Debunking Biases in Attention

Shijing Chen
University of New South Wales
arthur.chen@unsw.edu.au

Usman Naseem
University of Sydney
usman.naseem@sydney.edu.au

Imran Razzak
University of New South Wales
imran.razzak@unsw.edu.au

Abstract

Despite the remarkable performances in various applications, machine learning (ML) models could potentially discriminate. They may result in biasness in decision-making, leading to an impact negatively on individuals and society. Recently, various methods have been developed to mitigate biasness and achieve significant performance. Attention mechanisms are a fundamental component of many state-of-the-art ML models and may potentially impact the fairness of ML models. However, how they explicitly influence fairness has yet to be thoroughly explored. In this paper, we investigate how different attention mechanisms affect the fairness of ML models, focusing on models used in Natural Language Processing (NLP) models. We evaluate the performance of fairness of several models with and without different attention mechanisms on widely used benchmark datasets. Our results indicate that the majority of attention mechanisms that have been assessed can improve the fairness performance of Bidirectional Gated Recurrent Unit (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM) in all three datasets regarding religious and gender-sensitive groups, however, with varying degrees of trade-offs in accuracy measures. Our findings highlight the possibility of fairness being affected by adopting specific attention mechanisms in machine learning models for certain datasets. *Warnings: This paper contains offensive text samples*

1 Introduction

Recently, with the prosperity and popularity of large language models (LLM) all over different industries, they have achieved outstanding results with considerably high accuracy in various downstream tasks according to Naseem et al. [21]. However, with incredible advancements come new challenges, particularly in the realm of fairness and biasness. The study [22] demonstrated that Google Translate API, a popular and widely used machine translation system, exhibited a strong tendency towards male defaults, particularly in the field associated with stereotypes. As the LLMs are trained on large datasets, they have the potential to perpetuate or even amplify the bias inherent in the dataset [10]. This problem has sparked a growing interest in exploring the fairness nature of NLP models and how to mitigate the biases.

One of the most captivating research directions is using attention mechanisms. As the fundamental building block of the modern NLP paradigm, the attention mechanism was first introduced in 2014 in the machine translation domain [1]. They have been proven to promote performance in different downstream NLP tasks significantly. Despite that attention mechanisms can serve as post-processing debiasing techniques [19] [23], few pieces of research have been done investigating the potential for attention mechanisms to affect the fairness of models. According to our knowledge, how they explicitly influence fairness has not been thoroughly explored yet. In this paper, we explore the impact of the attention mechanism on fairness. The key contributions of this work are: we investigate how different attention mechanisms affect the fairness of two recurrent neural networks (RNN) based models i.e., BiGRU and BiLSTM with different attention mechanisms in terms of offensive language classification tasks. Our work studied the effects that attention mechanism can bring to BiGRU and BiLSTM on three different datasets, Jigsaw [6], Hate Speech Offensive Language (HSOL) [5] and HateXplain [18], in terms of fairness and biasness. More specifically, we investigate influencing gender and religious biases in comparison experiments involving BiGRU and BiLSTM with or without different attention mechanisms and using equalized odd metrics.

2 Background and Related Works

This section presents an overview of related work in attention mechanisms, including their developments and applications. Following that, we will discuss the researches and techniques that have been utilized in the field of fairness. Finally, we will examine the works and results from the intersection of attention mechanisms and fairness of the models.

2.1 Attention mechanism

The attention mechanism was first introduced into neural machine translation [1] aiming to solve the problem in machine translation due to the lack of word alignment, which caused focus to be spread over the whole sentence in the decoder. The formulation of this attention mechanism can be written as follow:

$$\begin{aligned} e_{ji} &= a(\mathbf{h_i^{in}}, \mathbf{h_j^{out}}) \\ \alpha_{ji} &= \frac{exp(e_{ji})}{\sum_i exp(e_{ji})} \\ \mathbf{c_j} &= \sum_i \alpha_{ji} \mathbf{h_i^{in}} \end{aligned} \tag{1}$$

Where a is the alignment function that measures the similarity between current hidden state $\mathbf{h_j^{out}}$ and annotation $\mathbf{h_i^{in}}$ by the dot product, the score α_{ji} is the attention score after the normalization using the Softmax function. The context vector $\mathbf{c_j}$ is the weighted sum of the product between the attention score α_{ji} and the annotation $\mathbf{h_i^{in}}$. This attention mechanism not only solved the problem of lack of focus on important parts of the input sentence but also solved the problem that RNN losing old information throughout the multiple times of propagation, as the attention score is calculated on behalf of every token in the input sentence.

This basic attention mechanism has been applied comprehensively across different NLP domains due to its simple and interpretable nature. In recent years different attention variants have been developed regarding more complex tasks. Such as the Hierarchical Attention that was constructed either in the bottom-up approach (word-level to sentence-level) [28] or in the top-down approach (word-level to character-level) [13], the Multi-dimensional Attention that was constructed to capture the attended representation from, for example, two different representation space [25] rather than just one dimension, and Memory-based Attention that was constructed based on soft memory addressing to solve the issue where the answer is indirectly related to the question in question answering problem domain [27].

In 2017, the landmark work by Vaswani [24] demonstrated the transformer model, which has revolutionized the field of NLP and Computer Vision (CV) and has been used to create state-of-the-art models for various tasks. The main crucial component of the transformer is Self Attention mechanism. The difference between Self Attention and basic attention we mentioned earlier is that for basic attention formulation in equation 1, the attention score is computed with external query vector ($\mathbf{h_j^{out}}$ in this case). On the contrary, the internal query is adopted to capture the intrinsic dependency between tokens in the input sentence.

$$\begin{aligned} e_i &= a(\mathbf{v_j}, \mathbf{v_i}) \\ \alpha_{ij} &= softmax(e_{ij}) \end{aligned} \tag{2}$$

Here $\mathbf{v_j}$ is the internal query chosen as each token in the input sequence to calculate the pairwise attention score for every pair of tokens within the input. In this way, the dependency and relation between any token with other tokens in the input can be easily captured and contributes to corresponding tasks.

2.2 Fairness

The concept of fairness in NLP often refers to the principle that models ought to abstain from creating or exacerbating societal biases and inequalities. The bias of the NLP system is generally divided into two categories, intrinsic and extrinsic. The intrinsic bias refers to the bias inherent in the representation, e.g., word embedding layer [2], and the extrinsic bias refers to the performance disparity shown in the specific downstream tasks and applications. Since intrinsic bias metrics do not correlate with extrinsic bias [9], we mainly focus on extrinsic bias metrics as intrinsic bias measure is not ideal for predicting the extrinsic biases in our context. There are different definitions of fairness in NLP, and each also refers to a measure used to measure the model to be fair or not. The three main definitions that are used:

- **Statistical Parities**. Let X denote the features used for prediction and Y denote the ground truth of the corresponding entry. Let $\hat{Y}$ be the

outcome variable. The outcome variable $\hat{Y}$ satisfies statistical parity if only $\hat{Y}$ and A are independent.

$$P(\hat{Y} = \hat{y}|A = a, X = x) = P(\hat{Y} = \hat{y}|X = x)$$

- **Equality of Opportunity**. The outcome variable $\hat{Y}$ satisfies equality of opportunity concerning class $y \in Y$ if $\hat{Y}$ and A are independent conditioned on $Y = y$.

$$P(\hat{Y} = \hat{y}|A = a, X = x, Y = y) = P(\hat{Y} = \hat{y}|X = x, Y = y)$$

  These metrics focus more on the true positive rate (TPR), which should be the same across different protected attributes under this criteria.

- **Equality of Odds**. The outcome variable $\hat{Y}$ satisfies equality of opportunity for class $y \in Y$ if $\hat{Y}$ and A are conditionally independent on Y

$$P(\hat{Y} = \hat{y}|A = a, X = x, Y) = P(\hat{Y} = \hat{y}|X = x, Y)$$

  These metrics focus more on the TPR and the false positive rate(FPR), which should be the same across different protected attributes under this criteria.

In this paper, Equalized Odds [11] is adopted, which uses the maximum between the absolute difference of TPR and FPR across different protected groups.

2.3 Combination

To the best of our knowledge, only a few works focused on the intersection of fairness and attention mechanism. Edelman et al. [7] presented a theoretical analysis of the inductive biases of self-attention models and found a phenomenon called *sparse variable creation*, which suggested bounded-norm Transformer layers create sparse variables and, therefore, sparsity bias. Mehrabi et al. [19] designed an attention intervention mechanism that leverages the attention mechanism and shows the effectiveness of this approach in terms of both fairness and accuracy. Qiang et al. [23] has developed a fairness-through-blindness approach called *Debiased Self-Attention* (DSA) which helps the vision transformer (ViT) to eliminate spurious features related to the sensitive attributes for bias mitigation.

3 Fairness in Attention

We investigated how the attention mechanism can affect group fairness across two different but homogeneous types of neural networks: BiLSTM [12] and BiGRU [4]. The reason for the choices of these two architectures is that as we want to investigate how attention mechanisms affect fairness performance, any self-attention-based architectures such as Transformers [24] become inappropriate choices. We chose to focus on text toxicity classification as our downstream tasks due to the relevance between the fairness performance of NLP models and the nature of text toxicity tasks. The definition of toxicity we incorporate here is from [3] stated as '*anything that is rude, disrespectful, or unreasonable that would make someone want to leave a conversation*.'

3.1 Dataset

To understand the impact of attention in fairness, we have used three datasets 1) Jigsaw, a large dataset released for the "Toxicity Classification" Kaggle competition [6] that contains online comments on news articles, and 2) HateXplain [18], a dataset recently introduced with the intent of studying explanations for offensive and hate speech in Twitter and Twitter-like data. 3) HSOL [5], a dataset that contains tweets that contain words and phrases from a hate speech lexicon.

3.2 Model Settings

The main two models that have been used here are BiGRU [12] and BiLSTM [4]. There are three different attention mechanisms that have been adopted, additive attention [17], dot product attention [1], and self-attention [24]. We used the same implementation of the self-attention mechanism in [26], where a randomly initialized vector is jointly learned as a query used to calculate the attention score. The choice of the optimizer is Adam [14] for all model settings, and 0.05 are chosen as the learning rate for all models. The five-fold cross-validation has been adopted to ensure accurate and precise experiment results.

3.3 Sensitive groups and Fairness Measure

Religion, race, and gender are considered the most common sensitive topics. In our work, we mainly focus on gender and religion as the bias originating from them is less concerned overall, but we believe they are equally harmful compared to race. Based

on the keyword searching technique, we categorized a data entry into the corresponding sensitive groups if they mentioned any related keyword in this topic. For each sensitive group, we randomly sample a small portion of data proportionally according to different labels from the sensitive group as a test set for protected attributes. We then sample the same amount of data with the same distributed labels outside of the sensitive group as a complementary test set, and then we compare the difference between the sensitive group test set and the complementary test set to investigate our questions. All models are trained on the other data that does not belong to either of the test set.

The metrics used here to measure the fairness performance of the models is the Equalized Odds [11] which is defined as:

$$EqOdd(\hat{y}, a, y) = \max_{a_i, a_j} \max_{y \in \{0,1\}} |P(\hat{y} = 1 | y = 1, a = a_i, y = y) - P(\hat{y} = 1 | y = 1, a = a_j, y = y)| \quad (3)$$

Where $\hat{y}$ is the prediction of the model, and y is the ground truth, and a_i represents the corresponding protected attributes (gender, religion, etc.). An equivalent way to calculate the equalized odd is the maximum of absolute true positive rate difference and false positive rate difference, where these differences are between a sensitive group and a complementary group.

4 Results

In this section, the results of the fairness comparison, the attention analysis, and the prediction analysis are reported. Further experimental results and diagrams are analyzed and discussed in the Appendix.

4.1 Fairness Comparison

For the fairness comparison test, the results suggested that attention mechanisms did impact the fairness performance of models no matter which model, which attention, and which dataset was chosen. However, under the different settings, the attention mechanism also affects the fairness performance differently, some of which came with a trade-off between accuracy and fairness measures. Throughout the experiments, the majority of attention mechanisms successfully improve the fairness performance on both models and sensitive groups in all datasets, with varying degrees of accuracy trade-offs.

Jigsaw. We investigate how the attention mechanism affects the fairness performance of BiGRU and BiLSTM on the Jigsaw dataset. In figure 1, the graph shows similar trends for two models in different sensitive groups. In religious groups, Additive attention with both models achieves the best results of fairness. However, it comes with the largest loss of accuracy as well. The basic dot product attention and self attention with BiGRU result in a loss in accuracy without any decrease in bias measures. The picture is different with BiLSTM as both attentions achieve a better fairness performance with trade-offs between accuracy. The Basic dot product attention with BiLSTM achieves the best result, significantly reducing the bais level with minimal loss in accuracy measures. In the gender group, the basic dot product attention for both models fails to improve fairness. The self and additive attention for both models improve the fairness for different degrees, and larger improvement comes with a larger trade-off between accuracy measures, with the self-attended BiGRU having the least bias mitigation, and the additive attended BiLSTM having the most.

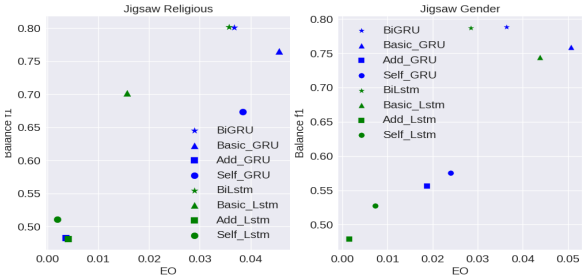

Figure 1: The accuracy and fairness of models in the Jigsaw dataset regarding religious and gender-sensitive groups. The y-axis Balanced f1 metrics are calculated by taking the average f1 scores on sensitive test sets and complementary test sets. The x-axis Equalized Odds (EO) is calculated by the maximum of the absolute true positive rate difference and false positive rate difference between the sensitive group and the complementary group.

Overall the BiLSTM with self attention and additive attention achieve the best results in terms of fairness measures in religious and gender groups, respectively, regarding table 1.

HateXplain. The trends on the HateXplain dataset are similar between the two sensitive groups. As shown in figure 2, all models with

Table 1: Fairness performance on Jigsaw dataset

Model	religious EO	gender EO
BiGRU w/o	0.0367	0.0364
BiGRU basic	0.0455	0.0506
BiGRU add	0.0036	0.0187
BiGRU self	0.0385	0.0240
BiLSTM w/o	0.0358	0.0284
BiLSTM basic	0.0157	0.0437
BiLSTM add	0.0041	**0.0016**
BiLSTM self	**0.0020**	0.0074

The table shows the fairness performance using bias measures Equalized Odds, which indicate the level of bias incorporated in the model. Throughout T-test, *p=0.031* for the best religious EO and *p=0.006* for the best gender EO. So both best results of EO are statistically significant. More detail can be found in Appendix A.

Table 2: Fairness performance on HateXplain dataset

Model	religious EO	gender EO
BiGRU w/o	0.0743	0.0379
BiGRU basic	0.0745	0.0291
BiGRU add	0.0736	0.0432
BiGRU self	0.676	0.0421
BiLSTM w/o	0.0863	0.0327
BiLSTM basic	**0.0481**	0.0258
BiLSTM add	0.0693	0.0233
BiLSTM self	0.0559	**0.0213**

The table shows the fairness performance using bias measures Equalized Odds, which indicate the level of bias incorporated in the model. Through the T-test, *p=0.012* for the best religious EO and *p=0.172* for the best gender EO. So the religious EO of BiLSTM with basic attention is statistically significant. More detail can be found in Appendix A.

attention successfully mitigate the bias with trade-offs in accuracy to different extents and greater mitigation with greater trade-offs, except that additive attended BiLSTM incurs the minimal loss of accuracy in religious groups. In the gender group, a similar trend persists apart from that BiGRU with self attention and additive attention fail to promote fairness measures in this experiment.

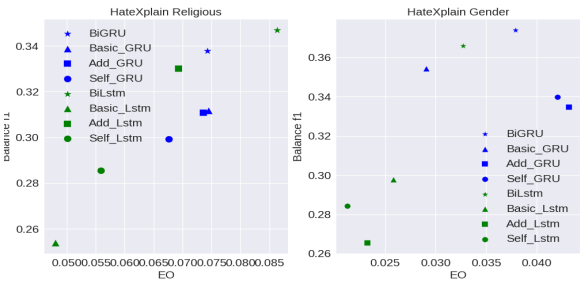

Figure 2: The accuracy and fairness performance of models in the HateXplain dataset regarding religious and gender-sensitive groups. The y-axis is the Balanced f1 score and the x-axis is EO.

From 2, the best models with the lowest bias are BiLSTM with basic dot attention for religious and BiLSTM with self attention for the gender group.

HSOL. The situation is lightly different from what was shown in the last two datasets, as shown in figure 3. In the religious group, the original BiGRU already achieves the highest accuracy with a relatively low level of baisness, except that other models and attentions persist the trend similar to that of the other two datasets. This abnormal phenomenon might originate in the fact that there are only about 200 data entries categorized in the religious group in this dataset. In contrast, thousands of entries are discovered as religious in the other two datasets and as gender groups in all datasets. And therefore, the small size of the test samples can be the reason for this outlier observation. In the gender group, all models and attentions, except additive attended BiLSTM, successfully reduced the level of bias to a similar significant extent. However, the trade-off they made varies, with basic attended BiGRU suffering from the least amount of loss in accuracy.

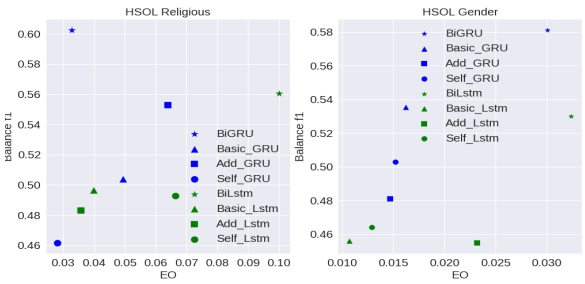

Figure 3: The accuracy and fairness of models in the HSOL dataset regarding religious and gender-sensitive groups. The y-axis is the Balanced f1 score, and the x-axis is EO.

According to table 3, the best model for HSOL came from BiGRU with self attention for the religious group and BiLSTM with basic dot product attention for the gender group.

4.2 Attention and Prediction Analysis

In this section, we report the analysis we carried out on attention mechanisms, mainly based on attention weight visualization and prediction analysis on

Table 3: Fairness performance on HSOL dataset

Model	religious EO	gender EO
BiGRU w/o	0.0328	0.0301
BiGRU basic	0.0494	0.0162
BiGRU add	0.0640	0.0147
BiGRU self	**0.0282**	0.0152
BiLSTM w/o	0.0999	0.0324
BiLSTM basic	0.0400	**0.0107**
BiLSTM add	0.0358	0.0232
BiLSTM self	0.0663	0.0129

The table shows the fairness performance using bias measures Equalized Odds, which indicate the level of bias incorporated in the model. Through the T-test, p=*0.169* for the best religious EO and p=*0.078* for the best gender EO, more detail can be found in Appendix A.

test samples. The model with significant improvement in fairness performance and minimal loss in accuracy is selected(BiGRU with basic attention mechanism). Considering the sequence length of input text, the analysis results of BiGRU with basic attention mechanism on gender group in HSOL dataset is shown in the following section. The other analysis results can be found in Appendix A.

```
----------Attention Text Focus----------
8220 dickfurari hell make malt liquor ads http co h0jgonf8
tooracist black guy school asked colored printers library
dxpperjay assholes referring girl bitch make dick bigger [
stepheeezyy nigga bitch 8220 187xo_ date female niggah wtf
abstractlife damn giggling offering tips shit heard bitch
boy make feel bitch [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [P
kejonasavage _chocgirl rich homie bird mansion lan tonight
lmfaoooo harleyyyquinn_ wiz perfect fame bitches typical t
rubs hands bird [PAD] [PAD] [PAD] [PAD] [PAD] [PAD] [PAD]
hberghattie snkscoyote wonder progs relegate young black g
girlfriendnotes reasons mermaid periods pants perfect hair
herman_nyrblog yankees bother show boston fold franchise c
```

Figure 4: BiGRU with Basic dot product attention on gender test set in HSOL. The color of the text reflects the weight that attention assigned to certain words, with red being the highest score and green being the lowest. This figure shows that while the attention mechanism captured the important information that might help the classification, it can also capture irrelevant sensitive words such as 'black', which might lead to amplifying the bias regarding the sensitive attributes

Attention Analysis. From the attention focusing on test text shown in figure 4, the attention has successfully targeted the words that can significantly contribute to the classification of the sentence and the heatmap of attention weights is shown in figure 5. However, all attention mechanisms in all experiment settings have contributed to losses in accuracy measures compared to the original Bi-GRU/BiLSTM in general. This may occur due to the complex nature of toxicity classification tasks which is also explained later in prediction analysis. Also, the over-reliance on the attention mechanism can be another reason why neural networks become over-fitted or over-specialized.

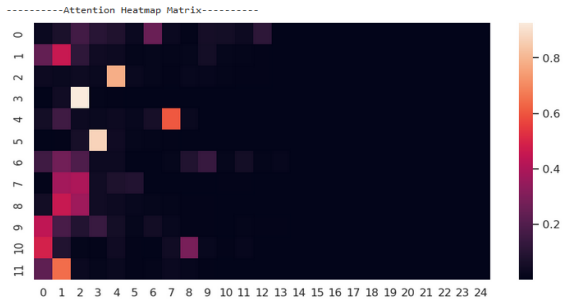

Figure 5: The Attention weigth heatmap of the BiGRU with Basic dot product attention on the gender test set in the HSOL

Prediction Analysis. From the prediction comparison of the same batch of test data that is used in attention analysis. As shown in figure 6, the model predicted precisely for 'neither' and 'offensive' labels with only one mispredicting in entry 6. However, the model predicted badly for the 'hatespeech' label. It predicted 3 'hatespeech' labeled test entries as 'offensive' and the other one as 'neither'. The result of this analysis shows that the indistinguishable label setting limited the performance of the models, and a clear definition of the difference between 'offensive' and 'hatespeech' needs to be incorporated.

	content	ground_truth	prediction
0	8220 dickfurari hell make malt liquor ads http...	hatespeech	offensive
1	tooracist black guy school asked colored print...	hatespeech	neither
2	dxpperjay assholes referring girl bitch make d...	hatespeech	offensive
3	stepheeezyy nigga bitch 8220 187xo_ date femal...	hatespeech	offensive
4	abstractlife damn giggling offering tips shit ...	offensive	offensive
5	boy make feel bitch	offensive	offensive
6	kejonasavage _chocgirl rich homie bird mansion...	offensive	neither
7	lmfaoooo harleyyyquinn_ wiz perfect fame bitch...	offensive	offensive
8	rubs hands bird	neither	neither
9	hberghattie snkscoyote wonder progs relegate y...	neither	neither
10	girlfriendnotes reasons mermaid periods pants ...	neither	neither
11	herman_nyrblog yankees bother show boston fold...	neither	neither

Figure 6: BiGRU with Basic dot product attention on gender test set prediction in HSOL

5 Discussions and Limitations

Our study covered three types of widely used single attention with different mechanisms of assigning attention weights. However, we did not cover some compound attention mechanisms such as dual attention mechanism [8] and Co-attention [27], which

might contain different patterns affecting the fairness of the models. Also, Transformer [24], the cornerstone of PLMs, should be considered in this study as it is composed of multiple self-attention modules, and the intersection impact of multiple attention mechanisms can be studied by incorporating this model. Apart from the classifier itself, the different word representation models, which are well discussed in Naseem et al. [21], can also be brought into scope since word embedding can also affect fairness. From the dataset aspect, the quality of text can be further improved with pre-processing techniques mentioned in Naseem et al. [20] to ensure better performance and reduce the effect of the irrelevant factors. Also, since the toxicity classification tasks are not easy even for a human, there are noisy data inside the chosen datasets since we found that we disagree with some of the human-annotated labels by manual checking. Furthermore, the HSOL and Jigsaw datasets are imbalanced in terms of distributions of different classes. Therefore, modifications can be made to the loss function in the same way as focal loss [16] or dice loss [15] to mitigate the influence of data imbalance.

6 Conclusion

In this work, we have investigated BiGRU and BiLSTM with three types of widely used attention mechanisms in three datasets regarding religious and gender-sensitive groups in terms of fairness performance as well as accuracy performance. The results demonstrate that all three types of attention mechanisms can mitigate the bias with a trade-off in accuracy in most scenarios of our experiments. These findings highlight that attention mechanisms, effective methods derived from human intuition of focusing, have the potential to be developed and incorporated as a debiasing methodology for bias mitigation in toxicity classification tasks.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[3] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[5] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515, 2017.

[6] Quan Do. Jigsaw unintended bias in toxicity classification. 2019.

[7] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[9] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*, 2020.

[10] Thilo Hagendorff, Leonie N Bossert, Yip Fai Tse, and Peter Singer. Speciesist bias in ai: how ai applications perpetuate discrimination and unfair outcomes against animals. *AI and Ethics*, pages 1–18, 2022.

[11] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. A nested attention neural hybrid model for grammatical error correction. *arXiv preprint arXiv:1707.02026*, 2017.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[17] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[18] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

[19] Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Attributing fair decisions with attention interventions. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 12–25, Seattle, U.S.A., July 2022. Association for Computational Linguistics.

[20] Usman Naseem, Imran Razzak, and Peter W Eklund. A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, 80:35239–35266, 2021.

[21] Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35, 2021.

[22] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381, 2020.

[23] Yao Qiang, Chengyin Li, Prashant Khanduri, and Dongxiao Zhu. Fairness-aware vision transformer via debiased self-attention. *arXiv preprint arXiv:2301.13803*, 2023.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[25] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[26] Jun Xie, Bo Chen, Xinglong Gu, Fengmei Liang, and Xinying Xu. Self-attention-based bilstm model for short text fine-grained sentiment classification. *IEEE Access*, 7:180558–180570, 2019.

[27] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406. PMLR, 2016.

[28] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

A Appendix

A.1 Analysis of BiGRU with basic attention on the religious group in Jigsaw dataset

Attention Analysis. Figure 7 shows that BiGRU with Basic dot product attention can also focus on the word important for toxicity classification. Figure 8 indicates that the attentions mainly focus on the first 40 tokens for this dataset when the sequence length of test samples is around 200.

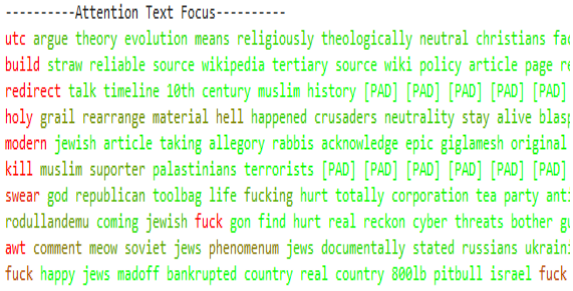

Figure 7: BiGRU with Basic dot product attention on the religious test set in Jigsaw dataset

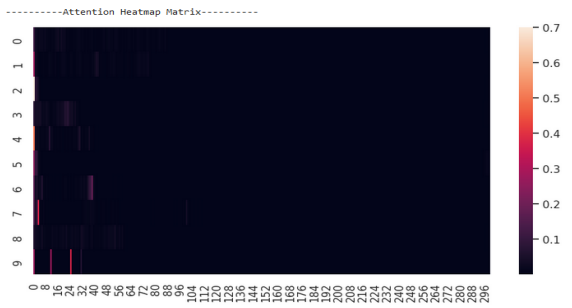

Figure 8: heatmap of attention weights of BiGRU with Basic dot product attention on the religious test set in Jigsaw dataset

Prediction Analysis. Figure 9 highlights good results of prediction that BiGRU with basic attention made on test samples. The model predicts all 'neutral' labeled data correctly and only 2 'toxic' data as 'neutral' data. The misprediction might

originate in the class imbalance of the dataset as the 'neutral' labeled data are nearly 10 times more than the 'toxic' labeled data.

	content	ground_truth	prediction
0	utc argue theory evolution means religiously t...	neutral	neutral
1	build straw reliable source wikipedia tertiary...	neutral	neutral
2	redirect talk timeline 10th century muslim his...	neutral	neutral
3	holy grail rearrange material hell happened cr...	neutral	neutral
4	modern jewish article taking allegory rabbis a...	neutral	neutral
5	kill muslim suporter palastinians terrorists	toxic	neutral
6	swear god republican toolbag life fucking hurt...	toxic	toxic
7	rodullandemu coming jewish fuck gon find hurt ...	toxic	toxic
8	awt comment meow soviet jews phenomenum jews d...	toxic	neutral
9	fuck happy jews madoff bankrupted country real...	toxic	toxic

Figure 9: Prediction of BiGRU with Basic dot product attention on the religious test set in Jigsaw dataset

A.2 Analysis of BiGRU with basic attention on gender group in HateXplain dataset

Attention Analysis. Figure 10 shows that BiGRU with Basic dot product attention not only focuses on the important word that is related to the sensitive topic, which is gender here but also captures words that are related to other sensitive attributes that refer to different ethnicity. Figure 11 indicates that the attention mainly assigns the highest weight to the first few tokens, but the tokens in the very last part can also be attended to some extent.

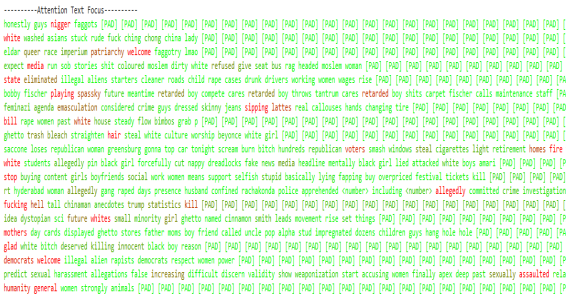

Figure 10: BiGRU with Basic dot product attention on the religious test set in Jigsaw dataset

Prediction Analysis. The model predicts badly overall in this dataset, as what is shown in figure 12. It tends to predict normal for nearly all data entries in test samples. The main reason is that for the HateXplain dataset, multiple annotators were used to labeling each data entry, and it is quite often that they contradicted each other, which results in a bad and imprecise label setting that further negatively affects the prediction as a form of noise.

Figure 11: heatmap of attention weights of BiGRU with Basic dot product attention on the religious test set in Jigsaw dataset

	content	ground_truth	prediction
0	honestly guys nigger faggots	hatespeeech	hatespeeech
1	white washed asians stuck rude fuck ching chon...	hatespeeech	normal
2	eldar queer race imperium patriarchy welcome f...	hatespeeech	normal
3	expect media run sob stories shit coloured mos...	hatespeeech	normal
4	state eliminated illegal aliens starters clean...	hatespeeech	normal
5	bobby fischer playing spassky future meantime ...	neutral	offensive
6	feminazi agenda emasculation considered crime ...	neutral	normal
7	bill rape women past white house steady flow b...	neutral	normal
8	ghetto trash bleach straighten hair steal whit...	neutral	normal
9	saccone loses republican woman greensburg gonn...	neutral	normal
10	white students allegedly pin black girl forcef...	normal	normal
11	stop buying content girls boyfriends social wo...	normal	normal
12	rt hyderabad woman allegedly gang raped days p...	normal	normal
13	fucking hell tall chinaman anecdotes trump sta...	normal	normal
14	idea dystopian sci future whites small minorit...	normal	normal
15	mothers day cards displayed ghetto stores fath...	offensive	normal
16	glad white bitch deserved killing innocent bla...	offensive	normal
17	democrats welcome illegal alien rapists democr...	offensive	normal
18	predict sexual harassment allegations false in...	offensive	normal
19	humanity general women strongly animals	offensive	normal

Figure 12: Prediction of BiGRU with Basic dot product attention on the religious test set in Jigsaw dataset

A.3 Significant T-test for all EO values compared to the results without attention mechanism

Significant test. Due to the small figure of EO metrics, it is necessary to carry out a significant test to ensure the difference is statistically significant. The double-sided T-test is adopted in a manner where each of the attended results is compared with the results without an attention mechanism. The raw data of five-fold cross-validation is used to calculate the t and p values for this test, the results are shown in the following tables.

Table 4: p value of results on Jigsaw dataset from T-test

Model	religious EO	gender EO
BiGRU basic	0.0626	0.2594
BiGRU add	0.004	0.2038
BiGRU self	0.283	0.2668
BiLSTM basic	0.6516	0.1018
BiLSTM add	0.6554	**0.0058**
BiLSTM self	**0.0314**	0.1034

The table shows the p-value for all results compared with the results from non-attended models. The values in bold font indicate the models that have the best EO results.

Table 5: p value of results on HateXplain dataset from T-test

Model	religious EO	gender EO
BiGRU basic	0.9822	0.7662
BiGRU add	0.9356	0.5031
BiGRU self	0.2162	0.5118
BiLSTM basic	**0.012**	0.5857
BiLSTM add	0.1582	0.3455
BiLSTM self	0.0541	**0.1718**

The table shows the p-value for all results compared with the results from non-attended models. The values in bold font indicate the models that have the best EO results.

Table 6: p value of results on HSOL dataset from T-test

Model	religious EO	gender EO
BiGRU basic	0.428	0.3067
BiGRU add	0.8501	0.256
BiGRU self	**0.1693**	0.2371
BiLSTM basic	0.125	**0.0784**
BiLSTM add	0.0231	0.2638
BiLSTM self	0.2652	0.0937

The table shows the p-value for all results compared with the results from non-attended models. The values in bold font indicate the models that have the best EO results.