Testing Improvements in Toxic Language Detection without Demographics through Adversarially Reweighted Training

Abstract—**Fairness in Machine learning (ML) has become a heated topic in recent years as ML systems have made rapid headway into critical socio-technical systems in recent years. Specifically, "Fairness without Demographics," the ability to debias ML models without access to protected groups, has become an important open challenge to solve. In our work, we analyze bias in Toxic Language Detection Models. We apply a novel approach called Adversarilly Reweighted Learning (ARL), described by Lahoti et al., to debias the models without providing access to more sensitive information such as the race and dialect from the dataset. Our work suggests that there are no significant improvements using ARL for debiasing our toxic language detection model both in overall metrics as well as in metrics from worst-performing groups due to the potential lack of computational identifiability of our data domain.**

INTRODUCTION

Machine learning has made rapid headway into socio-technical systems in recent years ranging from surveillance systems to automated resume screening[1]. As a result of these systems being increasingly used for decision-making in high-stake scenarios, there has been heightened public concern about the impact of digital technology on society. Specifically, research in recent years has cast doubts and concerns about ML fairness, where researchers discovered significant accuracy disparities across demographic groups in face detection[2], healthcare systems[3], and recommendation systems[4]. ML fairness has since been brought onto the center stage in recent years as a key challenge to overcome for the industry.

In this paper, we focus our attention on a specific ML Fairness problem in the Natural Language Understanding domain – bias in language toxicity detection. Current hate speech or toxic language detection systems exhibit bias towards minority languages and dialects[5]. Specifically, research has found that leading AI models for processing hate speech were one-and-a-half times more likely to flag tweets as offensive or hateful when they were written by African Americans and 2.2 times more likely to flag tweets written in African American English (which is commonly spoken by black people in the US)[5]. Since all social media platforms use AI models to assist its tasks of flagging and removing hate speech and toxic speeches, this bias in the model could potentially lead to more inequality for minorities.

We build on top of the work done by Zhou et al. [6] to understand how to de-bias models for the task of toxicity detection. Zhou et al. explore the then-current state-of-the-art debiasing methods for a biased tweet dataset and try to understand whether debiasing methods currently available are effective in removing the biases for toxicity detection. It analyzes two kinds of biases in language – lexical bias, which associates toxicity with the presence of certain words (e.g., profanity, identity mentions) [7], and dialectal bias, where toxicity is correlated with the surface marker of African-American English (AAE)[8]. The authors applied three types of techniques for debiasing, data filtering, building ensemble models with additional attributes identifying potential bias, and dataset correction that removes the bias from the dataset. Out of the three methods, the authors concluded that relabelling biased datasets is the best approach.

However, in the real world, data relabelling is a time-consuming process that's hard to conduct for multiple social, economic, and ethical reasons. On the other hand, acquiring additional attributes that identify potential bias could be difficult or illegal in some cases due to privacy, regulatory, and ethical reasons. For example, understanding whether a tweet comes from an African-American user could potentially correct the model from flagging their tweets as toxic due to their use of African-American English. But in reality, this attribute is rarely available.

Therefore we explore a new method proposed by Lahoti et al., which outlines a method named Adversarially Reweighted Learning (ARL) that improves the performance of models on worst case groups without the need for protected group attributes [9]. Lahoti et al.'s method allows models to improve without any additional data on top of the original datasets provided through an adversarial network, where the adversarial model dynamically rebalances the weights to have the primary model target performance of worst-performing groups.

As our contribution, we will show how an ARL-enabled model compares to models defined by Zhou et al. on debiasing the biased tweet dataset. We adapted the ARL structure onto a RoBERTA model used as the baseline model for Zhou et al. on toxicity detection tasks. The RoBERTA model served as the primary learner while a simple linear model was used as the adversary.

We hypothesize that the ARL structure, when adapted to the toxicity dataset, should improve the overall accuracy or the accuracy of a subgroup (African-American English) of the dataset. The original authors of the toxicity detection problem split the tweets into different dialects based on predictions of a dialect classification model [10], which suggest that there might be subgroups that are computationally identifiable within the dataset. Computationally identifiability is the key prerequisite of ARL.

Comparing the final results with our ARL model, we were not able to identify significant improvements using the ARL structure compared to the baseline outlined in Zhou et al. In the discussion section, we outlined our primary suspicion of why the ARL model did not improve the accuracy for this task, which is due to the potential lack of computational identifiability of the tweets.

Background

A. (Dialectal) Bias in Toxic Language Detection

Toxicity Language Detection is the task of detecting languages that are rude, offensive, toxic, or hateful. However, due to the highly complex and nuanced nature of language toxicity, current debiasing methods for traditional NLU tasks cannot be adopted to solve bias in toxic language detection. Unlike bias in traditional NLU tasks that are mostly generated during data creation, bias in toxic language detection is affected by their social dynamics of the world[11][12]. For example, regarding African-American English as a more toxic variant of English is a form of linguistic discrimination that is hard to be corrected by debiasing methods[13].

For our work, we will primarily focus on dialectal bias.

B. Dataset

We apply the same dataset on our ARL models as mentioned in Zhou et al, which is one adapted from a widely used hate speech English tweet dataset from Founta et al [14]. The original dataset consists of 100,000 tweets categorized by crowd-sourced annotators into different hate speech categories. Annotators only had access to the tweet itself, and no additional attributes (demographics of users, etc) were provided.

In Zhou et al, the authors modified the dataset in two ways. First, the paper only focused on tweets in the datasets that were *hateful*, *abusive*, or neither(*healthy*), and discarded other categories (*spam*). Hateful and Abusive tweets collected were relabelled as toxic. As a result, the dataset in use has 32k toxic tweets and 54k non-toxic tweets. Second, the paper trained a classification model that classified each tweet to be one of four dialects, based on the highest probability. The four dialects are *African-American English*, *White-aligned English*, *Hispanic*, and other. In our paper, we only focus on African-American English (AAE) and White-aligned English (WAE). Our work uses this modified dataset from Zhou et al for the sake of a fair comparison.

Other than the aforementioned modified dataset, Zhou et al. also explored relabelling the adapted toxic dataset through a AAE-WAE (White American English) Translator. The relabelled dataset result, which is the best amongst all results, is also shown in the table. Our ARL model is only tested on the normal toxic dataset without relabeling. For more detail of how the relabelling is done and the intuition, see the original paper from Zhou et al. Stats from the relabelled dataset are displayed only for comparison. We decided to not train our ARL model on the relabelled dataset as one of our assumptions is to understand whether ARL model is sufficient in replacing the relabelling efforts.

Finally, Zhou et al. also conducted data filtering methods on the same adapted datasets to produce Random, AFLITE, and DataMaps. Data Filtering methods were applied to reduce the bias during model training. We also applied the ARL models to the filtered datasets as well to understand the impact of the model with filtered datasets. For more detail on the datasets, see Zhou et al.

C. Protected Groups and lack of access to Protected Groups in Datasets

Protected groups, or protected features, are features that cannot be used as a reason to discriminate against someone [15]. In the United States, there are 8 such features: race, color, religion, sex, national origin, age, disability, and genetic information [15]. Protected attributes have been used by many effective de-biasing methods to reduce the bias of ML models at training [16] or at inference [17]. However, Protected groups, due to its privacy, legal, and regulatory concerns, are often precluded from datasets for ML algorithm training. Therefore, one of the biggest open challenges in the ML domain is to understand how to "address fairness without demographics". In our work, we investigate ARL to understand its ability to address the bias issue of toxicity detection without utilitizing protected attributes such as the race or sex of the twitter user writing the tweets.

D. Model Background

1) Rawlsian Max-Min fairness: Fairness is defined as the absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics [18][19]. In machine learning, three types of fairness have been defined[18]: Individual Fairness [20][21], Group Fairness [20][21], and Subgroup Fairness [22]. For Adversarially Weighted Learning, the paper authors Lahoti et al. use subgroup fairness, which will be the same utility function we use for our implementation of ARL on the toxic tweet dataset.

In order to quantify the improvement of subgroup fairness, Lahoti et al. uses Rawlsian Max-Min Fairness as the utility function of such fairness. The goal of Rawlsian Max-Min fairness is to maximize the benefit of the least-advantaged members of society. In practice, this equation chooses the worst performing group in a scenario based on the utility function, and improves the outcome of such a group. Equation of Rawlsian Max-Min Fairness is defined as:

$$J(\theta, \phi) = min_{\theta}\ max_{\phi}\ \Sigma_{i=1}^{n} \lambda_{s_i} \cdot l_{ce}(h_{\theta}(x_i), y_i)$$

$h(.)$ is a model that minimizes loss over the training data

$l(.)$ is some loss function

λ is a learned assignment of weights that maximizes the weighted loss of some group s

2) Computational-Identifiability: Lahoti et al., focus on the notion of computational identifiability to achieve fairness without explicit protected group features. This concept utilizes features in the input set that correlate with protected groups (which are often the cause for bias in the task) by targeting computationally-identifiable regions of errors for subgroups. A sub-group S is computationally identifiable if, for a given set of (binary) functions F, if there exists a function $f(x, y)$ which maps to 1 if and only if (x, y) ∈S.

3) Adversarially Reweighted Learning: ARL is modeled as a min-max objective between a learner and an adversary. The adversary is modeled to learn computationally-identifiable regions where the learner makes significant errors and is designed to return a higher value in higher loss regions by maximizing expected loss. The output is used inorder to obtain weights λ_i which are then rescaled to place a high weight on regions with a higher probability of errors [9].

$$J(\theta, \phi) = min_\theta \ max_\phi \ \Sigma_{i=1}^{n} \lambda_\phi(x_i, y_i) \cdot l_{ce}(h_\theta(x_i), y_i)$$

Lahoti et al., perform normalization of adversary outputs in order to prevent exploding gradients and constraint the optimization problem with $\lambda_i > 0$. Adding 1 ensures all the training examples are taken into consideration [9].

$$\lambda_\phi(x_i, y_i) = 1 + n \cdot \frac{f_\phi(x_i,y_i)}{\Sigma_{i=1}^{n} f_\phi(x_i,y_i)}$$

Hypothesis and Method

E. Hypothesis

Previous related work in debiasing toxicity detection models hints towards computational identifiability of dialectal subgroups [10], which forms our hypothesis towards testing whether such computational identifiability can be leveraged to utilize debiasing methods without the need for explicit protected features during training. For our experiments, we assume specific dialectal subgroups are computational identifiable, on the basis of which we examine the performance of ARL on natural language data for toxicity detection.

F. Implementation

Lahoti et al., test the original ARL on numeric and categorical data with standard feed-forward networks for both the models. In our experiment we extend the same to natural language using a binary classifier as the learner, trained to classify a particular tweet as toxic or not by minimizing the expected loss. For this, we fine-tune the RoBERTa-Large classifier for toxicity detection [23].

Our adversary model is a feedforward neural network with 682 and 341 hidden units. We tie the embeddings of dimension 1024 from our learner as adversary inputs after a defined number of pre-training steps. Although the adversary model shares the input embedding of the learner, gradients are only updated through the learner loss. Our experiments with more complex adversary models did not yield promising improvement in observed results and greatly increased the required training time due to limited compute resources.

Table I shows the hyperparameters used to train and evaluate the implementation of the vanilla RoBERTa model presented by Zhou et. al. Table II shows the hyperparameters used to train and evaluate the proposed ARL model.

Hyperparameter	Value
Number of Training Epochs	3
Learning Rate	1e-5
Training batch size	16
Evaluation batch size	16

Table I
TRAINING AND EVALUATION PARAMETERS FOR THE VANILLA ROBERTA MODEL.

Hyperparameter	Value	
	Learner	Adversary
Training Epochs	3	3
Learning Rate	1e-5	1e-5
Training batch size	16	16
Evaluation batch size	16	16

Table II
TRAINING AND EVALUATION PARAMETERS FOR THE ARL MODEL.

Results

G. Overall Performance Evaluation

The main results quantifying the performance of the base vanilla RoBERTa model defined in the paper by Zhou et al., as well as the proposed ARL model are seen in Table III, Table IV and Table V given below, and Table VI, and Table VII given in the Appendix.

	Test Data (12,893)	
Model	F1	FPR$_{aae}$
Vanilla (Zhou et al)[6]	92.33	16.84
Vanilla (Re-implementation)	92.44	17.25
ARL	**92.5**	17.00
LMixin-Dialect (Zhou et al)[6]	92.26	16.07
AAE-relabelled (Re-implementation)	91.58	**12.94**

Table III
COMPARISON BETWEEN F1 SCORES AND FPR (SPECIFICALLY FOR AAE DIALECT) FOR DIFFERENT MODELS TRAINED ON THE FULL TRAINING DATASET.

Table III provides a comparison of model performances using F1 score and FPR for African American English (AAE) dialect tweets. It is observed that the proposed ARL model has the highest F1 score, while the AAE-relabelled dataset implementation of Zhou et. al., has the least FPR.

Table IV compares the accuracy, F1 score and FPR, for the vanilla RoBERTa implementation given in the paper by Zhou et al, our implementation of the vanilla RoBERTa model, and

	Full Dataset (12893)		
	Acc	F1	FPR
Vanilla (Zhou et al)[6]	94.21	92.33	-
Vanilla (Re-implementation)	94.29	92.44	4.88
ARL	**94.36**	**92.50**	**4.81**

Table IV
COMPARISON BETWEEN ACCURACY, F1 SCORES AND FPR FOR DIFFERENT MODELS TRAINED ON THE FULL TRAINING DATASET.

the proposed ARL model on the full test dataset with 12,893 data samples.

From Table IV, it is observed that there is no significant improvement provided by the proposed ARL model as compared to the baseline RoBERTa model. The ARL model has the least FPR and most accuracy and F1 score on the full test dataset, but the improvements are not very significant.

		Full Test Dataset (12893)		
Dataset (33% of Train Data)	Model	Acc	F1	FPR
Random	Vanilla	94.31	92.48	4.98
	ARL	94.07	92.18	5.34
AFlite	Vanilla	93.81	91.87	5.7
	ARL	93.81	91.89	5.91
DataMap - Ambi	Vanilla	**94.42**	**92.55**	4.27
	ARL	94.38	92.52	4.48
DataMap - Easy	Vanilla	94.02	91.91	**3.8**
	ARL	94.15	92.15	4.14
DataMap - Hard	Vanilla	94.36	92.35	3.45
	ARL	94.31	92.30	3.54

Table V
COMPARISON OF ACCURACY, F1 SCORE AND FPR FOR VANILLA ROBERTA AND ARL MODEL TRAINED ON DIFFERENT DATASETS, AND TESTED ON THE FULL TEST DATASET.[6]

Table V compares the accuracy, F1 score and FPR between the vanilla RoBERTa model and the proposed ARL model trained on 5 different variations of the full train dataset. All 5 variations contain only 33% of the full train dataset. The variation of training data is part of the data filtering methods used by Zhou et al. to analyze if filtering aids in reducing spurious biases in data.

From Table V, we observe that the F1 scores and FPR for the ARL do not suggest any significant improvements. Tthe vanilla RoBERTa model trained on the Datamap - ambi dataset has the highest accuracy and F1 score for the complete test dataset, while the vanilla RoBERTa model trained on the Datamap - easy dataset has the least FPR. However, different models trained on different variations of 33% of training data can cause variations in measuring the stats, as seen in the variations in F1 score and FPR for all variations of the model and training data.

DISCUSSION

Our original hypothesis was to test if there is any significant increase in the performance of the model for specific subgroups, if the model is trained using the ARL methodology. The original hypothesis was made with the assumption that the sensitive fields like race or dialect can be drawn from the given data since they are computationally identifiable.

In our case, the use of ARL did not yield significant improvement over the existing architectures. The results show that the dialect groups might not be identifiable enough for ARL to make significant reweighting efforts to help improve the performance of the worst case group – tweets belonging to African American English Dialects. This is also supported by the dataset, where the mean of the probabilities that a given tweet belongs to the AAE dialect type is lesser than 50% and most data samples seem to have classified a tweet as a particular diet even if none of the 4 probabilities are higher than a given threshold for the multi-class classification. In the original ARL paper by Lahoti et al., the authors encountered a similar problem to what we saw with our dataset where one of their datasets, COMPAS, also had a weaker computationally identifiable group (race) in their dataset, which led to degradation in performance when ARL model is applied, compared to their other datasets. However, since the original paper did not outline the threshold of computationally identifiability or how to quantify computationally identifiability, it's difficult to predict the eligibility of our dataset without testing out using the actual ARL setup.

CONCLUSION

Improving model fairness without directly accessing protected attributes remains a difficult and under-studied challenge in machine learning. In our work, we explored the idea of Adversarially Reweighted Learning on improving the fairness of language toxicity detection without access to demographic information or dialectal information. The key prerequisite to adopting ARL is the computational identifiability of the data domain. The insignificant improvement with ARL in our experiment evidences that the nuance of natural language is far higher than the datasets used in the original ARL implementation and further modifications may be needed to experiment whether such adversarial methods are suitable for improving fairness without demographics information in NLP tasks. More specifically, our work shows that there needs to be more quantifiable ways to evaluate the prerequisite for ARL in various domains in order for future researchers to understand the feasibility of adopting ARL on their tasks based on computational identifiability of sub-groups in their dataset. Until then, the most effective debiasing methods for language toxicity detection will lie within access to protected attributes as well as improvements in annotating and relabelling procedures.

References

[1] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[2] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *FAT*, 2018.

[3] D. N. J. F. A. C. B. M. B. L. R. B. R. Hasnain-Wynia, D. W. Baker and J. S. Weissman., "Disparities in health care are driven by where minority patients seek care: examination of the hospital quality alliance measures," in *Archives of internal medicine*, 2007.

[4] I. M. A. J. D. E. O. A. D. M. M. D. Ekstrand, M. Tian and M. S. Pera, "All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness." in *FAT*, 2018.

[5] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1668–1678. [Online]. Available: https://aclanthology.org/P19-1163

[6] X. Zhou, M. Sap, S. Swayamdipta, N. A. Smith, and Y. Choi, "Challenges in automated debiasing for toxic language detection," *CoRR*, vol. abs/2102.00086, 2021. [Online]. Available: https://arxiv.org/abs/2102.00086

[7] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," 2018.

[8] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," in *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 25–35. [Online]. Available: https://aclanthology.org/W19-3504

[9] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. H. Chi, "Fairness without demographics through adversarially reweighted learning," *CoRR*, vol. abs/2006.13114, 2020. [Online]. Available: https://arxiv.org/abs/2006.13114

[10] S. L. Blodgett, L. Green, and B. O'Connor, "Demographic dialectal variation in social media: A case study of African-American English," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1119–1130. [Online]. Available: https://aclanthology.org/D16-1120

[11] A. Spears, *African-American Language Use: Ideology and So-called Obscenity*, 01 1998, pp. 226–250.

[12] G. Kasper, *Linguistic politeness: current research issues*, 1990.

[13] J. Rosa and N. Flores, "Unsettling race and language: Toward a raciolinguistic perspective," *Language in Society*, vol. 46, no. 5, p. 621–647, 2017.

[14] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press, 2018.

[15] "Protected characteristic." [Online]. Available: https://www.law.cornell.edu/wex/protected_characteristic

[16] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment &amp disparate impact," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, apr 2017. [Online]. Available: https://doi.org/10.1145%2F3038912.3052660

[17] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, 2016, pp. 3323–3331.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, jul 2021. [Online]. Available: https://doi.org/10.1145/3457607

[19] C. W. L. Weytingh, J. Mohazzab and B. Zaalberg., "Reimplementing the adversarially reweighted learning model by lahoti et al." 2021.

[20] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," 2011. [Online]. Available: https://arxiv.org/abs/1104.3913

[21] C. L. J. S. R. Kusner, MJ; Russell, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017.

[22] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," 2017. [Online]. Available: https://arxiv.org/abs/1711.05144

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

Appendix

	Full Dataset (12893)			NOI (602)		OI (553)		ONI (3236)	
	Acc	F1	FPR	F1	FPR	F1	FPR	F1	FPR
Vanilla (Zhou et al)	94.21	92.33		89.76	10.24 1.3	**98.84**	85.710.0	97.34	**64.72**
Vanilla (Re-implementation)	94.29	92.44	4.88	90.11	10.23	98.62	85.71	**97.44**	66.34
ARL	**94.36**	**92.50**	**4.81**	**90.19**	**9.21**	98.8	85.71	97.42	65.36

Table VI

Comparison between different models based on accuracy, F1 score and FPR for full test dataset and individual categories of toxic words in TOXTRIG word list

In Table VI and Table VII, nOI (Non-offensive minority identity mentions), OI (Possibly Offensive minority mentions), and OnI (Possibly offensive non-identity mentions) are three categories of toxic words defined in the TOXTRIG list of words [2].

Table VI VI provides insights on the performance of the vanilla RoBERTa model implemented by Zhou et. al, the vanilla RoBERTa implemented by us, and the proposed ARL method for the full test data set, as well as subsets of the test data consisting of those particular groups of toxic words. Comparing the F1 score and FPR of the three model implementations, it is observed that there is still no significant improvement provided by the proposed ARL method. By comparison, the ARL model has the best accuracy and FPR for the nOI group of toxic words, while the vanilla RoBERTa implementations perform better in the OI and OnI subgroups.

		Full Dataset (12893)			NOI (602)		OI (553)		ONI (3236)	
Dataset (33% of Train Data)	Model	Acc	F1	FPR	F1	FPR	F1	FPR	F1	FPR
Random	Vanilla	94.31	92.48	4.98	89.65	9.21	98.89	78.57	97.42	67.31
	ARL	94.07	92.18	5.34	89.36	9.89	98.89	78.57	97.37	68.78
AFlite	Vanilla	93.81	91.87	5.7	88.85	11.94	98.89	85.71	97.23	70.73
	ARL	93.81	91.89	5.91	**90.06**	11.60	98.89	85.71	97.23	70.73
DataMap - Ambi	Vanilla	**94.42**	**92.55**	4.27	89.66	7.5	98.80	85.71	97.61	61.46
	ARL	94.38	92.52	4.48	89.07	8.87	98.89	85.71	**97.63**	61.95
DataMap - Easy	Vanilla	94.02	91.91	**3.8**	85.15	**4.09**	98.71	85.71	97.13	63.41
	ARL	94.15	92.15	4.14	87.73	5.46	98.80	85.71	97.25	61.95
DataMap - Hard	Vanilla	94.36	92.35	3.45	88.01	6.14	98.80	85.71	97.19	**59.02**
	ARL	94.31	92.30	3.54	88.69	5.46	98.80	85.71	97.26	60.48

Table VII

Comparison between different models based on accuracy, F1 score and FPR for full dataset and datasets with data filtering methods applied

Table VII is similar in the observations provided, to Table V, but it provides additional information on the performance of the vanilla RoBERTa model implemented by us and the proposed ARL model, based on subsets of the test data focusing on the three categories of TOXTRIG words. For nOI toxic words, the ARL model trained using the AFlite dataset has the highest accuracy, and the vanilla RoBERTa model trained with the Datamap - easy dataset has the least FPR. For OI words, all the models trained on different training data have similar FPR and F1 score. For OnI words, the ARL model trained on Datamap-Ambi has the highest accuracy, while the vanilla RoBERTa trained on Datamap - Hard has the least FPR. Once again, there is no significant improvement provided by the ARL model across categories.