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 trainng 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 labelled as either 2 (hate), 1 (maybe hate), or 0 (not hate). The other datasets used to create this dataset are found in:

- 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 130,000 entries.

Neural Network Architecture The neural network used in this model consists of a

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`.