EmoHate: A Deep Learning approach To Detect Hate Speech using Emojis

Mez Mischa, Opreni Carlo, Spahr Guillaume
Group 26

Problem definition

- We define hate speech in accordance with the United Nations guidelines
- Enhancing Hate Speech Detection with an Ensemble Learning Approach Using Transformer-Based Models
- Develop a model capable of effectively classifying both text-only and emoji-based text simultaneously
- Contributing to safer and more harmonious online environments
- Build our work upon Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate

Key Related Works

- State-of-the-art : Transformer-based models, such as BERT, DistilBERT, and ELECTRA
- Transformer models have shown high performance in hate speech detection tasks
- Limitations : few benchmark datasets such as Davidson, Founta, and TSA ► miss diversity of hate speech in real-world scenarios, poor generalization to other domains
- Hatemoji Paper [1] : adversarial data generation with emojis ► enhanced robustness & accuracy in Hate Speech Detection

Method

- Three base models of DistillBERT trained on different datasets:
- One MLP that incorporates the ensembling
- Preprocessing, tokenization, target extraction
- Early stopping, weighted losses, downsampling, scheduler, use of Huggingface Trainer

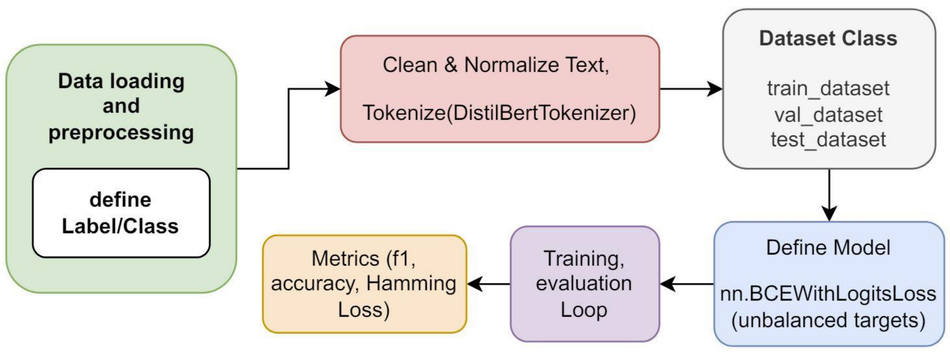


Fig 1. Training the model Diagram

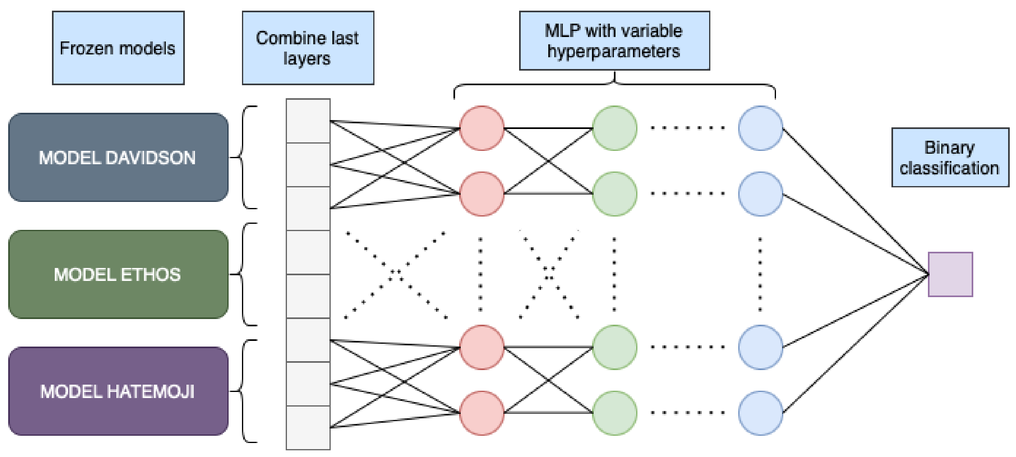


Fig 2. Structure of the MLP model

Dataset(s)

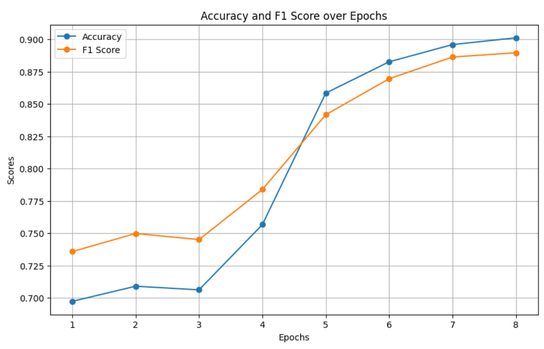
- HatemojiBuild and HatemojiCheck (emojis model)
- Ethos and Davidson and Dynamically Generated Hate Dataset (text-only model)

EE-559: Deep Learning, 2024

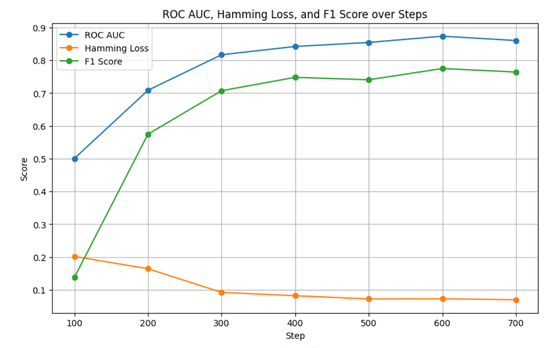
Validation

- For the training of the Emoji and Ethos model ► batch size of 8 per device for both training and evaluation, and a total of 5 epochs
- Multi-label classification : use of Hamming Loss, treats each label prediction independently and averages the error
- Overall good results for each separate model (Ethos [3], Hatemoji, Davidson [2]). Very good F1 score for Davidson, and good AUC ROC score for multilabel datasets

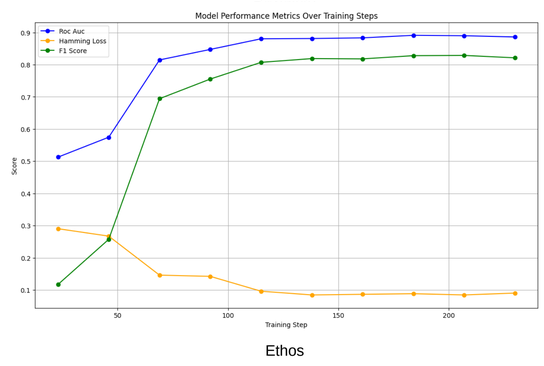
Dataset	Accuracy	Hamming	F1	AUC ROC
Hatemoji	-	0.09	0.77	0.87
Davidson	0.92	-	0.91	-
Ethos	-	0.07	0.82	0.88



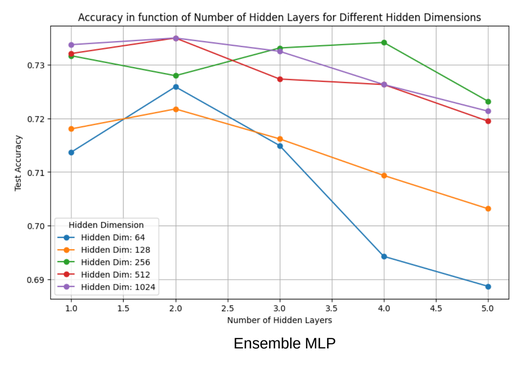
Davidson



Emoji



Ethos



Ensemble MLP

Models	Text based	Emoji based	Text-Emoji based
Davidson	0.59	0.56	0.58
Hatemoji binary	0.66	0.7	0.68
Ensemble MLP	0.74	0.74	0.73

Limitations

- Dataset size (often too small dataset available)
- Lot of preprocessing needed, dataset are not very clean
- Few balanced datasets available
- Bad generalization to text-only dataset when trained on emoji based texts

Conclusion

- Single Model performance is high
- We tried an intersting method to determine hate-speech in both text-only and emoij texts
- Our method achieves performance comparable to existing state-of-the-art techniques while not generalizing well with the ensembling MLP (possibilities: features incompatibility when combined, struggles finding patterns)
- Using diverse datasets and robust preprocessing methods to create a robust model

References:
[1] Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., & Hale, S. A. (2021). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. arXiv preprint arXiv:2108.05921.
[2] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 512-515).
[3] Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2020). Ethos: an online hate speech detection dataset. arXiv preprint arXiv:2006.08328.