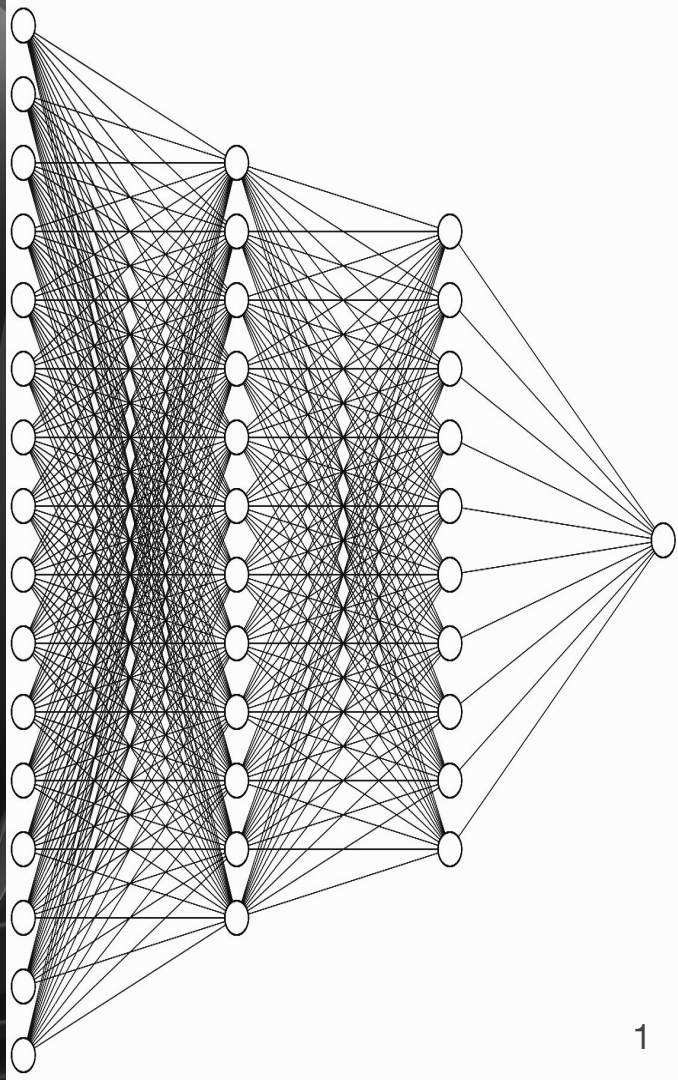
# AI Representation Gap:
## Building and Auditing a Hate Speech Detector For Demographic Fairness

**Bruna Macedo Porto** - Modeling Lead
**Prince Newman** - Data Lead
**Maya Silver** - Data Comms. Lead

December 2025



# Table of Content

# Domain Problem

The Promise and Problem of AI Content Moderation

**The Promise:**

- AI content moderation processes millions of comments daily
- Automates hate speech detection at a very large scale
- Protects communities from online harassment

**The Problem:**

- Does AI protect all communities equally??
- Can algorithms be biased even when
- Who is  protected? Who is silenced?

# Domain Problem

What does algorithmic bias in hate speech detection look like?

1. **Representation bias:** Unequal dataset representation
2. **Disparate Performance:** Different accuracy rates across groups
3. **Over-policing:** Disproportionate flagging of toxicity among certain communities
4. **Invisibility:** Missing attacks on marginalized groups

# Domain Problem

Why does this matter?

- Billions use AI-moderated platforms every day
- Unequal protection violates fairness principles
- Over-policing silences marginalized voices
- AI systems scale biases, not just reflects them

# Imports

**Libraries**

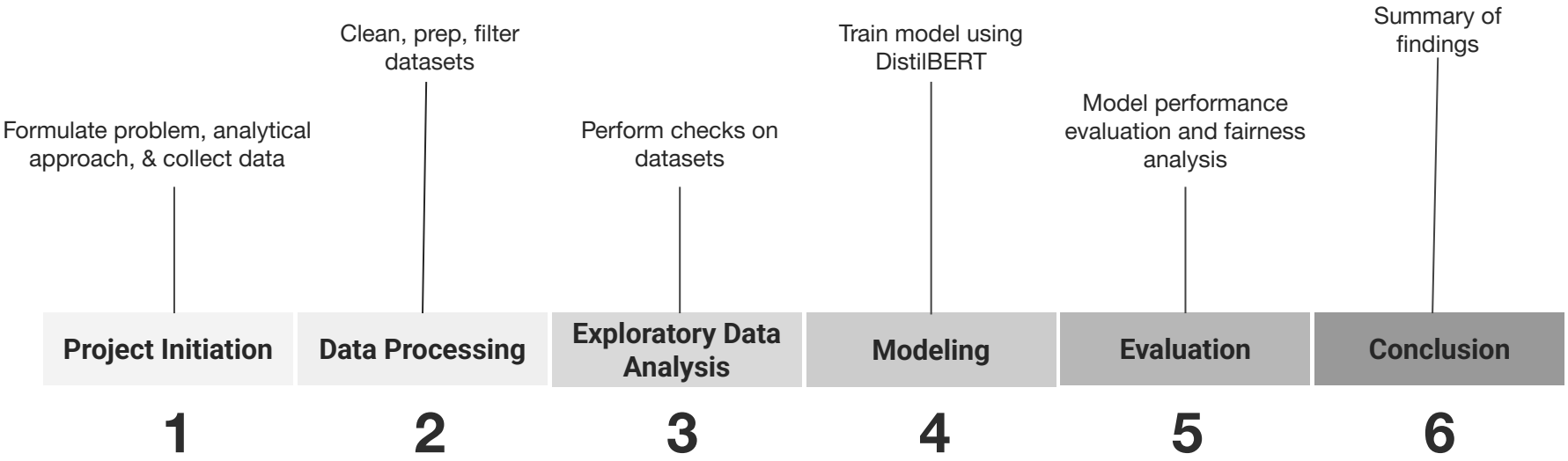
pandas \ numpy \ matplotlib \ seaborn
tqdm requests \ scikit-learn \ statsmodels
fairlearn \ transformers \ datasets torch \
plotly \ spacy \ nltk \ pyarrow \ fastparquet
\ joblib \ accelerate \ scipy

# Data Sources

**Jigsaw Civil Comments (Kaggle)**
Contains ~2 million online comments labeled for toxicity and identity mentions.
This dataset was used to train the hate speech detection model and analyze representation across demographic groups.

**HateCheck (Paul Röttger GitHub)**
A test suite with 3,728 cases designed to benchmark fairness in hate speech detection models.
Focuses on demographic variations and edge cases in language usage. This dataset was used to evaluate model fairness and identify potential biases in predictions.

# Project Structure



# Data Processing & Cleaning

**Missing Value Handling:** Removed all rows lacking comment text to ensure dataset integrity.

**Binary Label Generation:** Converted toxicity scores (0–1) into binary labels using a 0.5 threshold.

**Identity Field Processing:** Filled missing identity indicators and grouped them into aggregated categories.

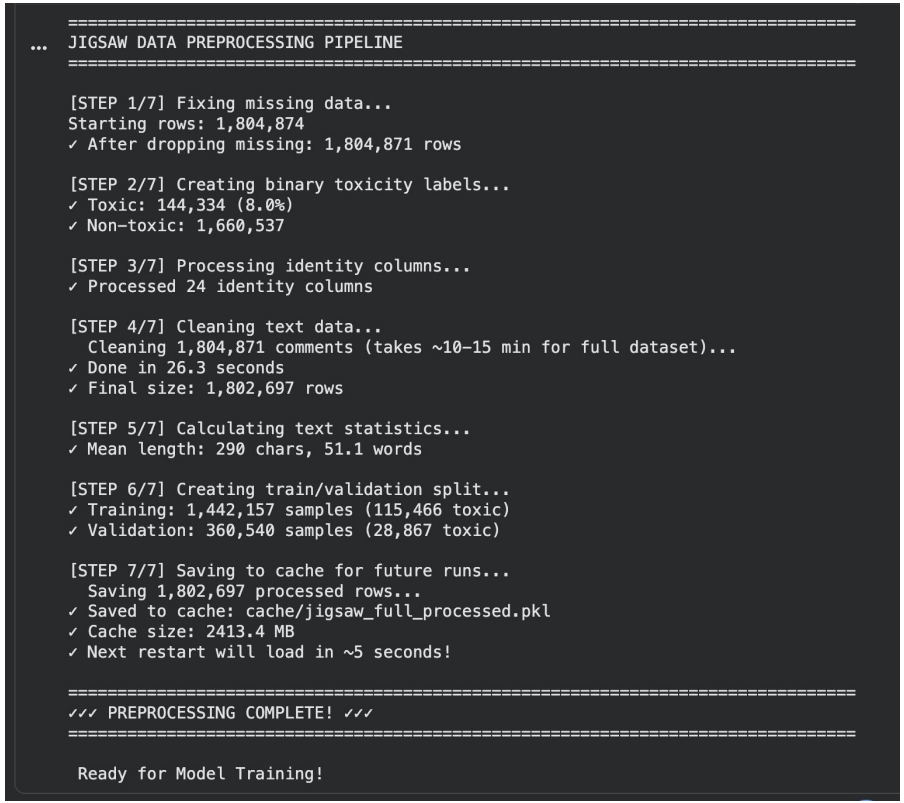
**Text Normalization:** Applied lowercasing, URL and HTML removal, special-character cleanup, and whitespace normalization.

**Feature Stats:** Computed text length and word-count features for downstream analysis.

**Stratified Split:** Performed an 80/20 train–validation split stratified by toxicity label.

**Caching:** Saved the processed dataset for faster subsequent loading.

**Output Summary:**
~1.8M cleaned comments; ~1.4M training samples; ~360k validation samples; class imbalance ≈ 11.5:1 (non-toxic : toxic).



# Exploratory Data Analysis

# EDA:

## Identity Group Representation

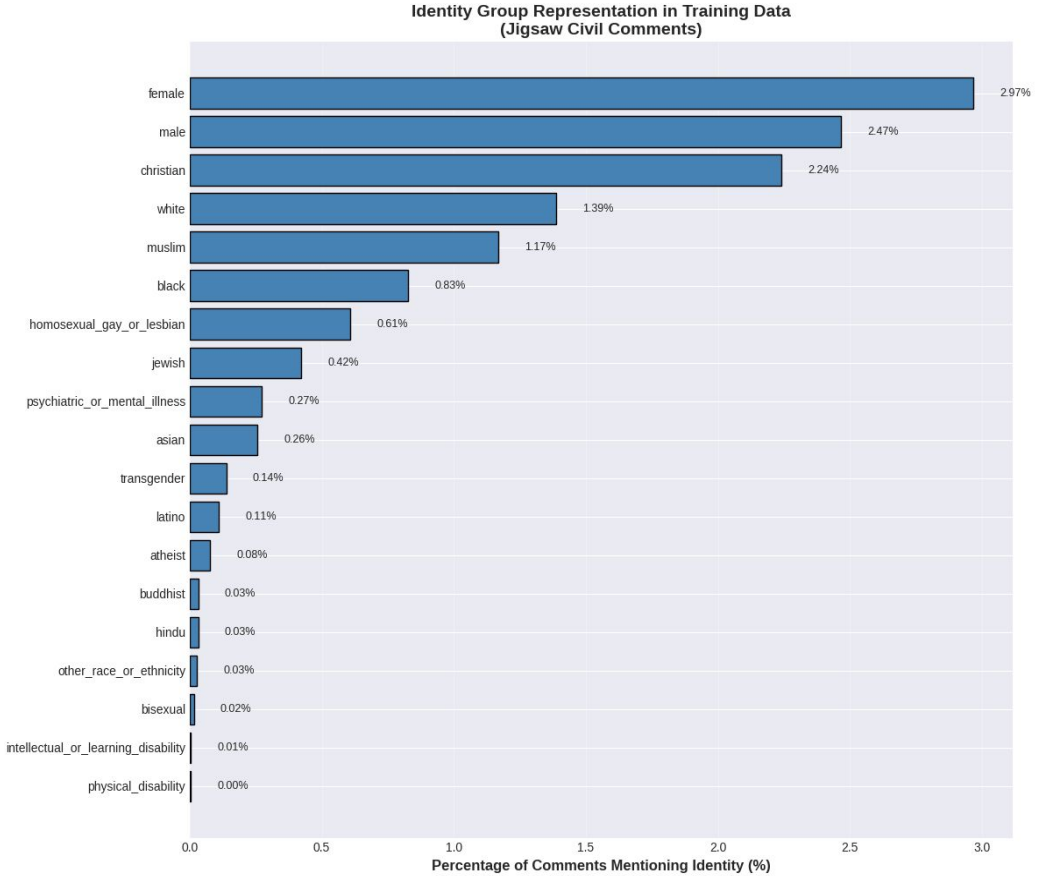
1. **Label Confirmation:** The chart correctly displays 19 identity groups with their respective percentages of mention in the Jigsaw Civil Comments training data, ordered by frequency.

2. **Representation Hierarchy:** The groups related to Gender ("female" at 2.97% and "male" at 2.47 are the most mentioned, followed by "christian" and "white".

3. **High-Level Representation:** The top four groups (female, male, christian, white) account for a disproportionately large share of the identity mentions, with each exceeding 1.3% representation**.**

4. **Low-Level Representation:** Many minority and disability groups, such as "bisexual" and "physical_disability", are mentioned in less than 0.03% of comments, indicating severe under-representation.

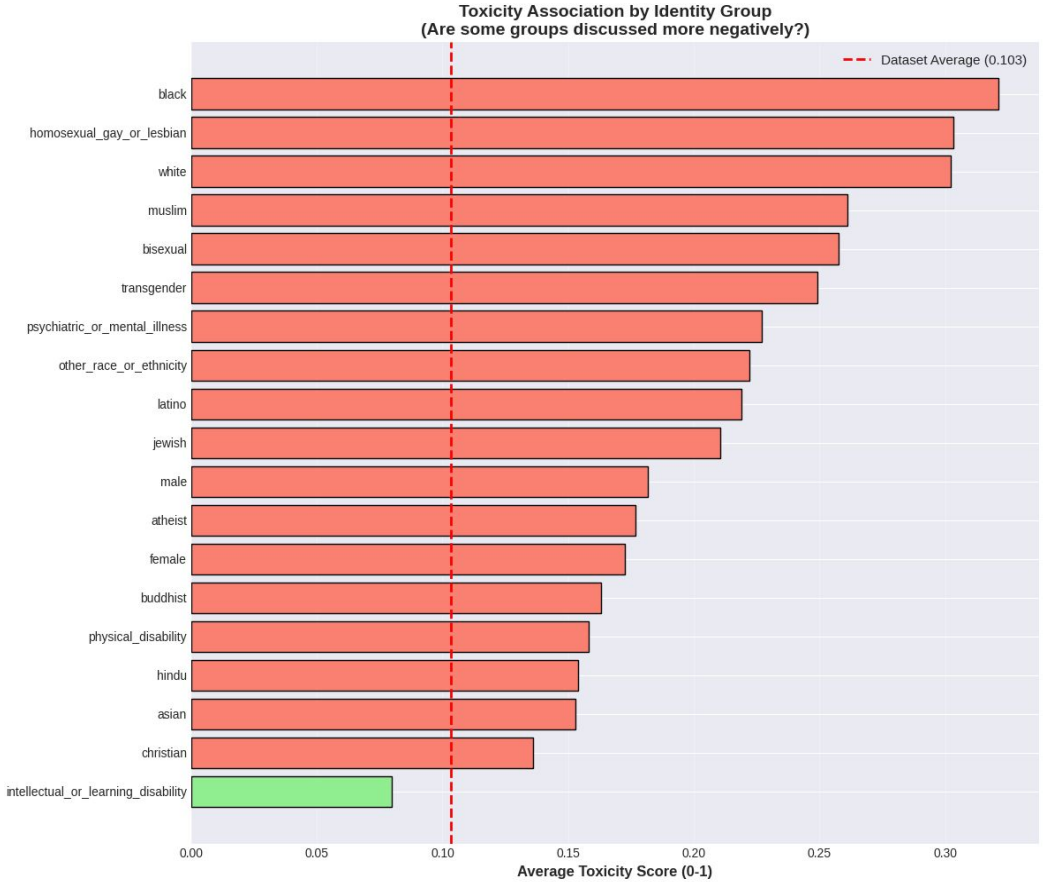
5. **Data Distribution:** The distribution is highly skewed, suggesting that the model training data has significantly more examples for common gender and majority religious/racial identities than for under-represented groups.



# EDA:

## Toxicity Score by Identity

1. **Metric Confirmation:** The chart visually represents the average toxicity score, where bars on the x-axis (0 to 1) show the mean toxicity of comments that mention the identity groups listed on the y-axis.
2. **Highest Toxicity:** Homosexual_gay_or_lesbian and transgender groups exhibit the highest average toxicity scores, suggesting they are discussed most negatively in the comments.
3. **Bias Observation:** The high toxicity scores for minority groups, especially LGBTQ+ and racial identities, indicate a problematic association where comments mentioning these groups are more likely to be labeled toxic, irrespective of the comment's true toxicity.
4. **Lowest Toxicity:** Majority groups like male, female, and christian generally show the lowest average toxicity scores, indicating that comments mentioning these identities are less often flagged as toxic.
5. **Representation vs. Toxicity:** There is a clear inverse relationship where the most underrepresented groups (low count, such as transgender) tend to have the highest average toxicity scores, highlighting the bias.
6. **Model Implication:** This dataset distribution causes models to develop unintended bias, incorrectly learning that the mere mention of a minority identity is a strong predictor of a high toxicity score.



# EDA:

## Toxicity Class Distribution

1. **Metric Confirmation:** The stacked bar chart illustrates the percentage breakdown of comments for each identity group that are classified as Toxic versus Non-Toxic.

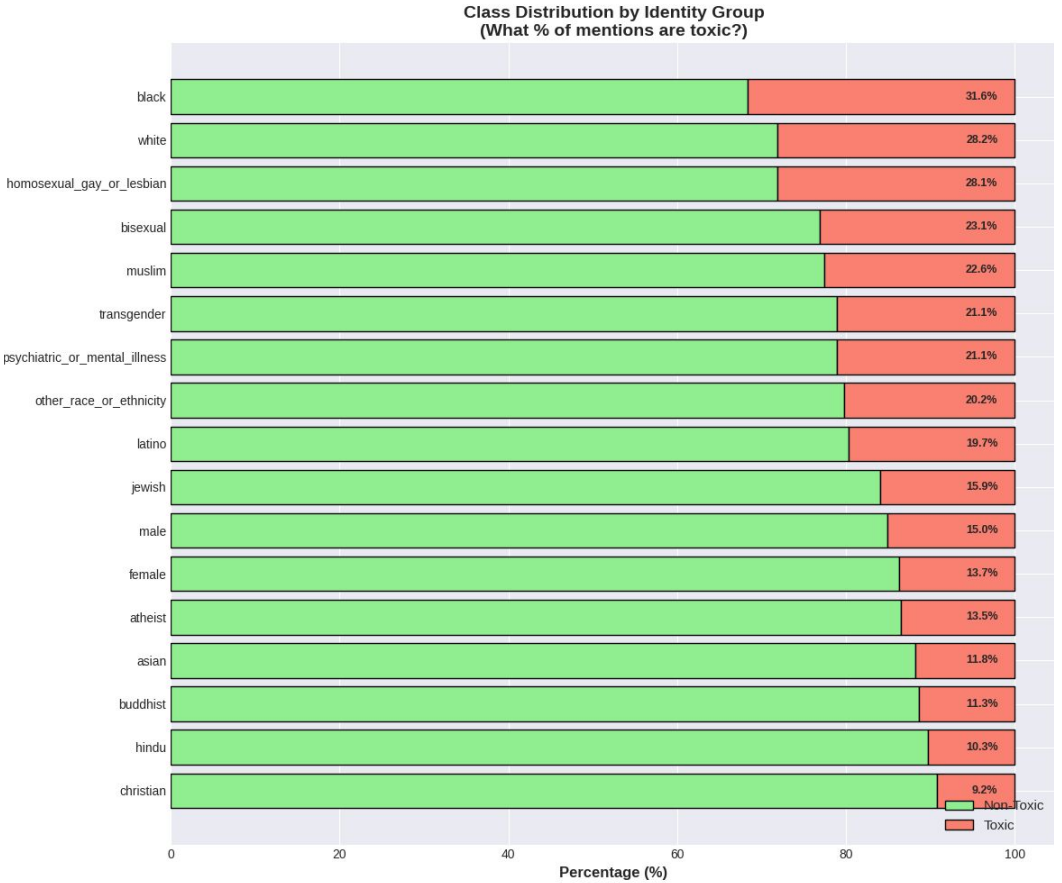
2. **Highest Toxicity:** The black identity group has the highest percentage of toxic mentions at 31.6%, meaning nearly one-third of comments mentioning "black" are toxic.

3. **Top Toxicity Cluster:** Groups including black, white, homosexual_gay_or_lesbian, and bisexual all have toxicity rates above 23%, forming the cluster most frequently associated with toxic comments.

4. **Lowest Toxicity:** The christian group shows the lowest toxicity rate at 9.2%, indicating it is the least likely identity to be mentioned in a toxic comment in this dataset.

5. **Minority Group Trend:** Most minority and marginalized groups (transgender, psychiatric_or_mental_illness, latino) show toxicity percentages above the overall dataset average, highlighting a negative conversational trend.

6. **Bias Reinforcement:** This distribution visually reinforces the un intended bias issue, as the data shows that comments mentioning certain identity groups are inherently more toxic than others.



# EDA:

## Sample Distribution By Identity

1. **Metric Confirmation:** The bar chart shows the total number of comments mentioning each identity group, utilizing a logarithmic scale on the x-axis for comparison.

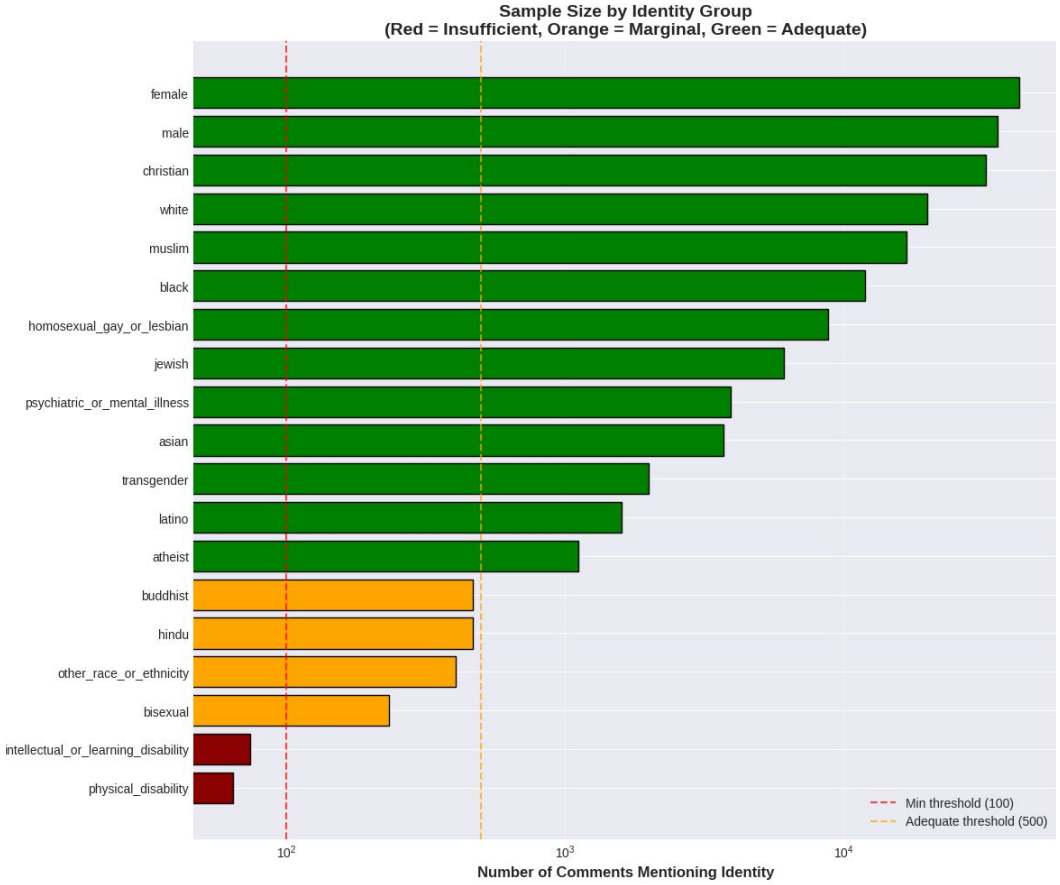
2. **Adequate Samples (Green):** Twelve groups, led by female and male, are classified as Adequate (Green, >= 500 comments), meaning they have sufficient data for training.

3. **Marginal Samples (Orange):** Four groups, including buddhist and bisexual, are classified as Marginal (Orange, between 100 and 500 comments), indicating they have barely enough data.

4. **Insufficient Samples (Red):** The groups intellectual_or_learning_disability and physical_disability are classified as Insufficient (Red, < 100 comments), making reliable model analysis difficult for these groups.

5. **Representation Disparity:** The highest count group (female) has over 10,000 comments, while the lowest (physical_disability) has fewer than 100 comments, illustrating an extreme sample size disparity.

6. **Log Scale Interpretation:** The logarithmic x-axis emphasizes that the top 8 groups have orders of magnitude more data than the bottom groups, compounding the potential for unintended bias.



# EDA:

## Co-occurrence Matrix

1. **Metric Confirmation:** The heatmap shows the raw count of comments where any two identity groups intersect, with a darker color indicating a higher frequency of co-occurrence.

2. **Highest Co-occurrence:** Male and Female are mentioned together most frequently, with a count of 42,794, reflecting high general discussion about gender.

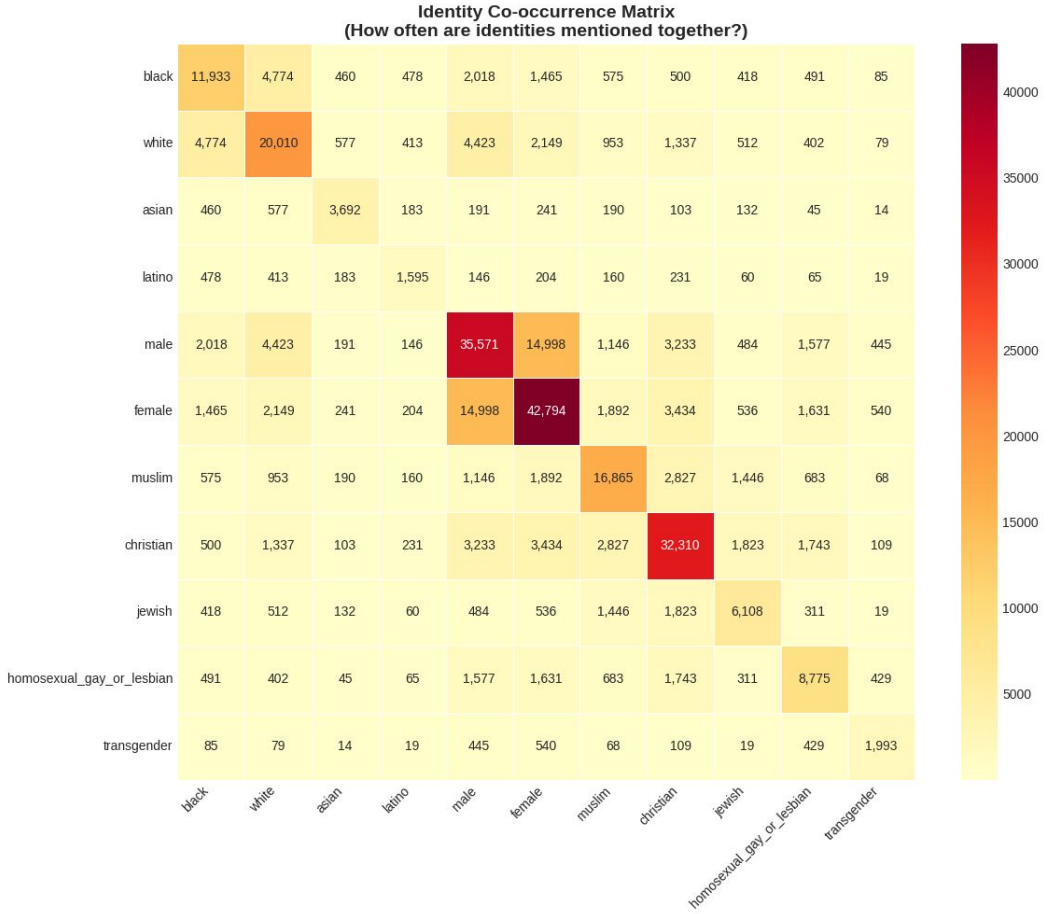
3. **High Intra-Group Mention:** The diagonal cells show high counts (e.g., 35,771 for male and 42,794 for female), indicating a high number of comments that mention the same group multiple times or use different terms for it.

4. **Racial Co-occurrence:** Black and White are frequently mentioned together (4,774 times), suggesting common comparative discussions or conflicts between these racial groups.

5. **Religious Co-occurrence:** Muslim and Christian are mentioned together 2,827 times, indicating a significant number of comments comparing or contrasting these two religions.

6. **Sexual Orientation Co-occurrence:** Homosexual_gay_or_lesbian is most often mentioned alongside male (1,577) and female (1,631), and also with transgender (1,993), suggesting discussions where sexual orientation and gender identity are linked.

7. **Low Co-occurrence:** Identity groups like Asian and Latino show very low co-occurrence counts with most other groups, indicating they are rarely the subject of comparison or paired discussion.



# EDA:

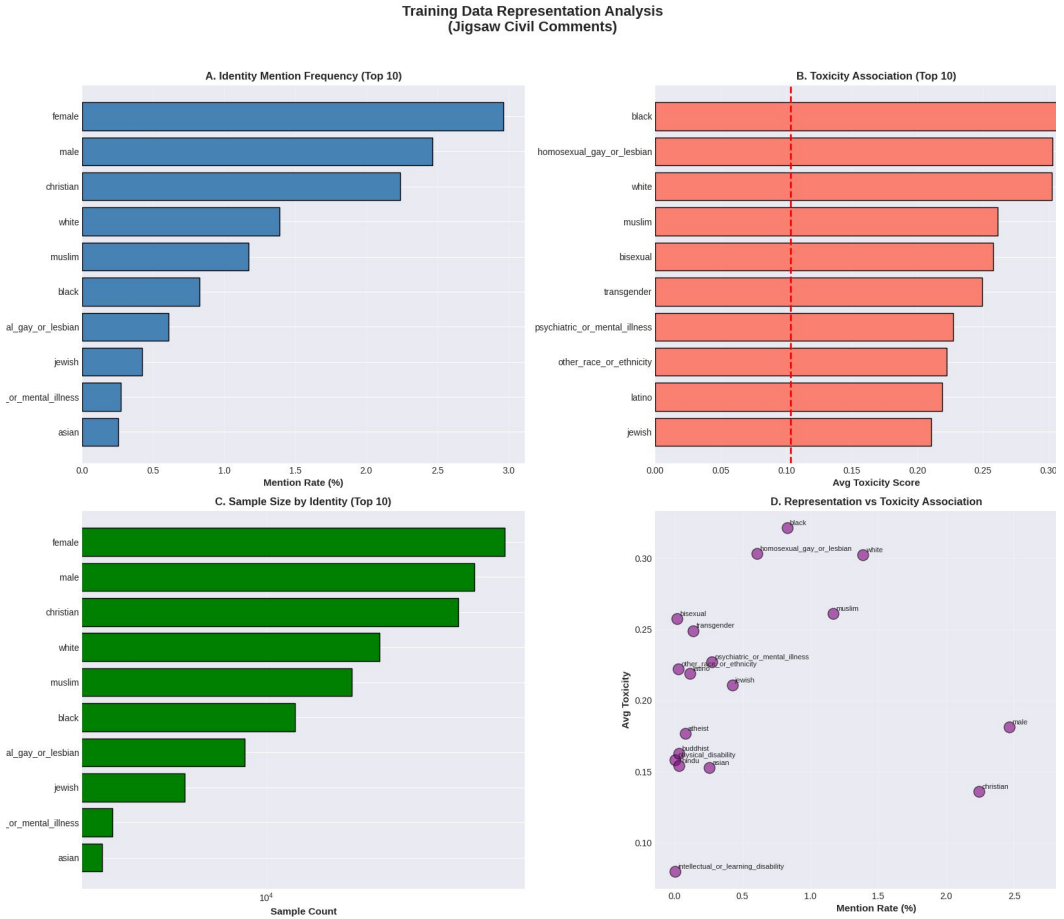
## Training Data Representation

A. **Identity Mention Frequency (Top 10):** Female and Male are the most frequently mentioned identities, establishing high representation for majority gender groups while showing a steep drop-off for minority groups.

B. **Toxicity Association (Top 10):** Black and homosexual_gay_or_lesbian exhibit the highest average toxicity scores, confirming that these groups are discussed most negatively compared to the dataset average.

C. **Sample Size by Identity (Top 10):** While most groups have Adequate sample sizes, the groups physical_disability and intellectual_or_learning_disability have Insufficient raw counts, limiting reliable analysis for those identities.

D. **Representation vs Toxicity Association:** This scatter plot demonstrates the unintended bias that groups with low representation (e.g., transgender) are overwhelmingly associated with high average toxicity scores, while high-representation groups (e.g., female) are associated with low scores.

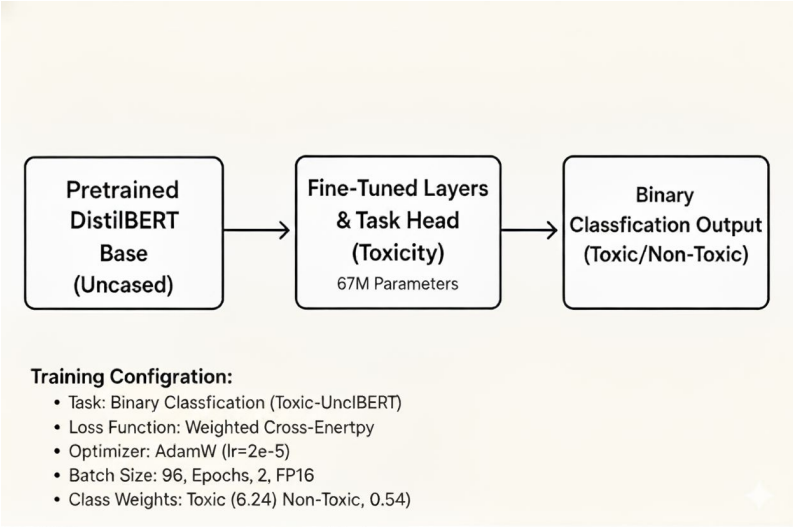


# Model Training & Analysis

# Methodology

| | | |
|---|---|---|
| Problem Type | | **Supervised Binary Classification:** Predict whether a comment is toxic (1) or non-toxic (0) |
| Analytical Framework | | 1) Train hate speech detector on Jigsaw data<br>2) Evaluate overall model performance<br>3) Calculate fairness metric by demographic group<br>4) Validate with HateCheck benchmark<br>5) Cross-dataset correlation analysis |
| Metrics | Performance Metrics | Accuracy, Precision, Recall, F1 |
| | Fairness Metrics | Detection Rate by group, False Positive Rate by group, Fairness Gap (max - min) |
| | Statistical Tests | Correlation |

# Model Architecture and Training




**The classification model was built using DistilBERT, a lightweight transformer pretrained on large-scale text corpora.**

**Approach:** Transfer learning was applied by adapting the pretrained DistilBERT model to the binary task of classifying text as toxic or non-toxic.

**Implementation:** The model was instantiated using the distilbert-base-uncased architecture for sequence classification, with class weighting incorporated to address dataset imbalance.

**Training:** The model was fine-tuned for two epochs using AdamW with weight decay, a learning rate of 2e-5, FP16 mixed precision, and batch size 96. Evaluation occurred every 5,000 steps, with checkpoints saved every 10,000 steps, and the best model selected based on F1 score.

# Overall Model Performance

PERFORMANCE METRICS

**91.5%**
Accuracy

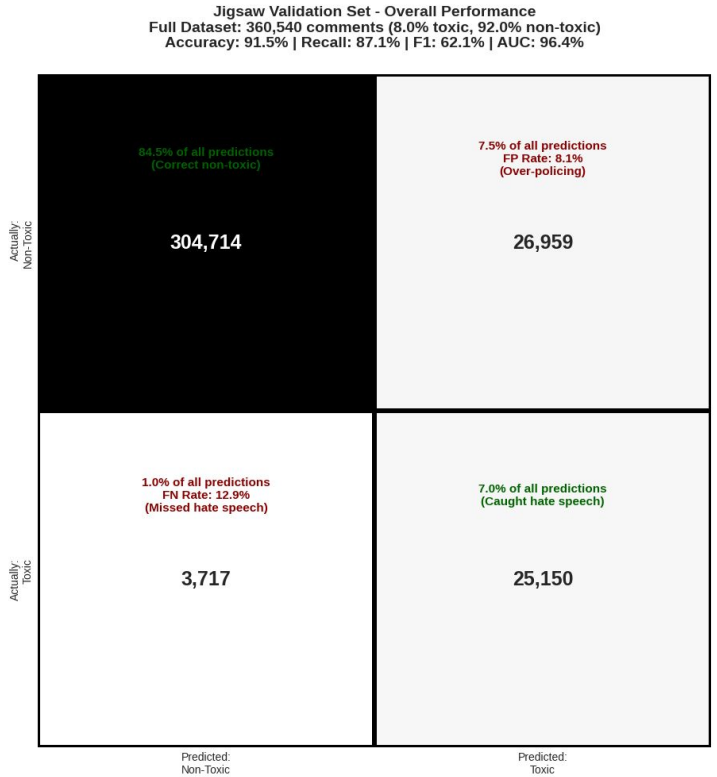
**87.0%**
Recall

**96.4%**
AUC-ROC

**62.1%**
F1 Score

Key Insight: Model prioritizes catching hate speech (87% recall) over precision due to class weighting

CONFUSION MATRIX ANALYSIS



| | Identity Group | Samples | Detection Rate | FP Rate | Recall | F1 |
|---|---|---|---|---|---|---|
| 3 | Latino People | 409 | 93.2% | 25.1% | 93.2% | 60.8% |
| 10 | Gay/Lesbian People | 2,222 | 93.1% | 49.4% | 93.1% | 59.7% |
| 0 | Black People | 2,968 | 93.1% | 51.9% | 93.1% | 59.9% |
| 1 | White People | 5,072 | 92.2% | 49.1% | 92.2% | 57.6% |
| 7 | Men | 8,912 | 88.7% | 20.8% | 88.7% | 57.8% |
| 9 | Trans People | 505 | 88.4% | 33.6% | 88.4% | 57.7% |
| 4 | Muslims | 4,135 | 88.0% | 36.1% | 88.0% | 57.6% |
| 8 | Women | 10,632 | 85.6% | 17.7% | 85.6% | 57.7% |
| 6 | Jewish People | 1,543 | 83.5% | 23.3% | 83.5% | 56.5% |
| 5 | Christians | 8,111 | 81.8% | 13.1% | 81.8% | 51.5% |
| 2 | Asian People | 886 | 80.8% | 14.9% | 80.8% | 60.3% |
| 12 | Intellectually Disabled | 19 | 50.0% | 17.6% | 50.0% | 33.3% |
| 11 | Physically Disabled | 18 | 0.0% | 47.1% | 0.0% | 0.0% |

# Model Fairness Analysis:

## Fairness Metrics by Identity Group

**Detection Rate = Protection Level**

- % of hate speech targeting a group that the model catches

**Well-Protected**

- Gay/Lesbian, Trans, Women, Muslims, Black, Jewish people

**Under-Protected**

- Disabled (Intellectually and Physically)

**Root cause**: Groups with fewer training examples and less toxic context receive less protection.

# The Protection-Surveillance Trade-off

HATECHECK RESULTS BY GROUP

**Black People**
Detection: 96.1% | FP Rate: **98.4%**

**Gay People**
Detection: 96.2% | FP Rate: **84.8%**

**Muslims**
Detection: 92.0% | FP Rate: **68.5%**

**Trans People**
Detection: 92.4% | FP Rate: **66.0%**

**Women**
Detection: 83.9% | FP Rate: **59.6%**

**Disabled People**
Detection: **74.0%** | FP Rate: **44.1%**

DOUBLE JEOPARDY PATTERN

Some groups face the worst of both worlds: under-protected AND over-policed.

**Extreme Over-Policing**

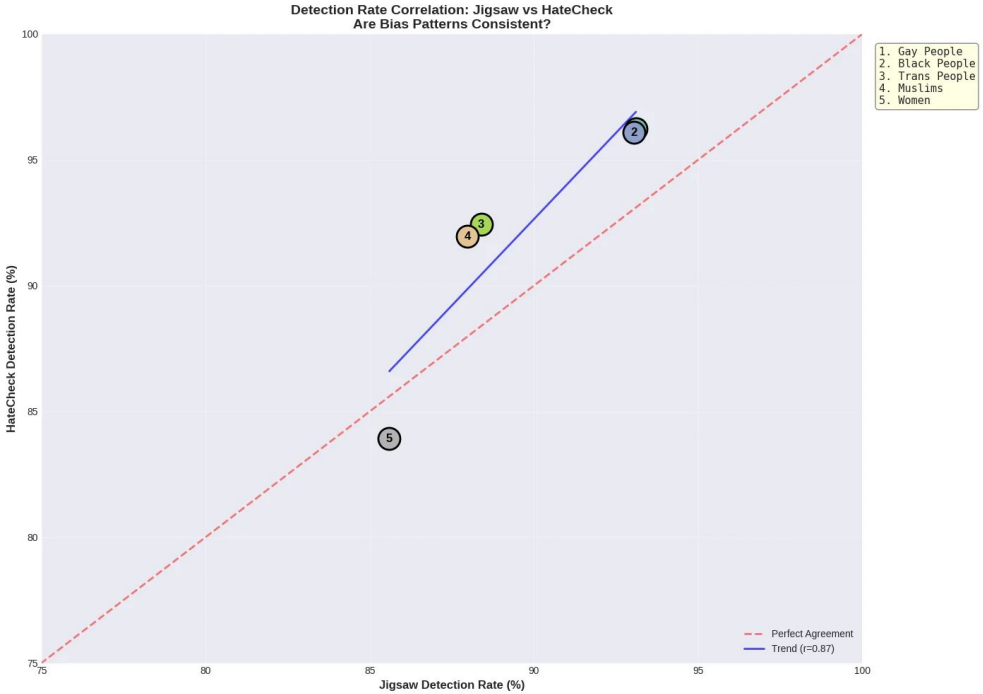
**Black People:** 98.4% false positive rate
**Gay People:** 84.8% false positive rate
High protection but severe over-surveillance

**Under-Protection**

**Disabled People:** 74.0% detection
26% of hate speech goes undetected

**Key insight:** High detection ≠ fairness. Groups can be well-protected yet severely over-policed.

# Cross-Dataset Validation




Detection Rate Correlation

**r = 0.875** p = 0.0099

Strong positive correlation - highly significant

### What This Proves

Groups under-protected in real-world data (Jigsaw) are also under-protected in controlled testing (HateCheck). The bias is **systematic**, not random.

WHY TWO DATASETS?

**Jigsaw:** Real-world distribution

**HateCheck:** Controlled tests, only target varies

**Together:** Matching patterns = model-embedded bias

# Results & Summary

# Key Results

**Research Question:**

Does a hate speech detector trained on real world data exhibit demographic fairness?

---

**Protection rate is relatively consistent:**

- 80.8% - 93.2% detection rates
- All groups receive adequate hate speech detection
- 12.4 percentage point gap across 11 identity groups

**Surveillance burden is highly unequal:**

- 13.2% - 51.0% false positive rates
- 37.8 percentage point gap
- Some communities face 4× more false alarms than others

# Implications for AI Content Moderation

| | |
|---|---|
| 01 | High accuracy ≠ Fair system |
| 02 | Training data shapes bias |
| 03 | Over-policing is as harmful as under-protection |

# Future Research Directions

| | |
|---|---|
| 01 | Fairness-aware training techniques |
| 02 | Intersectional analysis |
| 03 | Longitudinal monitoring |

# Conclusion

**What we showed:**

- Standard ML practices produce systematically biased hate speech detectors
- Patterns replicate across datasets
- High aggregate performance masks severe demographic disparities

---

**What we need:**

- Fairness-aware design
- Demographic-disaggregated evaluation
- Equitable treatment– not just high accuracy

# Thank you!