COMPARISON OF NLP MODELS ON CLASSIFYING TOXIC COMMENTS WHILE MINIMIZING LEARNED BIAS

Armanda Lewis, Lakshmi Menon, Tamar Novetsky
Center for Data Science
New York University

Sameen Reza
Steinhardt School of Culture, Education and Human Development
New York University

Abstract

In this project, we analyze the effects of bias in datasets and experiment with different language models to compare which is least vulnerable to inconsistencies in data. Specifically, we build models to analyze the toxicity level of a comment and evaluate its performance. Following this, we use different metrics to detect learned bias in the models and compare their results. This project suggests methods that can help improve both the validity of data as well as the performance of the model to estimate bias in text.

1 Introduction

Bias refers to information that can help decision making and is an essential part of all AI and machine learning systems. However, bias that arises out of cultural influence or preconceived notions towards specific genders, races or religion, violates fairness of systems and leads to harmful behavior (Recasens et al., 2013). It is important to detect, measure and reduce bias to maintain the integrity of these systems. There are two sources of bias in language models. Firstly, the data can contain an imbalance of statements relating to various identity groups, and models that use this data will implicitly learn to treat text that is associated with those identities differently. Another source of bias is the data itself being labeled incorrectly. Prior studies have focused on detecting biases in language models (e.g (Nangia et al., 2020)) or bias in datasets used to build the models (Dixon et al., 2018) but very few comparative studies are available that discern sources of bias in text and language models and their impact.

Our project analyzes the effects of bias in datasets and experiments with different language models to compare which is least vulnerable to such inconsistencies in data. Specifically, we build models to analyze the toxicity level of a comment and evaluate their performance, and then use different metrics to detect learned bias in the models and compare their results. Our project builds on previous studies to advance the use of machine learning algorithms to detect bias in online text-based communication.

2 Background

Machine learning models have been shown to learn "human-like semantic biases", often due to biases present in the training text corpora (Caliskan et al., 2017). This biased training data in turn encodes bias in word embeddings. Zhao et al (2019) found evidence of bias learned by ELMo word embeddings, which were used to create a state-of-the-art conference system which inherited the bias learned by the word embeddings (Zhao et al., 2019). The ConversationAI team also found that many models built to detect bias incorrectly learned to associate certain identities with toxicity due to the toxic contexts in which they often occurred (Jigsaw and ConversationAI, 2019). This is an example of how models are often trained to perform well only in the context of the data on which they are trained. In recent years, more researchers are realizing the importance of analysis and mitigation of bias in machine learning models, particularly for natural language processing. With this in mind, many different methods of testing and reducing bias have been developed, including datasets for measuring model bias (Nangia et al., 2020), metrics for measuring unintended bias (Borkan et al., 2019), and models for analyzing biased language (Recasens et al., 2013). In this study we use these metrics to compare learned bias among different models.

3 Data

Our data comes from an open dataset released by the Civil Comments platform (Bogdanoff, 2017) that was set up to improve the way that participants engaged in public dialogue around news topics and contains almost two million comments annotated in the following way (ConversationAI, 2020):

- 1,902,194 comments were coded (by humans) for toxicity, obscenity, sexual explicitness, threats, insults, and identity-based attacks, with codes not being mutually exclusive; and
- 360,000 comments were annotated for 24 identity-related categories. The most represented (over 500 examples in test set) identities included male, female, homosexual/gay/lesbian, christian, jewish, muslim, black, white, and psychiatric/mental illness.

Due to restrictions in computational resources, we used a subset of the 2M rows for this project.

4 Methodology

We conducted exploratory analysis of our data, and then built various models to do the following separate but interlinked tasks:

- How well does the model measure the toxicity of a given comment?
- How may we mitigate bias, i.e. minimize unintended learned bias?

In order to compare these, we experimented with NLP models using different word embeddings (GloVe and Common Crawl), and developed models using LSTM, BERT, and XLNet techniques. We trained each model separately using the Civil Comments training dataset and evaluated based on their performance on the Civil Comments test set. We describe these model architectures and their performances in the Models section. The best-performing configuration of each model was then optimized to reduce unintended bias, using methods suggested by Jieyu Zhao et al. (Zhao et al., 2019) and Lucas Dixon et al. (Dixon et al., 2018). We describe this technique in the Bias Mitigation section.

4.1 Bias Mitigation by Data Augmentation

Data augmentation has been found in previous research to perform well in reducing bias caused by imbalanced datasets (Zhao et al., 2019; Dixon et al., 2018). This method was found to be effective in cases where the model learned to associate certain identity terms with toxicity, as the dataset they were trained on contained more examples of these identities in a toxic context. In other words, the model learns that the identity itself is toxic, without considering the context it appears in. Data augmentation in this case involves adding enough non-toxic examples of each identity such that their proportion among the toxic comments is equivalent to their prior distribution in the overall dataset.

We experimented with two different methods for carrying out this data augmentation. In the first method, we started out with 50,000 random samples of the data added non-toxic samples of each identity according to the formula

$$n_{add} = \frac{p_{toxic} \cdot n_{total} - n_{identity}}{1 - p_{toxic}}$$

where
n_{add} is the number of non-toxic samples of a given identity to be added
n_{total} is the total number of comments
$p_{toxic} = \frac{n_{toxic\_identity}}{n_{toxic\_total}}$ is the proportion of the given identity in toxic comments.

This was done iteratively for each identity until the toxic distribution matched the prior distribution. Using this method, we were able to achieve a perfect match between the two distributions for nearly all identities.

However, as seen in Figure 1 in the appendix, the data already had a very large class imbalance between toxic and non-toxic comments, and adding more non-toxic comments further increased that imbalance. The resulting dataset contained 58,677 samples, of which only 5.63% were toxic. This resulted in poor performance by the models, as they tended to predict the majority class in all cases.

To overcome this, we attempted a second method of data augmentation. For this method, we started with a smaller, balanced dataset of 10,000 comments, of which 50% were toxic. We also relaxed the stopping condition for data augmentation, so that non-toxic samples were added until the identity was no longer oversampled in the toxic class, but we did allow it to be slightly undersampled. This resulted in a dataset of size 16,476, of

which 30.35% of samples were toxic. We were able to use this dataset to train our models, to see if the new data affected the models' performances. The distributions of subgroup proportions for both of these methods can be seen in Table 2 and Table 3 of the Appendix.

4.2 Toxicity Classification Models

4.2.1 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a language model that makes use of attention mechanism to learn the contextual relations between words based on all words to the left as well as the right of the target word. (Devlin et al., 2018). In this study, we used BERT as an encoder with an added classification layer to predict if the comment is toxic or not.

In the initial model, we used train, validation and test of sizes 10K, 1K and 2K comments respectively. Using pre trained BERT word embedding, each comment word was converted to its 768 dimension BERT vector. After conversion to vectors, a convolutional neural network was applied. The model was trained on three kernel sizes (2,3 and 4) using a batch size of 100, 10 epochs and optimizer as Adam. The lowest validation loss model was saved as the best model.

Model performance on validation and test dataset was evaluated. The dataset was augmented using a balanced number of toxic and non toxic comments. A new model was trained on the balanced dataset and the its performance was evaluated for comparison with the performance on unbalanced dataset.

4.2.2 LSTM

LSTMs build on the traditional RNN architecture and add a collection of gates that allow for selective carrying forward of information (Jurafsky and Martin, 2009). In our project, we experimented with the "vanilla" LSTM model, as well as one that was coupled with pre-trained word embeddings. We used GloVe embeddings based on Wikipedia (300), as well as the Common Crawl embeddings. Coupling these embeddings together was a best practice observed in literature (Pennington et al., 2014). Hyperparameter tuning was performed on the learning rate and our final configurations for our LSTM models included:

- LSTM (no GloVe embeddings): We had hyperparameters with a learning rate of .01; 3 epochs; a batch size of 32; max length of 200; and an architecture that includes an embedding layer with spatial dropout (.1); MaxPooling1D layer; 1 bidirectional LSTM layer (2 layers used for embedded version); Dense(50, activation="relu")

4.2.3 XLNet

XLNet is a generalized autoregressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order. (Yang et al., 2020) The XLNet model was built in reference to the model built by Josh Xin Jie Lee for a similar toxicity classification task (Xin Jie Lee, 2019). For this model, the input data was first tokenized, with sequence length limited to 300 tokens. The XLNet model was trained with a batch size of 8. It used the AdamW optimizer with a learning rate of 2e-5 and was trained for 3 epochs. The XLNet-Base-Cased architecture was used, which has 12 layers, 768 hidden layers, and 12 attention heads.

The initial model was trained on the randomly sampled 50K dataset, and then retrained on the 58K augmented dataset for comparison. The model did not perform well on the augmented dataset due to the class imbalance, and also ran into computational restrictions. Therefore, for later iterations, the 10K balanced dataset was used along with the 16K augmented dataset for comparison.

4.3 Analysis and Results

Three main metrics were used to evaluate the models. Overall performance on the toxicity classification task was measured using overall AUC scores. Bias between identities was measured using two variants of AUC for each subgroup. The BPSN AUC metric computes the AUC of the within-subgroup negative examples and the background positive examples. The BNSP AUC computes the AUC of the within-subgroup positive examples and the background negative examples. We hypothesized that a difference in subgroup AUC between groups indicates that although the model may have a high AUC overall, it is not performing equally among different subgroups. Furthermore, a higher value for BNSP AUC may indicate bias, as it implies that the model performs better when the subgroup is positive (i.e toxic). Similarly, if the BPSN AUC value is higher, it indicates that

the model performs better on that subgroup when it is negative (i.e. non-toxic).

Table 1 summarizes our training and test results in terms of AUC scores. The "Before" columns indicate model performance prior to data augmentation, and the "After" columns indicate model performance after data augmentation.

Model	Train		Test	
	Before	After	Before	After
LSTM	0.933	0.951	0.911	0.919
LSTM (GloVe)	0.931	0.932	0.939	0.937
BERT	0.90	0.99	0.675	0.95
XLNet	0.87	0.74	0.86	0.8

Table 1: Model Performance (AUC Scores)

4.3.1 BERT

The model performed fairly well on the train test (AUC = 0.90) but the performance on the test set was very low in comparison (0.675). This is likely both due to the model being trained on a small subset of data, as well as a sampling issue as the random sample drawn from the imbalanced dataset has a large class imbalance in terms of toxicity.

The model trained on the balanced data gave a better performance. The loss converged after 10 epochs [Figure 3] and the model gave an AUC score of 0.99 on the train dataset which reduced to 0.95 on the test [Table 1].

The results indicate that class imbalance affects the accuracy results drastically. Using a balanced dataset helped to train a model that gave reasonable results.

4.3.2 LSTM

The LSTM model that was trained with pre-trained embeddings performed better on the test set than the non-GloVe version. This is logical since the pre-trained embeddings help the model to identify better importance input sequences. The GloVe embeddings are based on word co-occurrence statistics and are generated from popular texts. Based on previously-mentioned scholarship, we combined two sets of GloVe embeddings (Wikipedia and Common Crawl) to increase the model's knowledge of co-occurrence.

Figure 4 (Appendix) shows the loss curves for the pre-trained embeddings/post-data augmentation version of the LSTM model, and indicates that the validation data was easier to fit than the training data. Both versions of LSTM performed relatively stably when looking at the before and after data augmentation runs.

4.3.3 XLNet

The XLNet model showed relatively stable performance before and after augmentation. The overall AUC values for train and test can be seen in the Model Performance (AUC Scores) table.

The initial XLNet model (i.e. 10K non-augmented samples) appeared to have less of a discrepancy between subgroups, suggesting that it may not be learning a strong bias associated with identities. After augmentation, the model showed more of a discrepancy between subgroups. To check whether this was merely a consequence of the increased data size, the model was trained again using a random, balanced dataset, with size matching that of the augmented dataset. This model produced results comparable to the performance after augmentation, suggesting that the augmented dataset did not largely impact the XL-Net model performance. The subgroup- AUC results table of these three configurations can be seen in Figures 5, 6, and 7 in the Appendix.

No_data_aug

	subgroup	subgroup_size	subgroup_auc	bpsn_auc	bnsp_auc
2	homosexual_gay_or_lesbian	45	0.387218	0.743217	0.549073
0	male	90	0.722500	0.798864	0.838488
7	white	46	0.750583	0.805236	0.863668
5	muslim	12	0.771429	0.805996	0.864494
3	christian	47	0.776423	0.836442	0.837509
6	black	33	0.796296	0.796354	0.914832
8	psychiatric_or_mental_illness	17	0.807692	0.869822	0.827877
1	female	129	0.816959	0.793329	0.889108
4	jewish	8	1.000000	0.884302	0.968167

Data_aug

	subgroup	subgroup_size	subgroup_auc	bpsn_auc	bnsp_auc
4	jewish	22	0.600000	0.805538	0.682847
5	male	195	0.609708	0.835772	0.630357
8	white	232	0.635406	0.802533	0.698075
3	homosexual_gay_or_lesbian	130	0.687548	0.801989	0.749204
0	black	135	0.693600	0.825464	0.711503
1	christian	131	0.794261	0.829469	0.798646
6	muslim	65	0.831967	0.811349	0.875091
2	female	286	0.856618	0.812805	0.877410
7	psychiatric_or_mental_illness	54	1.000000	0.802367	0.994626

Figure 1: Subgroup AUCs (LSTM model)

5 Discussion and Future Work

Due to computation issues, we worked with a subset of the original dataset, which placed a limit on the generalizability of our models and their results. This also made it difficult for us to detect bias in the models before or after data augmentation, and in at least one case made it seem like

the data augmentation worsened the bias. In future work, we would like to train and augment with the full dataset and see how this shapes the results.

There are several other next steps that can be taken with this work. In addition to making use of the full dataset, we also plan on running trained models on new datasets, such as Wikipedia comments and social media sites with robust comment features.

Since data augmentation is a viable method to mitigate bias in datasets, we plan on exploring ways to enhance the data augmentation process (Rastogi et al., 2020; Recasens et al., 2013). In our current study, we augmented our subset with rows from the larger dataset. A future study could look at the impact of data augmentation with external data, and we could reveal potential inherent bias issues with the Civil Comments dataset. We can also use additional word embeddings such as ELMo in appropriate models.

We looked at model performance on identity subgroups. One interesting topic is to look at intersectionality for combined sociocultural groups such as Black women and transgender LatinX youth. This interdisciplinary concept examines interactions between combination of groups, and any model that looks for bias should look into any effects (Cho et al., 2013; Nash, 2008).

Further, as an additional metric of model-introduced bias, we plan on evaluating the model based on its performance on the Crow-S Pairs dataset for measuring social biases in masked language models (Nangia et al., 2020).

References

[Bogdanoff2017] Aja Bogdanoff. 2017. Saying Goodbye to Civil Comments, December.

[Borkan et al.2019] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, pages 491–500, San Francisco, USA, May. Association for Computing Machinery.

[Caliskan et al.2017] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April.

[Cho et al.2013] Sumi Cho, Kimberlé Williams Crenshaw, and Leslie McCall. 2013. Toward a Field of Intersectionality Studies: Theory, Applications, and Praxis. *Signs: Journal of Women in Culture and Society*, 38(4):785–810, June.

[ConversationAI2020] ConversationAI. 2020. Conversation AI. Github repository (URL: https://conversationai.github.io/).

[Dixon et al.2018] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA, December. ACM.

[Jigsaw and ConversationAI2019] Jigsaw and ConversationAI. 2019. Jigsaw Unintended Bias in Toxicity Classification (URL: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview), March.

[Jurafsky and Martin2009] Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall, Upper Saddle River, N.J, 2nd ed edition. OCLC: 213375806.

[Nangia et al.2020] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. *arXiv:2010.00133 [cs]*, September. arXiv: 2010.00133.

[Nash2008] Jennifer C. Nash. 2008. Re-Thinking Intersectionality. *Feminist Review*, 89(1):1–15, June.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

[Rastogi et al.2020] Chetanya Rastogi, Nikka Mofid, and Fang-I. Hsiao. 2020. Can We Achieve More with Less? Exploring Data Augmentation for Toxic Comment Classification. *arXiv:2007.00875 [cs]*, July. arXiv: 2007.00875.

[Recasens et al.2013] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria, August. Association for Computational Linguistics.

[Xin Jie Lee2019] Josh Xin Jie Lee. 2019. Multi-Label Text Classification with XLNet, November.

[Yang et al.2020] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding.

[Zhao et al.2019] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. *arXiv:1904.03310 [cs]*, April. arXiv: 1904.03310.

6 Appendix

Identity	% of Toxic	% of Total	
		Before	After
asian	0.12	0.09	0.12
black	2.18	0.62	2.18
christian	2.15	1.02	2.15
female	5.12	2.57	5.12
homosexual	2.33	0.71	2.33
jewish	0.42	0.19	0.42
male	3.42	1.98	3.42
muslim	1.12	0.30	1.12
mental_illness	1.06	0.34	1.06
white	3.66	1.03	3.66

Table 2: Data Augmentation Method 1

Identity	% of Toxic	% of Total	
		Before	After
asian	0.58	0.53	0.58
black	11.40	7.22	11.40
christian	7.76	6.75	7.82
female	19.36	16.32	19.67
homosexual	5.94	4.97	6.05
jewish	1.90	1.46	2.14
male	15.14	12.53	17.01
muslim	5.74	3.66	6.12
mental_illness	2.58	2.16	2.59
white	15.62	10.26	17.23

Table 3: Data Augmentation Method 2

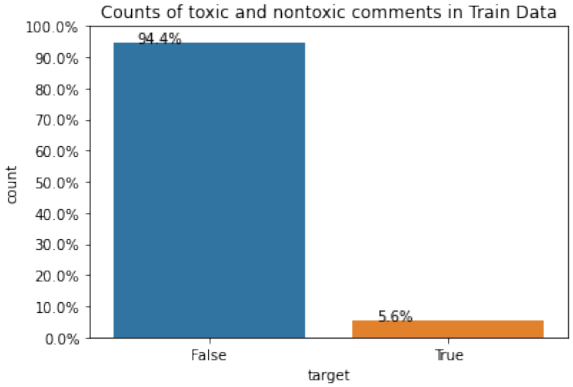

Figure 2: Toxic/Nontoxic data

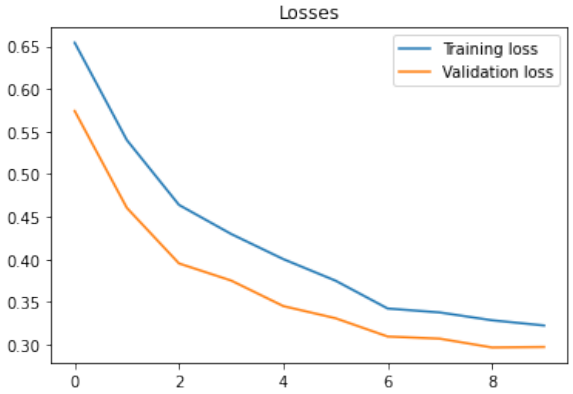

Figure 3: Train/Val loss for BERT (balanced dataset)

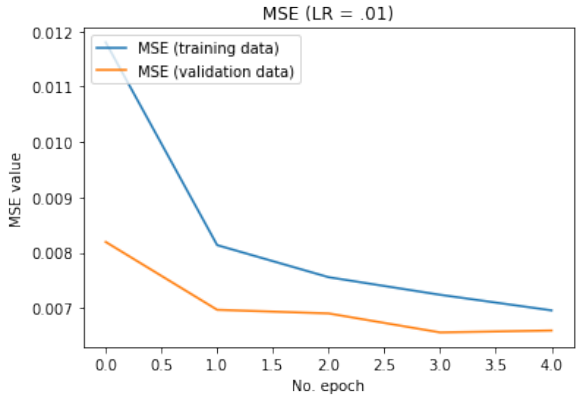

Figure 4: Train/Val loss for LSTM

	subgroup	subgroup_size	subgroup_auc	bpsn_auc	bnsp_auc
0	jewish	24	0.689474	0.694054	0.859961
1	muslim	45	0.663851	0.757762	0.772631
2	christian	115	0.872815	0.821370	0.914750
3	homosexual_gay_or_lesbian	26	0.856250	0.809601	0.909442
4	male	117	0.858333	0.802850	0.918974
5	white	46	0.698198	0.771035	0.793355
6	black	35	0.903846	0.806160	0.959572
7	female	139	0.877724	0.852336	0.888094
8	psychiatric_or_mental_illness	16	0.801587	0.849815	0.816429
9	asian	16	0.933333	0.837615	0.959342

Figure 5: XLNet Metrics (16K non-augmented)

	subgroup	subgroup_size	subgroup_auc	bpsn_auc	bnsp_auc
0	jewish	24	0.821053	0.790009	0.868907
1	muslim	45	0.570946	0.820305	0.593894
2	christian	115	0.784528	0.837147	0.786946
3	homosexual_gay_or_lesbian	26	0.868750	0.837307	0.868709
4	male	117	0.919048	0.784333	0.969869
5	white	46	0.725225	0.762072	0.802665
6	black	35	0.884615	0.750939	0.969187
7	female	139	0.914245	0.826586	0.921270
8	psychiatric_or_mental_illness	16	0.801587	0.814010	0.825900
9	asian	16	0.933333	0.802351	0.968825

Figure 6: XLNet Metrics (10K non-augmented)

	subgroup	subgroup_size	subgroup_auc	bpsn_auc	bnsp_auc
0	jewish	24	0.600000	0.819338	0.584399
1	muslim	45	0.611486	0.806999	0.609447
2	christian	115	0.767920	0.813022	0.756942
3	homosexual_gay_or_lesbian	26	0.750000	0.818299	0.734409
4	male	117	0.740476	0.809129	0.734323
5	white	46	0.653153	0.806538	0.651114
6	black	35	0.722222	0.818766	0.706597
7	female	139	0.801049	0.808502	0.793801
8	psychiatric_or_mental_illness	16	0.730159	0.761580	0.770256
9	asian	16	1.000000	0.816121	0.984412

Figure 7: XLNet Metrics (10K augmented)