**Shared Task – Report**

**Jessica & Rebecka**

**Preprocessing**

**Choice of Method: Transformer**

Transformer-based models are nowadays considered as the state-of-the-art models for various languages and have showed great success in sentiment analysis. However, the Arabic language imposes many challenges, also due to the various dialects and scarcity of resources and its morphological complexity.

**Hyperparameter settings**

We set the following Hyperparameters:

* **Set seed - 42**

This is just to set a randomizer that we get the same ‘randomization’ result

* **Optimizer – AdamW**

We used an adaptive optimizer, like Adam but an improved version called AdamW. Literatures suggest that AdamW yields better training loss and that models generalize much better than models trained with Adam.

(see: <https://towardsdatascience.com/why-adamw-matters-736223f31b5d> )

* + Learning rate for optimizer: **5e-5** got the best result: AUC 73%, with 1e-10 we got a lower accuracy
  + Epsilon value for optimizer: **1e-8** just needed some small value
* **Epochs - 4**

We first choose an epochs of 4 where we got a accuracy of 80,87%. Which was quite good.

* **Loss function – Cross Entropy loss**

We used Cross Entropy loss when adjusting model weights during training. With the Cross Entropy loss function the model learns how to map the input to the realistic probability output. In our case we have a classification problem, with known class labels (OFF=1|NOT=0), so the model needs to calculate the likelihood that a given example belongs to each class label. The difference of those two probability distributions can be calculated using cross-entropy.

* **BERT model – mini language**

We used the Huggingface pretrained BERT mini language model for Arabic (developed by asafaya) called ***asafaya/bert-mini-arabic*.** Arabic BERT is a set of BERT language models that consists of four models of different sizes trained using masked language modeling with whole word masking. Models with  
large, base, medium, and mini sizes were trained with the same data for 4M  
steps. We decided to use the***mini size***as training our model is shorter.

* **Tokenizer – Arabic tokenizer (using pre-trained model)**

Due to the rich and complex morphology in the Arabic language tokenization is quite challenging. Usually, one tokenizes by the space between letters (English). That’s why we used an Arabic tokenizer, and because we are using BERT-mini-arabic we used [***AutoTokenizer***](https://huggingface.co/docs/transformers/v4.20.0/en/model_doc/auto#transformers.AutoTokenizer)to load the ***pretrained model and its associated tokenizer.*** With AutoTonkenizer we instantiate the tokenizer with the same model name to ensure the tokenizer is using the same tokenization rules a model was pretrained with.

* **Batch size - 16**

For fine-tuning BERT, some authors recommend a batch size of 16 or 32. We decided to go for 16 as we are having a small model and also due to our small learning rate.

**Results on development data**

**Encountered problems**

Unsure how to deal with emojis

**Methods and steps for further result improvement**

For further result improvement I would recommend using maybe even a more specialized token, plus maybe it would be helpful to translate the emojis into sentences/words.