Domain Shifts in Civil Comments

Richard Huang, Aditya Retnanto, Aim Wonghirundacha

Abstract

Toxic comment classifiers suffer from unintended bias due to spurious correlations between toxicity and demographic information. In this project, we build a naive baseline model and then we apply various fairness and domain approaches to the model such as fairness through unawareness and IRM. The resulting accuracies were in line with expectations, where the applications lowered overall average accuracy but improved classification for toxic comments, and can be further enhanced with more advanced models.

1 Introduction

CivilComments is a dataset created after the closure of its website. Such data can be used to study the unintended bias in classification of toxic comments written on the internet. Automated toxic comment classification (such as rude, intolerable, or hateful speech, etc.) is important for moderating internet forums such as social media. However, previous instances of toxic comment classification were found to be biased in training. For identities that were frequently attacked online, the model was biased towards classifying non-toxic comments as toxic. This is explained by the fact that the model was picking up on spurious associations between the mentioning of certain racial/religious demographics and toxicity.

When classifying toxic comments, relying on spurious associations is undesirable because it would make the model unreliable when deployed. Suppose these models were used as part of an automated system to remove toxic comments. It would be undesirable for the system to remove comments based primarily on identity or demographic information. A model which relied on demographic information to predict toxicity would be especially unsuitable for an environment or forum with frequent discussions of identity. Considering that the comment toxicity classification could be used in a range of different environments, it is important to ensure that models accurately classify toxic comments without relying on demographic information. Unintended bias has negative consequences for out-of-distribution generalization.

We would like a classifier not to depend on demographic attributes, which can be formalized through different notions of fairness such as demographic parity, equalized odds, and predictive parity. However, since toxicity and identity are unlikely to be independent, and due to the impossibility of the multi-fairness theorem, only one of the distributional notions of fairness can hold. In this case, we think it is most important to maintain predictive parity to ensure that the model is able to correctly predict whether a comment is toxic or not without relying on demographic information. Other notions may be less appropriate. For example, demographic parity would require that the predictions are invariant to identity groups and to ensure that identity carried no information about the prediction. However, we know that this is not a good idea because we might expect toxic comments to be more directed at certain demographic groups than others. For each of our models, we check their accuracy for different groups.

Our goal for this project is to apply trustworthiness approaches to a classification model in order to reduce these spurious associations. A baseline, naive, model will be constructed to serve as a control, while two other models are attempts to alleviate the issue of a unintended bias due to spurious associations.

Preprint. Under review.

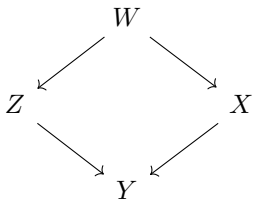

Figure 1: Causal DAG

2 Data

We utilize preprocessed Civil Comments [3] from the WILDS dataset [4]. The dataset is split into training (269,038), validation(45,180) and testing sets (133,782). Each row in the training data consists of the body of the comment, along with five metrics for toxicity: severe toxicity, obscene, identity attack, insult, and threat, measured between 0-1. 8 binary demographic metrics are also ascribed to each comment, which *male*, *female*, *LGBTQ*, *Christian*, *Muslim*, *other religions*, *Black*, and *White*. Our label of interest is a binary label of toxicity. The dataset is imbalanced, where a large proportion is considered toxic.

3 Approaches

3.1 Baseline Model

Our baseline model is a simple neural network that consists of an embedding layer, a GRU layer, and lastly a linear layer for the output. We preprocess our model by first finding the 50 - Dimensional GloVE [5] embeddings of an input text. The output of GloVE maps semantic similarity to numerical similarity. We further limit the number of words to 1000 and pad shorter sentences. The embeddings of the text are passed into the GRU layer, the output of which passes through the linear layer and is fed through a sigmoid layer where an output >= is classified as toxic. We utilize a batch size of 32 and train for 20 epochs. Additionally, we handle the data set imbalance by utilizing a weighted mean-square error. We anticipate that the baseline model will pick up spurious associations of mentions of protected identities to toxicity.

3.2 Fairness Through Unawareness (FTU)

Our second model utilizes the general structure of the previous baseline model, but instead, we implement fairness through unawareness by dropping any instances of identity in the training set.

Fairness through unawareness excludes protected features from the training data. This is in line with the idea that demographics should be treated as protected attributes and thus should not be considered in the learning of a model. However, fairness through unawareness frequently fails because there may be an association between the protected attribute and the label Y through other non-direct paths.

To predict toxic comments, the mention of identities is likely to give us some information about whether the comment is toxic or not. Toxic comments against particular demographic groups may be why the task is important in the first place. Therefore, we expect that an approach that tries to ensure fairness through unawareness by dropping instances of identity would have a lower overall accuracy.

Additionally, identity is still associated with other words in the comment. Figure 1 shows a causal directed acyclic graph (DAG) with Y being the perceived toxicity, Z being the demographic or identity terms, X being other words, and W being the writer of the comment. Even if we removed Z, there is still a path associating Z and Y through the writer. Since an association between the identity terms and the other features may still be present in the training data, fairness through unawareness in this case may still be relying on demographic information.

3.3 IRM

Our last model utilizes invariant risk minimization to classify the comments.

Invariant Risk Minimization was proposed as a solution for out-of-distribution generalization by assuming that the underlying causal structures across different environments stays the same [2]. The goal is to learn a data representation that is invariant across these groups, such that a predictor that minimizes risk in all groups should perform reliably out-of-distribution. This could help avoid scenarios where the model learns to rely on spurious correlations. Adragna et al. empirically validate IRM using the Civil Comments dataset by constructing two train environments and one test environment [1]. In the training environment, they induce a correlation between the sensitive attribute and the toxicity and reverse this correlation in the test set.

In this project, we try to use IRM in a different way in the context of toxic comment classification, by treating different demographic groups as different environments. This is not an unreasonable assumption because we could imagine that comments about a particular demographic group come from a discussion forum for that particular group. Framing the problem in this way, we would like our model to be able to classify toxic comments accurately in all these different environments. In particular, when trying to solve the IRM problem:

$\min_h \sum_{e \in E} \mathbb{E}_{(x,y)\ e}[L(h(x), y)]$
$E = \{e_1, e_2, e_3 ... e_n\}$

Each e_k is one of the protected demographics. By treating each demographic as a separate environment, the aim of the model is to find a generalized model that works well at classifying toxic comments in all environments, overrepresented or not. This is done by utilizing the 'CombinatorialGrouper' function from the WILDS library which yields 256 combinations of environments for 8 protected environments. We utilize a batch size of 128 proportionally sampled from 4 different environments.

All models were scored by both the overall accuracy score as well as individual demographic accuracy scores (classification accuracy on toxic comments belonging to a certain demographic). These were collected and are represented in the table below.

4 Results

	ERM Accuracy (%)	Unaware Accuracy (%)	IRM Accuracy (%)
Average Accuracy	68.61	55.40	55.23
Worst Group Accuracy	23.13	46.65	39.08
Not Toxic/Male	74.40	57.70	49.85
Toxic/Male	24.19	51.11	78.53
Not Toxic/Female	75.30	59.02	53.64
Toxic/Female	23.13	52.47	74.86
Not Toxic/LGBTQ	74.89	59.53	42.61
Toxic/LGBTQ	25.66	52.22	80.34
Not Toxic/Christian	72.76	61.66	67.69
Toxic/Christian	27.22	51.11	68.17
Not Toxic/Muslim	75.22	60.47	44.61
Toxic/Muslim	25.81	46.65	80.09
Not Toxic/Other Religions	74.43	61.68	56.64
Toxic/Other Religions	26.73	49.81	73.85
Not Toxic/Black	75.26	62.82	40.24
Toxic/Black	25.63	48.15	81.33
Not Toxic/White	72.88	57.96	39.09
Toxic/White	24.62	50.00	81.43

Figure 2: Accuracy table by model, separated by toxicity/demographic.

The performance of each model was evaluated based on two key metrics: Average Accuracy and Worst Group Accuracy. Additionally, we analyzed the performance of each model across various demographic groups and toxicity levels.

The Average Accuracy metric provides a measure of the overall performance of the model across all demographic groups and toxicity levels. The baseline model achieved the highest Average Accuracy of 68.61%, followed by the FTU model at 55.40%, and the IRM model at 55.23%. This suggests that the baseline model is the most effective at correctly classifying comments across all groups and toxicity levels. This result makes sense: the baseline model was designed to reduce empirical risk - loss over all the data, and thus should have the highest average accuracy score amongst the three.

When measuring by Worst Group Accuracy, the FTU model was better than the others. The Worst Group Accuracy metrics helps provide insight into how well the model can perform across diverse environments and is a good measure of the model's fairness. The FTU model was designed to ignore identity in training and thus could perform better.

Examining results by demographic and toxicity level, we see that the baseline model significantly outperforms the other models when classifying not toxic comments. The IRM model, despite having the lowest average accuracy, outperforms the others when classifying toxic comments.

Overall, each model had their own strength that was attributed to their purposeful design for this study. These three models illustrate the tradeoff between fairness and accuracy.

5 Conclusion

The application of trustworthiness for the CivilComments classification problem proved to have promising results. Both the FTU and IRM models performed reasonably well due to their respective designs and strengths.

The accuracies in our results were limited by the strength of our embedding model as well as neural network architecture. Although reasonably powerful, GloVE is known to fail when encoding out-of-vocabulary words. A lot of internet slang that humans can read as toxic is lost on embedding models to begin with, and sarcasm is also usually not picked up on for embedding models, especially for the simple one we used for this project. Transformer based models like BERT (the one used for the original study) are much more powerful at understanding context and are usually trained to capture and understand toxicity to begin with.

In the context of our study, the three models face several challenges in classifying text for toxicity. These challenges are inherent to the nature of language, the structure of the models, and the characteristics of the training data.

Language is inherently ambiguous and context-dependent, which poses a challenge for all three models. The baseline model may struggle with instances where the same phrase has different meanings in different contexts. The FTU model, which is designed to ignore protected demographics, may inadvertently ignore crucial context provided by these attributes. The IRM model, which learns invariant predictors across different environments, may struggle when the meaning of a word or phrase varies across contexts.

Bias in the training data is also a challenge. The baseline model is particularly susceptible to reproducing any biases present in the training data. The FTU model, by design, ignores the protected attributes and thus may perpetuate biases associated with these attributes. The IRM model may still learn biased predictors if these predictors are consistently useful across the environments represented in the training data. In this dataset, one specific limitation is that 256 environments have to be modeled. We can not assume that the interaction effects between protected identities are minimal. Certain combinations will be sparse and thus require bootstrapping.

The lack of representativeness in the training data can lead to poor performance in underrepresented groups. It is particularly problematic for the Unaware model, because if these attributes are correlated with underrepresentation in the training data, the Unaware model may perform particularly poorly on these groups.

While each model offers a unique approach to the challenge of toxicity classification, they all face significant challenges due to the complexity and diversity of human language. Addressing these challenges requires careful consideration of the training data and model design, as well as ongoing evaluation and adjustment of the model's performance.

6 Code

Code can be found at: https://github.com/aretnanto/TrustworthyProject
*Note that there is implementation of Distributionally Robust Optimization (DRO) - results and analysis were not reported due to performance issues.

References

[1] Robert Adragna, Elliot Creager, David Madras, and Richard Zemel. Fairness and robustness in invariant learning: A case study in toxicity classification, 2020.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[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*, 2019.

[4] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee,

Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2021.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. 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 2014. Association for Computational Linguistics.