SMART-mDeBERTa

Limozin Alexis [1] **Bodenstab Nina** [1] **Despature Joachim** [1]

Abstract

A Hate Speech classifier was implemented in the EPFL EE-559 "Deep Learning" course using advanced deep learning techniques. The project emphasized ethical considerations in addressing Hate Speech, defined by the UN as offensive discourse targeting individuals based on characteristics like race, religion, or gender. The classifier used the multilingual DeBERTa V3 model and SMART fine-tuning, achieving high performance using training datasets in English, French, German, and Arabic and demonstrating high zero-shot performance in Dutch. Preprocessing ensured balanced training and testing, while scalers and autocasting floats optimized computational efficiency.

Keywords: Hate classification, Multilingual, SMART fine-tuning, mDeBERTa v3

1. Introduction

In the context of EE-559 "Deep Learning," we developed a Hate Speech classifier, guided by research ethics, notably upheld by the Human Research Ethics Committee (HREC). Hate Speech, as defined by the UN, targets individuals or groups based on intrinsic traits, posing a threat to social harmony. It encompasses any communication: oral, written, or behavioral. It does so by using discriminatory language regarding religion, ethnicity, nationality, race, gender, or other identity factors. Recognizing its pervasive impact, we implemented a Hate Speech detector within the class to address its harmful effects.

2. Related Work

In our project, we develop a Hate Speech detector, diving into the realm of Natural Language Processing (NLP). A game-changer in NLP was Bidirectional Encoder Representations from Transformers (BERT)[1], which transforms text into numerical tokens. Its successor, RoBERTa [2], fine-tunes BERT's flaws, enhancing performance with dynamic Masked Language Modeling and optimized parameters. Other BERT variants like DistilBERT [3], XLM-Roberta [4], and ConvBERT [5] offer specialized advantages. DeBERTa [6] introduces "Disentangled attention" for better token prediction. Beyond BERT, models like ALBERT [7], ELECTRA [8], Switch Transformers [9], T5 [10], GPT series [11] [12], and BART [13] have also made notable contributions.

The literature review also highlighted the SMART technique, as described by Jiang et al. [5], which employs smoothness-inducing regularization and Bregman proximal point optimization to enhance the generalization of pre-trained models. This approach has demonstrated superior performance in the SST-2 sentiment analysis benchmark. Given its effectiveness, we adopted this technique for fine-tuning our model to improve Hate Speech classification.

3. Method

3.1. Data Preparation

Training and testing data, including text and labels, were sourced from .npy files and processed for tokenization using the DebertaV2Tokenizer, with a dataloader class ensuring proper sequence handling up to 256 tokens with a batch size of 32.

3.2. Model Architecture

The microsoft/mdeberta-v3-base model configured for binary sequence classification was utilized, augmented with SMART loss to improve robustness during fine-tuning.

3.3. Training Procedure

The training of the model utilized the Adam optimizer with a learning rate of 2e-5, with Automatic Mixed Precision (AMP) to dynamically cast tensors to lower precision during computations. This technique not only reduced the memory usage but also sped up the processing times by leveraging GPUs designed for lower precision operations. Additionally, the incorporation of gradient scaling addressed potential underflow issues in gradients by scaling them up during the backward pass and then appropriately scaling them down before the optimizer step. This was crucial in maintaining the stability of training in float16 precision. SMART loss integration—combining KL divergence and symmetric KL loss, weighted appropriately—further enhanced the robustness and convergence of the model. Training spanned 12 epochs with gradient accumulation over four steps to man-

age memory use effectively while updating model weights.

An overview of our model is seen in Figure 1.

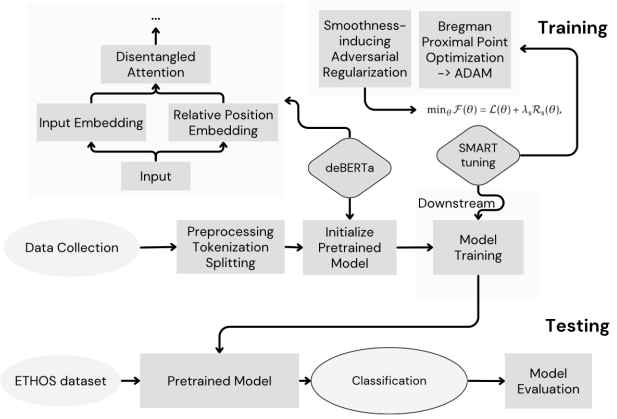

Figure 1. Overview of our SMART-mDeBERTa model when testing on ETHOS.

The choice of our datasets was also a significant focus of our project. Given that mDeBERTa is a multilingual model, with the right data, we hypothesized that it should be capable of transferring its Hate Speech detection capabilities from one language to another. With this in mind, we searched for relevant datasets not only in English but other languages such as German, French, and Arab as well.

3.4. Datasets

We selected the following datasets for our project and processed them. Following is a summarized list of modifications:

Hatemoji Dataset [14]: Used without any specific processing.
Toxygen Dataset [15]: Standard English dataset. Filtered to remove extraneous characters (e.g., b"..." and "-"), repetitions, URLs, and bracketed information.
Call Me Sexist but Dataset [16]: English dataset on implicit misogyny. Cleaned by removing instances of "MENTION number," "mkr," link references, empty rows, and "RT" (retweets).
Detecting Offensive Statements towards Foreigners in Social Media Dataset [17]: German dataset on xenophobia. Minimal cleaning required; usernames replaced with "Peter" and URLs removed.
Online Misogyny EACL2021 Dataset [18]: English dataset on misogyny. Removed URLs, rows shorter than five characters, lines beginning with "¿," and content moderation tags "[deleted]" and "[removed]."
en, fr, and ar datasets [19]: Standard English, french and Arab datasets. Cleaned URLs and user tags
To create our main dataset, we processed these datasets into a consistent format with two columns: one containing the text encoded as strings and the other containing their labels, where 0 for non-Hate Speech and 1 for Hate Speech. A data processing Python script was written to combine these different data files, check their formats, remove [NaN] rows, and split them into training and testing sets, adhering to a common 70/30 ratio—70% of the data was used for training and 30% for validation.

4. Validation

4.1. Performance and results with English data

We first chose to test our model on a benchmark English dataset for hate classification, the ETHOS dataset [20]. We chose this dataset as it was balanced and readily available for comparison. Training our model on different dataset combinations yielded the following results for this benchmark:

MODEL	ACCURACY	PRECISION	RECALL	F1
1	0.756	0.771	0.621	0.688
2	0.783	0.719	0.795	0.754
3	0.746	0.635	0.926	0.753
4	0.692	0.600	0.871	0.711
5	0.708	0.796	0.441	0.568
6	0.748	0.753	0.626	0.683
7	0.776	0.769	0.691	0.727
8	0.802	0.791	0.803	0.797

Table 1. PERFORMANCE OF MODELS TRAINED ON DIFFERENT DATASET COMBINATIONS

Datasets:
1. All, plus hatemoji
2. All
3. All, balanced
4. Without misogyny
5. Without en_dataset
6. Without German dataset (Detecting [...] media)
7. Without fr_dataset
8. Best model for metrics on ETHOS [21]

The stark difference on F1 score from the first and second tests, 0.688 vs. 0.754 indicate that the "Hatemoji" dataset significantly reduced our performance. Consequently, we decided to exclude the "Hatemoji" dataset and keep all the other datasets for the rest of the models. A possible explanation for this is that the embedding of emojis is too different from normal text and therefore negatively impacts overall performance.

The best results were obtained for model 2. It has a high accuracy with a balance between precision and recall, and the highest F1 score of 0.754. From models 4 and 5, we can see that the removal of English datasets decreases the

F1 score (0.711 and 0.568 respectively), meaning there is a good amount of datasets in model 2. We can also see from models 6 and 7 that the removal of datasets from other languages affects the F1 score on a English benchmark, which is promising for the capability of the model to transfer its performance on other languages.

Furthermore, the balanced dataset has a similar performance with an F1 score of 0.753 and, more importantly, scores well on the recall metric. Mollas et al. [20] emphasize that the manual verification of Hate Speech reports is costly. Consequently, optimizing for high recall and F1 score is crucial to minimize the need for human intervention in Hate Speech detection systems.

4.2. Zero-shot evaluation

We evaluated the zero-shot performance of the unbalanced and balanced models (2 and 3 from Table 4.1) on languages unseen during training to assess potential knowledge transfer from the embeddings for Hate Speech classification. These tests were conducted on the following datasets: "SWSR: A Chinese Dataset and Lexicon for Online Sexism Detection" [22], "K-MHaS: A Multi-label Hate Speech Detection Dataset in Korean Online News Comment" [23], and "DALC: the Dutch Abusive Language Corpus" [24]. The results can be seen in table 4.2.

MODEL	ACCURACY	PRECISION	RECALL	F1
1	0.678	0.703	0.571	0.630
2	0.711	0.569	0.667	0.614
3	0.744	0.813	0.816	0.814
4	0.731	0.844	0.540	0.658
5	0.730	0.697	0.382	0.494
6	0.666	0.917	0.566	0.700

Table 2. PERFORMANCE OF BALANCED/UNBALANCED MODELS ON UNTRAINED LANGUAGES

Model and Dataset:
1. Balanced, Korean
2. Balanced, Chinese
3. Balanced, Dutch
4. Unbalanced, Korean
5. Unbalanced, Chinese
6. Unbalanced, Dutch

The zero-shot evaluation showcased the model's cross-lingual generalization ability. Interestingly, the balanced model outperformed the unbalanced one on unseen languages, despite underperforming on English. This could stem from the training set's predominance of English non-hate instances, skewing the unbalanced model's English performance.

Unsurprisingly, Dutch, being semantically similar to the training languages like English, French, and German, achieved a higher F1 score of 0.814 compared to more distant languages such as Korean (0.630) and Chinese (0.614). Interestingly, Dutch surpassed the F1 scores of languages the model was trained on, potentially benefiting from the training on its neighboring languages (English, French, and German) while avoiding overfitting by keeping the learned parameters general.

4.3. GUI and Audio Recognition

To improve our front-end, a GUI was implemented (see Figure 2). A text is typed into the text-box and is then classified with a click on "Classify". Audio recognition was also implemented with the help of a python module "SpeechRecognition". The language can be selected at the bottom for the audio recognition.

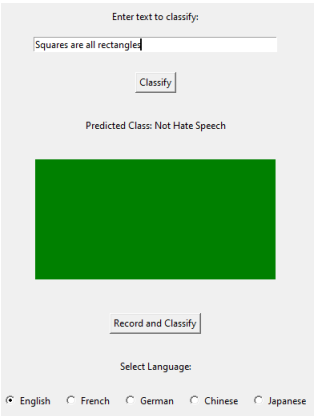

Figure 2. GUI

5. Conclusion

This study successfully implemented a Hate Speech classifier using the mDeBERTa model with SMART fine-tuning, achieving high accuracy across multiple languages. The project highlighted the ethical importance of addressing Hate Speech and employed advanced NLP techniques to enhance performance. Despite optimization efforts, the SMART fine-tuning method remains computationally intensive and time-consuming, posing challenges for developers with limited resources. Additionally, the lack of standardization across languages and datasets, along with dataset imbalances, complicates validation. Future work should focus on optimizing hyperparameters to improve performance, exploring methods to reduce computational demands, and establishing standardized multilingual datasets to ensure broader applicability and effectiveness in real-world scenarios.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "ROBERTA: A robustly optimized BERT pretraining approach," 7 2019.

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," 10 2019.

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 11 2019.

[5] Z. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan, "ConvBERT: Improving BERT with Span-based Dynamic Convolution," 8 2020.

[6] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," 6 2020.

[7] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[8] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[9] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv preprint arXiv:2101.03961*, 2021.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," 10 2019.

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," 5 2020.

[12] C. Alt, M. Hübner, and L. Hennig, "Fine-tuning Pre-Trained Transformer language models to distantly supervised relation extraction," 6 2019.

[13] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for natural language generation, Translation, and Comprehension," 10 2019.

[14] H. R. Kirk, B. Vidgen, P. Röttger, T. Thrush, and S. A. Hale, "Hatemoji: a test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate," *arXiv.org*, Aug. 12 2021. [Online]. Available: https://arxiv.org/abs/2108.05921.

[15] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "Toxigen: a large-scale machine-generated dataset for adversarial and implicit hate speech detection," *arXiv.org*, Mar. 17 2022. [Online]. Available: https://arxiv.org/abs/2203.09509.

[16] "View of 'call me sexist, but...': Revisiting sexism detection using psychological scales and adversarial samples," *ICWSM*, 2021. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/18085/17888.

[17] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 4619–4629, Association for Computational Linguistics, Jul. 2019.

[18] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, and H. Margetts, "An expert annotated dataset for the detection of online misogyny," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (P. Merlo, J. Tiedemann, and R. Tsarfaty, eds.), (Online), pp. 1336–1350, Association for Computational Linguistics, Apr. 2021.

[19] Vidhur2k, "Multilngual-hate-speech/data/all-processed at main · vidhur2k/multilngual-hate-speech." GitHub, 2023. [Online]. Available: https://github.com/vidhur2k/Multilngual-Hate-Speech/tree/main/data/all-processed.

[20] I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas, "ETHOS: a multi-label hate speech detection dataset," *Complex & Intelligent Systems*, Jan. 2022.

[21] G. Rajput, N. S. punn, S. K. Sonbhadra, and S. Agarwal, "Hate speech detection using static bert embeddings," *arXiv*, vol. arXiv:2106.15537v1, 2021.

[22] A. Jiang, X. Yang, Y. Liu, and A. Zubiaga, "SWSR: a chinese dataset and lexicon for online sexism detection." https://doi.org/10.5281/zenodo.4773875, May 2021. Zenodo (CERN European Organization for Nuclear Research).

[23] J. Lee, T. Lim, H. Lee, B. Jo, Y. Kim, H. Yoon, and S. C. Han, "K-MHaS: A multi-label hate speech detection dataset in Korean online news comment," in *Proceedings of the 29th International Conference on Computational Linguistics*, (Gyeongju, Republic of Korea), pp. 3530–3538, International Committee on Computational Linguistics, Oct. 2022.

[24] T. Caselli, A. Schelhaas, M. Weultjes, F. Leistra, H. van der Veen, G. Timmerman, and M. Nissim, "Dalc: the dutch abusive language corpus," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH)*, (online), Association for Computational Linguistics, Aug. 2021.