SMART-mDeBERTa

Nina Bodenstab, Alexis Limozin, Joachim Despature

Group 14

Problem definition

Hate speech represents a significant challenge to societal cohesion and personal dignity. To address this pressing issue, a hate speech detector was developed as part of EE-559 "Deep Learning". Leveraging sophisticated methodologies such as the deBERTa model with SMART fine-tuning, the project focuses on effectively identifying and addressing instances of hate speech. By doing so, it aims to foster a more inclusive and respectful digital discourse landscape, promoting the fundamental values of equality and tolerance in our interconnected world.

Key Related Works

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- DeBERTa: Decoding-enhanced BERT with Disentangled Attention
- SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization

Method

- Pretrain mDeBERTa NLP model using SMART fine-tuning
- Memory problems:
  - Gradient Accumulation
  - Scaler (type conversion)
  - Cache Clearing

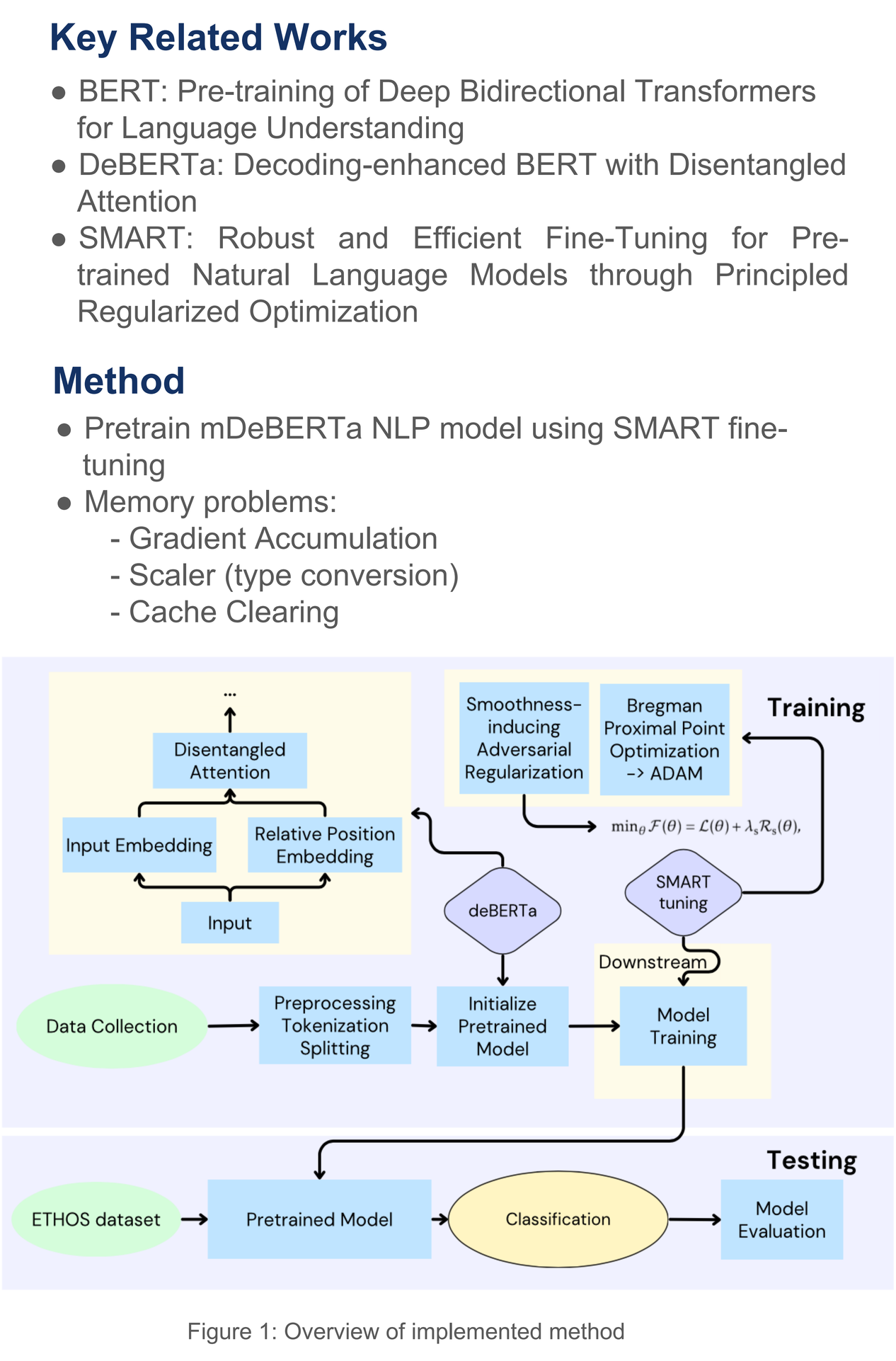


Figure 1: Overview of implemented method

Datasets

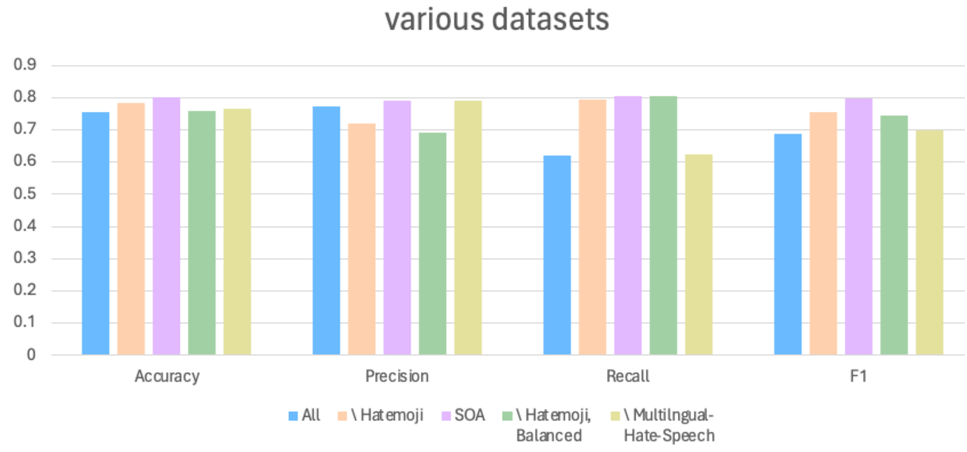
- Call Me Sexist but
- Online Misogyny
- Toxygen
- Detecting Offensive Statements towards Foreigners in
- Social Media Dataset
- En_dataset (MultiIngual-Hate-Speech)
- Fr_dataset (MultiIngual-Hate-Speech)
- Ar_dataset (MultiIngual-Hate-Speech)

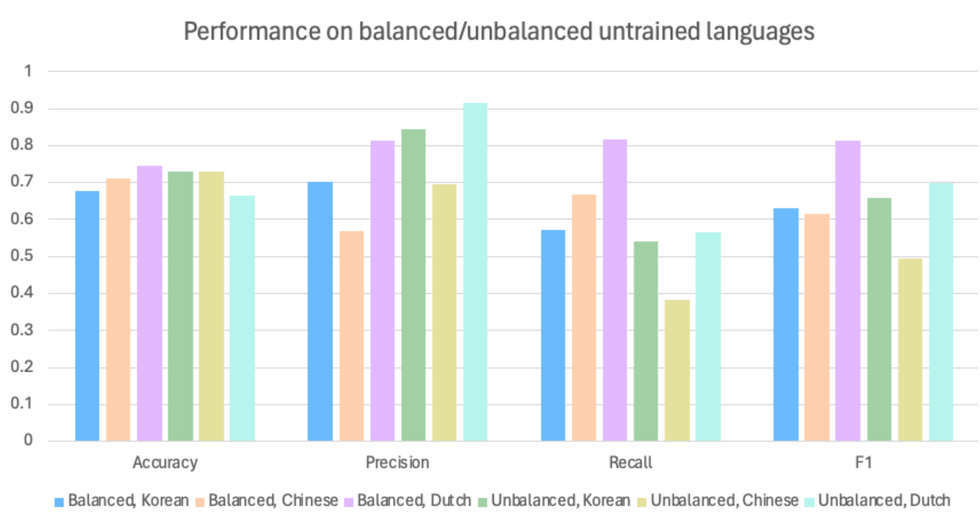
EE-559: Deep Learning, 2024

Validation

Method

- Dataset pre-processing: remove URLs, numbers, stopwords, empty rows, repeated character
- Performance testing on ETHOS dataset –> leaderboard
- Variation of training epochs to prevent overfitting
- Data balancing done by dropping excess non-hate data
- Performance metrics: Accuracy, Precision, Recall, F1
  →F1 most indicative





Discussion

- Best performance when pretraining on balanced datasets \hatemoji
- Very close performance to best existing model for ETHOS
- Remarkable at Dutch
- Knowledge transfer

Possible improvements

- Standardization of multilingual datasets
- Hyperparameter tuning
- Reduce computational cost

Limitations

- Computational Resources
- Data Imbalance and Labelling Bias
- Ethical Considerations (privacy and misuse)
- Adaptability of model
- Performance Evaluation
- Standardization of datasets across languages

Conclusion

We successfully developed a deep learning model that helps foster healthier online interactions by automatically identifying hate speech across diverse content formats. It is designed to prioritize accuracy and context comprehension, ensuring they differentiate between harmful hate speech and legitimate critical discourse or satire. This is done by combining the mDeBERTa NLP model with SMART finetuning. The F1 score obtained almost matches the performance of the SOA on ETHOS dataset.

References,

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[2] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," arXiv.org, Jun. 05, 2020. https://arxiv.org/abs/2006.03654

[3] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization," Cornell University, Jan. 2020, doi: 10.18653/v1/2020.acl-main.197.