Implicit and explicit hate speech detection with HateBERT

Jeremy Serillon, Océane Voland, Thomas Cirillo

Group 21

Problem definition

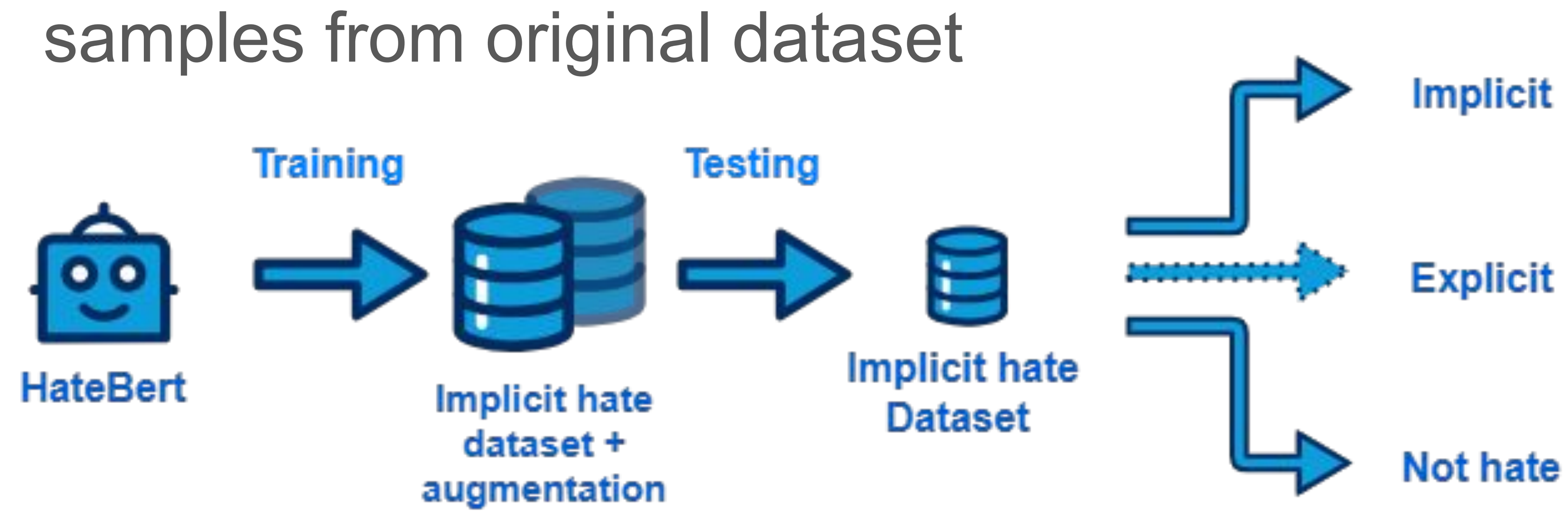
- **Implicit** and **explicit** hate speech detection
- Reproduce and subsequently trying to **improve** upon **existing work** in hate speech detection.
- Study performance of the pre-trained **HateBERT** model for **binary and multi-class** hate speech **classification**

Key Related Works

- Utilized **HateBERT** [1], a model retrained on **Reddit data**, as a base for abusive language detection.
- Employed the **Latent Hatred dataset** [2], **21'482 tweets** (from ***X,*** anciently ***Twitter***), to fine-tune HateBERT and compare findings with the original study's SVM/BERT classifiers.

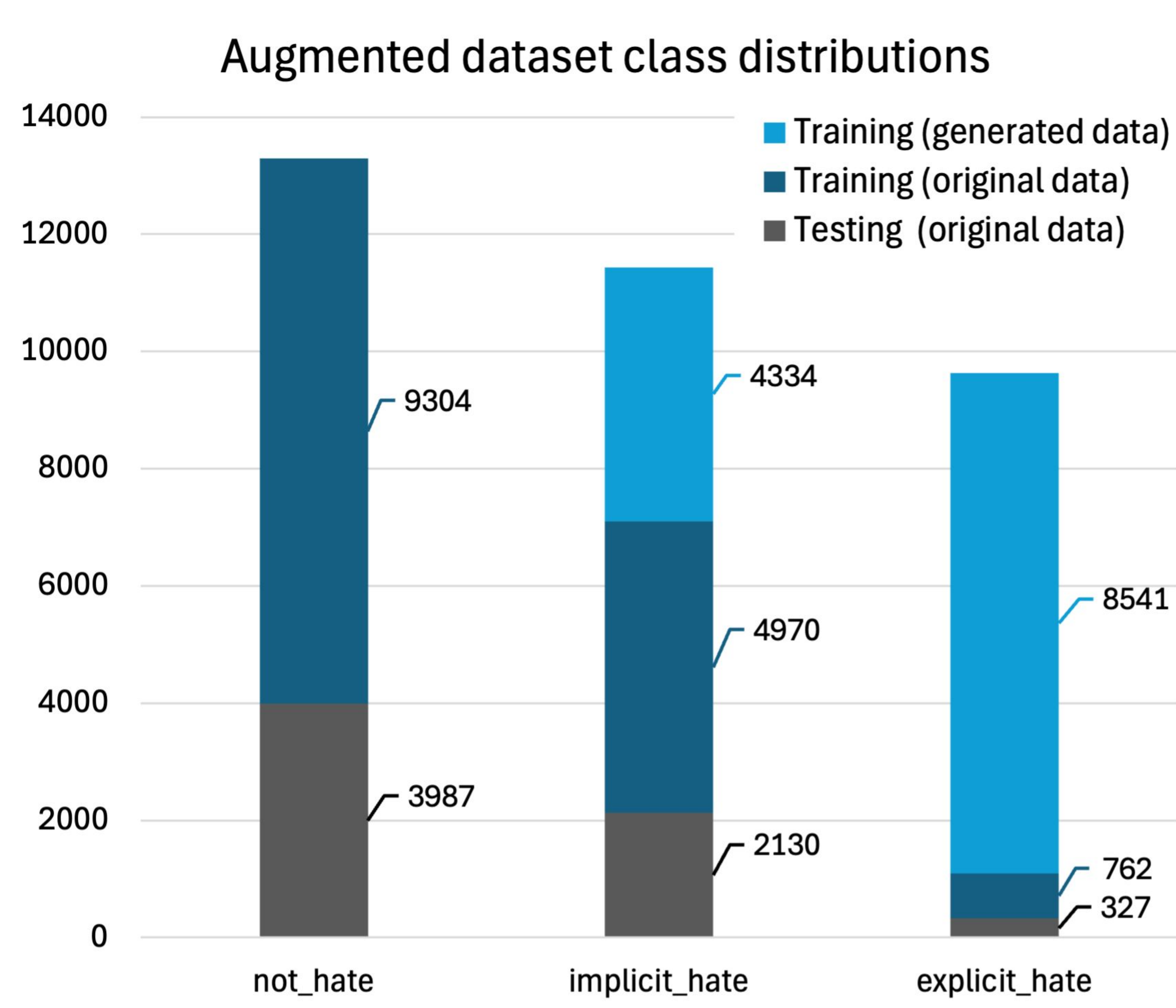
Method

- **Model** : Fine-tuned HateBERT
- **Split** : 60% training / 20% validation / 20% testing
- **Objective**: Binary (not_hate, implicit_hate) and multi-class classification (not_hate, implicit_hate, explicit_hate)
- **Regularization technique** : dropout, gradient clipping, cosinus learning rate scheduler
- **Data augmentation** : LLM-generated synthetic samples from original dataset



Datasets

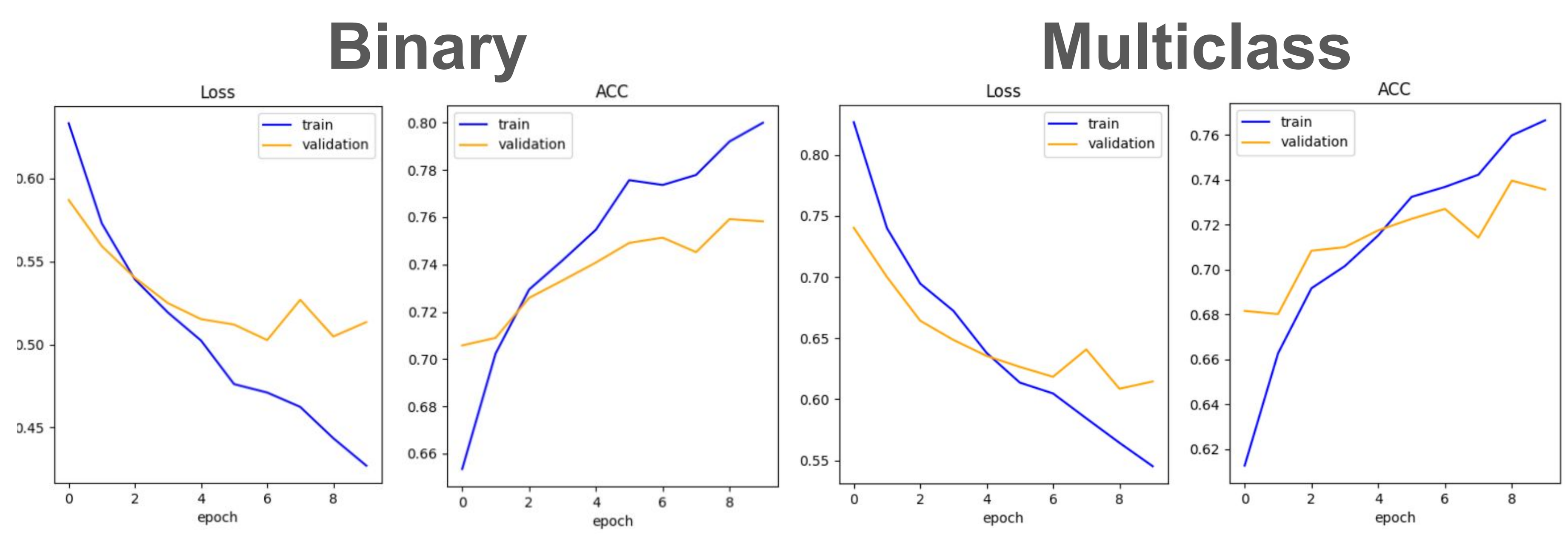
- **Implicit Hate Dataset / Initial dataset - (**MIT License) **[2]**
  Latent Hatred: A Benchmark for Understanding Implicit Hate Speech (2021)
  All samples: 13'291 not hate, 7'100 implicit and 1'089 explicit.
  Repository: https://github.com/SALT-NLP/implicit-hate
- **Artificially augmented Dataset - (**Using LLM Chatgpt o3 model)
  All samples : 13'291 not hate, 11'434 implicit and 9'630 explicit.
  Added samples: 0 not hate, 4'334 implicit and 8'541explicit.



EE-559: Deep Learning, 2025

Validation

Task	Binary	Binary	Binary	Three-Class	Three-Class
Source	Paper [2]	Our work	Our work	Our work	Our work
Dataset	Original	Original	Augmented	Original	Augmented
Batch Size	8	16			
Epoch	{1,2,3,4}	10			
Learning rate	{2e-5, 3e-5, 5e-5}	5e-6			
Weight Decay	-	0.05			
Dropout	-	0.3			
Precision	72.1	**73.3**	72.2	63.3	59.1
Recall	66.0	**74.3**	74.3	58.9	59.3
Accuracy	78.3	75.5	73.3	72.3	70
F1-score	68.9	**73.6**	72.3	60.1	58.7



Limitations

- **Dataset size** : Difficulty to find an exhaustive dataset classifying different hate speech forms. Therefore it is also hard to have sufficient high quality data to not quickly overfit the model.
- **Pretrained Model** : Saves a lot of computing power, but also restricts some parts of the training process, due limited flexibility in modifying the model architecture.
- **Augmented Data** : Quality of the generated samples (repetition, pattern,...)

Conclusion

- Despite a little lower accuracy compared to the paper's model [2], **our other metrics** for the binary classification without augmentation **outperform** their results. This is particularly encouraging given the challenges posed by the imbalanced datasets [3] and the sparse data for the implicit class.
- The data augmentation revealed to be very biased and leading to lower performances.
- Overfitting remains a challenge, potentially solvable through further parameter tuning, regularization methods implementation or adjustments to the model's size and layer structure.

References

[1] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "*HateBERT: Retraining BERT for abusive language detection in English,*" in *Proc. 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pp. 17–25, Association for Computational Linguistics, 2021. Aug. 2021

[2] *M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang, "Latent hatred: A benchmark for understanding implicit hate speech," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Online and Punta Cana, Dominican Republic, pp. 345–363, Association for Computational Linguistics, Nov. 2021.*

[3] Google Developers, "Accuracy, Precision, and Recall - Machine Learning Crash Course", [Online]. [Accessed: May 23, 2025].