Bias and Fairness in NLP, Benchmark Datasets as Evaluation Tools

CS499-002 Deep Learning Final Report

Ahmad AlSubaie
George Mason University

Leo Juarez
George Mason University

ABSTRACT

Machine learning algorithms are currently being used to classify individuals and groups of people. There are some serious ethical implications if these machine learning models carry out human biases, whether intended or not. There are many ways in which bias can creep into our model. To evaluate bias models, we explored fairness metrics in a natural language processing (NLP) setting using a BERT model. Specifically, in the context of hate speech classification. This allows us to explore the limitations of these metrics. Through experimental testing we compare the explanatory power of the balanced accuracy metric as compared to F1 scores and accuracy. We used hateXplain as our benchmarking dataset with 3 class labels (hate, offense, and normal).

1 Introduction

Machine learning algorithms are used to predict behaviors and classify individuals: these algorithms promise to improve decision-making across many domains, such as education, healthcare, finance.[1], [2] By considering thousands of factors and learning meaningful relationships from data. Some common machine learning model applications include determining an individual's credit rating, allocating housing, and aiding recruiters in the hiring process. There are some very serious ethical implications if these models go awry because when models and researchers are not bias aware. We have also come to realize that algorithms are not flawless; models that are built can be faithful to training data but for a number of reasons replicate or even amplify human biases.

This bias can creep into any process of machine learning: in preprocessing, the data can be imbalanced or have bias in the dataset. In-processing methods target the model itself, where the model's learning algorithm or architecture is modified altogether to eliminate discriminatory behavior. Finally, in post-processing, where the model's outputs are modified, transformed, or filtered. Our initial approach was to use an adversarial network as a foundation to measure different models with consistent datasets and list of metrics. Due to limitations in the compatibility of the dataset, the adversarial model was replaced with the BERT model. The task at hand is using the BERT model to classify if text given to the model is either offensive, hate speech, or normal speech. The goal is to improve hate speech classification by testing different metrics for their ability to explain the fairness and discrimination of the evaluated model. We hope to find metrics that will allow for the evaluation of fairness for any model.

2 Main technical sections

There were several challenges that were not anticipated during this project. We assumed that datasets can be applied to different models arbitrarily, this is not the case. Datasets are consistent about the different labels to use (in our context, speech is normal, offensive and hateful).[3] This limitation was the primary technical reason for abandoning adversarial debiasing[4] as an in-processing method. Adversarial debiasing required labels to be on the group identity of the target of the speech. This was not available in the dataset we chose. There are other factors and other data points that some models use and others don't, which reduces the ability to generalize these datasets across models.

Other reasons as to why we opted out of using an adversarial network for this task is due to its use of race, gender, and other group labels. It is illegal to discriminate against protected groups like race or class. So, if it is explicitly stated in code that protected groups are certain groups, it is a legal problem. Fundamentally, though, it is incompatible with our benchmark dataset.

The bias in the system enters the model through the dataset, and is exacerbated as it traverses the data pipeline. In order to reduce bias most effectively, debiasing must be considered before model creation and training. [5]

The procedure on how to test these models is an open issue. Depending on the context that a hate speech classifier might be implemented, it will need to be trained completely differently. For example, if you're using it as a component in a deep network to filter out only hateful text for a chatbot, the selection criteria can be very broadly defined. As compared to a classifier intended for flagging hateful and potentially offensive tweets. In this context, the model may not be capable of distinguishing subtleties of offensive speech that it was not trained on.

The ground truth labels in our context of hate speech classification is not completely trustworthy. This is due to the subjectivity of the labeling process. It was shown by Davidson et al, [6] that poorly labeled datasets would propagate and compound inherent biases. This was the motivating factor for the use of hateXplain [3] as our benchmarking dataset. We will later discuss that the

hateXplain dataset is an improvement from previous datasets for the same issue.

Current researched implementations are not well maintained, resulting in challenges to their testing and evaluations. Even if one can manage to get one working, it is hard to interpret the results because of the lack of fundamental theory of machine learning; each research paper uses its own definition of fairness, and most of these definitions are totally incompatible to our context.

Mitigating bias begins before the model is even created, where we look at the dataset. To evaluate our models, we use a dataset called HateXplain. This dataset addresses the biases of the labels by having a team of subject matter experts each labeling the speech from extremist forums. This is an improvement over previous datasets where they had annotators who were not subject matter experts and whose native language was not English. This means that for previous datasets, they were missing professionals who understood all of the cultural and linguistic subtleties. The HateXplain dataset is also verified through metrics like model plausibility and faithfulness. This is much better than previous datasets which were verified experimentally only after they were released and contained problems. So, all of these factors make HateXplain's dataset a great benchmark dataset because it is one of the best labeled datasets for hate speech since it was labeled by professionals fit for the task of classifying speech as well as being verified through explainability metrics.

For our experiment, we wanted to compare the results of two models that were trained on different datasets to see if mitigating bias really does come from using a better benchmark dataset. First, a BERT model is trained from the UCI hate speech dataset. It was tested against the benchmark dataset. At the same time, we tested this against a model that was actually provided by the HateXplain team. This model provided by the HateXplain team is also a BERT model. Our hypothesis is that the HateXplain dataset will lead to more informative metrics.

3 Implementation

We trained a BERT model with parameters L = 12, H = 768, A=12 on the UCI hate speech dataset.[7] The model was trained for 18 epochs reaching a loss of 0.17 on categorical cross entropy. The model was then evaluated on the hateXplain dataset. Metrics measures are accuracy, balanced accuracy[8], and F1.

The hateXplain models was provided through hugging face's transformers. [9]

4 Evaluations

Model	acc	balanced acc	F1
BERT	0.352	0.356	0.351
hateXplain	0.683	0.679	0.679

Fig1, summary of evaluation metrics using accuracy, balanced accuracy, and F1 scores

0 - hate speech, 1- normal, 2 - offensive

BERT

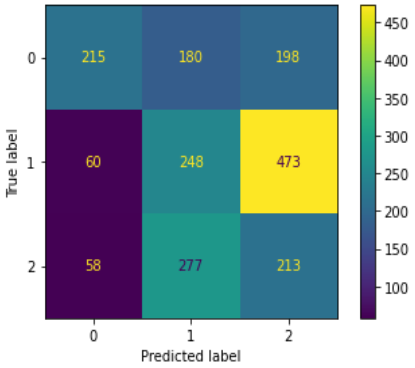

Fig2, confusion matrix of BERT model

hateXplain

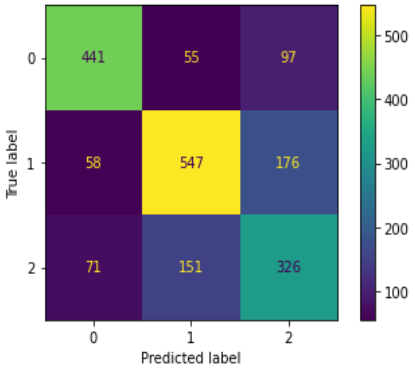

Fig3, confusion matrix of hateXplain model

5 Results

The reported accuracy of the hateXplain model was 0.698. Our reproduced results show an accuracy of 0.683. Comparing the performance of our BERT model, we observe the following. 1) Balanced accuracy is insufficient of a metric to be used in this context. It is expected that balanced accuracy would account for the imbalance in class labels in the dataset, thus reducing the accuracy score accordingly. In this instance the difference between the scores of balanced accuracy and accuracy was not statistically significant enough to reveal that underlying imbalance. 2) Poorly trained models are more likely to

mislabel normal speech as offensive speech, but are less likely to mislabel normal or offensive speech as hateful. This indicates that these models are correctly distinguishing *hate speech*, but are challenged by *normal,* and *offensive* speech. This can be seen in the confusion matrix, the low error rate of the 0 column (the hate speech column). 3) Datasets that are bias aware and apply sound methodology to their creation results in benchmarking datasets that can improve model interpretability.

6 Related work and discussion

So now that we explored evaluating fairness in a model, there are a couple ways to go moving forward. Standardized metrics are perhaps one of the most important ways forward. In order for metrics to be standardized, metrics need to be tested across hundreds of different models, not just two models. Of course, large-scale testing like this is outside the scope of what we could explore for this project and we left for future research. Other such models include HateBERT[10]

Relatedly, a risk management AI framework released by the National Institute of Standards and Technology (NIST)[11] was recently published. NIST is an agency of the United States Department of Commerce. This was released just a few weeks ago during implementation of this project, and is a huge first step of many towards a standardization of metrics and practices. Despite it being one of the most thorough and legitimate frameworks for this field, it's currently in its nascent stages, and has plenty of comments and criticisms. Much of the discussion taking place is about people making suggestions to improve the framework based on their own experiences in AI. The important takeaway is that it is being worked on and important discussion on fairness and bias mitigation in AI is taking place.

Another improvement for deep learning is datasheets for datasets. Datasets should contain metadata that contains important information about the dataset and its possible applications. For example, a dataset could contain metadata that explains if it is meant for training, what demographics it covers, and most importantly, information on how the dataset was labeled and who labeled it. In the context of a hate speech classifier, the hateXplain data was labeled by professionals on the subject matter, which in turn, improved the models who used that dataset.

Finally, some open problems in this area include figuring out which groups are underrepresented and making benchmarks for each of those cases. Another open problem is applicability across modalities. In this project, we explored fairness in the context of text classification. Some other applications where bias is present include image classification and voice recognition.

7 Conclusion

We have shown experimentally that BERT can correctly distinguish hate speech from non-hate speech, but struggle in distinguishing non-hate speech as non-hateful. This indicates the need for more examples of offensive, and normal language in datasets.

Fairness is much harder to implement in practice as it is highly contextual, and requires finding the compatible metrics for your specific models as well as applications.

further research should include professionals in fields of law and social sciences as well as machine learning to overcome these challenges.

Standardization of metrics for evaluating fairness is an open question that the use of benchmark datasets can solve.

8 Research Artifacts

https://github.com/Ahmad-AlSubaie/CS499-DL-debaising

Training information such as model implementations are available here.

REFERENCES

[1] J. A. Mattu Jeff Larson,Lauren Kirchner,Surya, "Machine Bias," *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=WXm9z79e0T7-hUjzwsfdbQxryFu6WxDL (accessed Mar. 04, 2022).

[2] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 469–481, Jan. 2020, doi: 10.1145/3351095.3372828.

[3] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," *arXiv:2012.10289 [cs]*, Apr. 2022, Accessed: May 09, 2022. [Online]. Available: http://arxiv.org/abs/2012.10289

[4] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New Orleans LA USA, Dec. 2018, pp. 335–340. doi: 10.1145/3278721.3278779.

[5] S. K. B, A. Chandrabose, and B. R. Chakravarthi, "An Overview of Fairness in Data – Illuminating the Bias in Data Pipeline," in *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, Kyiv, Apr. 2021, pp. 34–45. Accessed: Apr. 18, 2022. [Online]. Available: https://aclanthology.org/2021.ltedi-1.5

[6] T. Davidson, D. Bhattacharya, and I. Weber, "Racial Bias in Hate Speech and Abusive Language Detection Datasets," *arXiv:1905.12516 [cs]*, May 2019, Accessed: Apr. 18, 2022. [Online]. Available: http://arxiv.org/abs/1905.12516

[7] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated Hate Speech Detection and the Problem of Offensive Language," *arXiv:1703.04009 [cs]*, Mar. 2017, Accessed: Apr. 18, 2022. [Online]. Available: http://arxiv.org/abs/1703.04009

[8] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *Proceedings of the 2010 20th International Conference on Pattern Recognition*, USA, Aug. 2010, pp. 3121–3124. doi: 10.1109/ICPR.2010.764.

[9] "Hate-speech-CNERG/bert-base-uncased-hatexplain · Hugging Face." https://huggingface.co/Hate-speech-CNERG/bert-base-uncased-hatexplain (accessed May 09, 2022).

[10] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "HateBERT: Retraining BERT for Abusive Language Detection in English," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, Online, Aug. 2021, pp. 17–25. doi: 10.18653/v1/2021.woah-1.3.

[11] E. (Fed) Tabassi, "AI Risk Management Framework: Initial Draft - March 17, 2022," p. 23.