Preliminary results for building counterspeech classifiers

The counterspeech models are intended to classify a piece of text into 3 different classes:

1. Hateful (hateful/not hateful/unclassifiable hateful)
2. Constructive (not constructive/constructive/unclassifiable constructive)
3. Agree (agree/disagree/unclear)

Two different models are built. A model without reference (a baseline model that uses a single piece of text, a reply tweet, as input) and a model with reference (which also includes the reference tweet to the reply tweet as input)

The dataset consists of 6480 English tweets pairs, divided into a 80/20 training/test split.

The xlm-roberta-large model is used as a base model.

Models are trained on a batch size of 32, with Adam optimizer, a learning rate of 2e-5 for 5 epochs.

For the model without reference, accuracy for ‘Hateful’ is approx. 83%, for ‘Constructive’ 86 and for ‘Agree’ 66, yielding average accuracy of 78 (Varying slightly pr. epoch and for each run)

Et billede, der indeholder diagram, linje/række, Kurve, tekst

Automatisk genereret beskrivelse

For the model with reference, accuracy for ‘Hateful’ is approx. 83, for ‘Constructive’ 88 and for ‘Agree’ 70, yielding average accuracy of 78.

Et billede, der indeholder diagram, linje/række, Kurve, tekst

Automatisk genereret beskrivelse

Thus, including the reference tweets as a predictor does, as expected, improve performance for the ‘agree’ class.

Models starts overfitting after approx. 4 epochs

The models are better at correctly assigning the disagree label in the ‘agree’ class than agree and unclear. Similarly, the models are also often slightly better at correctly assigning the not hateful label in the ‘hateful class.

When applied to data, performance seems decent:

Et billede, der indeholder skærmbillede, tekst

Automatisk genereret beskrivelse

More GPU is needed for finetuning the models on Danish data, as well as test other base models or improve models otherwise.