Hate Classification using Machine Learning

Joseph Adams

Overview This model is a ternary classification model that classifies text as either 'hate', 'maybe hate', or 'not hate'. I decided to make it as a personal project to learn more about machine learning (as well as just for fun). None of the data used in the training of this model is owned by me.

Dataset The model is trained on a dataset comprising of selected entries from 5 other datasets. Each entry is labeled as either 2 (hate), 1 (maybe hate), or 0 (not hate). The data used to create this dataset was taken from the following sources:

- Automated Hate Speech Detection and the Problem of Offensive Language [1]
- Hate Speech Dataset from a White Supremacy Forum [2]
- Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application [3]
- HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection [4]
- Large-Scale Hate Speech Detection with Cross-Domain Transfer [5]

The dataset comprises 137,343 entries. A stratified version of k-Folds cross-validation (where the class distribution remained consistent for each fold) was used to split the dataset into $k = 10$ folds for training and testing. The stratified version was used to address the class imbalance, where there are 84,016 'not hate', 37,827 'maybe hate', and 15,500 'hate' entries. Each entry in the dataset was normalised with `normalise.py`, which made the text lowercase, removed most special characters (punctuation, emojis, unnecessary whitespace, etc.), and replaced various text segments to keep the data consistent. The text data was also written to individual files for each entry, for each dataset using `build.py` if it does not already exist, in order to provide a hard copy for the tokenisation process. The assembled data object created in `model.py` is stored as *data.pkl* for faster access.

Tokenization The Hugging Face tokenizer library's *Byte-Pair-Encoding* tokenizer was used. It generated an overall vocabulary of 30,000.

Neural Network Architecture The model's neural network consists of an input layer; an embedding layer of 64 nodes; two fully connected layers, the first with 128 nodes and the second with 64; and a fully connected output layer. For the two hidden fully connected layers, each layer involved batch normalisation, a *leaky* ReLU activation function to provide nonlinearity, and dropout to help prevent overfitting. *Xavier initialisation* was used to initialise the weights of the network.

Performance The model's performance was evaluated using cross entropy loss. Over each fold, the model had an average validation accuracy of 0.780, as well as an average validation loss of 0.566. The accuracies and losses across training and validation for each fold are shown in the following plots:

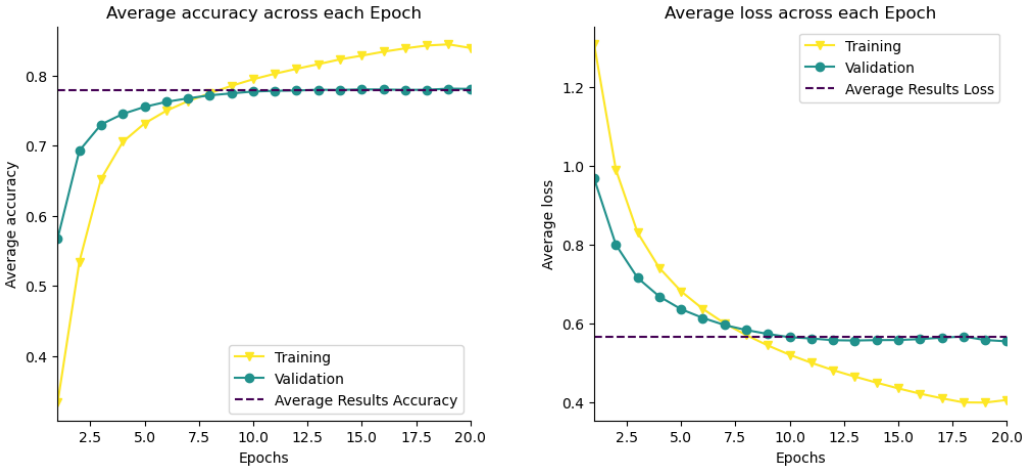

Figure 1: The average accuracy and loss across each epoch.

The number of epochs the model was trained across varied with each fold. This was due to early stopping when the validation loss did not improve past $\delta = 0.001$ of the previous best validation loss. On an *Intel i5-9400* CPU and an *Nvidia GTX 1650* GPU, the model took an average of 55 minutes to train and validate across all folds. The AdamW optimizer was used to train the model, with a learning rate of $1e$-5, and a weight decay of $1e$-4. A scheduler was used that reduced the learning rate by a factor of 0.1 when it started to plateau.

Evaluation & Next Steps The model's classification abilities are satisfactory, although there is room for improvement. The model is overfitting, and the validation loss is quite high. In future I plan to better tune the hyperparameters and add more data, especially data that diverges more from typical online speech. I also plan to experiment with bidirectional LSTM layers, as well as other neural network architectures.

References

[1] Thomas Davidson et al. "Automated Hate Speech Detection and the Problem of Offensive Language". In: *Proceedings of the 11th International AAAI Conference on Web and Social Media*. ICWSM '17. Montreal, Canada, 2017, pp. 512–515.

[2] Ona de Gibert et al. "Hate Speech Dataset from a White Supremacy Forum". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 11–20. DOI: 10.18653/v1/W18-5102. URL: https://www.aclweb.org/anthology/W18-5102.

[3] Chris J Kennedy et al. "Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application". In: *arXiv preprint arXiv:2009.10277* (2020).

[4] Binny Mathew et al. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 17. 2021, pp. 14867–14875.

[5] Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. "Large-Scale Hate Speech Detection with Cross-Domain Transfer". In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 2215–2225. URL: https://aclanthology.org/2022.lrec-1.238.