Evaluating catastrophic forgetting of state-of-the-art NLP models in predicting moral values.

CSE3000 Research Project by Florentin Arsene
f.arsene@student.tudelft.nl

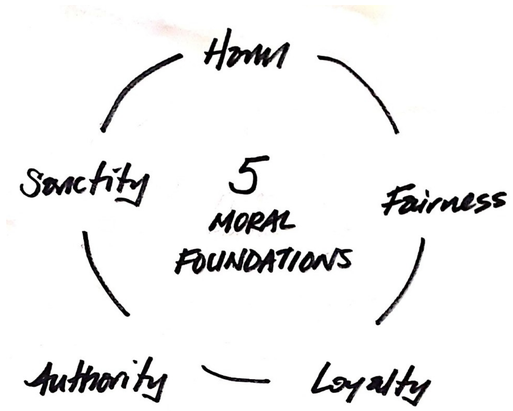
Contact details:

Background

Why would we want to predict moral values?

1. Personal moral values drive people's day-to-day actions.
2. Improve collaboration between AI and humans.

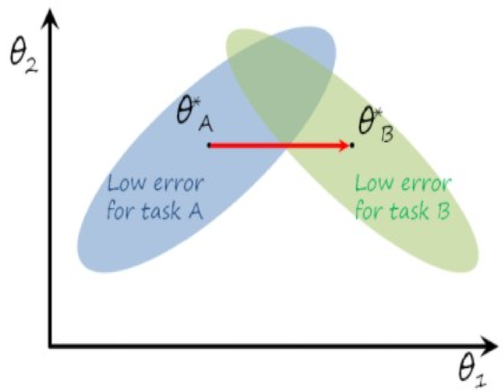
The moral values are analyzed according to the Moral Foundation Theory



Research Question

Why evaluate catastrophic forgetting?

1. Big weakness of Deep Neural Networks.
2. Useful in real world applications aimed at sequentially learning new tasks, while not forgetting to perform old tasks.



Methodology

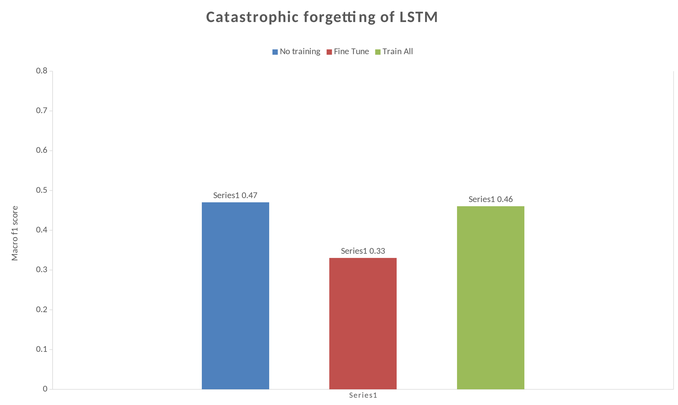
- Preprocess 7 datasets of tweets, corresponding to 7 socially relevant domains: ALM, BLM, Baltimore Protests, hate speech(Davidson), 2016 Presidential election, #MeToo movement and Hurricane Sandy.
- Implement 3 models:
  - LSTM
  - fastText
  - BERT
- Pre-train the models on 6 datasets, then sequentially train on the 7th dataset. Repeat this 7 times, each of the 7 datasets being, in turn, the new dataset.
- Evaluate catastrophic forgetting of the models.

Experiments

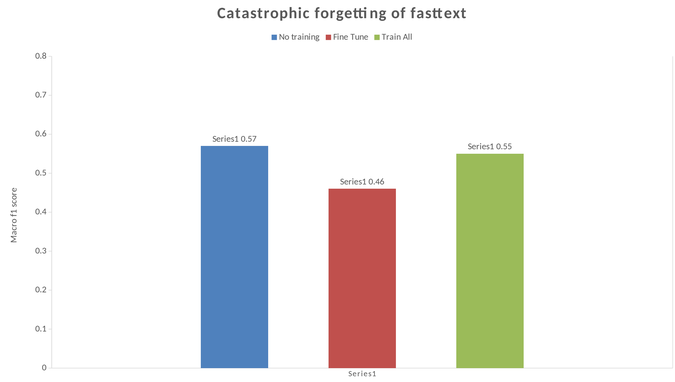
Experiments assume a model is pre-trained on 6 datasets and received a new dataset that it wants to train on. We have 3 types of experiments:

- No training: The model does not train on the new dataset.
- Fine Tune: The model trains on the new dataset.
- Train All: The model trains on a combination of the new dataset and an amount of old pre-training data equal to the new dataset size.

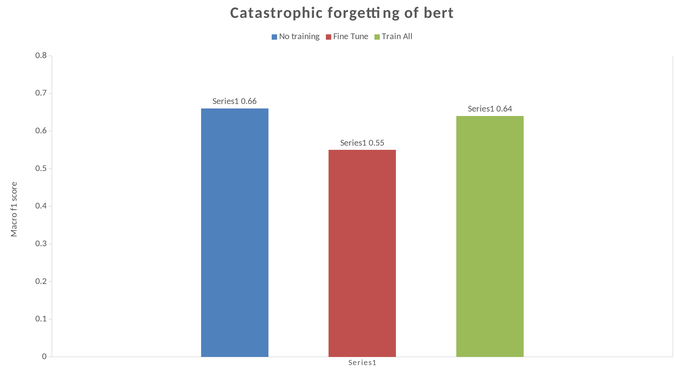
LSTM results



fastText results



BERT results



Conclusion

- BERT outperforms all the models in terms of performance, taking a lot of time to train.
- fastText is the fastest in terms of training time, however it is hard to configure the model.
- Catastrophic forgetting occurs irrespective of the model.
- Catastrophic forgetting can be mitigated by training on a combination of old data with new data.

Supervised by Enrico Liscio and Pradeep Murukannaiah

Contact details: {e.liscio, p.k.murukannaiah}@tudelft.nl

TU Delft