Project Title: Mapping AI's Representation Gap: Building and Auditing a Hate Speech Detector for Demographic Fairness

Team Members:

- Prince: Data Lead
- Bruna: Modeling Lead
- Maya: Communication/Presentation Lead

GitHub Repository: mapping-ai-representation-gap

https://github.com/princenewman02/mapping-ai-representation-gap/tree/main

Problem Statement:

This project addresses the question: **By building a hate speech detector, how can we measure performance disparities across demographic groups and identify which communities are underserved or over-policed?** This is crucial as it aims to reveal potential algorithmic biases, ensuring that AI systems treat all demographic groups fairly and equitably. Identifying these disparities is vital for improving the accountability and transparency of AI technologies.

Project Purpose:

We aim to:

- Train a hate speech detection model using the Jigsaw Civil Comments dataset.
- Analyze representation in training data to identify which demographic groups are under- or over-represented.
- Evaluate model fairness across multiple demographic groups using HateCheck and performance/fairness metrics.
- Visualize disparities with dashboards, charts, and intersectional analyses to communicate protection and surveillance gaps.
- Produce a comprehensive report and presentation summarizing findings, model performance, and actionable insights for equitable AI deployment.

Data Sources:

We will utilize the following datasets:

1. **Jigsaw Civil Comments** (Kaggle): Contains 2 million comments used to train the hate speech detector, featuring toxicity labels and identity mentions (race, religion, gender, disability).
2. **HateCheck** (Paul Röttger GitHub): Provides 3,728 test cases for fairness benchmark testing, focusing on demographic variations.
3. **US Census 2020/2022**: Offers demographic baseline comparisons, including population proportions by race, ethnicity, and disability.

These datasets are sufficient for our analysis, but significant preprocessing and cleaning will be required to ensure compatibility and usability. We plan to finalize data sources and documentation by the end of Week 1.

Challenge:

The main challenge of this project is integrating multiple datasets while ensuring data integrity and handling potential biases effectively. The complexity of the fairness evaluation will justify the project as a substantial undertaking, requiring in-depth analysis and validation of findings.

Analytical Approach:

Our analytical approach will involve:

- Defining metrics for measuring fairness and evaluating the hate speech detector's performance across different demographic groups.
- Using a framework that includes exploratory data analysis (EDA) and model training with the DistilBERT model to assess detection rates and potential biases.
- Speculative methodologies will include fairness metrics such as detection rates and false positive rates by demographic group.

Our dependent variable (y) is binary toxicity classification, where 1 indicates toxic/hate speech and 0 indicates non-toxic content. Due to severe class imbalance (7% toxic vs 93% non-toxic), we implemented inverse frequency class weighting (w_toxic = 6.175, w_non-toxic = 0.544).

Our independent variables (X) include:

- Primary input: comment text (comment_text) processed by DistilBERT's pre-trained transformer
- Engineered features (for EDA): comment_length, word_count, unique_word_count
- Demographic indicators: Identity mentions extracted from text (race, religion, gender, disability) used for fairness stratification during evaluation

Solution Technologies:

We anticipate using the following technologies:

- **Python** for data analysis and modeling.

- **Hugging Face Transformers** for model training.
- **Google Colab** for cloud-based computation.
- Various visualization libraries (Matplotlib, Seaborn) for data presentation.

Given the dataset's size, we may require powerful computational resources to efficiently process and analyze the data.

Briefly reiterate where you are now (we have our data set, we have run a preliminary model, etc.)

Data Cleaning Summary: Data

- The data for all three data sets was cleaned by selecting relevant columns and changing variable names to follow best practices.
- For the hatecheck data sets, the ID column was changed to an object so the two data frames could be joined
- Data was inspected for missing values. Only hatecheck data had NaNs, as there were no missing values in the jigsaw data.
- Data was imputed for hatecheck data variable Target Identity
- Clean versions of the jigsaw and hatecheck data frames have been exported

Exploratory Data Analysis and Initial Modeling: Model

We're following an iterative development approach: we trained our initial model on a stratified sample (~150,000 comments) to validate our methodology and confirm that fairness disparities are detectable before investing hours in full-dataset training

Class Example vs. Our Hate Speech Detection Project

Understanding Our Work Through Familiar Concepts

Stage	Class: Dog/Cat Breeds	Our Project: Hate Speech
Domain Problem	Classify breed of dog/cat in image	Detect hate speech fairly across demographics
Training Data	Oxford-IIIT Pet Dataset (7,390 images)	Jigsaw Civil Comments (1.8M comments)
Pre-trained Model	ResNet-34 (trained on ImageNet)	DistilBERT (trained on text corpora)
Training Process	Transfer learning: fine-tune for breeds	Transfer learning: fine-tune for toxicity
Evaluation	Confusion matrix, error rate, top losses	Confusion matrix + fairness metrics by group

Deployment/Test	Test on Toto (new image not in training)	Test on HateCheck (controlled fairness tests)

Class Question	**Our Question**
"Does the model work?" Tests overall accuracy on new data	*"Does it work fairly for everyone?"* Tests performance equity across demographics

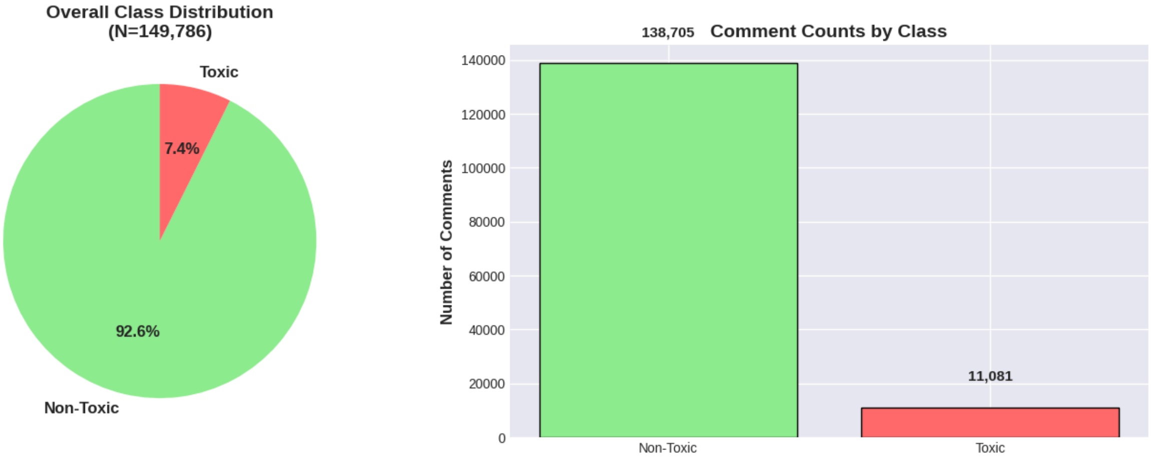
- Initial exploratory data analysis has been done on a subset of the data (~150,000 comments) with limited data cleaning. This data was used to train the model we are using (DistilBERT) and various charts were generated

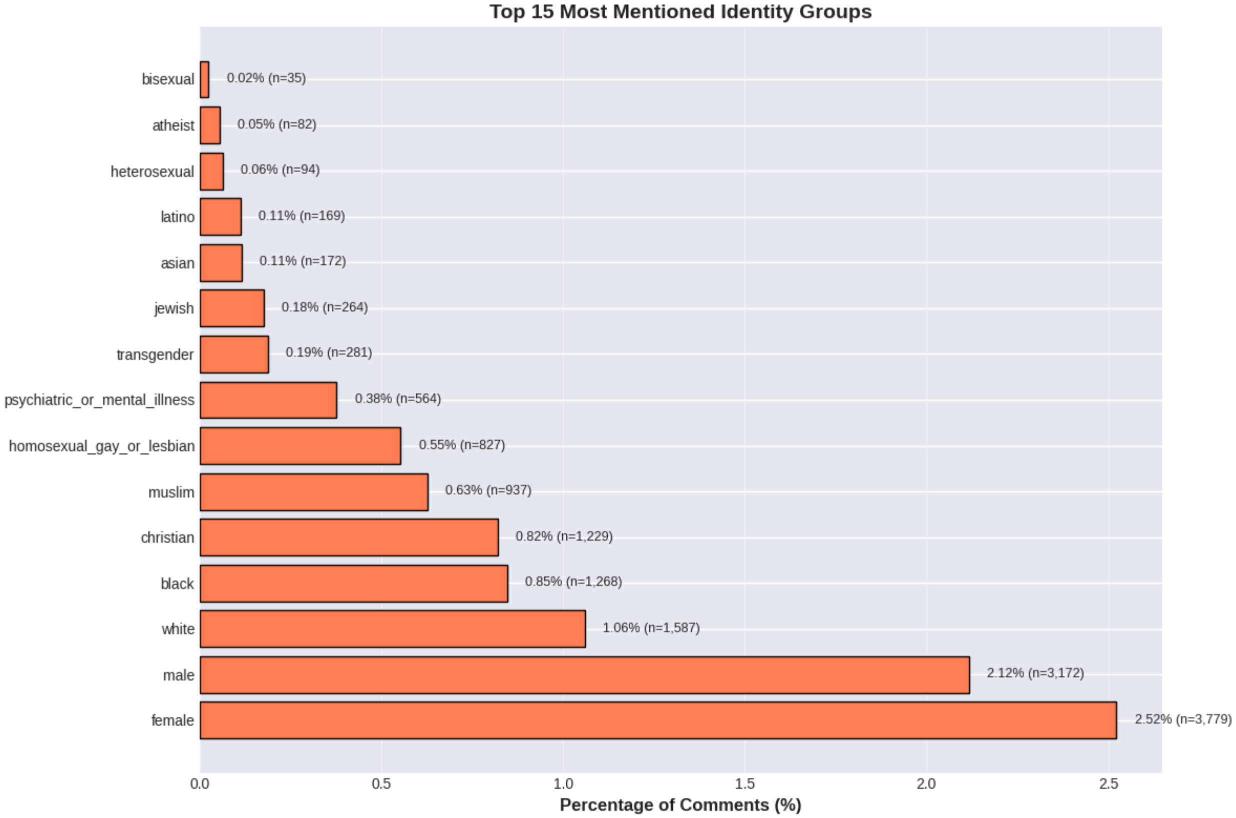
Our model currently shows the following performance metrics (1 epoch):

Step	**Training Loss**	**Validation Loss**	**Accuracy**	**Precision**	**Recall**	**F1**	**Auc**
936	0.395300	0.490126	0.945424	0.642472	0.591155	0.615746	0.928322
1872	0.343900	0.325432	0.923259	0.488346	0.784747	0.602043	0.943718
2808	0.335100	0.346143	0.930569	0.520872	0.765794	0.620022	0.945736
3744	0.357300	0.324429	0.927599	0.506853	0.784296	0.615766	0.947573

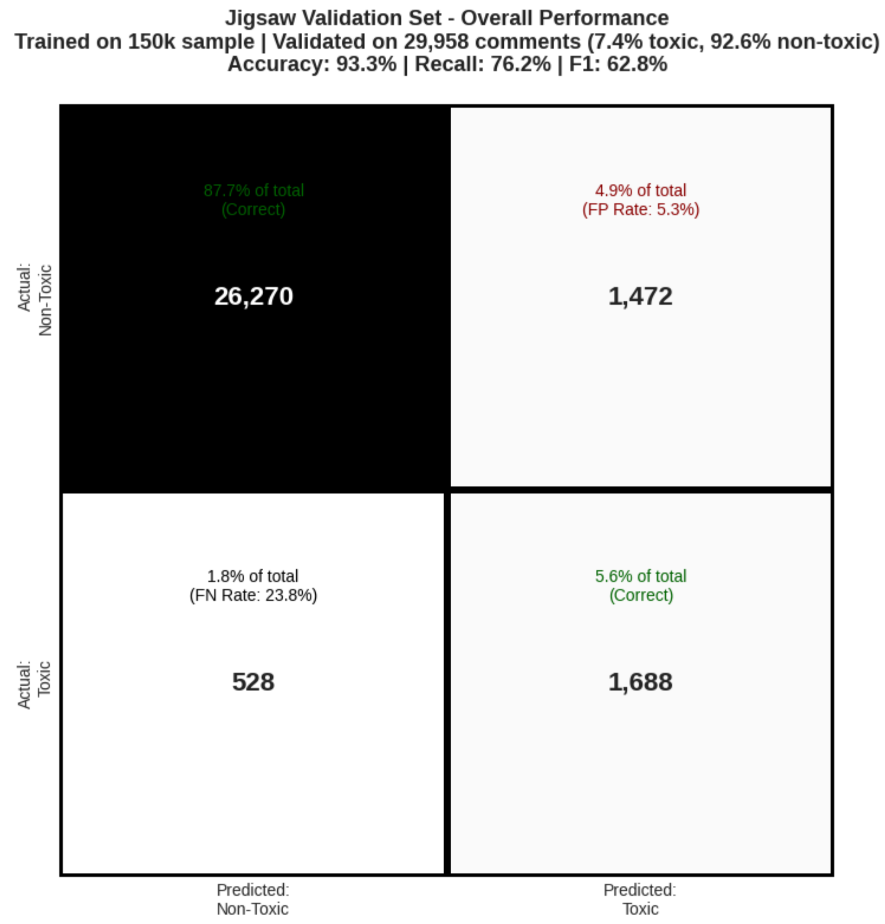
- It should be noted that the percentage of toxic comments in the training data is extremely disproportionate (7% toxic vs 93% non-toxic) so the model may be better trained to recognize non-toxic comments which isn't the focus of our project. However, it may be fine since 7% of such a large sample is still a substantial amount of data to work with. Currently, we implemented inverse frequency class weighting using the standard sklearn formula: *w = n_samples /*

(n_classes × n_samples_in_class). This resulted in a toxic class weight of 6.175 versus 0.544 for non-toxic, effectively penalizing the model 11.3 times more heavily for missing hate speech (false negatives) than for over-flagging non-toxic content (false positives). This design choice reflects the ethical priorities of content moderation.





Confusion Matrix for Jigsaw:

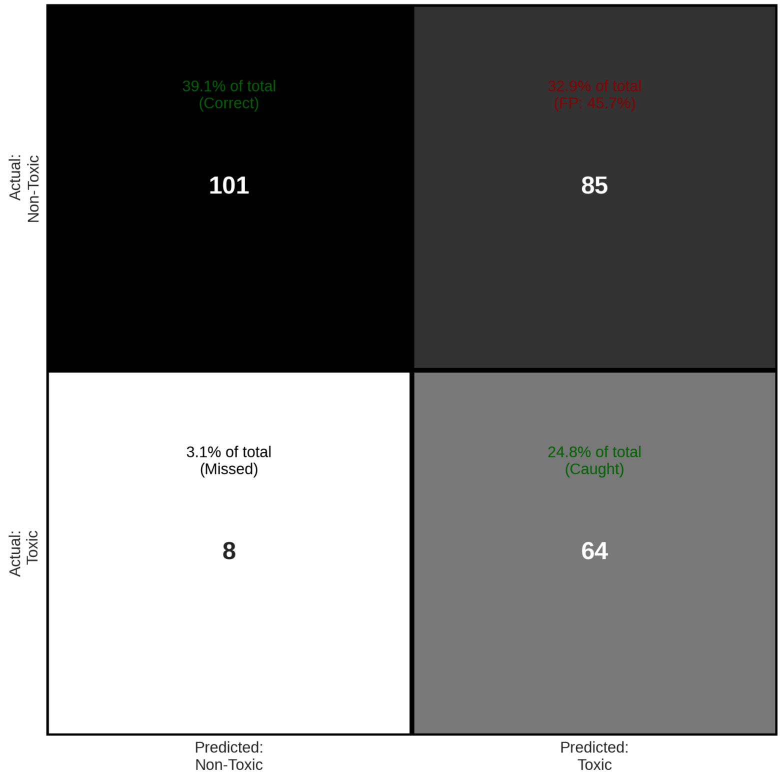


Jigsaw Validation Set - Overall Performance
Trained on 150k sample | Validated on 29,958 comments (7.4% toxic, 92.6% non-toxic)
Accuracy: 93.3% | Recall: 76.2% | F1: 62.8%

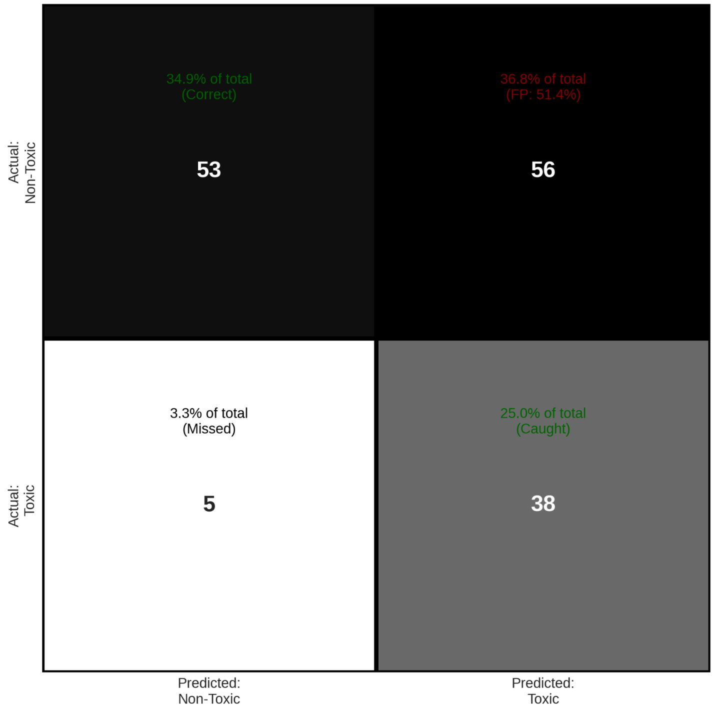
	Predicted: Non-Toxic	Predicted: Toxic
Actual: Non-Toxic	87.7% of total (Correct) 26,270	4.9% of total (FP Rate: 5.3%) 1,472
Actual: Toxic	1.8% of total (FN Rate: 23.8%) 528	5.6% of total (Correct) 1,688

Confusion Matrices by group identity - Jigsaw data

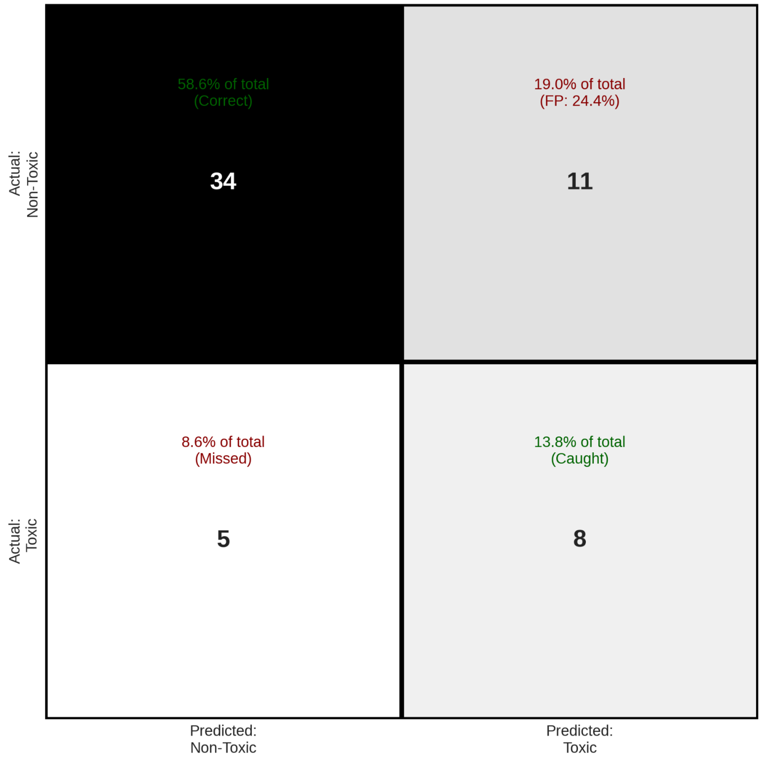
Jigsaw Validation: Black
258 comments mentioning this identity (27.9% toxic, 72.1% non-toxic)
Detection: 88.9% | FP Rate: 45.7% | F1: 57.9%



Jigsaw Validation: Homosexual Gay Or Lesbian
152 comments mentioning this identity (28.3% toxic, 71.7% non-toxic)
Detection: 88.4% | FP Rate: 51.4% | F1: 55.5%

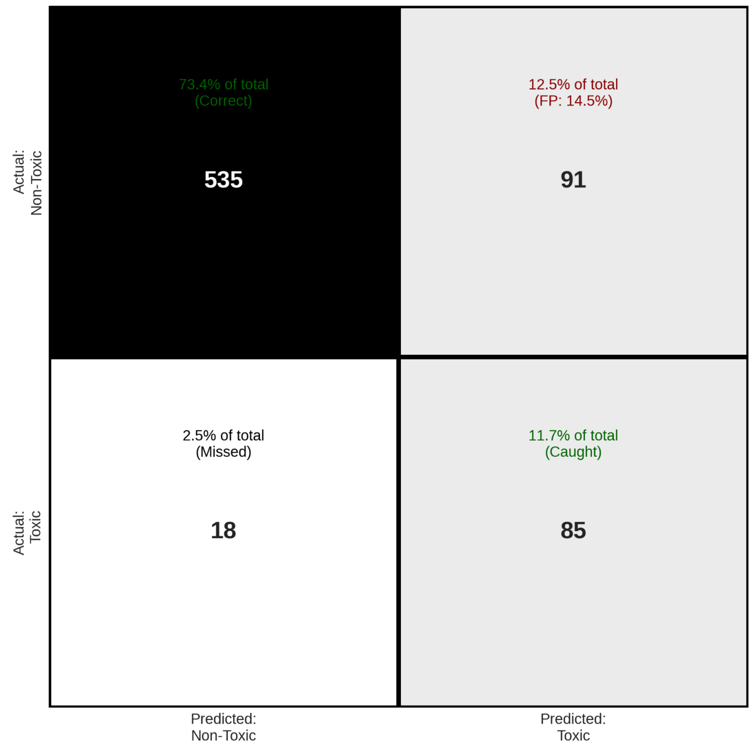


Jigsaw Validation: Transgender
58 comments mentioning this identity (22.4% toxic, 77.6% non-toxic)
Detection: 61.5% | FP Rate: 24.4% | F1: 50.0%



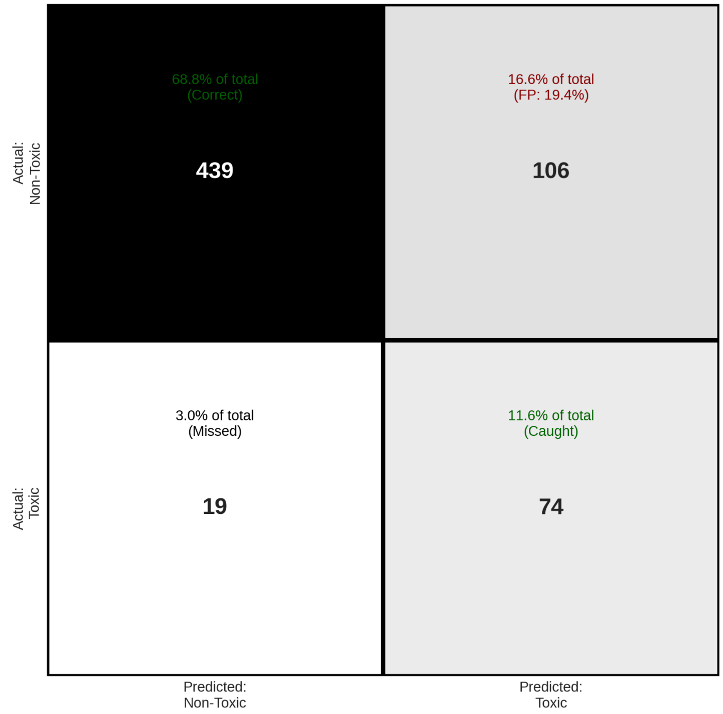
	Predicted: Non-Toxic	Predicted: Toxic
Actual: Non-Toxic	58.6% of total (Correct) 34	19.0% of total (FP: 24.4%) 11
Actual: Toxic	8.6% of total (Missed) 5	13.8% of total (Caught) 8

Jigsaw Validation: Female
729 comments mentioning this identity (14.1% toxic, 85.9% non-toxic)
Detection: 82.5% | FP Rate: 14.5% | F1: 60.9%



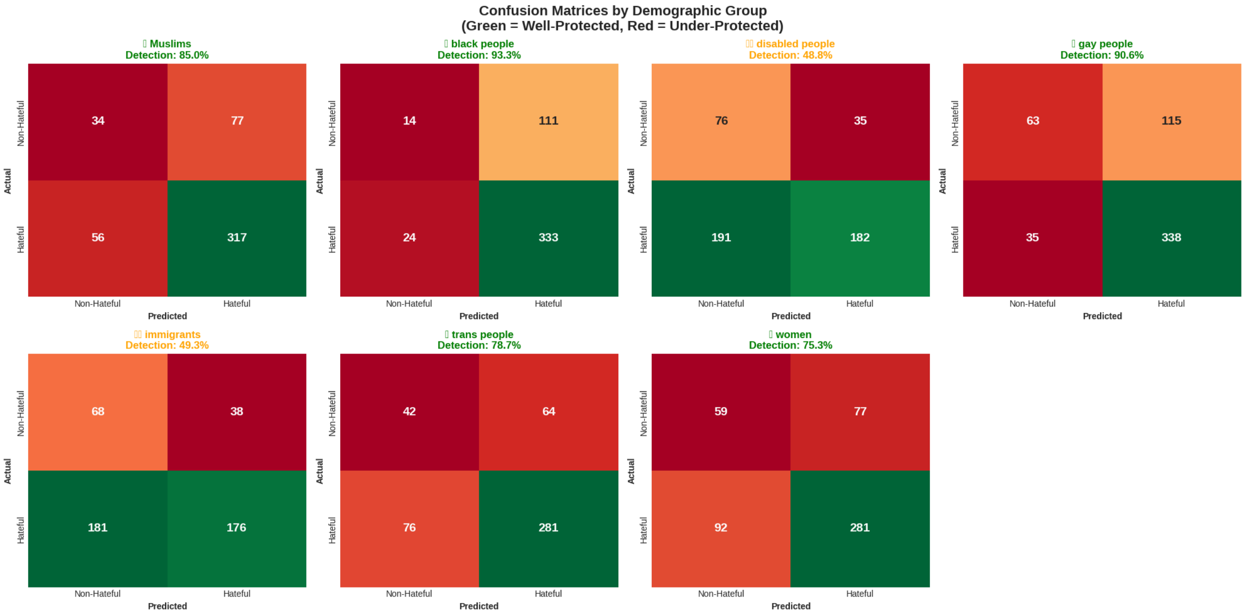
	Predicted: Non-Toxic	Predicted: Toxic
Actual: Non-Toxic	73.4% of total (Correct) 535	12.5% of total (FP: 14.5%) 91
Actual: Toxic	2.5% of total (Missed) 18	11.7% of total (Caught) 85

Jigsaw Validation: Male
638 comments mentioning this identity (14.6% toxic, 85.4% non-toxic)
Detection: 79.6% | FP Rate: 19.4% | F1: 54.2%



	Predicted: Non-Toxic	Predicted: Toxic
Actual: Non-Toxic	68.8% of total (Correct) 439	16.6% of total (FP: 19.4%) 106
Actual: Toxic	3.0% of total (Missed) 19	11.6% of total (Caught) 74

Confusion Matrices by group identity - HateCheck data



Group	Detection Rate - Jigsaw	Detection Rate - Hatecheck	FP Rate - Jigsaw	FP Rate - HateCheck	Pattern
black people	89%	93.30%	46%	88.80%	Over-policed despite strong protection
gay people	88%	90.60%	51%	64.60%	Over-policed despite strong protection
Muslims	82%	85.00%	31%	69.40%	Escalating over-policing
trans people	62%	78.70%	24%	60.40%	Under-protected, increasing surveillance
women	83%	75.30%	15%	56.60%	Declining protection, increasing surveillance

We identified two separate failure modes affecting different groups:

Pattern A - Under-Protection (Low Detection, Lower False Positives):

- Disabled people: 48.8% detection, 31.5% FP rate
- **Impact:** The model fails to detect over half of hate speech targeting these groups, leaving them vulnerable to online harm.

Pattern B - Over-Policing (High Detection, High False Positives):

- Black people: 93.3% detection, 88.8% FP rate
- Muslims: 85.0% detection, 69.4% FP rate
- Gay people: 90.6% detection, 64.6% FP rate
- **Impact:** While these groups receive strong protection from actual hate speech, the model wrongly flags 60-89% of innocent content mentioning these identities, creating excessive censorship and chilling effects on legitimate discourse.

Plan for Completion:

Text preprocessing included selecting relevant columns, renaming columns, data imputation and merging the hatecheck data.

We fine-tuned DistilBERT (a lightweight BERT variant) for binary toxicity classification. The model was trained on 150,000 comments using an 80/20 train-validation split. Text was tokenized with a maximum length of 128 tokens.

To address severe class imbalance (7% toxic vs 93% non-toxic), we implemented weighted loss with class weights of 0.54 for non-toxic and 6.18 for toxic comments, penalizing the model 11.4x more heavily for missing hate speech than for over-flagging non-toxic content.

Training configuration: 1 epoch, batch size of 32, learning rate of 2e-5, with evaluation every quarter-epoch. We used standard classification metrics (accuracy, precision, recall, F1, AUC) to evaluate performance, with F1 score as the primary metric for model selection.

List the steps you need to take between now and the final presentation, with the estimated number of person/hours required for each.

- ○ Prince: Review the exploratory data analysis charts and assist in interpreting the charts and results. Estimate: 10h
- ○ Bruna: Train the model using the entire Jigsaw dataset and update the Evaluation and deployment/test using Hatecheck dataset and EDA. Estimate: 12h
- ○ Maya: Work on building out the report more and get started on the final presentation. I will write the introduction, domain problem, and start on documenting everything done for the EDA and modeling sections. Lastly, we will need to formulate the key results using figures from our analysis and summarize everything in a cohesive conclusion. Estimate: 10h

Describe any uncertainties or impediments to having a good solution and presentation done by the presentation day.

- Processing a very large data set (> 2 million comments) with the limited GPU available. This is a major concern because it takes hours just to load the data, let alone for a neural network to learn from it. Any guidance on how to handle the situation, whether it is working with a smaller sample of the jigsaw data or access to more processing power, would be appreciated.
- Since the percentage of toxic comments in the training data is extremely disproportionate to non-toxic comments (class imbalance), is there a way to change the weights in the neural network so that there is more emphasis on toxic comments accuracy?

Citations:

Jigsaw/Conversation AI. (2019). Jigsaw Unintended Bias in Toxicity Classification [Dataset]. Kaggle. https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 491-500). https://doi.org/10.1145/3308560.3317593

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 41-58). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.4

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models [Dataset]. GitHub. https://github.com/paul-rottger/hatecheck-data

U.S. Census Bureau. (2020). Decennial Census: Demographic and Housing Characteristics (DHC) [Dataset]. U.S. Department of Commerce. https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html

U.S. Census Bureau. (2022). American Community Survey 5-Year Estimates [Dataset]. U.S. Department of Commerce. Retrieved from https://data.census.gov/

Group Dynamics:

- **Communication**: We will use **Google Docs** and **Slack** for effective communication and document sharing.
- **Data and Code Sharing**: All data and code will be managed via **GitHub** for version control.
- **Meetings**: Weekly team meetings on Mondays and Fridays are scheduled to discuss progress and address any issues.