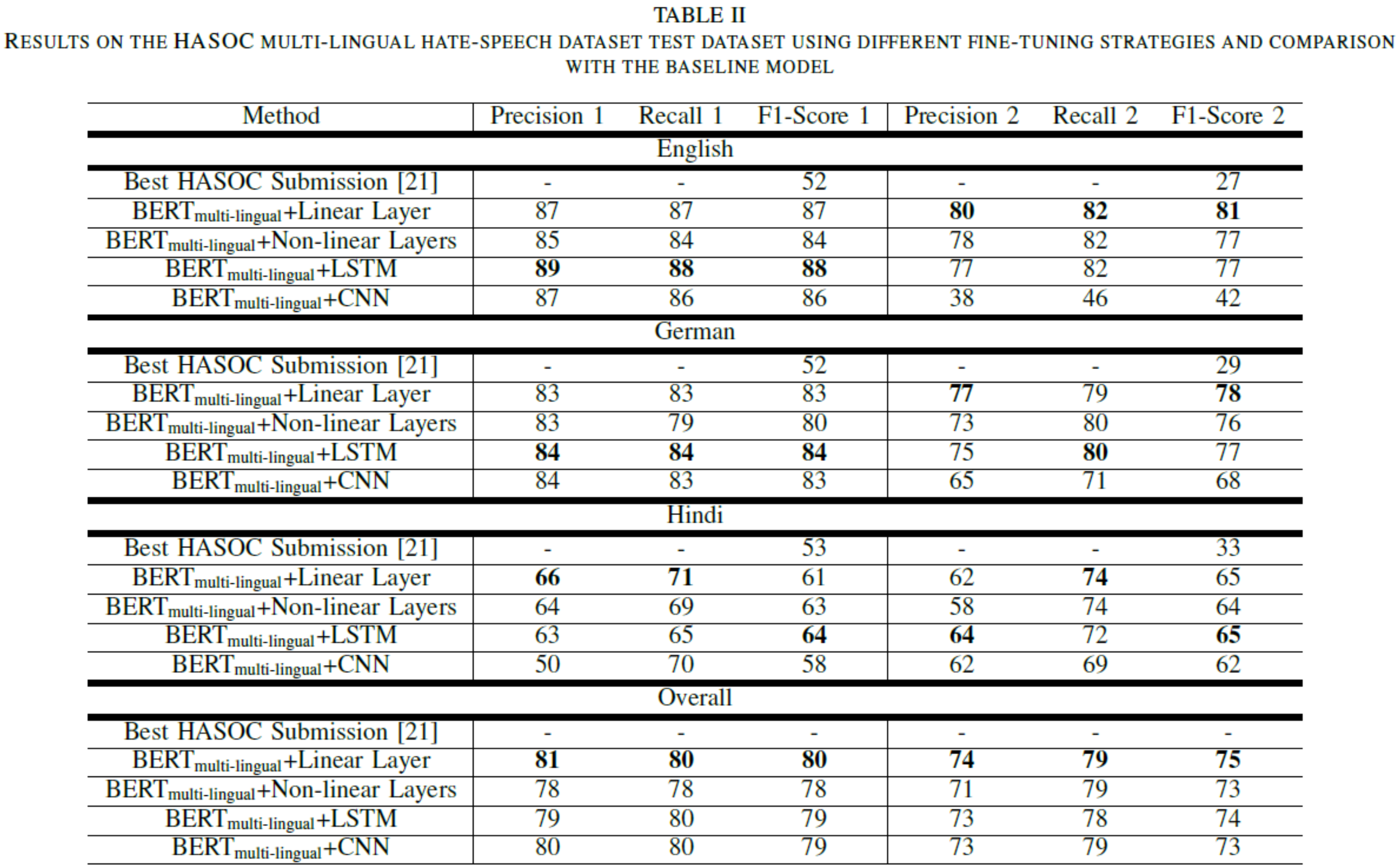
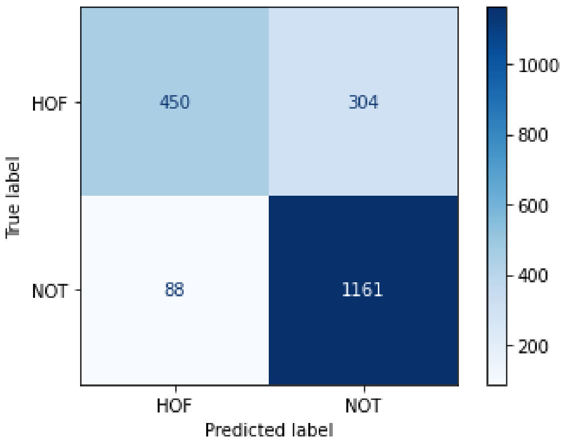


TABLE II

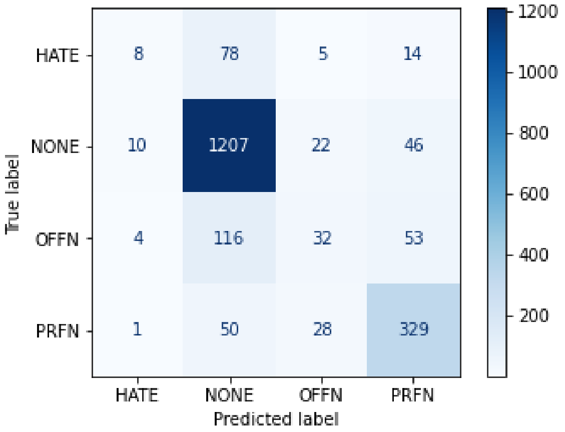
RESULTS ON THE HASOC MULTI-LINGUAL HATE-SPEECH DATASET TEST DATASET USING DIFFERENT FINE-TUNING STRATEGIES AND COMPARISON WITH THE BASELINE MODEL

Method	Precision 1	Recall 1	F1-Score 1	Precision 2	Recall 2	F1-Score 2
English						
Best HASOC Submission [21]	-	-	52	-	-	27
BERT$_{\text{multi-lingual}}$+Linear Layer	87	87	87	**80**	**82**	**81**
BERT$_{\text{multi-lingual}}$+Non-linear Layers	85	84	84	78	82	77
BERT$_{\text{multi-lingual}}$+LSTM	**89**	**88**	**88**	77	82	77
BERT$_{\text{multi-lingual}}$+CNN	87	86	86	38	46	42
German						
Best HASOC Submission [21]	-	-	52	-	-	29
BERT$_{\text{multi-lingual}}$+Linear Layer	83	83	83	**77**	79	**78**
BERT$_{\text{multi-lingual}}$+Non-linear Layers	83	79	80	73	80	76
BERT$_{\text{multi-lingual}}$+LSTM	**84**	**84**	**84**	75	**80**	77
BERT$_{\text{multi-lingual}}$+CNN	84	83	83	65	71	68
Hindi						
Best HASOC Submission [21]	-	-	53	-	-	33
BERT$_{\text{multi-lingual}}$+Linear Layer	**66**	**71**	61	62	**74**	65
BERT$_{\text{multi-lingual}}$+Non-linear Layers	64	69	63	58	74	64
BERT$_{\text{multi-lingual}}$+LSTM	63	65	**64**	**64**	72	**65**
BERT$_{\text{multi-lingual}}$+CNN	50	70	58	62	69	62
Overall						
Best HASOC Submission [21]	-	-	-	-	-	-
BERT$_{\text{multi-lingual}}$+Linear Layer	**81**	**80**	**80**	**74**	**79**	**75**
BERT$_{\text{multi-lingual}}$+Non-linear Layers	78	78	78	71	79	73
BERT$_{\text{multi-lingual}}$+LSTM	79	80	79	73	78	74
BERT$_{\text{multi-lingual}}$+CNN	80	80	79	73	79	73

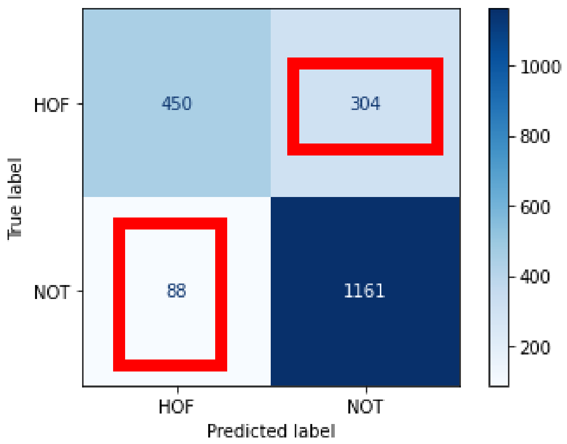
Sub-task 1

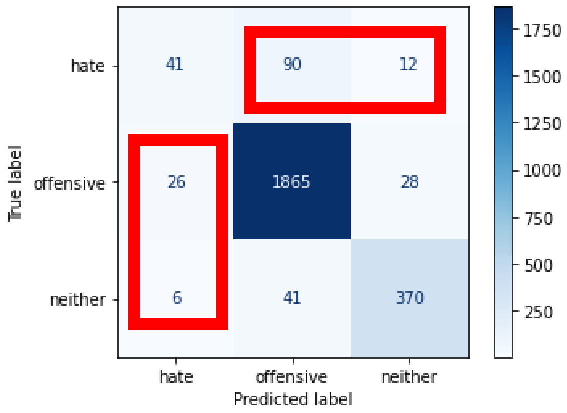


Sub-task 2



Discussion – The Problem with a High False Negative and False Positive Rate





- The large number of false positives (FP) and false negatives (FN) for the hate class can be problematic.
- The presence of FP samples would fringe upon the users’ freedom of expression
- The presence of FN samples may allow hateful comments to remain and propagate through the social network.
- To create a clearer picture, we performed a manual inspection of some of the misclassified samples in the English sub-set of the HASOC dataset.

Discussion – Manual Inspection of Misclassified Samples

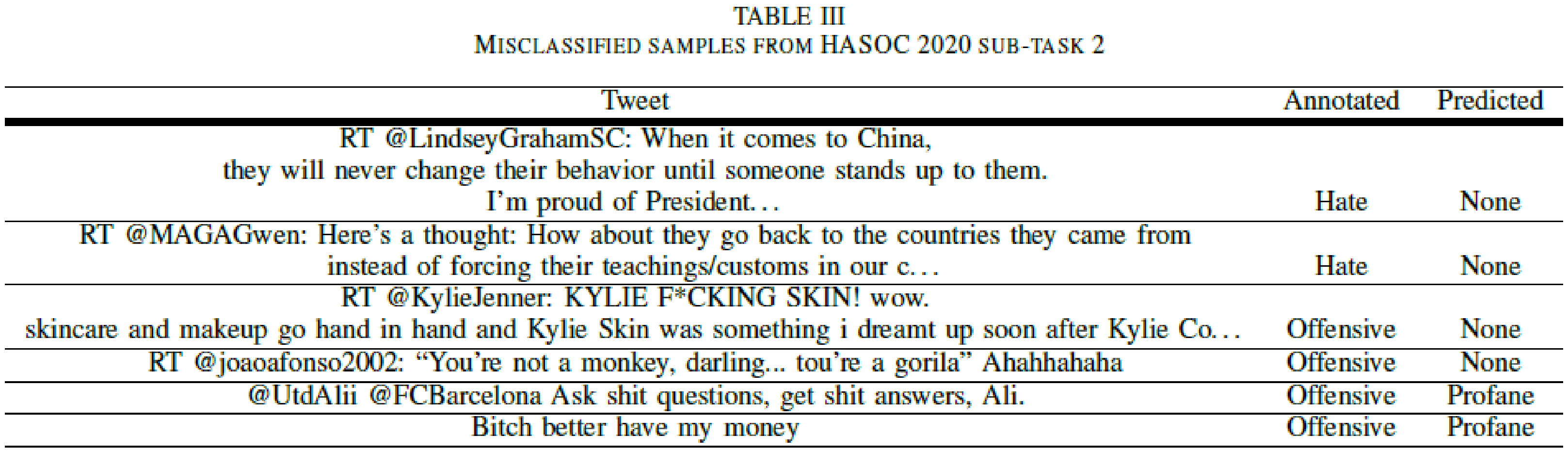


TABLE III
MISCLASSIFIED SAMPLES FROM HASOC 2020 SUB-TASK 2

Tweet	Annotated	Predicted
RT @LindseyGrahamSC: When it comes to China, they will never change their behavior until someone stands up to them. I'm proud of President…	Hate	None
RT @MAGAGwen: Here's a thought: How about they go back to the countries they came from instead of forcing their teachings/customs in our c…	Hate	None
RT @KylieJenner: KYLIE F*CKING SKIN! wow. skincare and makeup go hand in hand and Kylie Skin was something i dreamt up soon after Kylie Co…	Offensive	None
RT @joaoafonso2002: "You're not a monkey, darling... tou're a gorila" Ahahhahaha	Offensive	None
@UtdAlii @FCBarcelona Ask shit questions, get shit answers, Ali.	Offensive	Profane
Bitch better have my money	Offensive	Profane

- In rows one, two, and four, the model has failed to capture the complex semantics that led to desired annotation. There are no hateful or offensive words used in these tweets, and the model should rely entirely on the context and semantics to detect the right class. **Possible solution: a semantically richer training set**
- In the third sample, the swear word covered with the star may not have been recognized by the language model which resulted in a false prediction. **Possible solution: more obscured offensive words (such as covered with stars) in samples may help the model**
- Some of the tweets could belong to two or more classes (such as the last two rows). Therefore, the language model may sometimes predict a correct label that may be different from the annotation due to the limitations in the dataset's labeling mechanism. **Possible solution: multi-label samples**

Conclusion

- We reproduced one of the first works that utilized transformer-based architectures and fine-tuning strategies for the problem of hate speech detection.
- We also showed that these proposed fine-tuning techniques can be easily applied to other sub-tasks of text classification and achieve a great performance, which shows the robustness of the proposed fine-tuning strategies.
- The BERT backbone in the models considered can also be substituted with its multi-lingual version to address multi-lingual hate speech detection tasks.
- Finally, we performed a short error analysis using confusion matrices and manual inspection of misclassified samples.

Thank You!