# Analyzing bias in supervised learning algorithms: Challenges in comment toxicity classification on minority groups

Anonymous

## I. Introduction

Social media applications use classification algorithms to predict various factors and to improve the performance of the application with time. [1]–[3] The outcomes of the algorithm may be used to determine issues or problems in a society/region [4]. In such scenarios, it is essential to examine the fairness and accountability of the classification model. In this paper, we examine how fair a classification model that predicts the toxicity of Twitter comments is for minority groups. This work uses a data set [5] that consists of Twitter comments, identities of each comment (that denotes if a comment contains information about a social class/category like religion, race, gender etc.) and categorization of the comments (as toxic or not). We aim to critically analyze the effects that data gaps in minority communities bring to the model. We also look at the performance of the model as a whole, and then compare it with the performance of different identities of a social class. Do we need to re-imagine classification algorithms while dealing with minority groups in a society - this question is what drives the discussion of this paper.

## II. Literature Review

The analysis by P. S. Br Ginting et al. [6] provides a valuable insight for our research question. The authors use Logistic Regression as their classification model to detect hate speech on Twitter. The feature vectors of the comments are in the form of Term Frequency Inverse Document Frequency (TF IDF) [7]. The experimental results produced using TF IDF vectors illustrate a significant increase in performance. This is why we drew inspiration from their methods to use TF IDF vectors in our classification models.

There is a plethora of research related to evaluating inherent bias in machine learning algorithms [8], [9]. The related work by Ayanna Howard et al. [10] examines the challenges in recognizing biases in machine learning systems, and their data sets. In an image recognition model, the authors examine the problems of using training data that is not representative of a fair population. The discussion also includes the challenges faced in maintaining generalization for prediction algorithms, while improving the performance for minority classes. The steps taken to identify the minority classes and decrease the overall bias of the system has been carefully studied and leveraged for the scope of our research.

## III. Method

### A. Classification model creation

Our results are produced by first building a classification algorithm that predicts if a given tweet comment is toxic or not. Supervised learning algorithms are chosen in the scope of this project because we need a form of validation using true labels for the predictions, based on the accuracy, precision and recall scores of the models. The data set we are dealing with consists of only numerical values. Further, the classification is binary - Toxic or Not Toxic. Hence, the two types of classifiers were finalized and developed for this work:

- Gaussian Naive Bayes Classifier
- Logistic Regression Model for Classification

Both these classifiers were provided inputs as Twitter comments in the form of 1000 dimensional TF IDF vectors [11]. The dimension of the vector is the 1000 most valuable words in the vocabulary

of the data set. Each vector forms a feature vector for classification in our analysis. Hence, we use the TF IDF values of the words present in a comment to determine the toxicity of the comment. No other parameter (like the region from where the comment is made, or the time when the comment was posted) is utilized for classification.

### B. Study of Data Distribution

Now that the models are built, the next step was to study the data distribution of the training data, on which our classifiers were built. For the scope of this research, we chose two prominent social categories to study: *Religion* and *Gender*. The identities annotated under these categories for the comments are:

| Category | Identities |
|---|---|
| Religion | Atheist, Buddhist, Christian, Hindu, Jewish, Muslim, Other religion |
| Gender | Male, Female, Transgender, Other gender |

TABLE I. Annotated identities considered for this research.

For each category, training data was filtered to fetch comments that fall under any one of the identities in the category. Then we extracted the distribution of each identity in a category to understand the majority and minority groups in the data.

### C. Comparisons of model performances on majority and minority groups

In conjunction with data analysis, each classifier was tested on the following sub-group in each category (Religion and Gender):

- Training data (comments) that were linked to the category.
- Training data annotated with one identity in a category

We analyzed the patterns produced by different evaluation metrics on all the groups of data by the classification models. These patterns conveyed key themes regarding the fairness of the models for minority classes. Key Evaluation Metrics [12] used to contextualise the results:

- Accuracy Score
- Precision Score (using Confusion Matrix): This measurement was vital to check how many comments were actually toxic, of the ones that were labeled as toxic. The following method was used to measure the precision of the interesting class (Toxic):

$$\frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)} \quad (1)$$

- Recall Score (using Confusion Matrix): Of the comments that were toxic, how many were correctly labelled as toxic is measured by this metric. From the confusion matrix, recall score was calculated as:

$$\frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)} \quad (2)$$

Where:
$TP$ = An actual toxic comment that is labelled toxic
$FN$ = A comment that is toxic, but is classified by the algorithm as not toxic
$FP$ = A comment that is not toxic but gets classified by the algorithm as toxic

Recall is the most valuable metric for us, since classifying a toxic comment as not toxic is what can deteriorate the main goal of our classifier (i.e., to reduce the toxicity in the comments of the platform).

## IV. Results

The following results (see Table II) were obtained on computing the accuracy score for the classifiers, after testing them on the development data.

| Model | Accuracy in development data |
|---|---|
| Gaussian Naive Bayes | 63.68% |
| Logistic Regression | 82.76% |

TABLE II. Accuracy score of classification models on development data

The next steps involve the analysis of the models with a focus on minority groups.

### A. Data Distribution of Training Data

*1) Dealing with comments linked to Religion:* The results extrapolated from the training data, on the basis of religion are illustrated in Table III.

| Total tweets that are linked to a Religion | 54,086 |
|---|---|
| Tweets about Atheists | 1080 |
| Tweets about Buddhists | 482 |
| Tweets about Christians | 34,163 |
| Tweets about Hindus | 486 |
| Tweets about Jewish | 6714 |
| Tweets about Muslims | 18,249 |
| Tweets about Other religions | 271 |

TABLE III. Number of Twitter comments linked to different religions

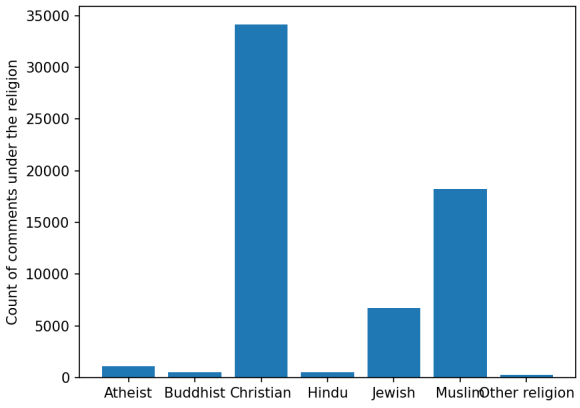



Fig. 1. Distribution of Twitter comments that are linked to a religion

Figure 1 provides a clear picture of the existence of a majority religion and several minority groups in the training data. Our models are trained to classify comments using information where the dominant religion followed or talked about is Christianity. Minority religions (like Hinduism, Buddhism, Atheism, etc) form a significantly low fraction.

*2) Dealing with comments linked to Gender:* Likewise, an extrapolation was done on the basis of Gender for the Twitter comments used to train the model:

| Total tweets that are linked to a Gender | 65,541 |
|---|---|
| Tweets about Males | 35,998 |
| Tweets about Females | 43,069 |
| Tweets about Transgenders | 2200 |
| Tweets about Other genders | 9 |

TABLE IV. Number of twitter comments linked to different genders

A similar pattern (see Table IV and Figure 2) is present in the data set in the context of gender.

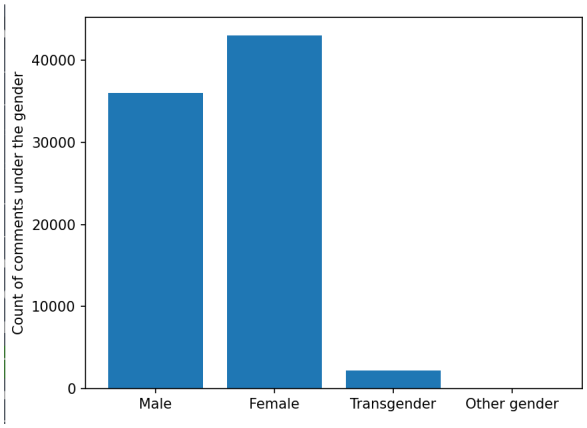



Fig. 2. Distribution of Twitter comments that are linked to a gender

The models are trained to determine toxicity while primarily dealing with two dominant genders - Male and Female.

These results showcase a clear *data gap* for minority classes of both categories in the data used for building the model. It is important to note at this point that the training algorithms **do not consider the identities** while determining toxicity.
In the next section, we produce findings to understand how considerate our model is towards minority data classes.

### B. Accuracy of the models on majority and minority classes

For both categories, we first calculated the accuracy of the Gaussian Naive Bayes model and Logistic Regression Classifier for each identity belonging to the category. This yielded the following observations:

| **Identity of comment** | **Accuracy of Gaussian Naive Bayes Model** | **Accuracy of Logistic Regression** |
|---|---|---|
| Atheist | 0.72151 | 0.86075 |
| Buddhist | 0.48979 | 0.81632 |
| Christian | 0.80313 | 0.91145 |
| Hindu | 0.61702 | 0.95744 |
| Jewish | 0.60515 | 0.83047 |
| Muslim | 0.53164 | 0.77938 |
| Other religion | 0.66666 | 0.77777 |

TABLE V. Performance of different models on different subsets of religions

| Identity of comment | Accuracy of Gaussian Naive Bayes Model | Accuracy of Logistic Regression |
|---|---|---|
| Male | 0.62376 | 0.84488 |
| Female | 0.65556 | 0.85717 |
| Transgender | 0.58598 | 0.78343 |
| Other gender[1] | 0.0 | 1.0 |

TABLE VI. Performance of different models on different subsets of genders

Three critical findings from our experiments are:

- A high variation in accuracy can be observed for a given model for different sub-classes of data. For example, in the context of religion, the Gaussian Naive Bayes shows a 80.31% accuracy for comments linked to Christianity. However, the accuracy drops to 48.97% while classifying comments on Buddhism.
- The models work much better on the identity that is dominant in the category. In the context of gender, both models showcase the highest accuracy while dealing with comments related to Male and Female.
- The Gaussian Naive Bayes model does not perform as well when tested on a minority group, like comments related to Buddhism or Hinduism under religion or comments related to transgenders under gender. Comparatively, the Logistic Regression model classifies data better when tested on minority classes.

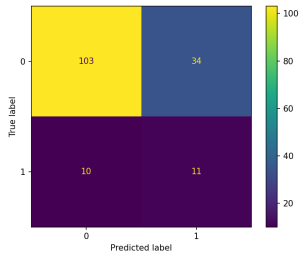
### C. Confusion Matrix as an Evaluation Metric

Having discussed how a given model performs better for a dominant identity that it has been trained on, it is important to note that there is an inherent bias in the algorithms. We then proceed to analyze the fairness of the algorithms by evaluating the confusion matrix of the performance in some of the minority classes of the religion category.

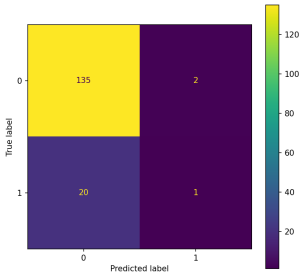
*Note:* In our scope, 1 implies toxicity. Hence, a true value of 1 means that a comment is toxic. A predicted value of 1 means that the comment is labelled toxic by our classifier.

The confusion matrices provide a key insight in our research.

[1]The accuracy score for this class is considered as unreliable, since the count of comments in this sub-class is too low to provide generalised evaluation.

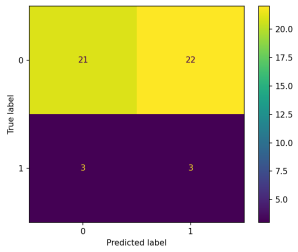



(a) Gaussian Naive Bayes

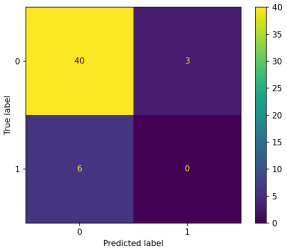



(b) Logistic Regression

Fig. 3. Confusion Matrix for comments annotated with Atheism



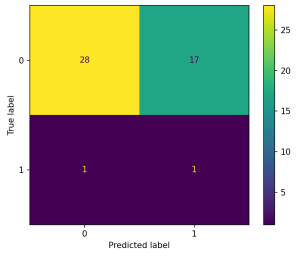

(a) Gaussian Naive Bayes



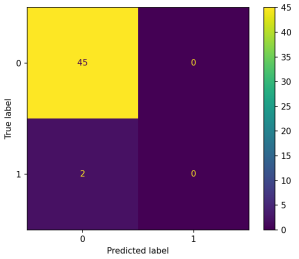

(b) Logistic Regression

Fig. 4. Confusion Matrix for comments annotated with Buddhism

- For all the three minority classes, we observed (from (a) of Figures 3,4 and 5), that the Gaussian Naive Bayes Model produces **more False Positives, and lesser False Negatives**. This indicates a **low Precision value, and a high Recall value**

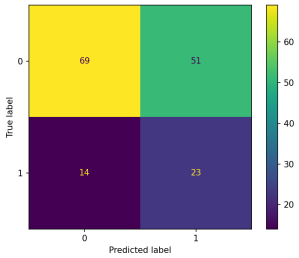



(a) Gaussian Naive Bayes

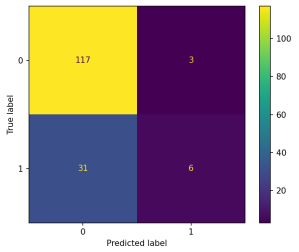



(b) Logistic Regression

Fig. 5. Confusion Matrix for comments annotated with Hinduism




(a) Gaussian Naive Bayes



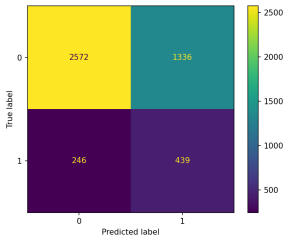

(b) Logistic Regression

Fig. 6. Confusion Matrix for comments annotated with Transgender

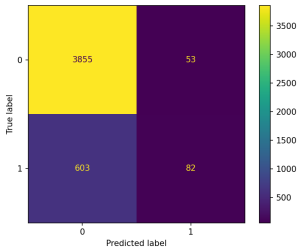
of the model.

- On the other hand, the Logistic Regression model produces **lesser False Positives and more False Negatives**. (See (b) of Figures 3, 4 and 5) This indicates a **high Precision and a low Recall value** of the model.

A very similar pattern (See Figures 6, 7 ) is showcased by the models when tested under Gender. Here, we generated the confusion matrix for the sub-class of transgender identity, a prominently small minority. We also generated the matrix for female, a comparatively larger sub-class. For both these minority classes, Gaussian Naive Bayes model shows a higher precision, (lesser false positives) whereas Logistic Regression models have a higher recall (lesser false negatives).




(a) Gaussian Naive Bayes




(b) Logistic Regression

Fig. 7. Confusion Matrix for comments annotated with Female

## V. Discussion

In Section 4, we demonstrated the disparities in two supervised learning models when performed for minority groups of a category.

The first observation that introduced a possible bias in the Machine Learning (ML) algorithms is the uneven data distribution. While dealing with a widely used application (Twitter), data sets are often seen as reliable representations of a population [13]. However, the distribution of data collected for training displays a clear data gap in minority groups of the society. Hence, the models learn to classify a comment as toxic or not, based on features that indicate toxicity for comments related to the

majority identity of a society.

The accuracy scores provides proof to our presumption that the data gap with respect to minority groups introduces a bias in the learning system, irrespective of the model. The number of comments incorrectly classified rises in minority sections of a social category. This drop in accuracy occurs even though the identity of a comment is not considered as an attribute in the feature vector used for classification. However, there is an inconsistency with the notion that a drop in accuracy in certain sub-classes degrade the fairness of a classifier. The overall generalisation, which includes the outcome rate of the majority class, should also account for the inclusiveness of a classifier.

While the accuracy scores proved that a given model performs worse for minority classes as compared to the dominant identity, the confusion matrices point out **how both algorithms err for minority groups**.
Gaussian Naive Bayes model has a lower precision and higher recall, implying that an incorrect prediction is more likely to mean that a non-toxic comment was labelled as toxic. However, the Logistic Regression model showcases a higher precision and lower recall. Hence, an incorrect prediction is most likely because a toxic comment is classified as non-toxic, which increases the number of false negatives.

Consequently, though Logistic Regression shows a higher accuracy for all categories, it can prove to be a more problematic model to use. A lower recall implies that this model is likely to predict a toxic comment as non-toxic for a minority group. Whereas in Gaussian Naive Bayes model, when a comment is wrongly classified, it is likely that a non-toxic comment is classified as a toxic one. Toxicity towards a section of a community in an ecosystem is an indication of hatred and oppression against the section. Unquestionably, it is more vital to ensure that the toxic comments are correctly classified. Hence, while contextualising meaningful fairness in the algorithm, "Toxic" is a more interesting class label than "Not toxic". Incorrectly classifying non-toxic comments can be penalised less as compared to incorrectly classifying a toxic comment.

## VI. Conclusions

The study of the performance of two benchmark classification models - Gaussian Naive Bayes and Logistic Regression, shows that an imbalanced representation of different groups can introduce limitations in the original goal of the classifier. From analysing the training data, we can argue that a bias can occur whenever we have a dominant (majority) class along with several minority classes that are under-represented.

Using accuracy as a sole metric was shown to be a potential concern, since it fails to take the interesting class into account during incorrect predictions. The confusion matrix provides a more accurate measure of the models' inconsistency with minority group performances. The precision and recall derived from the matrix gives a more comprehensive view of the weakness of the algorithms when associated with minority classes.

Future work in this domain includes improving the feature selection for classification in minority groups, and validating our approach using a different training algorithm (like unsupervised learning).

## References


[1] A. Bermingham and A. Smeaton, "On using twitter to monitor political sentiment and predict election results," in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, 2011, pp. 2–10.

[2] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. Silenzio, and M. De Choudhury, "A taxonomy of ethical tensions in inferring mental health states from social media," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 79–88.

[3] V. Maslej-Krešňáková, M. Sarnovskỳ, P. Butka, and K. Machová, "Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification," *Applied Sciences*, vol. 10, no. 23, p. 8631, 2020.

[4] P. Ravi, G. S. Hari Narayana Batta, and S. Yaseen, "Toxic comment classification," *International Journal of Trend in Scientific Research and Development (IJTSRD)*, 2019.

[5] J. AI, "Jigsaw unintended bias in toxicity classification," https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification, accessed: July, 2022.

[6] P. S. Br Ginting, B. Irawan, and C. Setianingsih, "Hate speech detection on twitter using multinomial logistic regression classification method," in *2019 IEEE International*

*Conference on Internet of Things and Intelligence System (IoTaIS)*, 2019, pp. 105–111.

[7] A. M. Ramadhani and H. S. Goo, "Twitter sentiment analysis using deep learning methods," in *2017 7th International annual engineering seminar (InAES)*. IEEE, 2017, pp. 1–4.

[8] S. Wachter, B. Mittelstadt, and C. Russell, "Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law," *W. Va. L. Rev.*, vol. 123, p. 735, 2020.

[9] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining algorithmic fairness in india and beyond," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 315–328.

[10] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 2017, pp. 1–7.

[11] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[12] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, S. Huang, M. Brooks, M. J. Lee, and H. Asadi, "Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods," *American Journal of Roentgenology*, vol. 212, no. 1, pp. 38–43, 2019.

[13] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.