Minimizing Un-intended Bias in Toxic Text Classification

Prabhnoor Singh [1]

Abstract

The domain of classifying toxic comments in online conversations is a very sensitive one. Online bullying can severely impact someone's mental health and can have very harsh consequences. The need for systems that can accurately filter online toxicity increases with the increase in the ease of access to the internet. This is a delicate domain to work in where un-intended bias has been observed to be a hindrance in building successful systems that can differentiate between toxic and non-toxic conversations. Thus, using global metrics such as accuracy, f1-score or AUC might not be ideal. Further, the use of basic Machine Learning models is just not enough. It is important to have the right set of data, modeling techniques, and scoring metrics to build accurate models. This work explores the use of a modified AUC metric that accounts for the performance of the model not just globally, but also incorporates the performance on individual identity subgroups. This work also compares performance using TfIdf and contextual representation of text on Logistic Regression, Random Forest, LSTM inspired architecture, and transformer-based architecture: BERT.

1. Introduction

As the ease of access to the internet increases, the interactions among people on social media platforms and conversational websites proliferate as well. It is a sad fact that not all conversations on social media are respectable or even humane in nature. Online abuse and bullying have been commonplace for a long time now. There have been countless instances wherein the mental health of the abusee was affected severely leading to harsh consequences in certain cases. Online abuse can occur in various forms such as hate crime, targeting of a minority, racial abuse, cyberbullying, and so on. A small difference of opinion or just the spite of an abuser can trigger such behavior.

Although this behavior has been established to be inappropriate and morally incorrect, still some people continue to practice this behind the wall of security offered by their computer screens. A very practical solution is to identify such comments and block them even before they are posted on the websites. A combination of Machine Learning and Psychologically inspired solutions have seen some progress in tackling this problem. Many online platforms have identified and accepted the fact that they must prevent such cases of online abuse on their platforms and for that they constantly look to improve their systems as the current systems work only up to a certain extent and are far from being perfect. Moreover, miscreant users continue to bypass the systems and find some alternate ways to beat the existing approaches and practice online bullying through toxic comments. Thus, there is a need to continuously update and reinforce the current systems with novel advancements in the relevant fields.

There has been significant work done in the field and the problem has been formalized as a toxicity classification problem where the task is to identify whether a textual comment is toxic or non-toxic. Numerous techniques ranging from linear models such as Logistic Regression to neural approaches such as Multi-Layered Perceptrons, CNNs, and LSTMs have been used to solve this text classification task. These techniques have seen state-of-the-art results on the benchmark datasets. However, a couple of problems have been identified that add hindrance in deploying these systems in production. As this is a sensitive problem, with potentially disastrous consequences, the room for error is very less.

A major issue has been observed in the existing solutions, the cause for which has been traced to the bias in the data. In many instances, certain identity subgroups such as ”mental_illness” have the majority of toxic examples associated with them in the dataset. So, the systems that are trained on such datasets, learn the evident co-relation in the data and classify every mention of such subgroups as toxic. However, there are quite a few comments that might be informational in nature and mentions the subgroups with good intent. The comment: ”Mental illness is no joke. Please show some

[1] Department of Electrical Engineering and Computer Science, York University.

Project Report for EECS5327 (Winter 2021), Department of Electrical Engineering and Computer Science, York University.

sensitivity and try to at least be a little empathic." is not at all a toxic comment but, due to the nature of data which might have the majority of the examples having the similar vocabulary to be toxic in nature, the trained model would be biased.

This un-intended bias needs to be first measured and then made sure to be dealt with while training the models before deploying them in production. To measure the un-intended bias, this work explores the use of a modified Receiver Operating Characteristic - Area Under Curve (AUC) metric which is a weighted average of traditional AUC score and some "bias" AUC metrics. The bias AUC metrics can be used as a stand-alone metric to observe the scores for individual identity subgroups and these can also be utilized by a single metric for final model comparisons. This work also explores the use of frequency-based text representation techniques such as term frequency-inverse document frequency (TfIdf) and contextual embeddings. In terms of modeling techniques, this work explores the performance of the mentioned metrics using Logistic Regression, Random forest, LSTM with auxiliary objective (discussed in the methodology section), and the transformer-based BERT architecture.

This paper is structured such that the "Literature Review" section discusses the advancements made in this domain and then mentions the inspirations behind the novel techniques used in this work. The "Methodology" section formalizes the methods, techniques, and architectures used in this work. "Results" sections discuss the results and scores achieved on the mentioned techniques. Finally, the "Conclusion" section summarizes this work and talks about other problems that can be seen as a scope for future work.

2. Literature Review

There has been a lot of work done related to the identification and classification of toxicity in conversations online, primarily textual in nature. In the work, (Chakrabarty, 2019), the use of Logistic Regression and a simple neural architecture was explored for toxic comment classification. Along the same lines (Georgakopoulos et al., 2018) compared the use of linear models on Bag of Words representation with the use of Convolutional Neural Networks and showed significantly better results with neural approaches. (Saif et al., 2018) also compared the use of Logistic Regression for toxic comment classification with the neural-based models for the same task and reached similar conclusions. (Sharma & Patel, 2018) also worked with deep neural approaches to predict different classes of toxicity: toxicity, severe toxicity, obscenity, threat, insult, or identity hate. They also relied on deep neural approaches for this task. The work by (van Aken et al., 2018) compares the use of deep and shallow networks and proposes the use of an ensemble of models for the same task and does in-depth error analysis. In the survey work, (Andročec, 2020) covered most of the work done till now in the field. This work shows that 90% of the work focuses on using linear models or LSTMs for the task of predicting toxicity in conversational texts. However, the most popular choice of metrics in all of these works was accuracy, f1 score, or ROC-AUC score. These metrics are global in nature and do not take into account different identity subgroups while calculating the scores.

The work by (Vasserman et al., 2018) observed that while building such models, there was an un-intended bias for more frequently targeted groups (e.g. words like "black", "muslim", "feminist", "woman", "gay" etc). The sentence "I am a gay woman" was wrongly classified as toxic by many advanced neural approaches due to the mentioned bias. This work proposed several solutions such as user feedback, balancing training samples, pinned-AUC score, and collaboration of datasets. Further, (Borkan et al., 2019) performed an in-depth study of the contribution of metrics for this un-intended bias that propagates due to the use of simple metrics such as accuracy. They propose the use of threshold-agnostic metrics to counter the bias.

In terms of model-based strategy, the attention mechanism (Vaswani et al., 2017) followed by language model-based pre-training and fine-tuning transformer architectures: GPT-2: (Radford et al., 2018) and BERT: (Devlin et al., 2018) rocked the NLP world. These transformer-based techniques have shown better results than the traditional sequence-based RNN techniques.

Inspired by the need for better metrics to incorporate the performance on the identity subgroups and better modeling strategies that could capture context as well as sequence, this work explores the use of bias AUC metrics (subgroup AUC, BPSN AUC, BNSP AUC) alongside traditional AUC and also does a comparative study of performance of various modeling approaches, such as linear models, ensemble models, LSTM models, and transformer models, on these metrics.

3. Methodology

3.1. Dataset Description

The dataset used in this work contains several textual comments from online platforms with a target label, specifying toxicity in the comments. Further, information regarding the mention of specific identity subgroups are also given as a floating point score for each of the comments. The identity subgroups are as follows: *male*, *feamale*, *transgender*, *other_gender*, *heterosexual*, *homosexual_gay_or_lesbian*, *bisexual*, *other_sexual_orientation*, *christian*, *jewish*, *muslim*, *hindu*, *buddhist*, *atheist*, *other_religion*, *black*, *white*, *asian*, *latino*, *other_race_or_ethnicity*, *physical_disability*, *intellec-*

tual_or_learning_disability, *psychiatric_or_mental_illness*, *other_disability*. The target value being greater than or equal to 0.5 represents the classification label. The identity subgroups don't contribute to the classification task, rather they are used for the final metric calculation. In this work 200,000 such training examples were used and 100,000 samples were used for reporting of results at evaluation time.

The labeling of the target label: toxicity and the identity subgroups was performed by trained annotators (refer "Software and Data" section for more details). The scores were assigned as the probabilities of the results from multiple annotators, for instance, if 7 out of 10 annotators mentioned that a certain subgroup was present in a text, then that subgroup was assigned a score of 0.7.

Some example instances from the data are as follows:

Comment: "I picture this clown yelling "fathisht" with a massive lisp, what being on the Autism spectrum & all.."

- **Toxicity Label:** 1
- **Subgroups:** intellectual_or_learning_disability (0.7), other_disability (0.1), psychiatric_or_mental_illness (0.2) (all others: 0)

Comment: "Why did you assume all the nurses in the hospital were females?"

- **Toxicity Label:** 0
- **Subgroups:** female (0.8) (all others: 0)

Comment: "Your efforts will eventually pay off. Stand strong LGBT community."

- **Toxicity Label:** 0
- **Subgroups:** homosexual_gay_or_lesbian: 1, bisexual: 0.5, transgender: 0.5 (all others: 0)

The dataset also has meta-data regarding the comment authorship and annotation details which are not discussed in this section as they were not utilized in this work.

3.2. Metrics

To measure the un-intended bias, a new metric, which is a weighted combination of some sub-metrics, is explored. The final metric is a weighted average of "simple AUC" and "biased AUCs". This helps to monitor bias for toxic and non-toxic comments on individually targeted subgroups mentioned above. The simple AUC score is quite popular and needs no explanation, however, the concept of subgroup biased AUCs is novel and interesting and needs further explanation.

The simple AUC metric captures the performance on the classification task only. It doesn't take into account whether performance on certain sub-groups is worse than others or whether there exist some sub-groups that have toxic and non-toxic comments which are very difficult to differentiate from each other by the trained model. The goal of introducing subgroup biased AUCs is to able to measure performance on each of the subgroups as well as have one single metric which takes into account this performance on each of the subgroups.

3.2.1. Subgroup AUC

This sub-metric is calculated by considering only toxic and non-toxic comments mentioning a specific identity subgroup. A high score of subgroup AUC indicates that the toxic and non-toxic comments pertaining to the specific subgroup are easily differentiable by the model.

3.2.2. Background Positive Subgroup Negative (BPSN) AUC

This sub-metric is calculated by considering only toxic comments not mentioning a certain identity subgroup and the non-toxic comments mentioning the same identity subgroup. This metric will have a low score if the model predicts high toxicity for non-toxic comments mentioning a certain identity subgroup.

3.2.3. Background Negative Subgroup Positive (BNSP) AUC

This sub-metric is calculated by considering only toxic comments mentioning a certain identity subgroup and the non-toxic comments not mentioning the same identity sub-group. This metric will have a low score if the model predicts low toxicity for toxic comments mentioning a certain identity subgroup.

3.2.4. Overall Metric

The formula for the biased AUCs is given as:

$$M_p(m_s) = (\frac{1}{N_s}\sum_{s=1}^{N_s} m_s^p)^{(1/p)}$$

Where:

$M_p =$ Mean function of power p
$N_s =$ Number of identity subgroups
$m_s =$ the biased AUC calculated as subgroup/BPSN/BNSP

The final metric is the weighted sum of the "simple AUC" and biased AUCs given as:

$$\text{Overall Score} = w_{simple}AUC_{simple} + \sum_{b=1}^{B} w_b M_p(m_s, b)$$

Where:

$$\begin{aligned} w_{simple} &= \text{Weightage to simple AUC} \\ w_b &= \text{Weightage to biased AUC} \\ B &= \text{Number of biased AUCs} \\ (m_s, b) &= \text{the biased AUC for b} \in \text{(subgroup/BPSN/BNSP)} \end{aligned}$$

The overall metric is a single metric score that takes into account simple AUC as well as biased AUCs. Further, the biased AUC scores can be used to measure the model performance on the subgroups individually.

3.3. Machine Learning Framework

In the training and evaluation for the given problem, the following data representation and modeling techniques were used.

3.3.1. Frequency Based Representation

To represent text into a numerical vector format such that it can be fed into the machine learning models, the first approach to be used was Term-Frequency Inverse Document-Frequency (TfIdf). In this, each text comment is represented by a fixed-size vector of TfIdf scores of the words that appeared in the piece of text. The machine learning models that were utilized for this task were:

- Logistic Regression
- Random Forest Classification

3.3.2. Word Embeddings

To capture the contextual information, Glove word vectors were used. Each text comment was represented by the mean of the word embeddings of all the words that appeared in the text. Logistic Regression and Random Forest classification were used for training and evaluation for this representation as well.

3.3.3. LSTMs

For the toxicity classification task, the context and sequence of words in the text seemed to be extremely relevant. So, to capture the sequence of words appearing in the comment text, LSTMs were also experimented within this work. A concatenation of Glove embeddings and FastText Crawl embeddings was used to initialize the embedding layer in the LSTM based architecture used for the toxicity classification. In addition to the primary objective of toxicity classification, an auxiliary objective of predicting the strength of the identity subgroups was also introduced for the LSTM based architecture. The multi-task objective of toxicity classification and subgroup prediction would help the model train better. The architecture used is shown in figure 1.

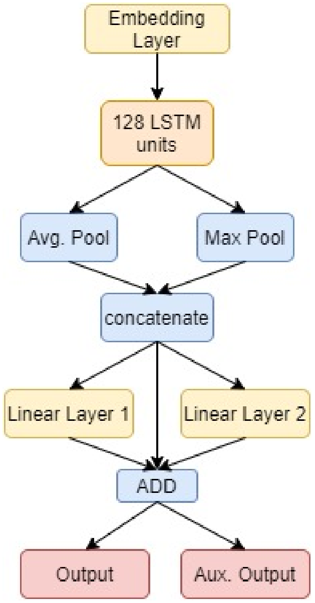


Figure 1. LSTM based architecture with auxiliary objective

3.3.4. Transformers

Although LSTMs can capture information in a sequence, attention-based mechanisms such as Transformers can capture the whole context very naturally as they enable every word to peek at every other word in the sequence, and with the assistance of positional embeddings, the sequential information is also captured. Due to the mentioned benefits of transformers over LSTMs, an extremely popular transformer-based model - BERT (Bidirectional Encoder Representations from Transformers) was also utilized for the classification task. The BERT model was fine-tuned from the pre-trained BASE model rather than training the model from scratch. The architecture is given in figure 2.

The results for all of the mentioned mechanisms and techniques are covered in the Experiments section.

4. Experiments

The results table in the figure 3 depicts the Overall AUC score (weighted sum of simple AUC and bias AUCs) obtained from all the methods explained in the ”Methodology” section.

The TfIdf feature representation yielded ”Overall AUC”

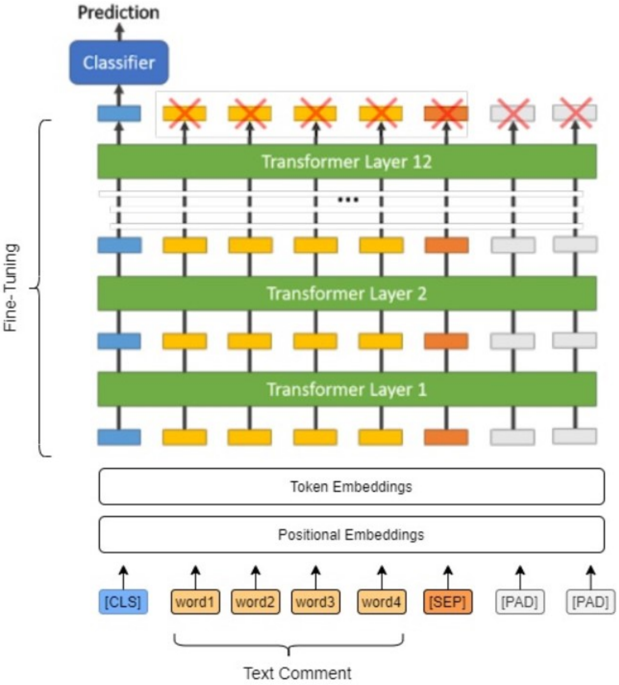
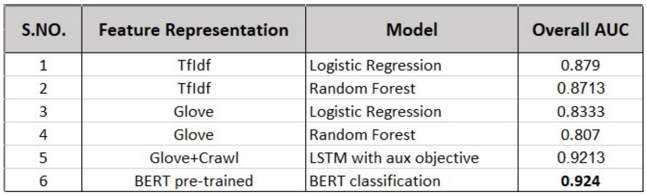
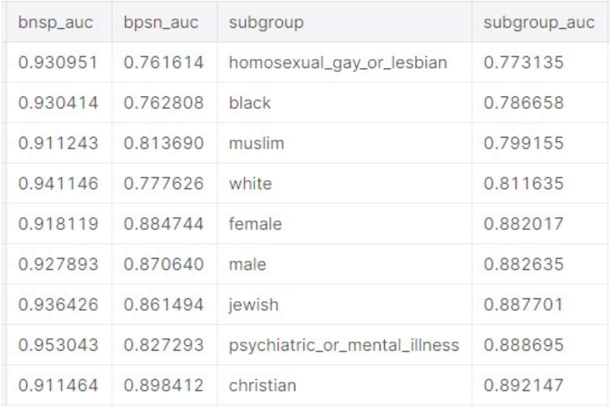


Figure 2. BERT fine-tuning



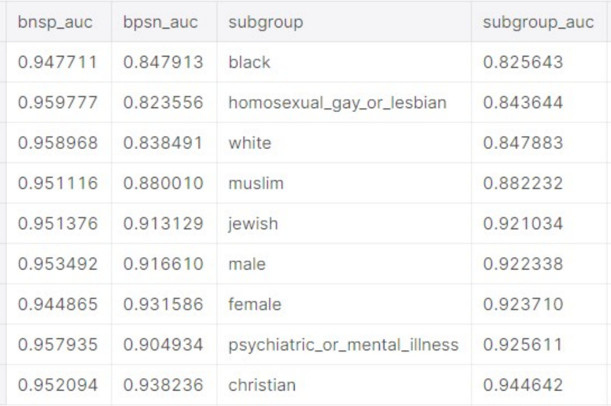
S.NO.	Feature Representation	Model	Overall AUC
1	TfIdf	Logistic Regression	0.879
2	TfIdf	Random Forest	0.8713
3	Glove	Logistic Regression	0.8333
4	Glove	Random Forest	0.807
5	Glove+Crawl	LSTM with aux objective	0.9213
6	BERT pre-trained	BERT classification	**0.924**

Figure 3. Comparison of different models



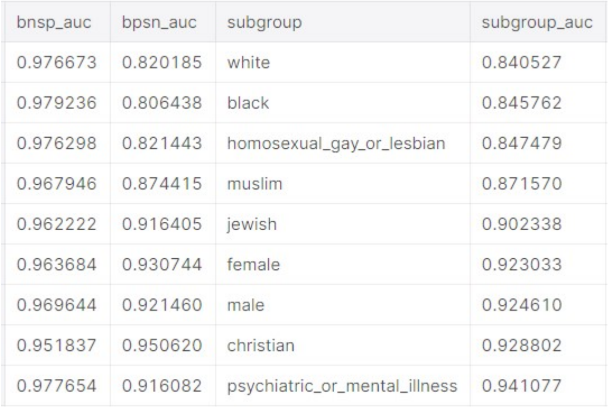
bnsp_auc	bpsn_auc	subgroup	subgroup_auc
0.930951	0.761614	homosexual_gay_or_lesbian	0.773135
0.930414	0.762808	black	0.786658
0.911243	0.813690	muslim	0.799155
0.941146	0.777626	white	0.811635
0.918119	0.884744	female	0.882017
0.927893	0.870640	male	0.882635
0.936426	0.861494	jewish	0.887701
0.953043	0.827293	psychiatric_or_mental_illness	0.888695
0.911464	0.898412	christian	0.892147

Figure 4. Identity subgroup scores (Logistic Regression)



bnsp_auc	bpsn_auc	subgroup	subgroup_auc
0.947711	0.847913	black	0.825643
0.959777	0.823556	homosexual_gay_or_lesbian	0.843644
0.958968	0.838491	white	0.847883
0.951116	0.880010	muslim	0.882232
0.951376	0.913129	jewish	0.921034
0.953492	0.916610	male	0.922338
0.944865	0.931586	female	0.923710
0.957935	0.904934	psychiatric_or_mental_illness	0.925611
0.952094	0.938236	christian	0.944642

Figure 5. Identity subgroup scores (LSTM)



bnsp_auc	bpsn_auc	subgroup	subgroup_auc
0.976673	0.820185	white	0.840527
0.979236	0.806438	black	0.845762
0.976298	0.821443	homosexual_gay_or_lesbian	0.847479
0.967946	0.874415	muslim	0.871570
0.962222	0.916405	jewish	0.902338
0.963684	0.930744	female	0.923033
0.969644	0.921460	male	0.924610
0.951837	0.950620	christian	0.928802
0.977654	0.916082	psychiatric_or_mental_illness	0.941077

Figure 6. Identity subgroup scores (BERT)

scores of 0.879 and 0.8713 for Logistic Regression and Random Forest respectively, which were significantly higher than the results from the same models trained on Glove embeddings. It appears context without sequence is not close to the ideal solution. A major increase in the metric scores was observed when sequence models were used. The LSTM based architecture with auxiliary training objective was able to get the metric score of 0.9213. Finally, the transformer-based BERT model when fine-tuned the toxicity data achieved the highest score of 0.924.

The figures 4, 5, and 6 show the bias AUC scores (subgroup, BNSP, BPSN) for some of the identity subgroups for Logistic Regression (with TfIdf), LSTM based model and BERT model respectively. As per the expected behavior, the scores for Logistic Regression are clearly much worse as compared to the scores from the sequence model and the transformer model. Interestingly, the results are not as obvious for the LSTM and BERT models. The BERT model achieved a higher ”Overall AUC” score as compared to the LSTM model. However, there are quite a few sub-metrics where LSTM based results are better than the BERT. The BERT model seemed to do better on the BNSP AUC, while the LSTM model performed better on the BPSN metric. For instance, in the identity subgroup ”female”, the LSTM based model achieved better scores on BPSN AUC as well as the subgroup AUC as compared to BERT results. Therefore, the individual bias scores (figures 4, 5, and 6) can assist in choosing the models for deployment in production or hyper-parameter tuning based on a specific task dedicated to the use-case (such as having the requirement for high BPSN scores for ”female” identity subgroup).

5. Conclusion

In this work, a new metric: ”Overall AUC” was explored for catering to the need of having a metric that could account for the performance of toxicity classification for individual identity subgroups such as ”white”, ”male”, ”female”, ”jewish” and so on. The ”Overall AUC” score is comprised of traditional AUC and three bias AUCs (subgroup AUC, BPSN AUC, BNSP AUC). The traditional AUC is responsible for measuring the performance of the models globally, i.e. all subgroups together. And, the bias AUCs are responsible for measuring the performance of the model on individual identity subgroups. The formalization of the combination of traditional AUC and bias AUCs along with its practical use-case was demonstrated in this work.

In terms of representing text in numerical format, TfIdf, and contextual embedding techniques: Glove and FastText Crawl were explored. In terms of modeling, Logistic Regression, Random Forest, LSTM based architecture with an auxiliary objective of predicting the subgroups, and transformer-based BERT models were explored. LSTM based model and BERT produced impressive results of 0.9213 and 0.924 on the ”Overall AUC” metric. Further, a limited qualitative analysis showed the importance of observing the individual bias AUC metrics in driving certain use-cases based on the identity subgroup-inspired tasks.

Scope for Future Work

This work assumes that all the toxic comments would use the words and characters from the vocabulary that we know of. However, it has been observed by many platforms that miscreant users have exploited the systems and used alternate ways to communicate their hate. The alternate ways include tricks such as using special characters instead of the alphabets that would fool the machine, but would be easily understood by humans. Some examples include the use of the symbol ”*” to mask out some obvious characters or the use of character ”ğ” instead of ”g”, and so on. Some people draw obscene diagrams using the letters of the alphabet to propagate hate. These challenges are not handled by the use of vocabulary-oriented models as used in this work and require further work.

Software and Data

Experiments

All the experiments were run on free GPU resources provided by Kaggle and Google Colab. No model used in this work took more than 5 hours of train time on the freely available GPU resources. So, this work can be easily reproduced using the resources made available by Kaggle and Google.

The code and results are compiled in form of Jupyter Notebooks in the Github Repository: link. To reproduce the results mentioned in this work, refer the Readme.md file and follow along the cells of the individual notebooks.

Data

The data was made public in a Kaggle challenge: Jigsaw Toxicity Challenge by Jigsaw, a unit within Google.

Tools & Libraries

Deep learning architectures were trained using PyTorch and Transformers library. NLTK and Scikit-Learn have been used for text-based feature extraction and basic model training. Glove embeddings and FastText Crawl embeddings were used as contextual word embeddings. Numpy was used for generic vector-related tasks.

References

Andročec, D. Machine learning methods for toxic comment classification: a systematic review. *AActa Univ. Sapientiae Informatica 12 DOI: 10.2478/ausi-2020-0012*, pp. 205–216, 2020.

Borkan, D., Dixon, L., Sorensen, J., and Thain, N. Nuanced metrics for measuring unintended bias with real data for text classification. *pre-print*, 2019.

Chakrabarty, N. A machine learning approach to comment toxicity classification. *Arxiv*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171–4186, 2018.

Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., and Plagianakos, V. P. Convolutional neural networks for toxic comment classification. *SETN ’18: Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, (35):1–6, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2018.

Saif, M. A., Medvedev, S. A. N., Medvedev, M. A., and Atanasova, T. Classification of online toxic comments using the logistic regression and neural networks models. *AIP Conference Proceedings 2048, 060011*, 2018.

Sharma, R. and Patel, M. Toxic comment classification using neural networks and machine learning. *International Advanced Research Journal in Science, Engineering and Technology*, 5(9), 2018.

van Aken, B., Risch, J., and Krestel, R. Challenges for toxic comment classification: An in-depth error analysis. *Proceedings of the 2nd Workshop on Abusive Language Online (co-located with EMNLP)*, 2018.

Vasserman, L., Li, J., Adams, C., and Dixon, L. Unintended bias in creating identity-based models. Technical report, Jigsaw, Google, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Aidan N. Gomez, K., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.