Harnessing Text Analytics to Shield Children from Online Toxicity: A Moderator Extension

Margot Gerard, Richard El Chaar, Atharva Vyas, Hazel Foo

10 Feb 2025

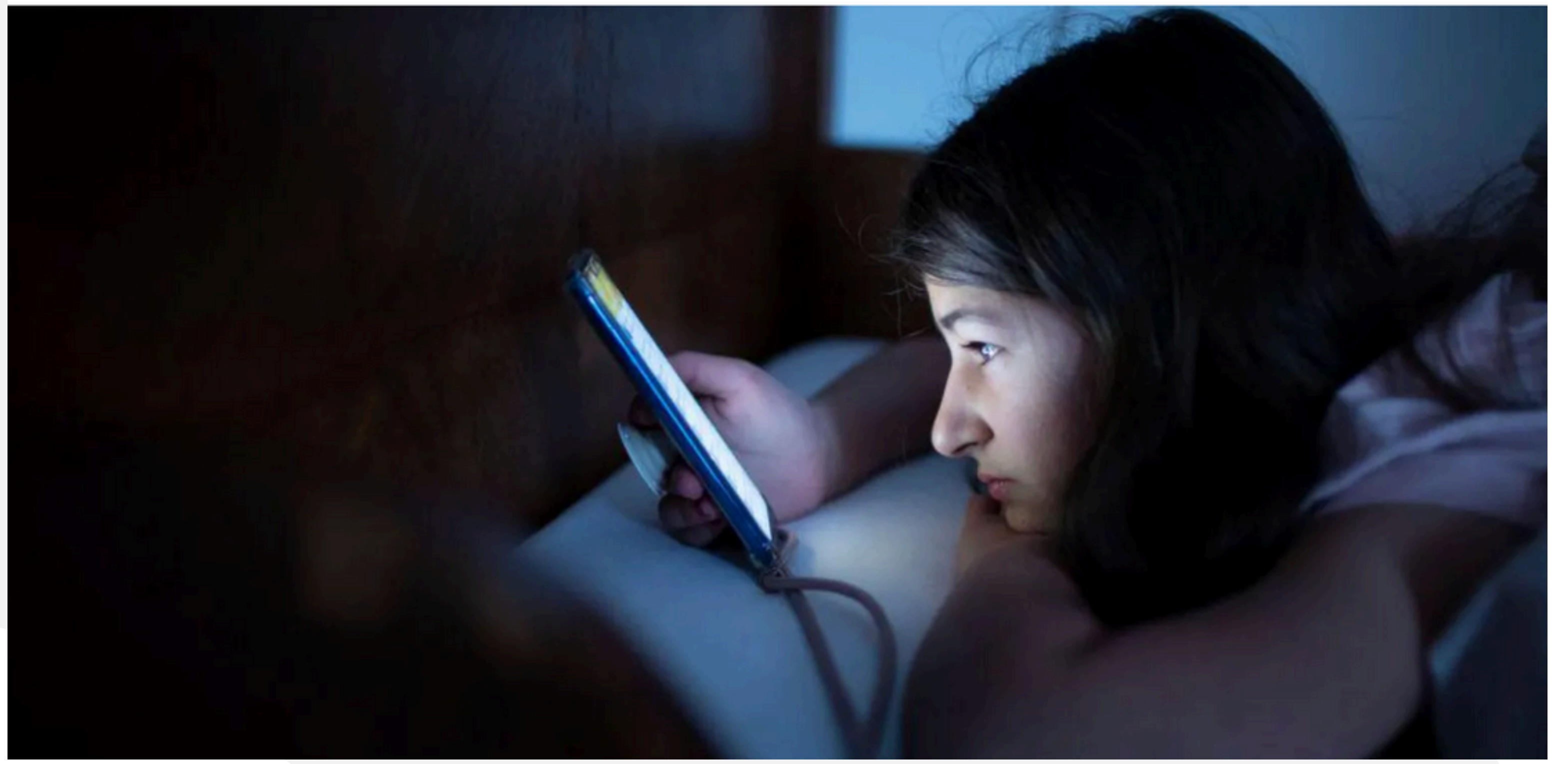
SOCIAL MEDIA AGE

Australia approves social media ban on under-16s

28 November 2024

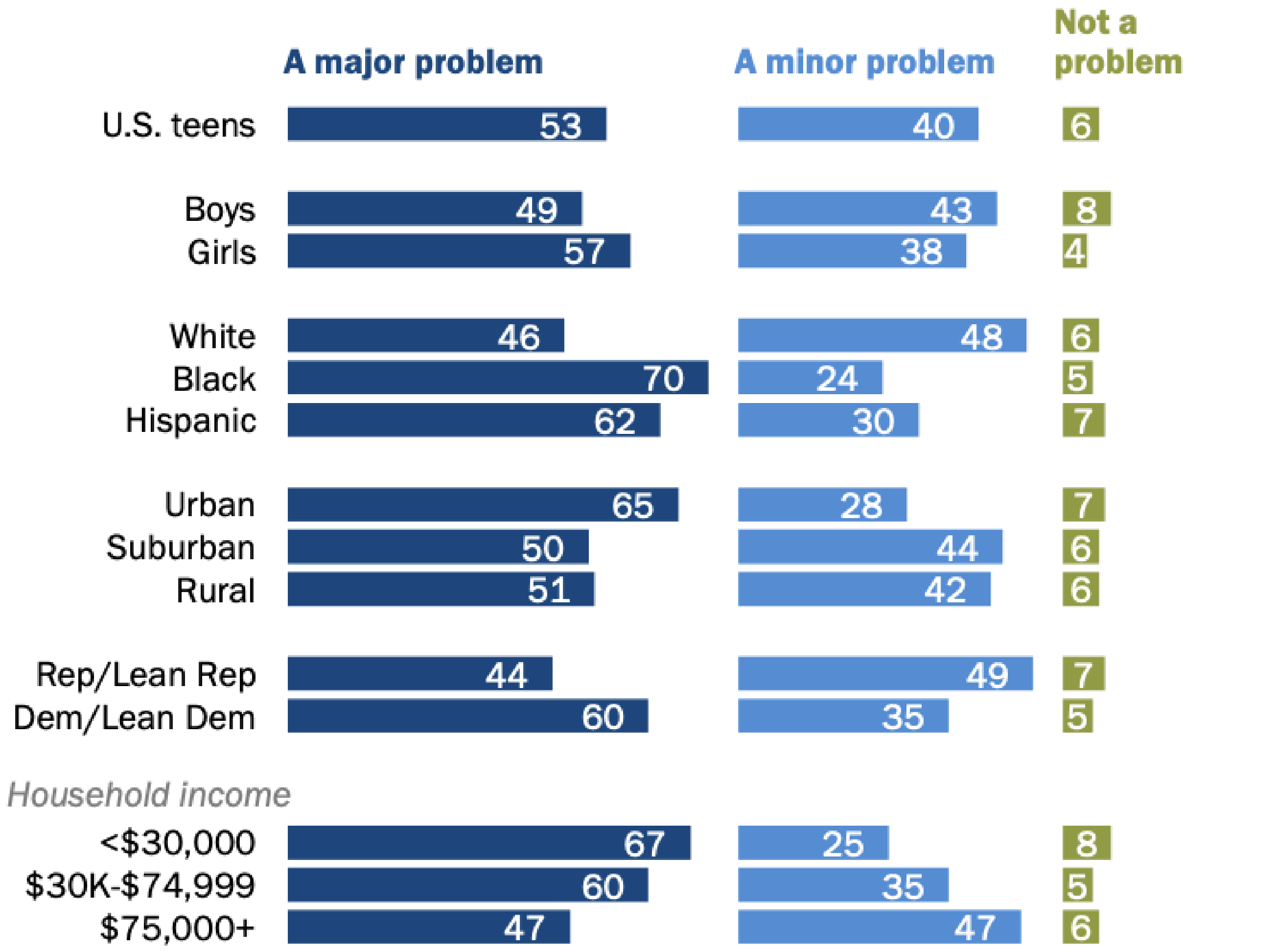
Share Save

Hannah Ritchie
BBC News, Sydney



Black or Hispanic teens are far more likely than White teens to say online harassment and bullying are a major problem for people their age

% of U.S. teens who say online harassment and online bullying are __ for people their age



Note: Teens are those ages 13 to 17. White and Black teens include those who report being only one race and are not Hispanic. Hispanic teens are of any race. Those who did not give an answer are not shown.
Source: Survey conducted April 14-May 4, 2022.
"Teens and Cyberbullying 2022"

CONTENT MODERATION TODAY

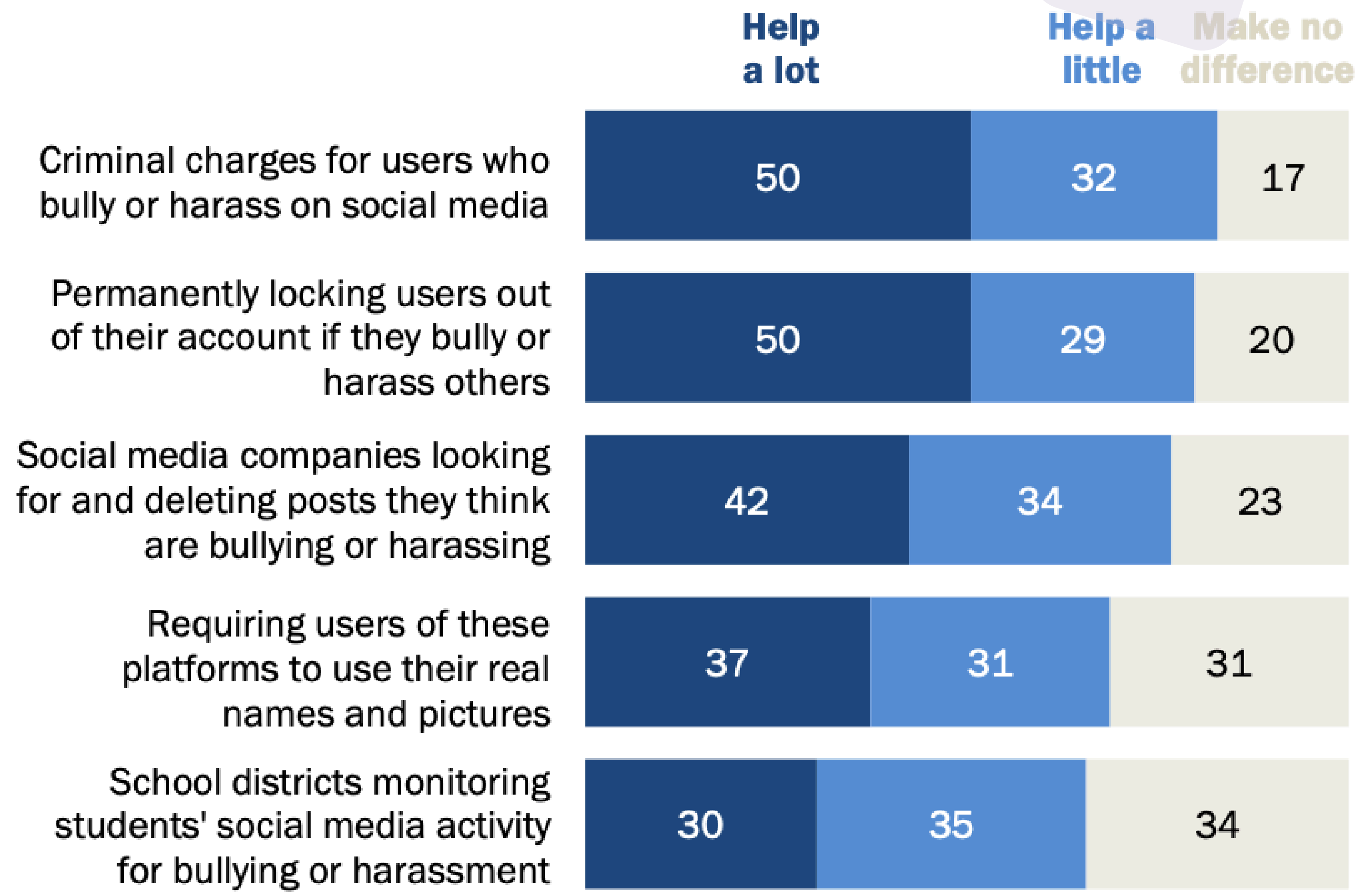
Social media companies "shamefully far" from tackling illegal and dangerous content

1 May 2017

The Home Affairs Committee has strongly criticised social media companies for failing to take down and take sufficiently seriously illegal content – saying they are "shamefully far" from taking sufficient action to tackle hate and dangerous content on their sites.

Half of teens think banning users who bully or criminal charges against them would help a lot in reducing the cyberbullying teens may face on social media

% of U.S. teens who say each of the following would __ in reducing the amount of harassment and bullying that teens may face on social media



Note: Teens are those ages 13 to 17. Those who did not give an answer are not shown.
Source: Survey conducted April 14-May 4, 2022.
"Teens and Cyberbullying 2022"

PEW RESEARCH CENTER

TEXT ANALYSIS APPROACH

STEP 1

Build text classifiers using training data from Kaggle

STEP 2

Finetune model parameters & choose the best performing classifier

STEP 3

Validate model performances on external web posts and comments

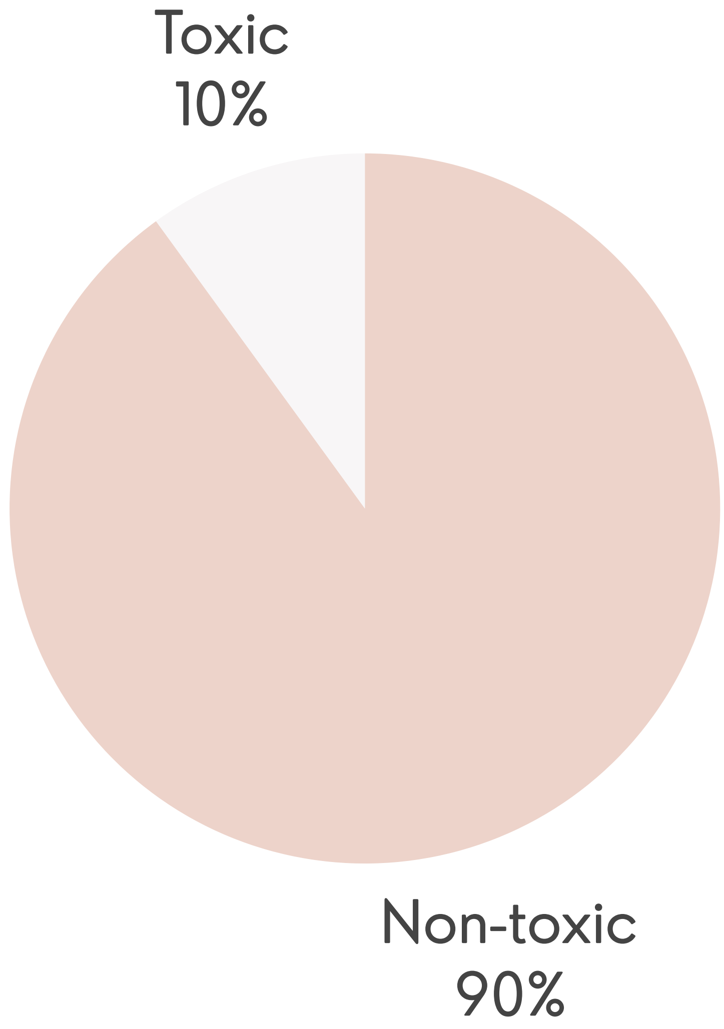
Data Description

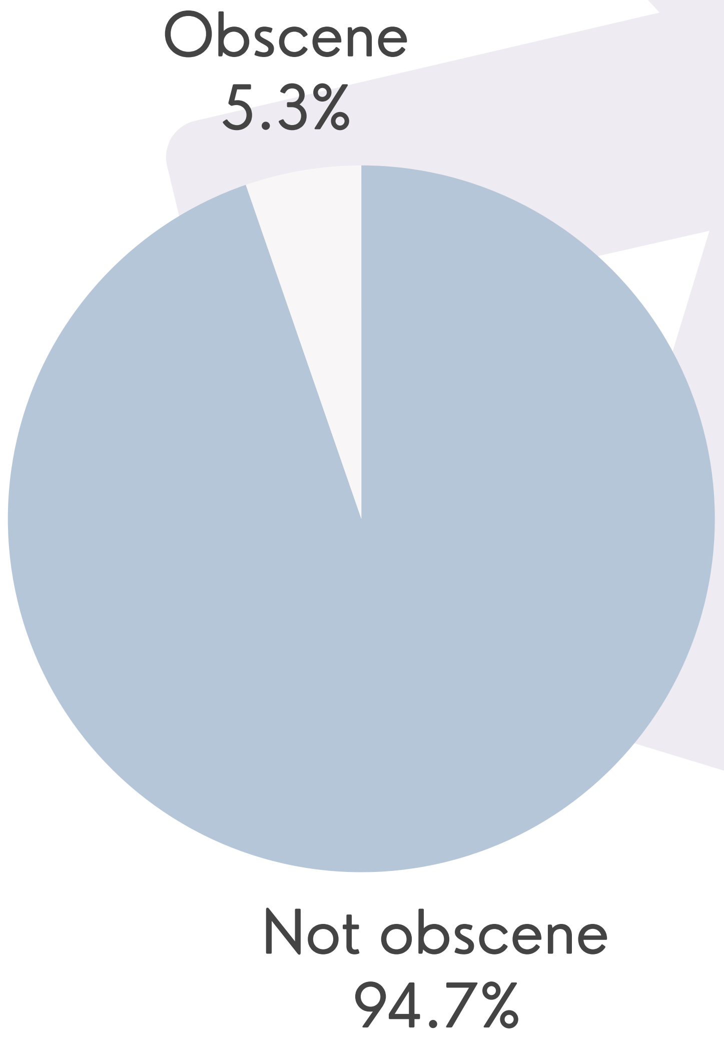
- 159450 comments from Wikipedia's talk page
- Variables:
  - Comment ID
  - Comment content
  - Binary variables for 6 different types of toxicity
    - Toxic
    - Severe toxic
    - Obscenity
    - Identity hate
    - Insult
    - Threat

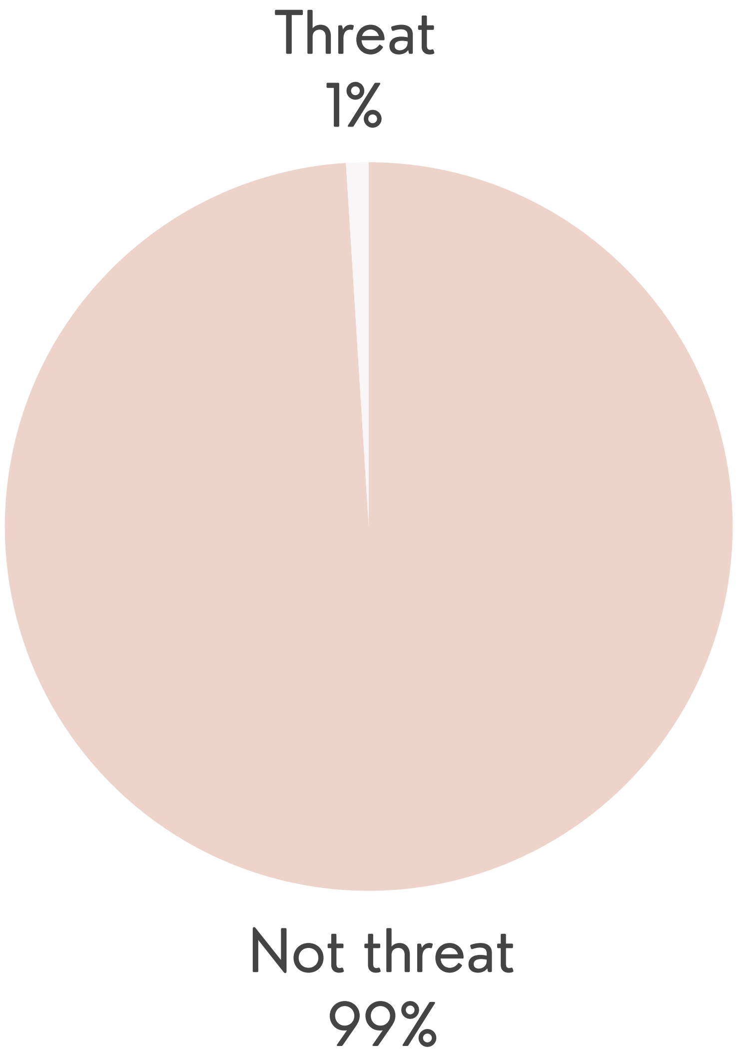
cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. Toxic Comment Classification Challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge, 2017. Kaggle.

Step 1: Build text classifier

Toxicity type breakdown

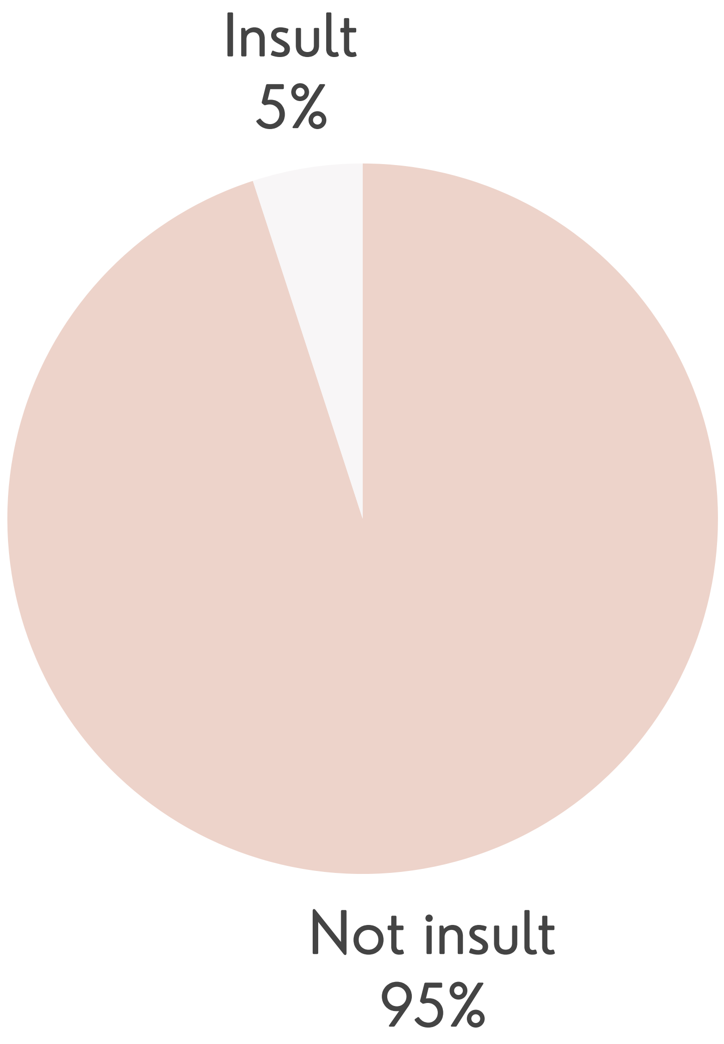


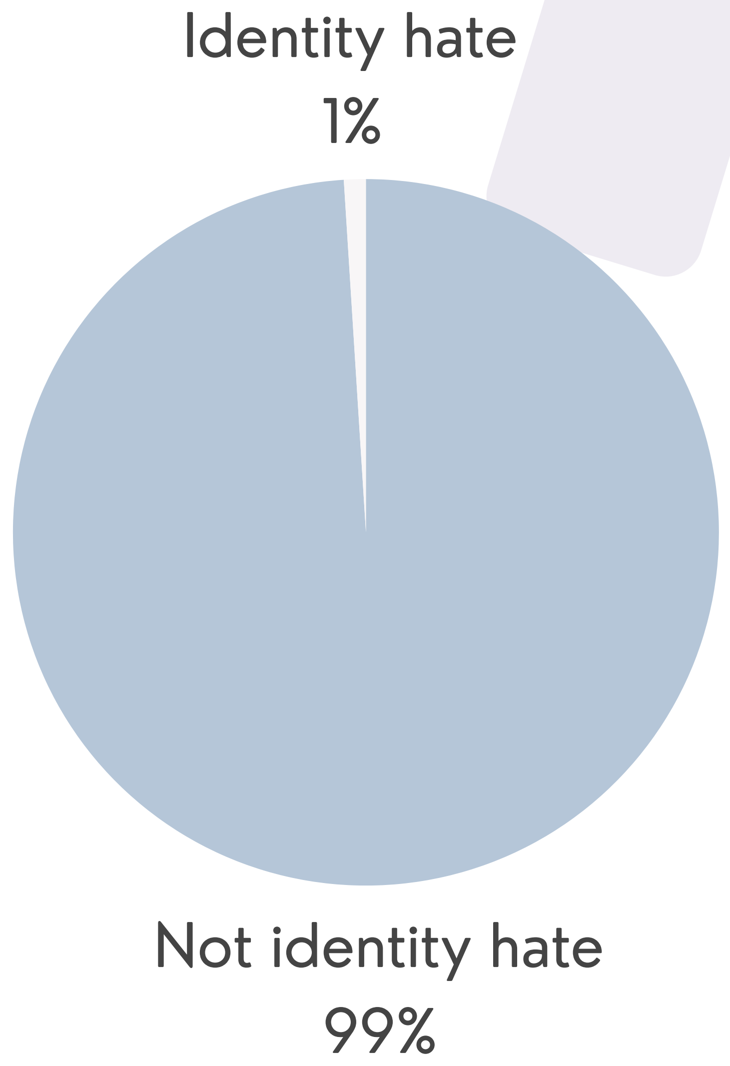




Step 1: Build text classifier

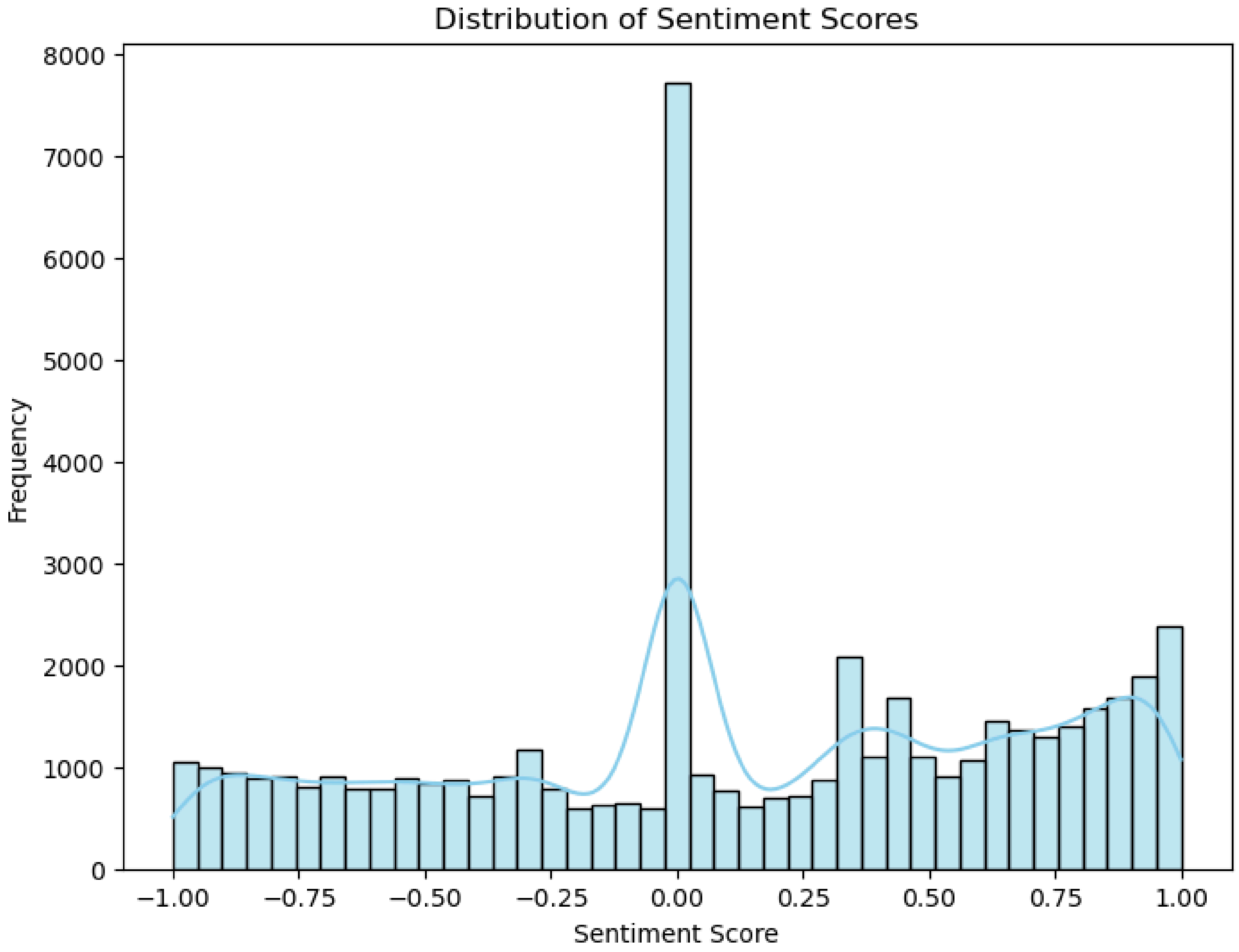
Toxicity type breakdown





50000 comments were randomly chosen for training, while maintaining the proportion of each toxicity type

Pre-Processing & Exploratory Results

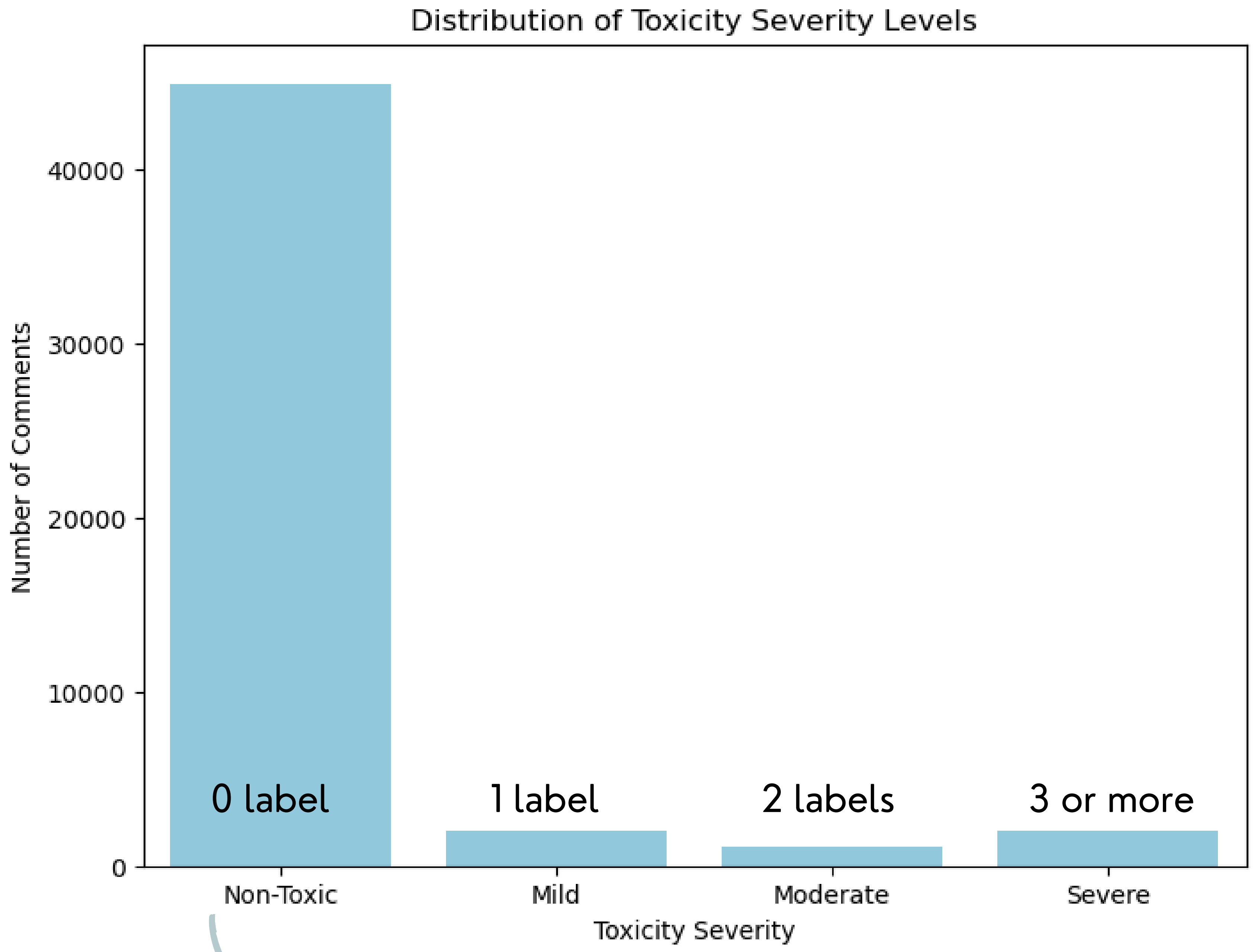


VADER Sentiment Analysis
Overall mean score: 0.119

Step 1: Build text classifier

Pre-Processing & Exploratory Results

Categorised toxicity into 4 bins



Non-toxic comments dominate

Pre-Processing & Exploratory Results

Word Cloud for Mild Comments

Step 1: Build text classifier

Pre-Processing & Exploratory Results

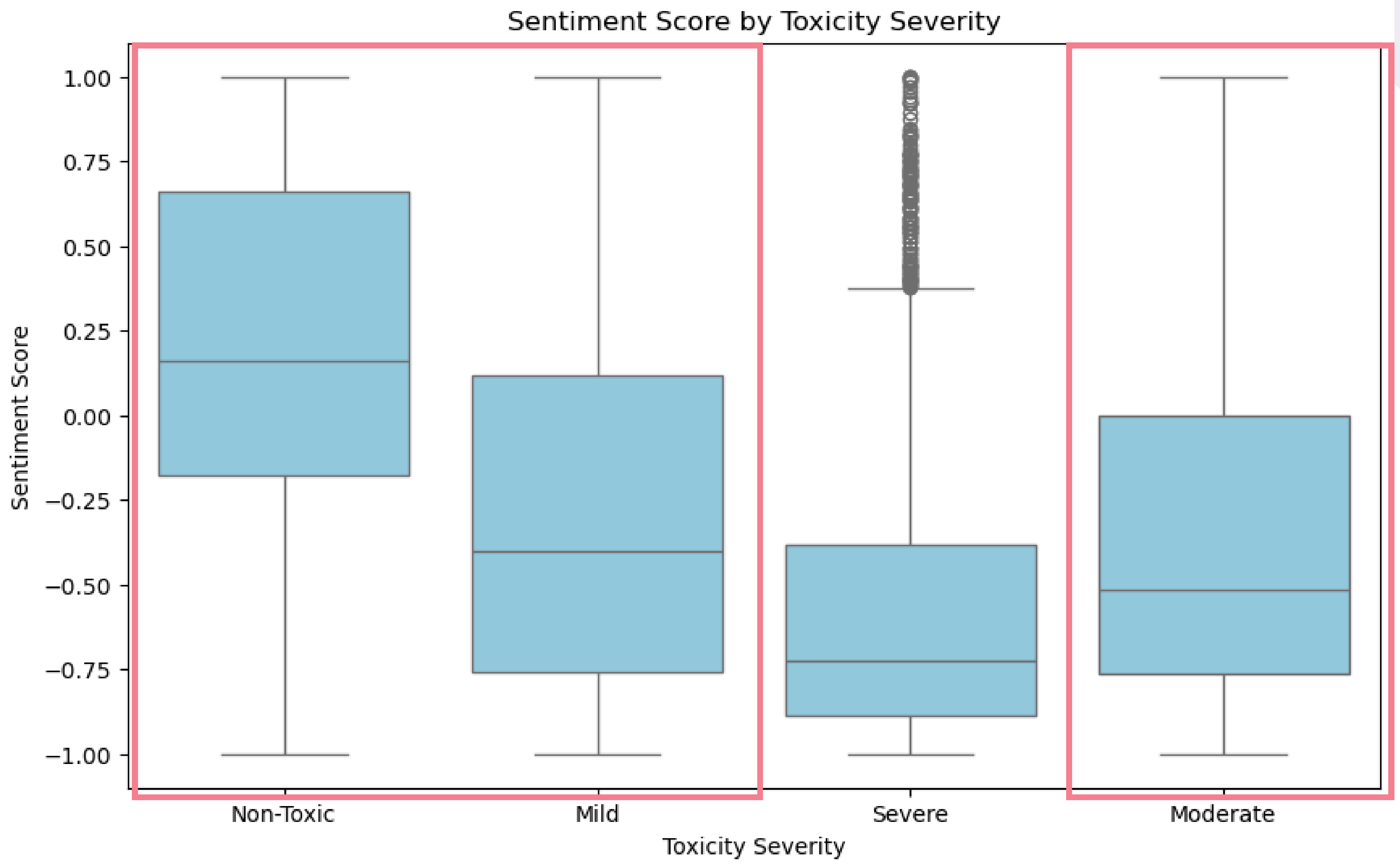
Word Cloud for Moderate Comments

Pre-Processing & Exploratory Results

Word Cloud for Severe Comments

Step 1: Build text classifier

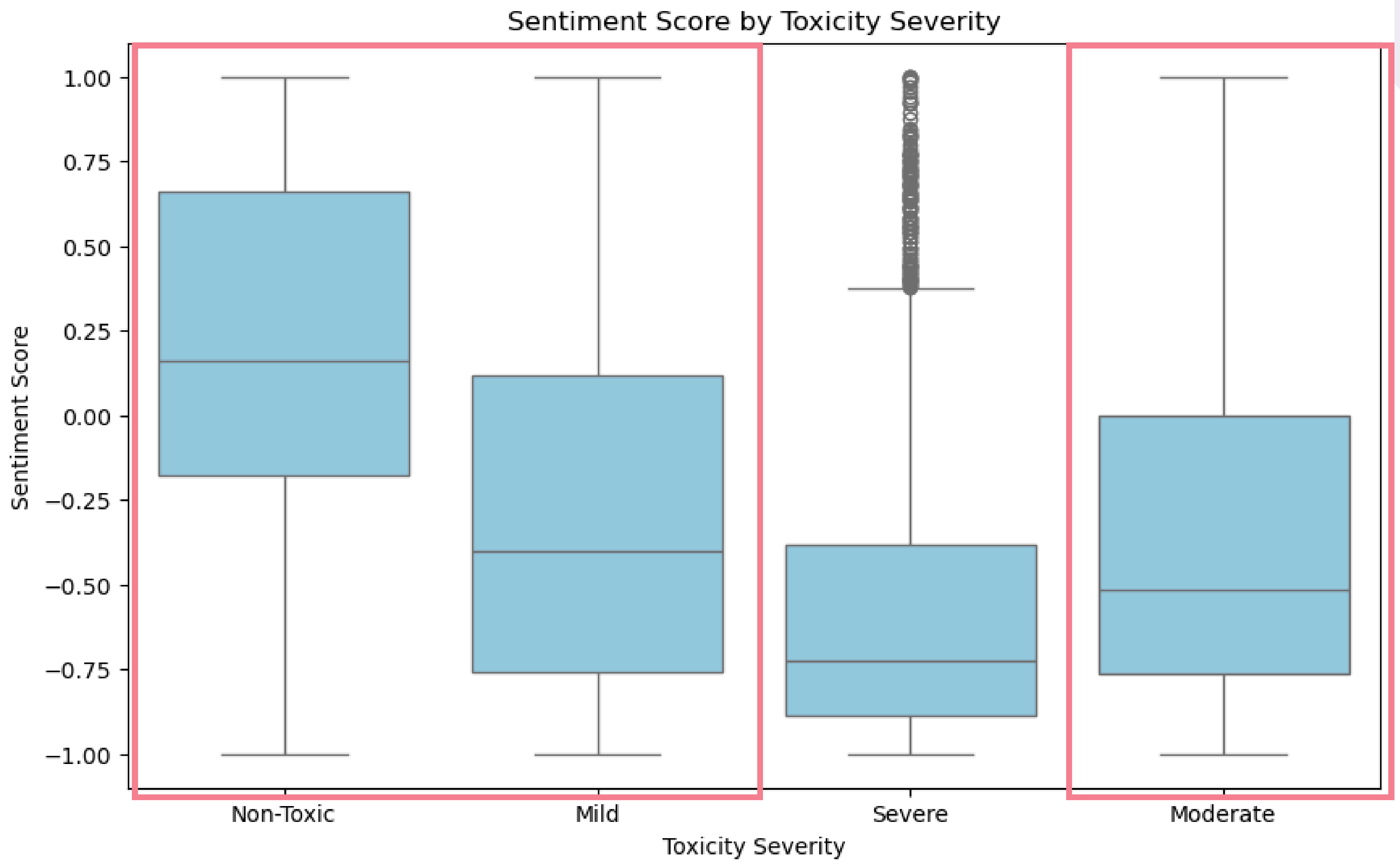
Pre-Processing & Exploratory Results



- Average score decreases as toxicity increases
- Scores are ineffective at differentiating between the nuances of mild vs moderate vs severe toxic comments

Step 1: Build text classifier

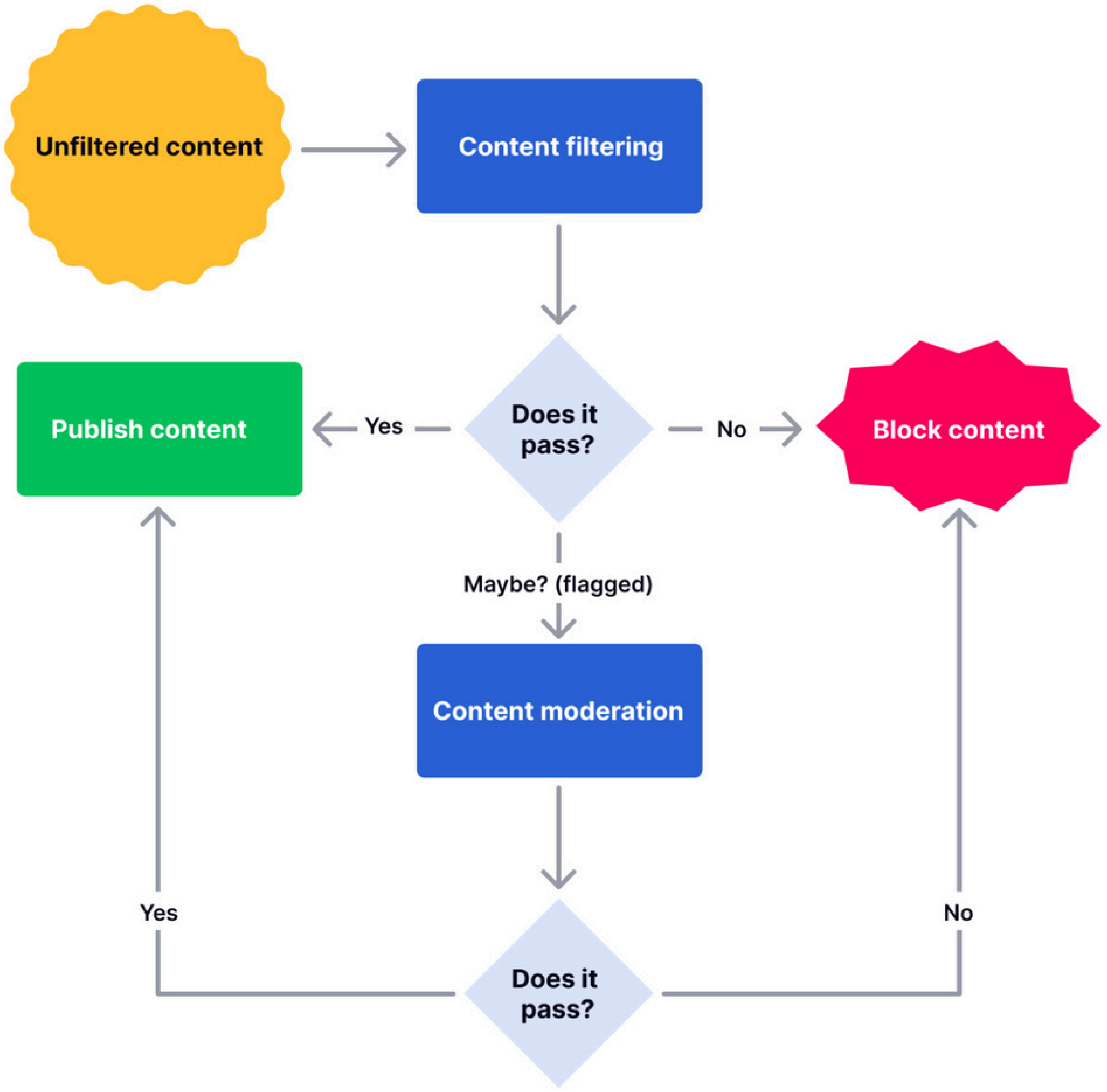
Pre-Processing & Exploratory Results



- Average score decreases as toxicity increases
- Scores are ineffective at differentiating between the nuances of mild vs moderate vs severe toxic comments

Goal: Binary classification of toxic vs not-toxic using TF-IDF & count vectorizer

Content Moderation Flowchart



PRE-PROCESSING TEXT

TOKENIZE

Splits text into individual words for more granular processing, enabling the model to analyze each token's contribution to toxicity

LEMMATIZE

Reduces words to their root form, minimizing redundant variants and improving model consistency.

VADER

Generates sentiment scores to capture emotional tone, helping identify and quantify toxic sentiments.

Step 2: Choose best model

Model Selection

Randomized Grid Search CV & Grid Search CV

Option 1: Multinomial NB

- Straightforward & efficient
- Tuned for class and fit prior -- learn class probabilities from the data, laplace (alpha) which assigns probability to words even if they are absent

Option 2: Gradient Boosting

- Captures complex interactions between words
- Tuned for learning rate, max depth, max features, min sample leaf & split, number of estimators, subsample

Option 3: SVM

- Robust to overfitting
- Tuned for kernel, C -- regularisation parameter that maximises the margin between classes, gamma -- influence of training sample, class weight -- adjusts importance of each class

Option 4: MLPClassifier

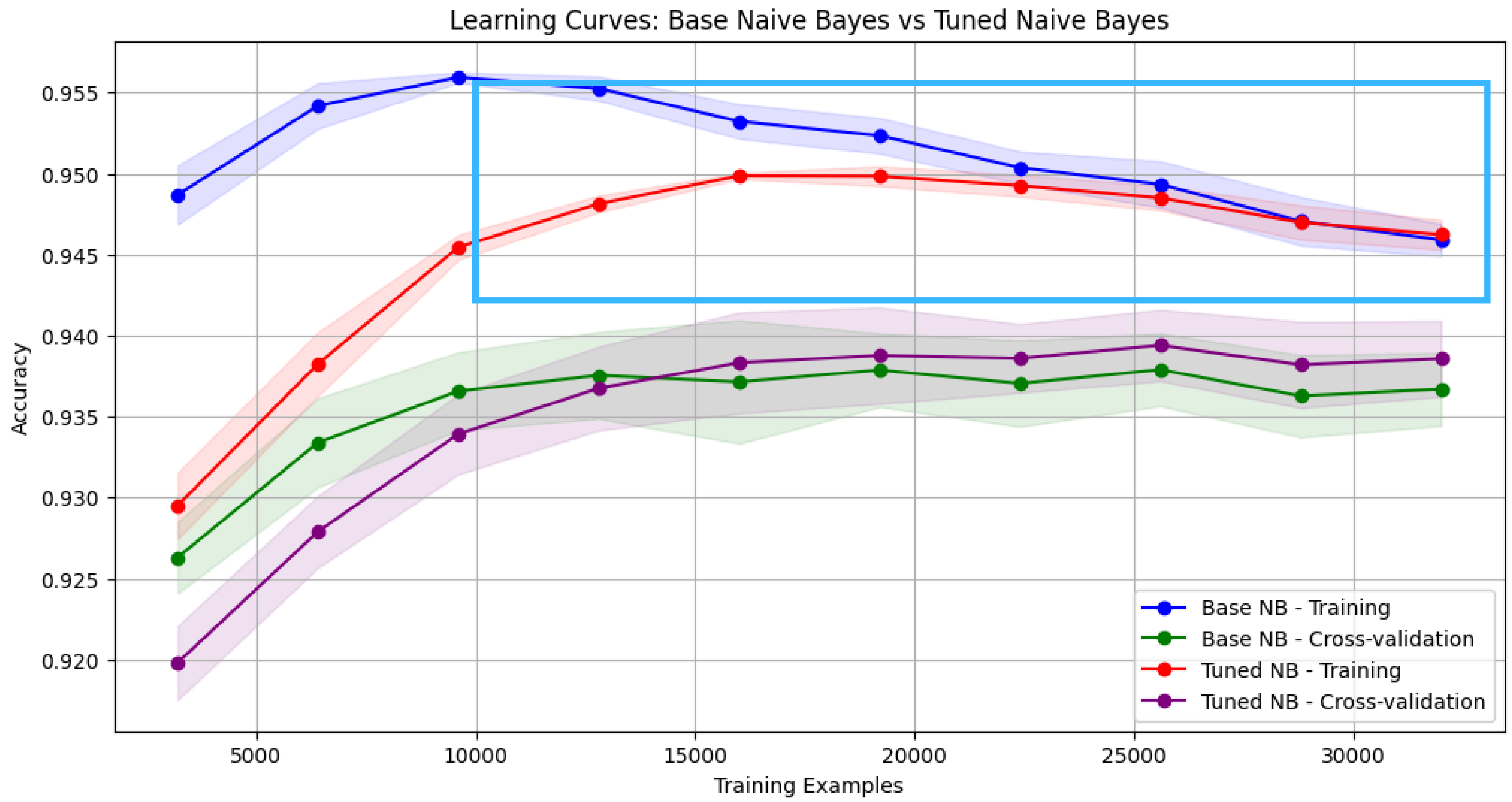
- More effective at learning subtleties in word meanings
- Tuned for learning rate, hidden layer size, alpha, activation function

Step 2: Choose best model

Naïve Bayes

Best parameters for Naïve Bayes:
{'alpha': 2.0, ‘fit_prior': True, 'class_prior': No}

Potential risk: model likely to favour the majority class, may perform unfavourably on unseen data



- Both models dropped in accuracy as training size increased
- Training accuracy of tuned model generally lower but CV accuracy was generally higher
- Gap between training and CV curve for tuned model is smaller than the base model

Step 2: Choose best model

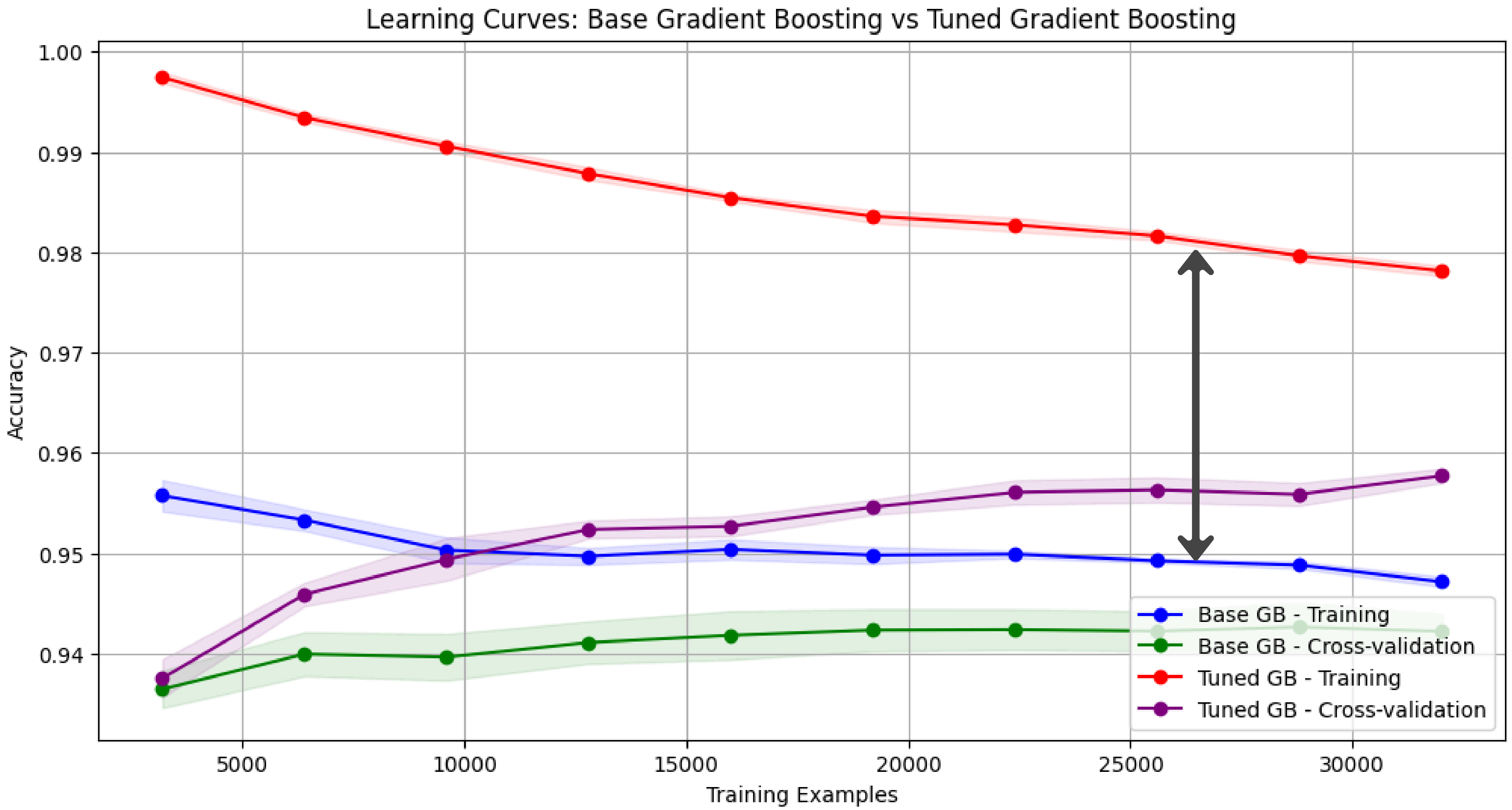
Naïve Bayes

Metric	Total	Class 0	Class 1
Accuracy	0.93		
F1 Score	0.93		
AUC-ROC	0.90		
Precision		0.97	0.66
Recall		0.96	0.69

Step 2: Choose best model

Gradient Boosting

Best parameters for Gradient Boosting Classifier:
{'learning_rate': np.float64(0.16045488840615987), 'max_depth': 7, 'max_features': 'sqrt', 'min_samples_leaf': 8, 'min_samples_split': 15, 'n_estimators': 499, 'subsample': np.float64(0.6203074124157587)}



- Performance of tuned model markedly better than the base model
- CV performance of tuned model diverged from base model as training size increased
- Error rate of tuned model reduced
- Gap between training and CV larger for tuned model

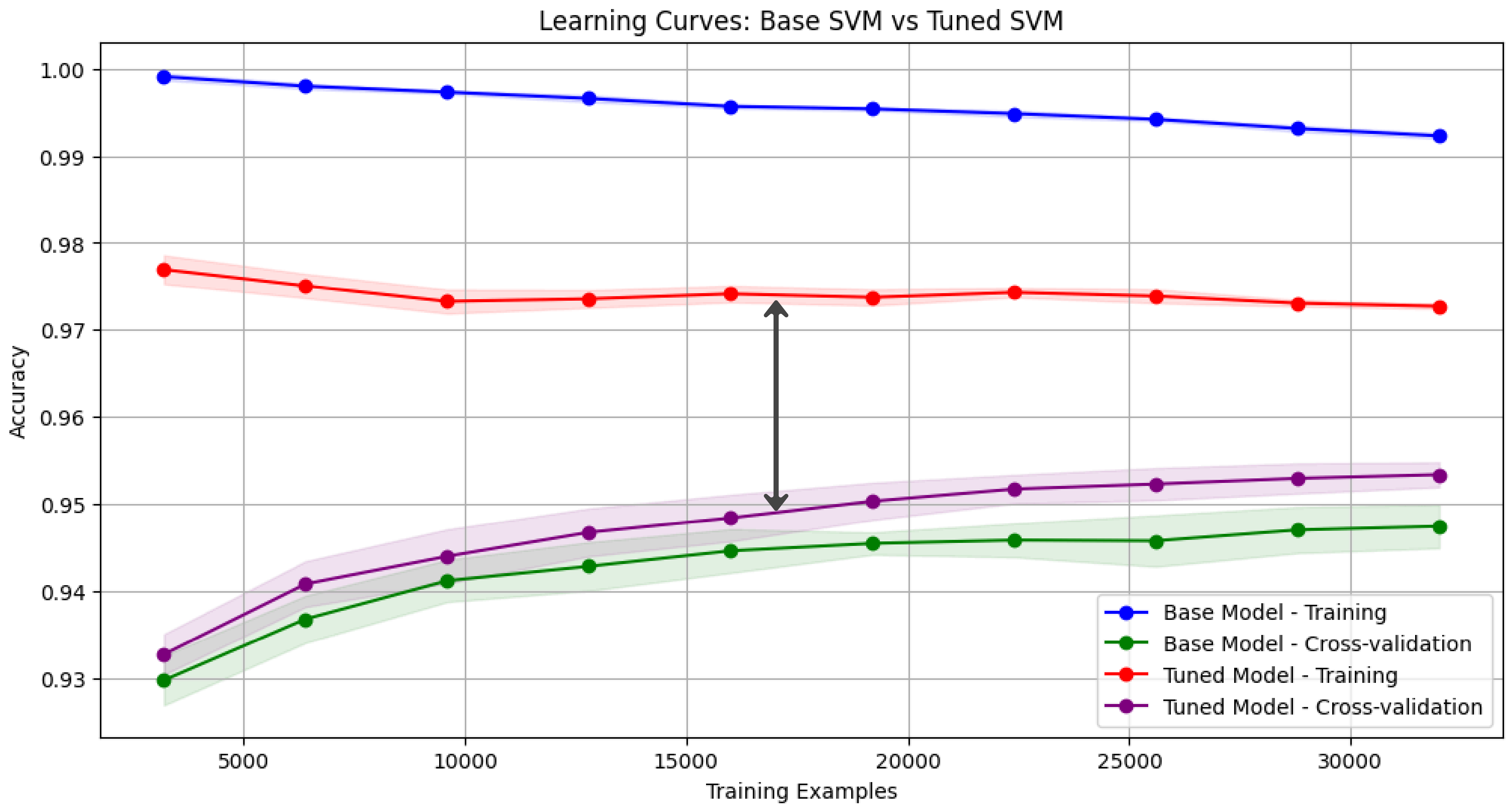
Step 2: Choose best model

Gradient Boosting

Metric	Total	Class 0	Class 1
Accuracy	0.96		
F1 Score	0.95		
AUC-ROC	0.96		
Precision		0.96	0.89
Recall		0.99	0.66

Step 2: Choose best model

Support Vector Machine



Best parameters for Support Vector Machine:
{'C': 0.1, 'kernel': 'rbf', 'gamma': 'scale', 'class_weight': None}

Potential risk: model likely to favour the majority class, may perform unfavourably on unseen data

- Training accuracy of tuned model lower but higher CV accuracy
- Gap between training and CV curve for tuned model is smaller
- Training error for tuned model is smaller

Step 2: Choose best model

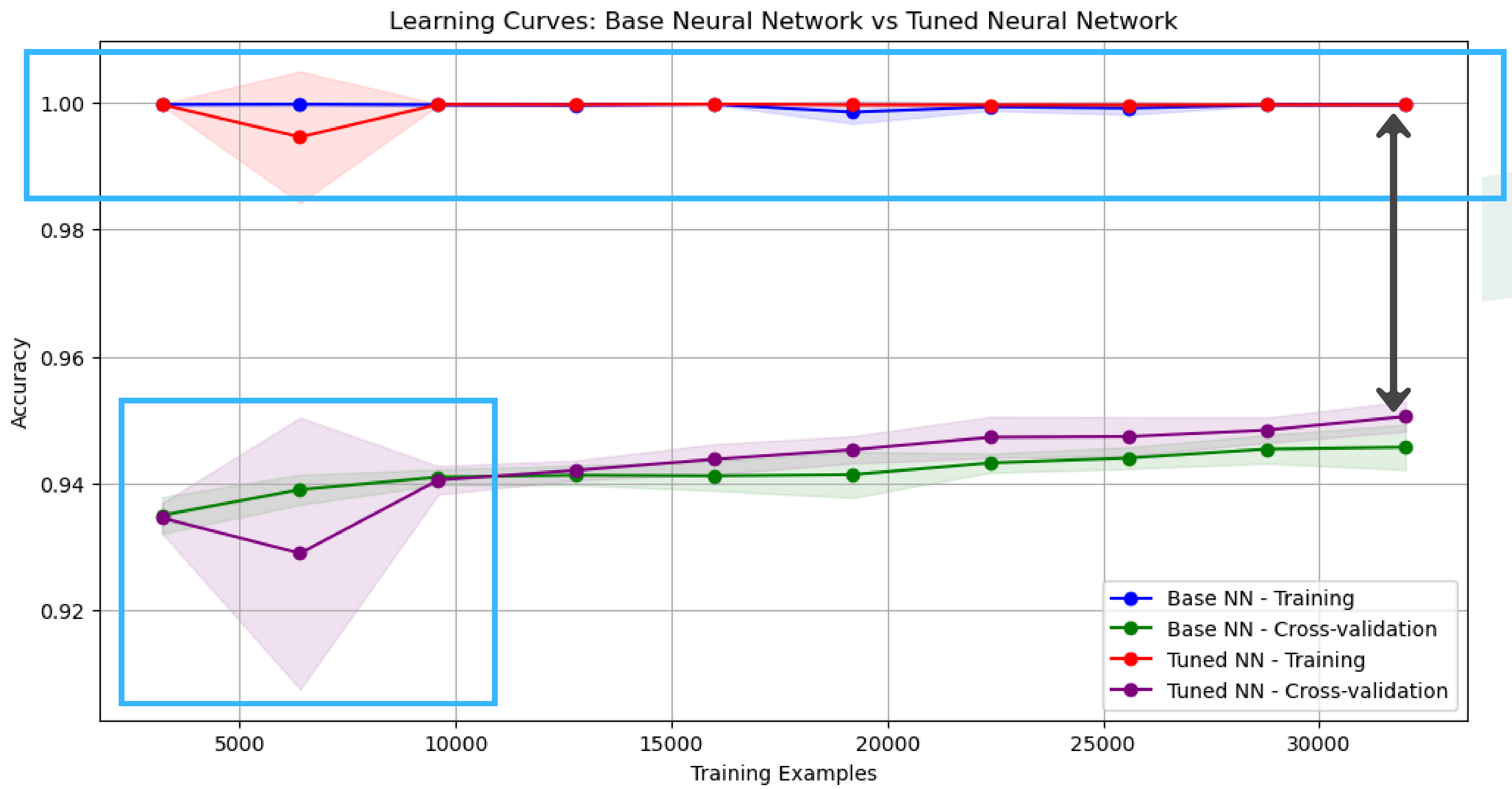
Support Vector Machine

Metric	Total	Class 0	Class 1
Accuracy	0.95		
F1 Score	0.95		
AUC-ROC	0.95		
Precision		0.96	0.88
Recall		0.99	0.63

Step 2: Choose best model

MLP Classifier

Best parameters for Neural Network:
{'learning_rate_init': 0.001, ‘hidden_layer_sizes': (100, 50), 'alpha': 0.001, ‘activation': ‘ReLU'}



- Training set accuracy consistent high irregardless of tuning
- Gap between training and CV curve relatively large
- Training error for tuned model significantly larger with smaller training set

Step 2: Choose best model

MLP Classifier

Metric	Total	Class 0	Class 1
Accuracy	0.95		
F1 Score	0.95		
AUC-ROC	0.94		
Precision		0.97	0.77
Recall		0.98	0.71

Step 2: Choose best model

MLP Classifier

Metric	Total	Class 0	Class 1
Accuracy	0.95		
F1 Score	0.95		
AUC-ROC	0.94		
Precision		0.97	0.77
Recall		0.98	0.71

Preliminary result: MLP Classifier seems to perform best based on performance metrics

EVALUATION ON EXTERNAL DATA: REDDIT

Testing Models on External Reddit Data

Why Kaggle?

Large-scale labeled data is rare.

- Toxicity detection requires extensive labeled data, but such datasets are difficult to find.
- Kaggle provides a 50,000-row labeled dataset, which makes training effective.

But Real-World Data is Messy...

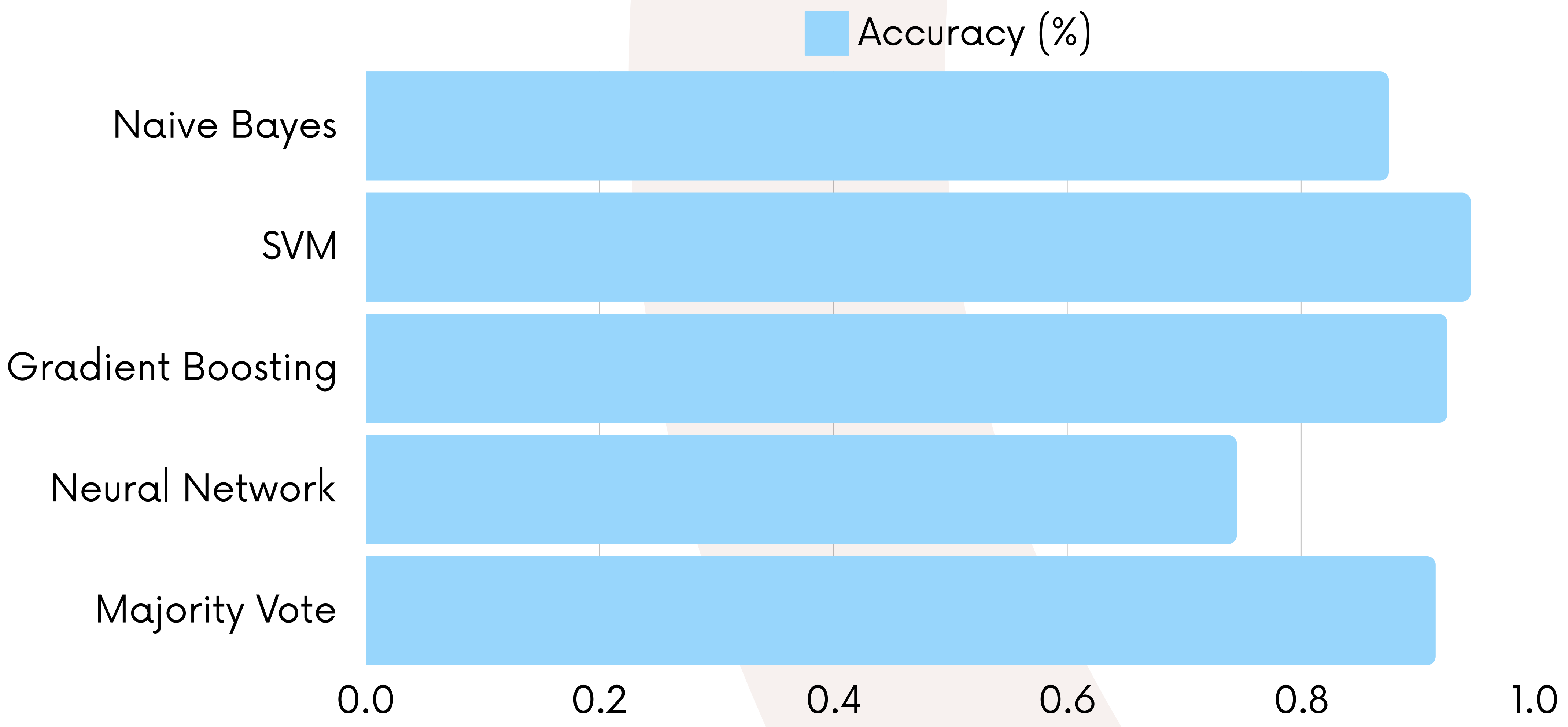
- Social media content is unpredictable—it includes sarcasm, hidden toxicity, evolving slang, and nuanced intent.
- Kaggle data may not fully prepare models for real-world applications.

Testing on External Reddit Data

What We Tested

📌 Dataset: 200 Reddit posts & comments manually labeled for toxicity.
📌 Source: **r/politics** – A controversial subreddit with high engagement & diverse opinions.
📌 Labeling Approach:

- Each post/comment was labeled toxic (1) or non-toxic (0) using GPT-4 and manual verification.
- This allows us to evaluate model performance on real-world content.



Step 3: Validate with External Data

Model Performance

How Well Do These Models Detect Toxic Content?

- ✅ Minimize false positives (not block safe content).
- ✅ Minimize false negatives (catch all toxic content).
- ✅ Balance precision & recall to ensure effectiveness.

Metric	Naive Bayes	SVM	Gradient Boosting	Neural Network	Majority Vote
F1 Score	0.39	**0.72**	0.55	**0.16**	0.45
AUC-ROC	0.64	**0.79**	0.69	**0.51**	0.65
Precision	0.47	0.93	**1.00**	**0.14**	**1.00**
Recall	0.33	**0.58**	0.38	**0.21**	0.29

SVM performs best overall → Balanced precision, recall, and F1 score
Gradient Boosting & Majority Vote have high precision but low recall
Neural Network fails on this dataset → Weak recall and poor precision

NEXT STEPS

The model is already fine-tuned, but we can improve it further..

How Do We Move to a Chrome Extension?

🚀 **Steps to Deployment**

1️⃣ Expand Training Data → Use other labeled datasets to improve detection across different types of websites.

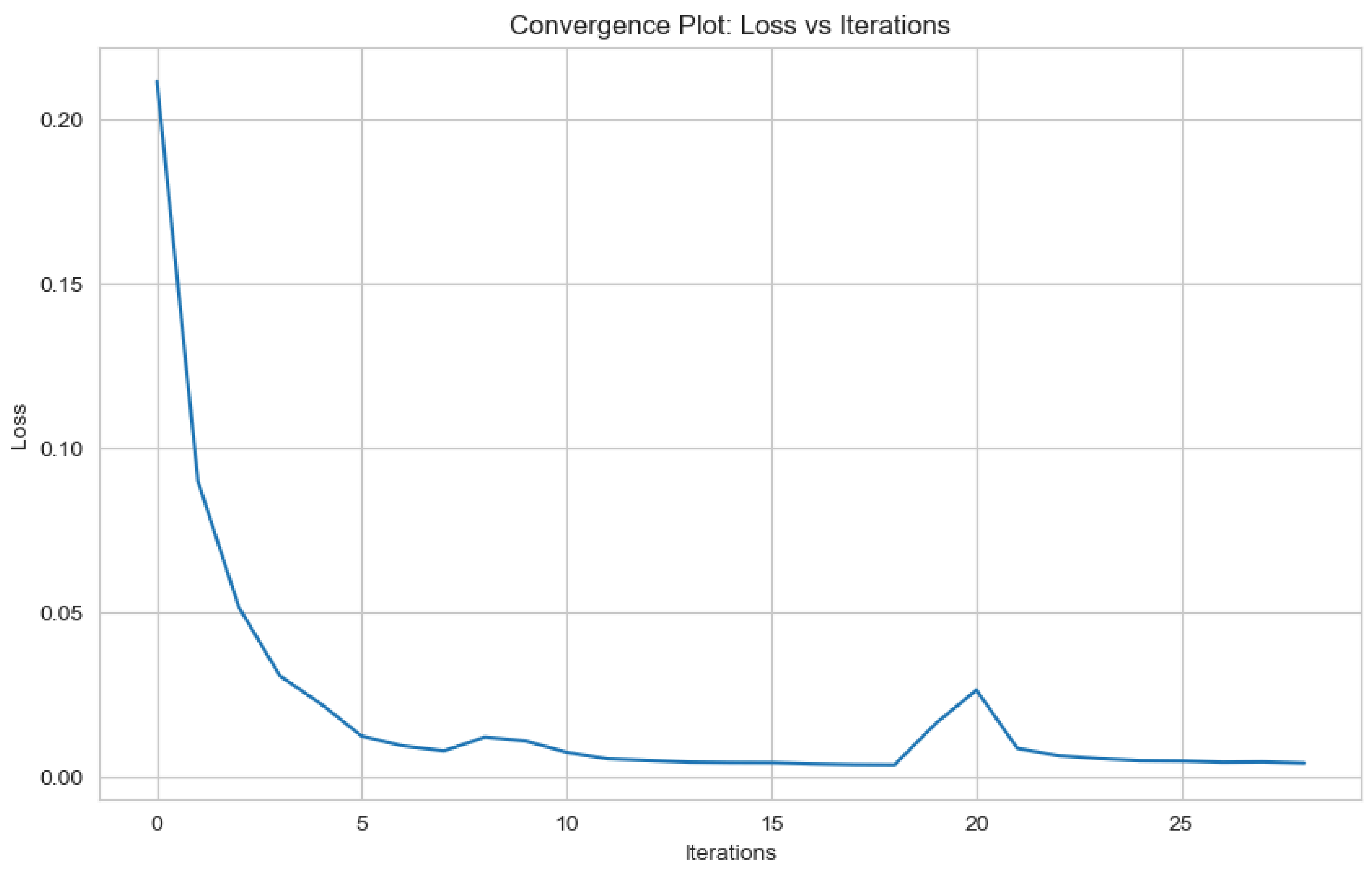
2️⃣ Develop a Chrome Extension → Integrate the trained model into a real-time browser plugin.

3️⃣ Enable Parent Feedback for Adaptive Learning →

- Parents review flagged content (approve/reject).
- Monitor False Positives & False Negatives → The model learns from mistakes over time.

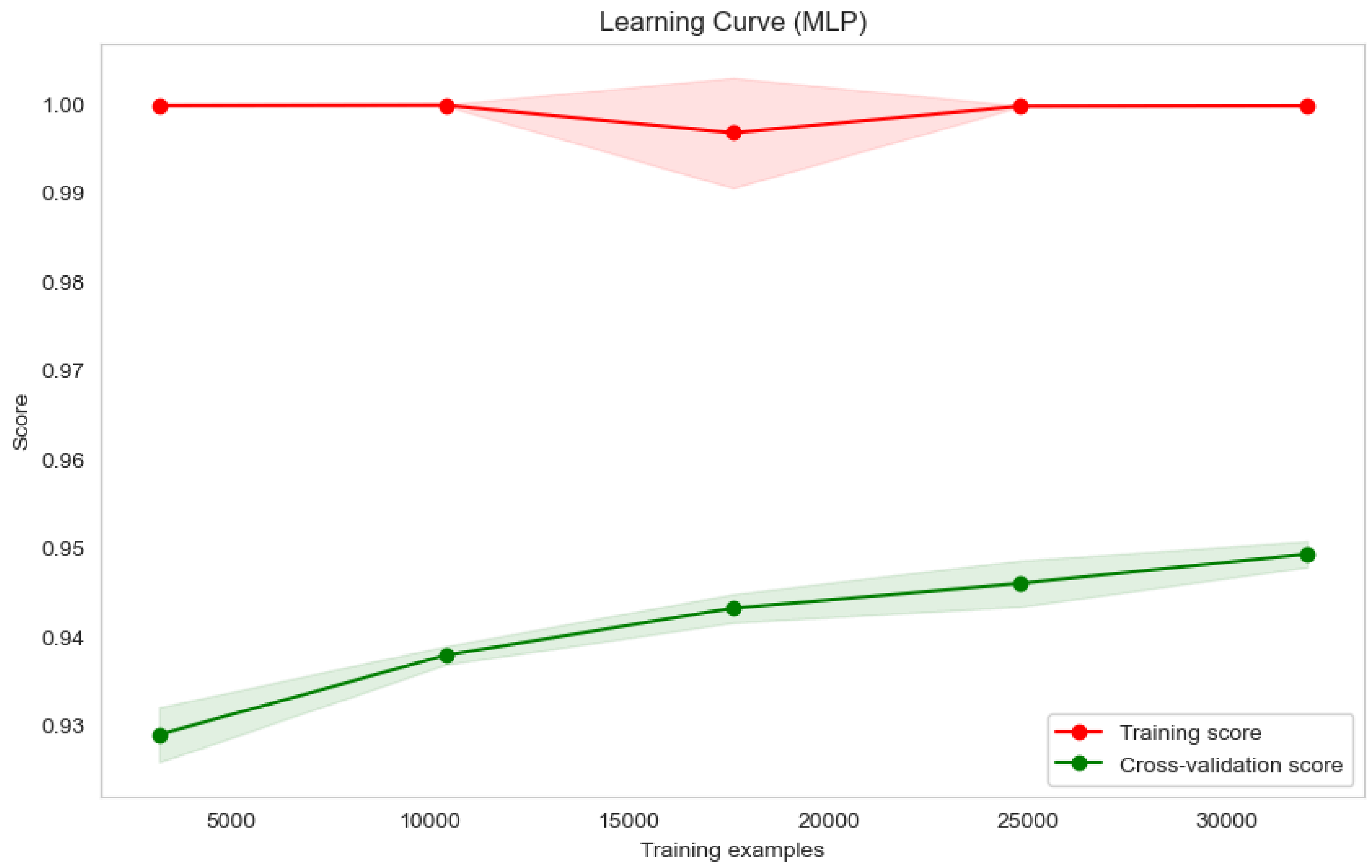
TH@#K YOU

Neural Network

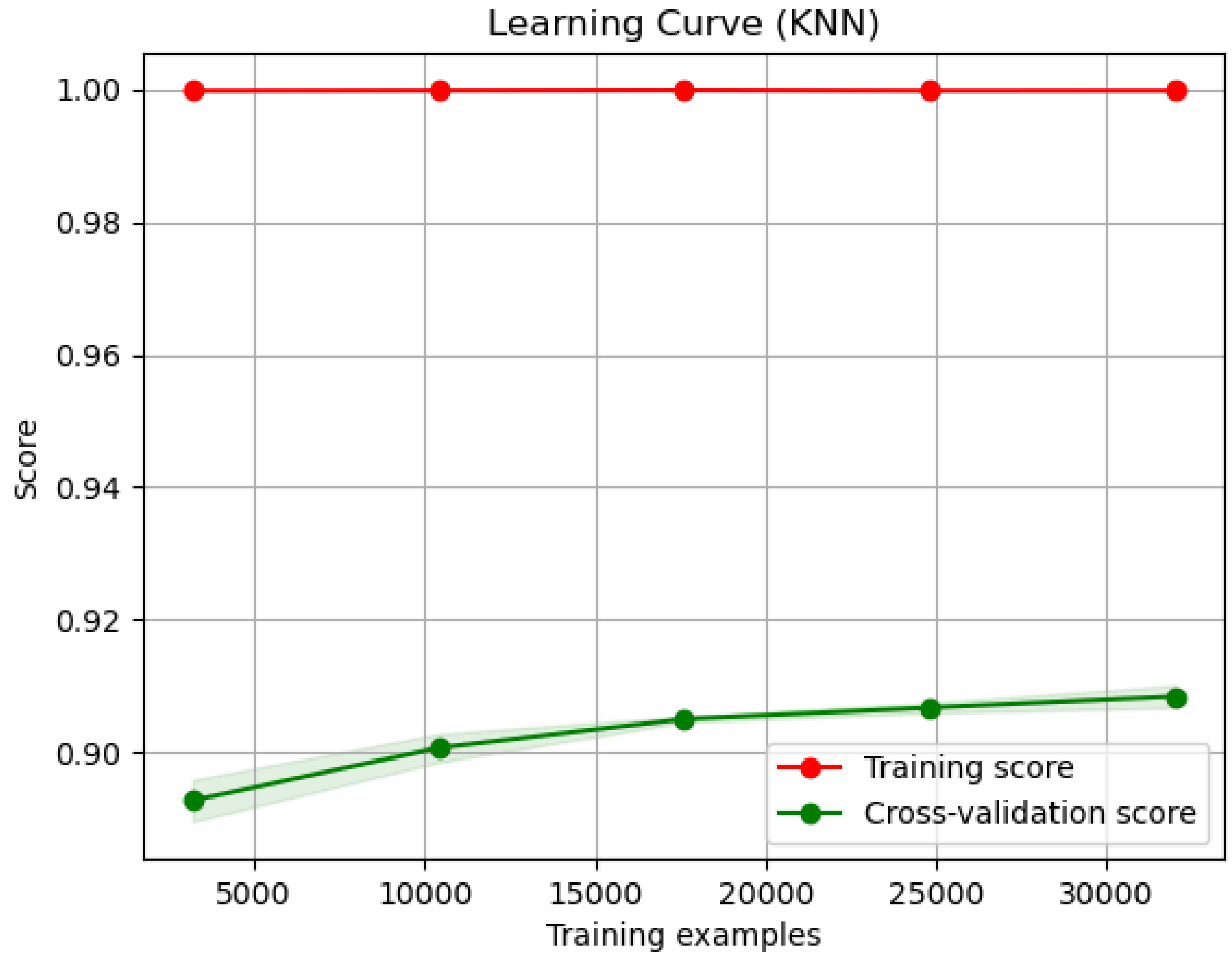


best Hypermarameters:

- learning_rate_init: 0.001
- hidden_layer_sizes: (100, 50)
- alpha: 0.001
- activation: ReLU



K-Nearest Neighbor



Best parameters:

weights: distance, n_neighbors: 7, metric: cosine, algorithm: auto