COMP90049 – Introduction to Machine Learning

Assessment 3: The Effect of Balancing Techniques on Performance and Fairness in Toxicity Classification

Words: 2172

1. Introduction

In recent years, Machine Learning (ML) models have been employed to make online spaces safer by filtering toxic comments. A toxic comment is "anything rude, disrespectful, or otherwise likely to make someone leave a discussion" [1]. Because toxicity is directed *at* individuals/groups, it manifests differently in terms of identity. As the marginalised already feel the burden of discrimination, a successful classifier must generalise en-masse. Ethics is thus intrinsically tied to performance. In this paper I ask: *how do dataset balancing techniques in toxicity classification affect performance and fairness?*

To explore this question, I use a dataset of comments provided by The Conversation AI team, a research initiative founded by Jigsaw/Google [1]. I limit myself to a single dataset of 155,000 labelled instances of non/toxic comments. Within each instance, twenty-four identity labels have been included, in categories of Religion, Race, Sexuality, Gender, and Disability, each indicating an identity's presence.[1]

I approach my question over five sections. First, I conduct a review of the intersection between the literatures of toxicity classification, imbalanced datasets, and algorithmic fairness. Second, I analyse the dataset, generating my modelling and evaluative approach from its imbalance, choosing to critique fairness and performance through a comparison of Threshold-Adjustment and Synthetic-Resampling. Thirdly, I deliver my results, which demonstrate both data-balancing techniques fail to adequately generalise in toxicity classification, leading to insufficient improvements in performance and fairness. I finish by considering the cause of this failure.

2. Literature Review

Toxicity Classification is not a new problem. Within the literature, most acknowledge the problem of data imbalance, but reply with algorithmically founded solutions [2], [3], [4]. How instead, might a balancing approach proceed?

Since Japkowicz and Stephen's seminal paper on the problem of unbalanced datasets, two distinct approaches have been adopted [5]. Krawczyk terms these "Data-Level", which balances the dataset through under/over-sampling, and "Algorithm-Level" which modifies learning algorithms to alleviate majority bias through cost mechanisms [6]. Dixon et al. exemplifies the former, mitigating unintended bias through the addition of third-party and generic templated data [7]. Whilst Krawczyk and

[1] An unlabelled dataset of 200,000 instances is also available.

Wozniak consider the latter, proposing cost-adjustments through thresholds derived from Receiver-Operating-Characteristic (ROC) curve analysis [8]. I implement a variation of each, discussed in the next section.

How should we evaluate the ethics of these performance enhancing techniques? The Algorithmic Fairness literature has emerged in recent years with the ideal that "algorithmic systems should behave… without discrimination on the grounds of *sensitive characteristics*" [9, p. 2]. Considering Madaio et al.'s framework of harm, toxicity classification runs the risk of *quality-of-service* harm - where the system disproportionately fails for certain identities - and *representation* harm - where the development of algorithmic systems over/under-represents certain identities [10].[2]

Two main approaches exist for evaluating and mitigating fairness [12]. *Anti-Classification*, where protected identities and their proxies are left out of the learning process. Since toxicity classification relies on some *intended* bias towards protected attributes, this method is untenable. Alternatively, *Classification-Parity* (CP) is defined in terms of equal errors in classification across identity groups.[3] For the purposes of toxicity classification, the goal is to minimise *unintended* bias across different groups. I employ and evaluate two metrics of CP in my results.

3. Method

My development process stems from data analysis. Toxic comments represent 16.1% of the training set, indicating strongly imbalanced data (Fig.1). Base rates for group membership are largely divergent across identity categories, with at least one group per category extremely underrepresented pointing to pre-existing representation bias (Fig.2/3).

In maintaining the scope of my project, a single Sentence-Bert (S-BERT) embedded dataset has been adopted, which maps comment sentiment to a 384-dimensional space [14]. Moreover, to evaluate fairness, I only consider a single category, *Gender*. I have chosen this category as it has two similarly represented groups ('Male', 'Female') and two under-represented ('Transgender', 'Other gender').

As a foundation I model a suite of learning algorithms. Utilising Sci-Kit Learn libraries my suite includes: K-Nearest-Neighbours with Euclidean-distance; Bernoulli Naïve Bayes; Logistic Regression with Newton-Conjugate-Gradient descent; Multi-Layer Perceptron optimised through Grid-Cross-Validation [15].

[2] The motivation for the Conversation AI team's development of this dataset comes from wanting to alleviate representation harm effected in a previous challenge. See [11].

[3] There also exists another variant unsuitable for this context, *individual fairness*, which defines fairness as *similar individuals being treated similarly*, see Dwork et al. [13].

Fig.1: Class Imbalance

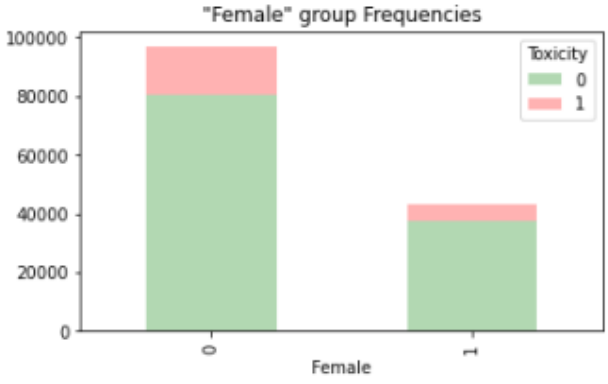


Fig.2: 'Female' Cross-Tabulation

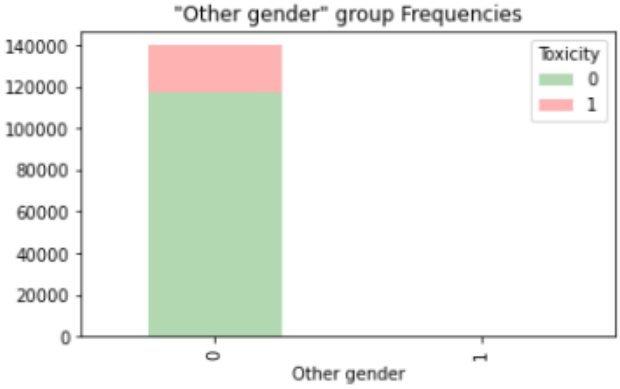


Fig.3: 'Other Gender' Cross-Tabulation

To test data balancing techniques, this suite is compared over three iterations. First, an *initial* run evaluates the base performance and fairness of the classifier suite. Second, an *algorithm-level* run applies thresholds derived from ROC curves to *probability* predictions of each model, giving more weight to toxicity classification [8]. Third, a *data-level* run, where Synthetic-Minority-Oversampling-Technique (SMOTE) balances the data distribution (50:50) before modelling by generating synthetic instances as a convex combination of random samples from the minority class and one of its k-nearest neighbours [16].

To evaluate the *performance*, a holdout strategy is applied, with each model tested on validation (15,000 instances) and training (140,000) sets to control for fitness. A baseline Zero-Rule Classifier (0R) is adopted, classifying all instances as the majority (non-toxic). To capture minority labelling effectiveness, its precision and recall, as well as macro-F1 and ROC scores are employed.

To evaluate *fairness* I adopt two metrics through the Fairlearn API [17]. First, Demographic-Parity (DP), which computes the difference between *selection* rates for toxicity across different groups [18, p. 4]. Where S represents the sensitive group:

$$|P[Y = 1|S = 1] - P\left[\hat{Y} = 1\middle|S \neq 1\right] \leq \epsilon$$

A score of 0 indicates *equal* selection rate. Second, Equalised-Odds (EO), which computes the ratio of the difference between different groups false-positive and false-negative rates [18, p. 5]:

$$\left|P[Y=1|S=1,Y=0]-P\left[\hat{Y}=1\middle|S\neq 1,Y=0\right]\right|\leq\epsilon$$
$$\left|P[Y=1|S=1,Y=1]-P\left[\hat{Y}=1\middle|S\neq 1,Y=1\right]\right|\leq\epsilon$$

A score of 1 indicates *equal* fairness between groups and across classes.

Considering all metrics, a successful model is one which has greater majority *and* minority class accuracy than 0R overall, whilst increasing fairness.

4. Results

4.1 Performance

Four developmental results form the basis of my outcomes. First, to narrow datasets, the initial run resulted in greater *peak* performance in S-BERT over an alternative TFIDF dataset: accuracy (0.83-0.82), precision (0.65-0.64), recall (0.65-0.53), ROC (0.68-0.66). Second, MLP tuning became ad-hoc. Despite using GCV over varying combinations of layers [250-200-150-10], solvers ['lbfgs'-'adam'], and iterations [100-200-300], overfitting often resulted, leading to great accuracy but low ROC in validation. Through brute testing, and to maximise ROC, I settled on (200-100-5) layers, 'lbfgs' solver, and 200 iterations. For similar reasons, the same structure is applied in SMOTE. Third, for Threshold-Adjustment, the threshold's obtained are defined alongside each model (Fig.5), these were derived from trial and error. Fourth, because KNN-5 produced the best ROC results of [1-3-5-10] neighbour variations, this number was used for SMOTE generation.

Model	*Acc*	*1-Pre*	*1-Rec*	*F1*	*ROC*
Zero-R	0.81	0	0	0.45	0.5
KNN (5)	0.8069	0.47	0.21	0.59	0.67 (T) 0.58 (V)
NB	0.6923	0.34	0.65	0.62	0.67 (T) 0.68 (V)
L-Reg	0.8337	0.65	0.26	0.64	0.61 (T) 0.61 (V)
MLP	0.8333	0.59	0.37	0.68	0.71 (T) 0.66 (V)

Fig.4: Embedded-Naïve

Model	*Acc*	*1-Pre*	*1-Rec*	*F1*	*ROC*
KNN (5) (0.3)	0.7493	0.37	0.46	0.62	0.79 (T) 0.64 (V)
NB (0.4)	0.683	0.33	0.67	0.61	0.68 (T) 0.68 (V)
L-Reg (0.2)	0.7539	0.41	0.69	0.67	0.74 (T) 0.73 (V)
MLP (0.17)	0.7507	0.41	0.71	0.68	0.81 (T) 0.74 (V)

Fig.5: Embedded-Threshold-Adjusted

Model	*Acc*	*1-Pre*	*1-Rec*	*F1*	*ROC*
KNN (5)	0.3596	0.22	0.96	0.36	0.69 (T) 0.59 (V)
NB	0.6735	0.33	0.68	0.60	0.71 (T) 0.68 (V)
L-Reg	0.7277	0.39	0.74	0.66	0.78 (T) 0.73 (V)
MLP	0.756	0.40	0.61	0.66	0.84 (T) 0.70 (V)

Fig.6: Embedded-SMOTE

Figures 4-5-6 present my *performance* findings. A general trend emerges in both data-balancing techniques. Both lead to decreased overall accuracy and precision juxtaposed with greater recall and ROC. For example, considering MLP between runs Initial->Threshold->SMOTE: accuracy (0.83->0.75->0.75), precision (0.59->0.41->0.4), recall (0.37->0.71->61), F1 (0.68->0.68->0.66), ROC validation (0.66->0.74->0.7) (see Figures 7/8/9). Overall,

Threshold-Adjustment performs slightly better than SMOTE, with the latter having greatest recall overall, at the expense of correctness. One exception is of note. Naïve-Bayes remains relatively constant across each run: accuracy (0.69->0.68->0.67), precision (0.34->0.33->0.33), recall (0.65->67->0.68), F1 (0.68->0.61->0.6), ROC (0.68->0.68->0.68).

The best models are as follows. In terms of accuracy, *initial* Logistic-Regression and MLP (0.834-0.833) beat the 0R baseline (0.81) but still express low minority class recall (0.26-0.37) and ROC (0.61-0.66). Giving greater weight to the *task* of toxicity classification, Threshold-Adjusted MLP performs best overall, with decent accuracy (0.75), recall (0.71), and ROC (0.74). However, no model successfully beat OR across majority *and* minority classification, indicating failure under the defined definition.

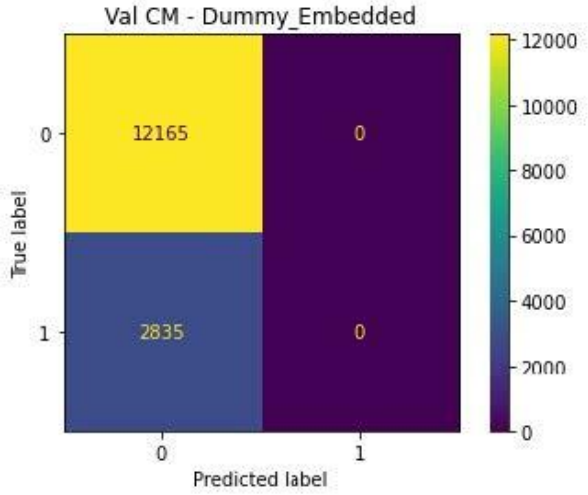


Fig.7: 0R for Comparison

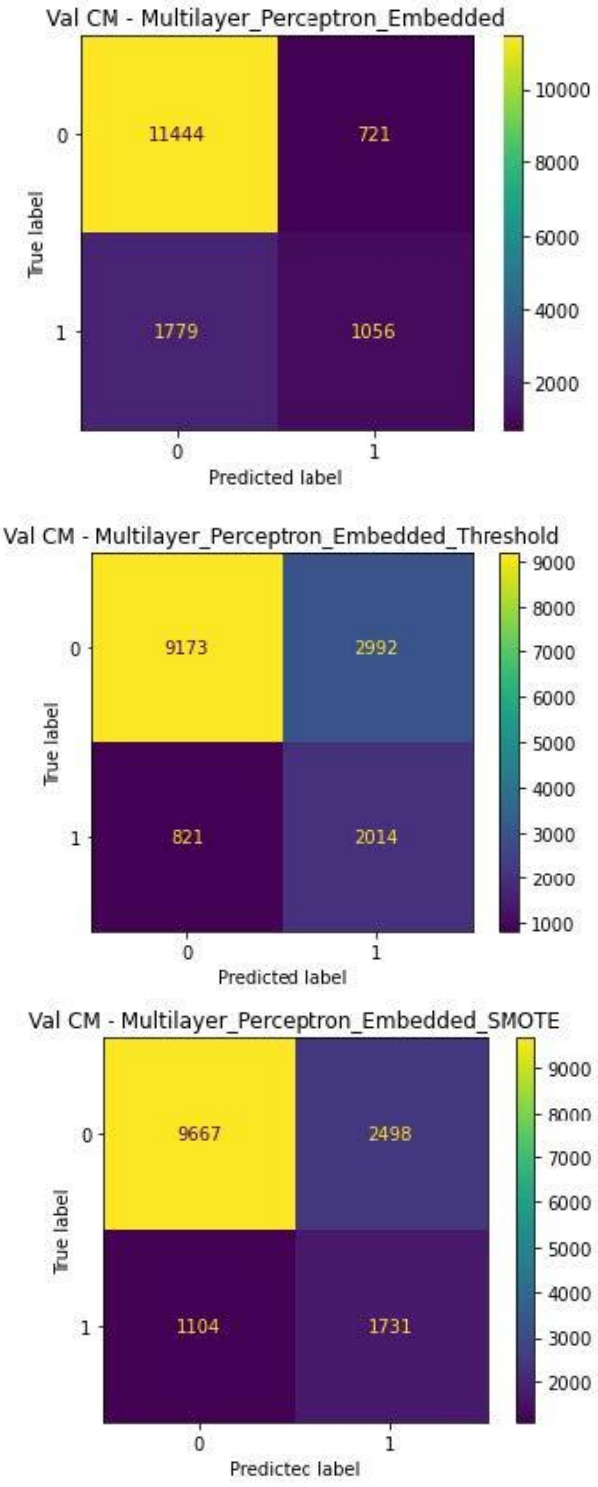
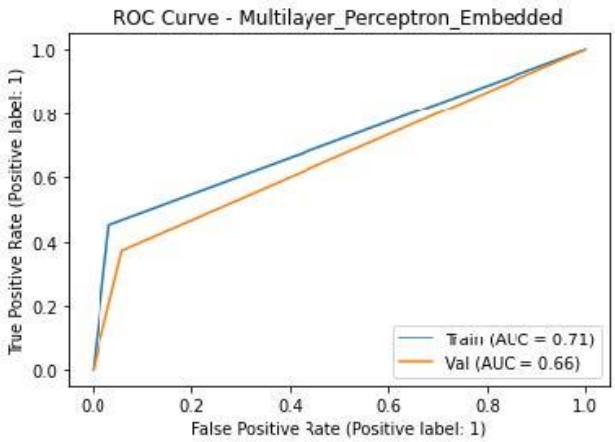
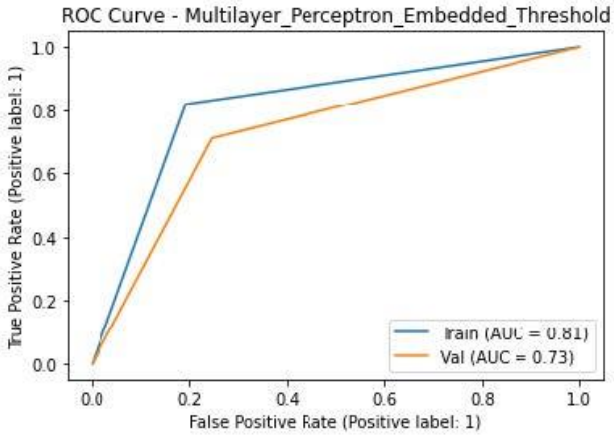
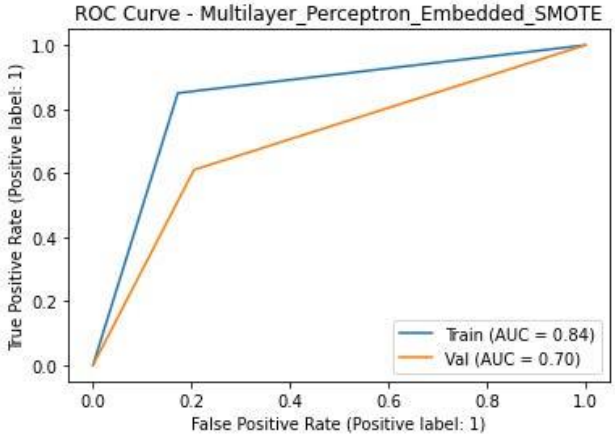


Fig.8: MLP Confusion Matrices







Figs.9: MLP ROC Curves

4.2 Fairness

Figure 10 presents my fairness findings. Briefly, two notable but illusory characteristics present themselves. First, considering the *initial* run, fairness results appear contradictory across each identity group, as DP indicates fairness (close to 0), whilst EO indicates unfairness (far from 1). Again, Naïve-Bayes proves the exception, with worse DP and better EO across groups. Considering initial Naïve-Bayes' recall is 200% greater than the rest of its suite, DP's 'fairness' is merely a reflection of poor minority class prediction. Thus, fairness should be understood only from EO. Second, across each iteration, 'Other gender' delivers drastically better EO, with a notable high score of 0.9612 in SMOTE. Further analysis revealed that only three instances of this group exist in the validation set, none of which are toxic, delivering a false-positive rate of 0. Thus, this illusion of fairness is a distortion generated by representation bias.

To ease my analysis, I consider each model separately under EO. As with performance, Naive-Bayes' results remain consistent across data-balancing techniques deviating less than 20% in each case. Similarly, KNN's fairness remains consistent, however it never deviates from the initial contradiction. We can ascribe this to low recall in Threshold-Adjustment (0.46), and (despite high recall) low accuracy in SMOTE (0.35). The remaining models see improvements across iterations. [4] Notably, for well-represented classes,

[4] Comparisons are made in form (<Threshold>-<SMOTE>).

Logistic-Regression performs better with Threshold-Adjustment compared to SMOTE (0.1355-0.1071 for 'Female'). Whilst the opposite, greater EO in SMOTE, occurs for our remaining under-represented class 'Transgender' (0.117-0.1848). Similarly, under MLP our under-represented class fares better in SMOTE (0.0687-0.1779). And well-represented classes perform better in Threshold-Adjustment (0.1171-0.109 for 'Female').

The best fairness results come from Threshold-Adjusted-Naïve-Bayes, also registering the highest single EO score overall ('Trasngender':0.3497).[5] The latter can be explained considering Figures 11/12/13, which juxtapose results on the 'Transgender' group alongside the remaining dataset over each iteration. Under Threshold-Adjustment, best results come from a relatively lower selection rate, leading to greater precision. However, recall still remains stagnant, meaning these best results remain unfair. Consequently, I conclude that both balancing techniques only minorly improve fairness.

5. Discussion

From the preceding section, three points need further analysis. First, why did each balanced model's performance fail to generalise? This can be explained by considering the commonality between the applied balancing techniques. Although different means, both generate their effects through *projecting* the minority class. Whilst in SMOTE this is obvious, under Threshold-Adjustment, the projection is more minute, applied only to the output function. Accordingly, each strives to achieve parity by magnifying the underlying distribution of the perceived minority class. SMOTE's more explicit projection likely causes its worse results compared to Threshold-Adjustment. Overall, the failure to truly generalise reflects large aleatoric uncertainty within the minority sample.

Moreover, whilst minority classification improves, it does so at the cost of majority selection. Given the underlying distribution of the problem, if majority class effectiveness is impacted, this will necessarily have a greater effect on accuracy. An ideal model thus needs to improve minority selection without carry on effects. This could be achieved by adding more non-synthetic data, and/or generating more complex models.

Second, how does performance failure affect fairness? The initial suite drew most of its power from the majority class, leading to poor EO fairness results. When underlying representation-bias is disregarded, both balancing techniques saw increased fairness across all groups. Yet, even the best results remained fundamentally unfair. Again, elaborating on the 'Transgender' group in Threshold-Adjusted Naïve-Bayes (Figures 11/12/13), high precision means non-toxic 'Transgender' related comments are filtered less, allowing the community to have a conversation. However, recall remains low, resulting in unfiltered toxicity, and a quality-of-service harm. Consequently, poor fairness results from the inability of each model to capture the underlying distribution.

[5] Excluding 'Other gender'.

		Naïve		Threshold-Adjusted		SMOTE Fairness	
M	*G*	*DP*	*EO*	*Ditto*	*DP*	*Ditto*	*DP*	*EO*
KNN	Male	0.0168	0.0635		0.0288	0.0744		0.0109	0.0073
	Female	0.0429	0.0666		0.0981	0.088		0.0352	0.0374
	Transgender	0.0252	0.0725		0.002	0.1324		0.0842	0.0953
	Other gender	0.0823	0.2067		0.2337	0.455		0.1853	0.9612
NB	Male	0.1047	0.1237		0.1007	0.1184		0.1092	0.1288
	Female	0.2289	0.2599		0.2279	0.2514		0.2092	0.2096
	Transgender	0.1241	0.3585		0.1212	0.3497		0.0441	0.2495
	Other gender	0.3648	0.6511		0.3811	0.6695		0.3940	0.6787
L-Reg	Male	0.0251	0.0474		0.0732	0.0836		0.0708	0.0684
	Female	0.0388	0.0720		0.1307	0.1355		0.1202	0.1071
	Transgender	0.0048	0.1414		0.1166	0.1117		0.1734	0.1848
	Other gender	0.7480	0.2575		0.3178	0.6896		0.3634	0.7411
MLP	Male	0.0416	0.0648		0.0686	0.0783		0.0534	0.043
	Female	0.0575	0.0946		0.1234	0.1171		0.1073	0.109
	Transgender	0.0296	0.1583		0.0661	0.0687		0.1713	0.1779
	Other gender	0.1185	0.3725		0.6706	0.7598		0.3783	0.7920

Fig.10: Fairness Results Across Iterations

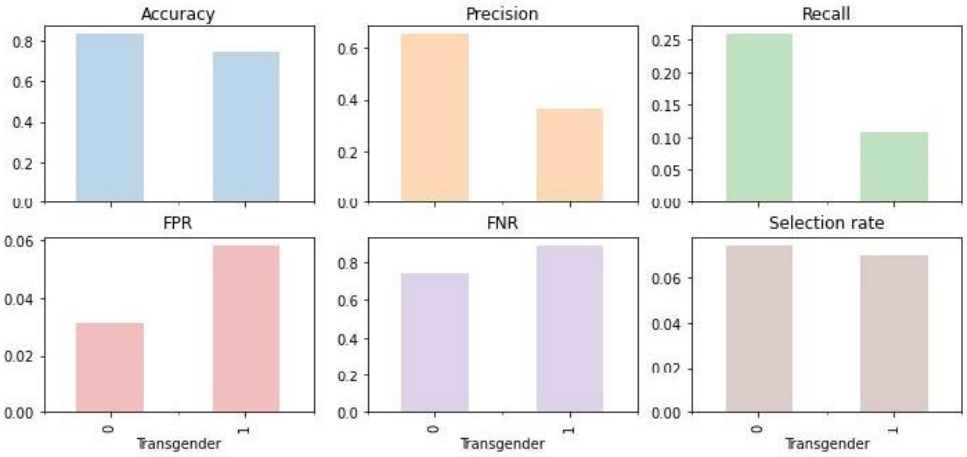


Fig.11: Initial Naïve-Bayes 'Trasngender' Metrics

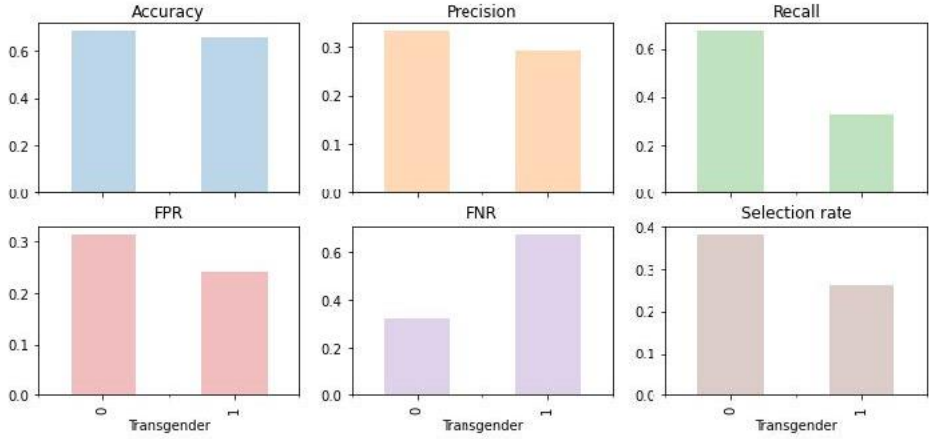


Fig.12: Threshold-Adjusted Naïve-Bayes 'Transgender' Metrics

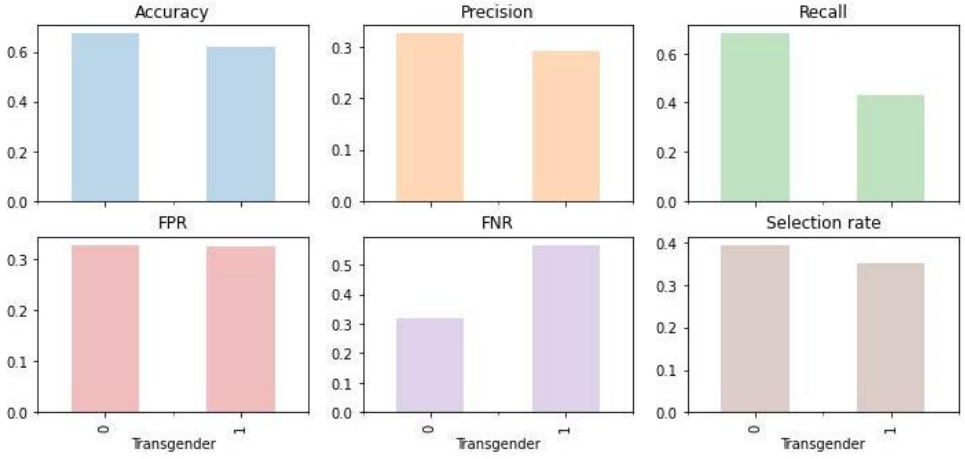


Fig.13: SMOTE Naïve-Bayes 'Transgender' Metrics

Lastly, why the discrepancy in fairness metrics? Fairness isn't something which can be mathematically reduced. Context remains paramount. As DP measures *equality* over selection between groups, it is inappropriate for the task of toxicity classification, because toxicity is only understood by identifying *differences* between groups. [6] Indeed, concerning EO, the question remains open as to whether equality of errors between groups is ethical. As Fazelpour et al. point out, fairness under this definition is idealised, conceived from a utilitarian perception of justice [19]. If the goal is to keep people in the conversation, then as Davis et al. argue, surely greater emphasis should be given to mitigating harm against those who are already silenced under structural oppression [20]. Thus, the question of the correct fairness metric remains open.

6. Conclusion

Ultimately, the two demonstrated data-balancing techniques - Threshold-Adjustment and SMOTE - failed to generalise in the task of toxicity classification. I drew this conclusion from an iterative process of applying each technique to a suite of models, aiming for equal minority and majority class classification and improved fairness. Given the constraints of this problem, I recommend using these techniques only in combination with more minority class data or more complex modelling.

The high threshold for success comes from the potential harm in misclassifying toxicity. Questions remain over how to characterize fairness in this context. Considering my discussion, further research is necessary to develop an approach which centres those already marginalised, focusing on equity rather than equality.

[6] Instead, it would be useful for AI in hiring.

References

[1] "Jigsaw Unintended Bias in Toxicity Classification." https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification (accessed Oct. 02, 2022).

[2] P. Ravi, H. Narayana Batta, S. Greeshma, and S. Yaseen, "Toxic Comment Classification," *Int. J. Trend Sci. Res. Dev. IJTSRD*, vol. 3, no. 4, pp. 24–27, Jun. 2019.

[3] S. Zaheri, J. Leath, and D. Stroud, "Toxic Comment Classification," *SMU Data Sci. Rev.*, vol. 3, no. 1, Apr. 2020, [Online]. Available: https://scholar.smu.edu/datasciencereview/vol3/iss1/13

[4] B. Van Aken, J. Risch, R. Krestel, and A. Löser, "Challenges for Toxic Comment Classification: An In-Depth Error Analysis," Sep. 2018, doi: 10.48550/arXiv.1809.07572.

[5] N. Japkowicz and S. Stephen, "The Class Imbalance Problem: Significance and Strategies," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, 2002.

[6] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016, doi: 10.1007/s13748-016-0094-0.

[7] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and Mitigating Unintended Bias in Text Classification," in *Proceedings of the 2018 AAAI*, New York, NY, USA, Dec. 2018, pp. 67–73. doi: 10.1145/3278721.3278729.

[8] B. Krawczyk and M. Woźniak, "Cost-Sensitive Neural Network with ROC-Based Moving Threshold for Imbalanced Classification," in *Intelligent Data Engineering and Automated Learning – IDEAL 2015*, Cham, 2015, pp. 45–52. doi: 10.1007/978-3-319-24834-9_6.

[9] H. J. P. Weerts, "An Introduction to Algorithmic Fairness," pp. 1–18, May 2021, doi: 10.48550/arXiv.2105.05595.

[10] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2020, pp. 1–14. doi: 10.1145/3313831.3376445.

[11] "Unintended Bias and Identity Terms," *Jigsaw*, Oct. 12, 2021. https://medium.com/jigsaw/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23 (accessed Oct. 03, 2022).

[12] S. Corbett-Davies and S. Goel, "The Measure and Mismeasure of Fairness: A Critical Review of Fair

Machine Learning.” arXiv, Aug. 14, 2018. doi: 10.48550/arXiv.1808.00023.
[13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness Through Awareness.” arXiv, Nov. 28, 2011. doi: 10.48550/arXiv.1104.3913.
[14] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.
[15] “Sci-kit Learn API,” *scikit-learn*. https://scikit-learn/stable/modules/classes.html (accessed Oct. 04, 2022).
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
[17] “API Docs — Fairlearn 0.8.0.dev0 documentation.” https://fairlearn.org/main/api_reference/index.html (accessed Oct. 04, 2022).
[18] D. Pessach and E. Shmueli, “A Review on Fairness in Machine Learning,” *ACM Comput. Surv.*, vol. 55, no. 3, p. 51:1-51:44, Feb. 2022, doi: 10.1145/3494672.
[19] S. Fazelpour and Z. C. Lipon, “Algorithmic Fairness from a Non-ideal Perspective,” in *AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 2020, pp. 57–63. doi: https://doi.org/10.1145/3375627.3375828.
[20] J. L. Davis, A. Williams, and M. W. Yang, “Algorithmic reparation,” *Big Data Soc.*, vol. 8, no. 2, Jul. 2021, doi: 10.1177/20539517211044808.