CYBER BULLYING DETECTION

USING MACHINE LEARNING ALGORITHMS

SRINIVASA REDDY DHONTHIREDDY (U00956207)

NANDITHA MIRIYAM (U00970508)

SAKETH DONTHA (U00956936)

SAI RAM PADALA (U00933894)

ABSTRACT

Cyberbullying in the digital era pervades all social media platforms, affecting the mental health and safety of individuals. This project introduces machine learning to identify and classify cyberbullying from social media text data. A six-category dataset of bullying instances (age, gender, religion, ethnicity, not cyberbullying, other cyberbullying) was preprocessed and analyzed. Four classification models such as Random Forest, XGBoost, Gradient Boosting, and ANN were designed, tested, and optimized. The best results were produced by the optimized XGBoost model with 81% accuracy and a macro F1-score of 0.81. The algorithm identifies clear and subtle cyberbullying with success, showcasing the promise of machine learning to create real-time, scalable solutions to online safety.

Keywords: cyberbullying, random forest, XGboost, Gradient Boosting, ANN, etc.

1. INTRODUCTION

In an increasingly digital world, social media and other online communication channels have become central to communication, particularly among younger generations. Anonymity and breadth have also made social media platforms hotbeds for cyberbullying—the harassment, threat, or humiliation of an individual using electronic means [1]. Detection of such harmful acts in real-time is crucial to prevent psychological harm, ensure online security, and allow for timely intervention.

The purpose of this project is to design an automated system for detecting cyberbullying through the application of machine learning (ML) methods to social media-based text data. The aim is to classify if the inputted text has cyberbullying content or not, and if it does, to classify the type such as age-based, gender-based, religion-based, etc.

1.1. Challenges

The main challenges in this application are:

- **Textual Variability:** The language used in cyberbullying differs greatly in style, slang, and subtext, which makes it hard to find with rule-based or keyword methods [2].
- **Multi-class Classification:** The dataset has multiple kinds of bullying, making it necessary for multi-class classification instead of binary classification.
- **Class Imbalance:** Though the dataset here is fairly balanced, classifying between similar classes (e.g., ‘not_cyberbullying’ vs ‘other_cyberbullying’) is not easy [3].
- **Noisy Data:** Tweets include hashtags, emojis, mentions, and URLs that aren't necessarily tied to bullying intent but influence model performance if not effectively preprocessed.

2. SOLUTION

For solving the presented challenges, we designed an end-to-end machine learning pipeline including data cleansing, feature extraction, model training, hyperparameter optimization, and performance assessment.

2.1. Preprocessing Texts and Selecting Features

We're assuming that contextual textual content (not metadata) is reliable enough to spot cyberbullying. The raw dataset text (tweet_text) was cleaned utilizing Natural Language Toolkit (NLTK) tools:

- Lowercasing
- Removal of URLs, mentions, hashtags, & punctuation

- Remove stopwords
- Tokenisation and lemmatisation (where relevant)

The cleaned data (clean_text) was further converted to numerical features by applying CountVectorizer with an upper limit on 500 features (bigrams and unigrams) to extract word frequencies indicating bullying patterns.

Label encoding

Target variable cyberbullying_type, with six class values, was converted to integers using LabelEncoder for enabling multi-class classification.

Handling Data Skewness or Inconsistencies

No missing data were detected in the dataset, and all classes were distributed very evenly. Therefore, neither SMOTE nor undersampling was needed.

2.2. Machine Learning Models and evaluation

Four models were utilized, all with their own distinguishing strengths in text classification:

a) Random Forest Classifier

Random Forest is an ensemble learning approach that constructs multiple decision trees and aggregates their outputs for better accuracy and less overfitting [4]. It

- Deals with noisy data and non-linear relationships effectively
- Resistant to overfitting by virtue of bootstrapped sampling
- Was tuned with parameters such as n_estimators, max_depth, and min_samples_leaf

b) XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is an efficient high-performance implementation of gradient boosting decision trees [5]. It:

- Utilizes a boosting approach in which new models adjust for errors made by previous models
- It is renowned for speed and accuracy.
- Was tuned with learning_rate, max_depth, subsample, colsample_bytree

c) Gradient Boosting Classifier

Gradient Boosting constructs an ensemble of weak learners (decision trees) sequentially that try to minimize their predecessor's error [6].

- Focuses on hard-to-predict examples
- Provides excellent control over bias-variance tradeoff
- Was tuned like XGBoost but with slower learning rate and smaller number of trees to prevent overfitting

d) Artificial Neural Network (ANN)

ANNs were modeled on the human brain and prove to be very efficient in learning sophisticated patterns in data [7]. Our architecture consisted of:

- Input layer consisting of 512 neurons
- Two hidden layers with 256 neurons and 128 neurons employing ReLU activation
- Dropout layers (0.4 and 0.3) for

- Output layer using softmax activation for multiclass prediction
- Categorical cross-entropy as loss function and Adam optimizer

2.3. Model tuning

Each was trained initially on default parameters but subsequently fine-tuned through hyperparameter optimization to perform better. Some examples include:

- For Random Forest, we tune the parameters n_estimators, max_depth, min_samples_split,
- **For XGBoost:** tuning with respect to n_estimators, max_depth, learning_rate,
- **For Gradient Boosting:** adjusting n_estimators, learning_rate, max_depth, min_samples_split, and min_samples_leaf
- **For ANN:** modifying layer numbers and units, Dropout rates, number of epochs, batch size and number of epochs

2.4. Evaluation Metrics

For measuring performance on models, we employed:

- Accuracy – for correctness in general
- Precision, Recall, F1 Score – to measure effectiveness for all classes
- Confusion Matrix – to visualize prediction distribution and misclassification

The models were tested on an 80-20 stratified train-test split to have samples from all classes represented. ANN was employed with one-hot label encoding and softmax output as well as categorical cross-entropy loss.

3. EMPIRICAL EXPERIMENTS

The success of the developed system of cyberbullying detection could be measured by looking at the class distribution and also the accuracy of the used machine learning models across the classes. The data set had a relatively equitable spread over six categories in the range of about 7,800 to 8,000 samples per class

- **Religion-based bullying**: 7,998 samples
- **Age-related bullying**: 7,992 samples
- **Gender-based bullying**: 7,973 samples
- **Ethnicity-related bullying**: 7,961 samples
- **Not cyberbullying**: 7,945 samples
- **Other forms of bullying**: 7,823 samples

3.1. Visual analysis

Figure 1: Wordcloud representation of bullying tweets

- **Not Cyberbullying**

Common Words: "rt", "school", "love", "people", "thing", "day", "good", "thank"

They primarily consist of positive or neutral remarks. "Love", "school", and "thank" imply general conversation or individual expression without offensive or malicious intention

- **Religious cyberbullying**

Common Words: "muslim", "idiot", "terrorist", "islam", "jihad", "slam", "christian", "radical"

Tweets within this class involve hate speech and derogatory language aimed at religious identities, particularly Muslims. "Terrorist" and "radical" are inappropriately used, indicating religious stereotyping and abuse.

- **Gender-based cyberbullying**

Common Words: "gay", "rape", "bitch", "women", "joke", "female", "think", "sexist"

This indicates offensive speech aimed at someone based on gender or sexual orientation. The use of the word "rape," "bitch," and the term "sexist" indicates aggressive or insulting language towards women and LGBTQ+ individuals.

- **Age-based cyberbullying**

Common Words: "high school", "kid", "bullied", "middle school", "

This subcategory revolves around age-based insults or bullying, primarily of school-age people. "High school" and "bullied" references imply peer mockery or bullying of younger age classes.

- **Other cyberbullying**

Common Words: "rt", "thing", "bully", "idiot", "make", "want", "dont", "twitter", "fucking"

This encompasses general bullying not falling under defined categories. It involves offensive language ("fucking", "idiot") and bullying-related vocabulary and also expresses hostility in interactions.

- **Ethnicity based cyber bullying**

Common Words: "nigger", "dumb", "fuck", "obama", "ass", "racism"

This class includes some of the most overt and extreme hate speech. Racial slurs and insults figure high, reflective of targeted harassment on the basis of ethnicity or race.

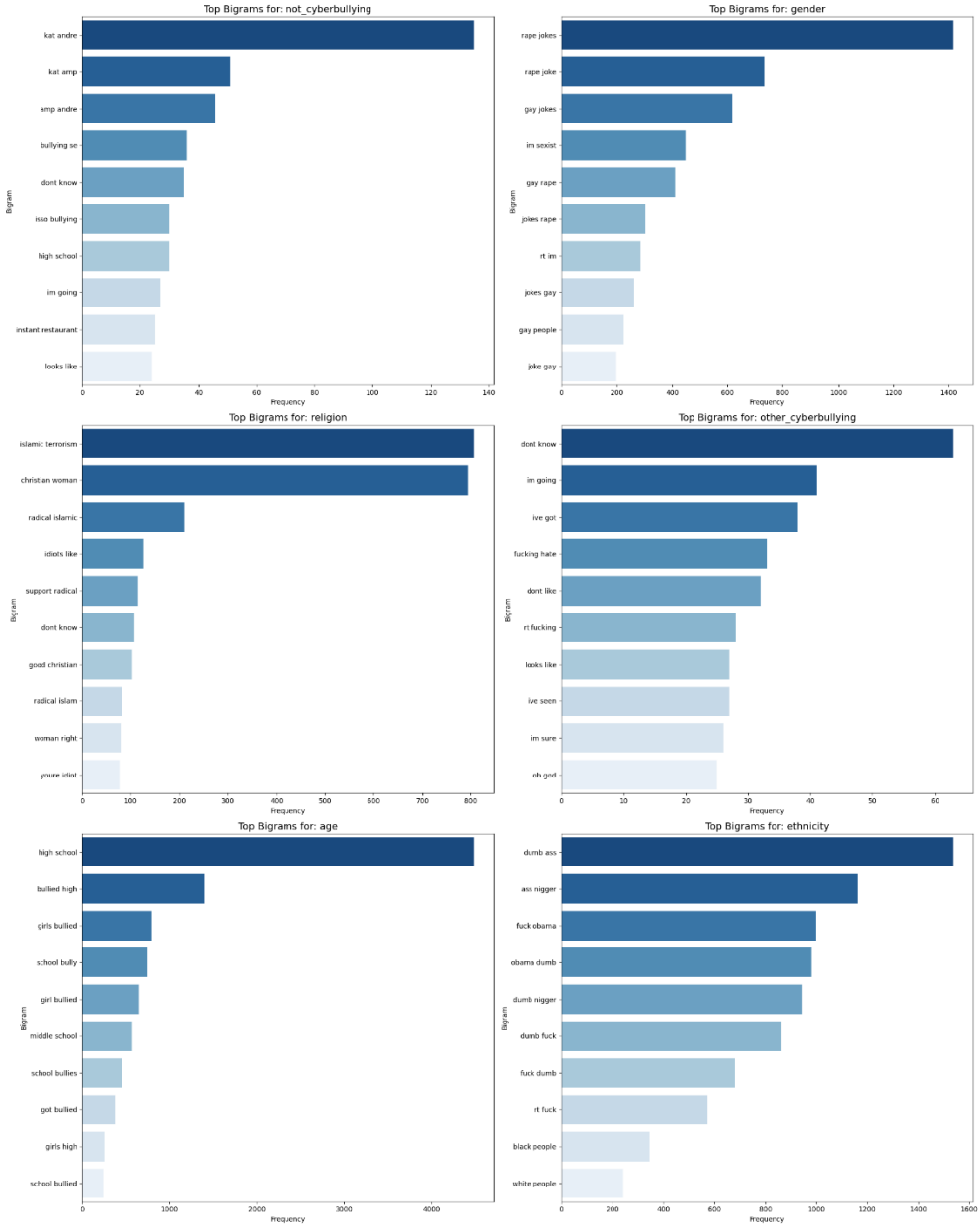


Figure 2: Bigram representation of cyberbullying tweets

- **Not Cyberbullying**

These statements seem unrelated to bullying and probably arose from innocuous or pop culture discourse. This indicates the model's capability to spot non-bullying content.

- **Gender-based cyberbullying**

They encompass rape-related material and homophobic insults, justifying that such a class should be strictly monitored and classified as extremely sensitive.

- **Religious cyberbullying**

The bigrams reveal targeted attacks on religious communities, particularly Muslims and Christians, in language such as "terrorism" and "radical," usually in a derogatory or accusing manner.

- **Other cyberbullying**

These words and phrases are less strongly aimed but still convey aggressive tone or general bullying, in particular "fucking hate". This class has mixed, less organized abuse.

- **Age-based cyberbullying**

The terminology used exhibits direct targeting of school-aged individuals, primarily high school and middle school students. The use of such terms as "bullied" and "school bully" points to traditional examples of age-based cyber harassment.

- **Ethnic-based cyberbullying**

This class is filled to the brim with blatant racial slurs and hate speech, with expressions of extreme racism and ethnic insults. Phrases such as "fuck obama" and the routine use of the word "N-word" suggest targeted racist attacks.

3.2. Model performance

a) Random Forest (Default)

In this case, all the models are trained with default parameters and evaluated using confusion matrix and other metrics like precision, recall and F1 score.

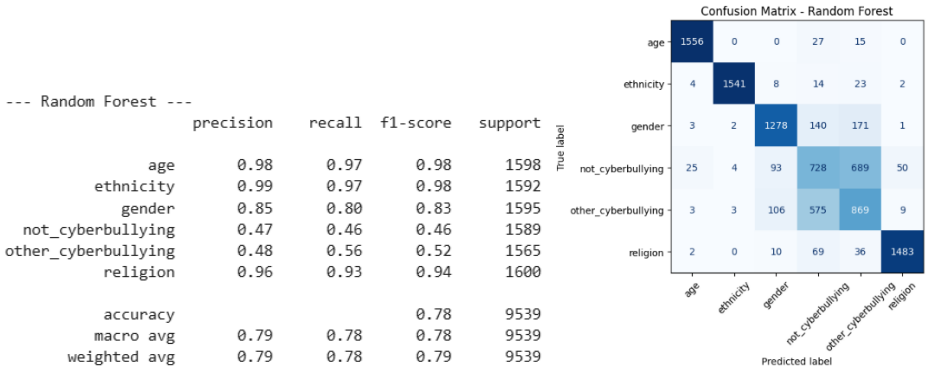
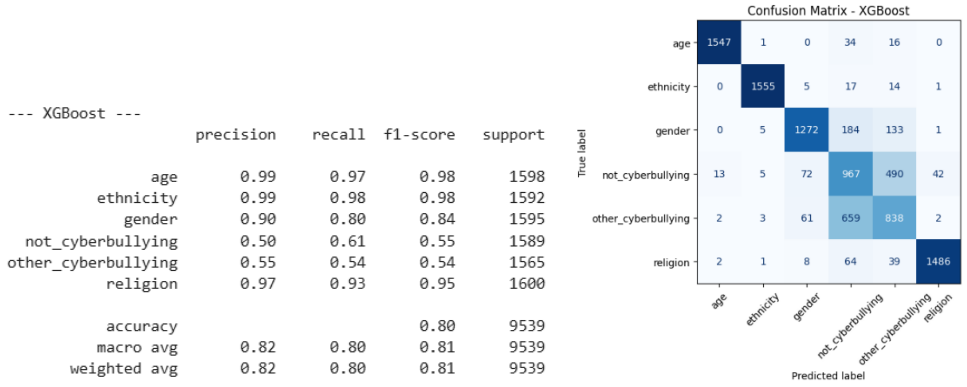


Figure 3: Default random forest performance

The Random Forest classifier had a total accuracy of 78% and high F1-scores in age with a score of 0.98, ethnicity with a score of 0.98, and religion with a score of 0.94 but lower for not cyberbullying with an F1 of 0.46 and other cyberbullying with an F1 of 0.52 as it tended to misclassify them repeatedly. Although it has a macro average F1-score of 0.78, it performs poorly in subtle content and evidently requires more context-sensitive methods.

b) XGBoost (Default)



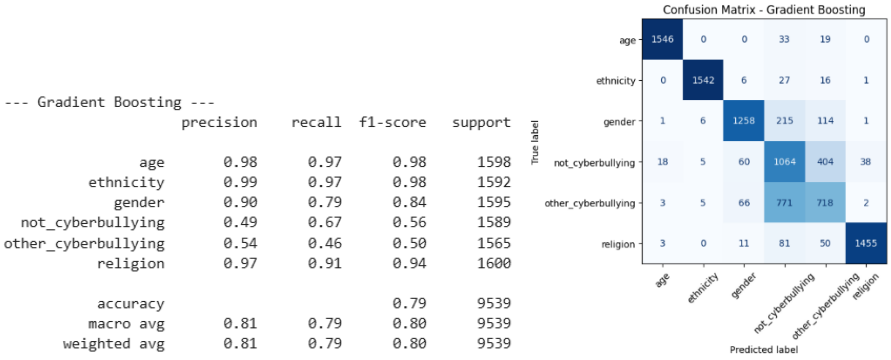
--- XGBoost ---

	precision	recall	f1-score	support
age	0.99	0.97	0.98	1598
ethnicity	0.99	0.98	0.98	1592
gender	0.90	0.80	0.84	1595
not_cyberbullying	0.50	0.61	0.55	1589
other_cyberbullying	0.55	0.54	0.54	1565
religion	0.97	0.93	0.95	1600
accuracy			0.80	9539
macro avg	0.82	0.80	0.81	9539
weighted avg	0.82	0.80	0.81	9539

Figure 4: Default XGBoost performance

The XGBoost model has a total accuracy of 80% and high F1-scores for age with an F1 of 0.98, ethnicity with an F1 of 0.98, and religion with an F1 of 0.95. The XGBoost outperformed Random Forest in not cyberbullying with an F1 of 0.55 and other cyberbullying with an F1 of 0.54, although there is confusion between the two classes. The gender class was also at F1 = 0.84, indicating moderate confusion with related classes. XGBoost emerges as a better and more balanced model with a macro average F1 of 0.81, particularly in the classification of class overlaps.

c) Gradient Boosting (Default)



--- Gradient Boosting ---

	precision	recall	f1-score	support
age	0.98	0.97	0.98	1598
ethnicity	0.99	0.97	0.98	1592
gender	0.90	0.79	0.84	1595
not_cyberbullying	0.49	0.67	0.56	1589
other_cyberbullying	0.54	0.46	0.50	1565
religion	0.97	0.91	0.94	1600
accuracy			0.79	9539
macro avg	0.81	0.79	0.80	9539
weighted avg	0.81	0.79	0.80	9539

Figure 5: Default Gradient Boosting performance

Gradient Boosting achieved a general accuracy of 79% with high F1-scores of age with an F1 of 0.98, ethnicity with an F1 of 0.98, and religion with an F1 of 0.94. It distinguished better between not cyberbullying with an F1 of 0.56, with improved recall compared to previous models. Yet, other cyberbullying continued to be challenging, with a decreased F1-score of 0.50 due to the overlap with related classes such as gender and not cyberbullying. The gender class performed an F1-score of 0.84, with little misclassifications. With a macro F1-

score of 0.80, the model performs steadily and accurately for sharply defined categories but needs improvement in subtle cases.

d) ANN (Default)

```
--- ANN ---
                    precision    recall  f1-score   support

               age       0.98      0.97      0.98      1598
         ethnicity       0.99      0.97      0.98      1592
            gender       0.85      0.82      0.83      1595
  not_cyberbullying       0.49      0.63      0.55      1589
other_cyberbullying       0.55      0.44      0.49      1565
          religion       0.95      0.93      0.94      1600

          accuracy                           0.80      9539
         macro avg       0.80      0.79      0.80      9539
      weighted avg       0.80      0.80      0.80      9539
```

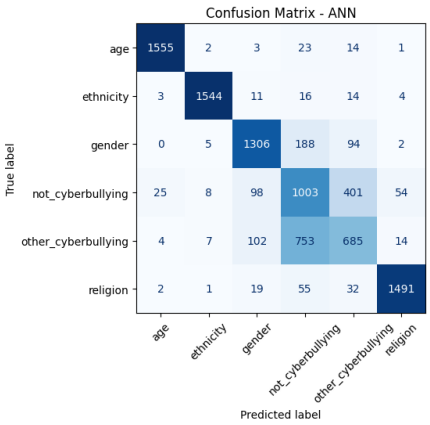


Figure 6: Default ANN performance

The ANN model demonstrated a general accuracy of 80%, with high F1-scores for age with F1 of 0.98, ethnicity with a score of 0.98, and religion with a score of 0.94, indicating stable performance in flagging well-delineated cyberbullying. The gender class offered an F1-score of 0.83, with some confusion with other classes. For not cyberbullying, the model performed fairly with F1 score of 0.55, indicating better recall (0.63), but continued to confuse some with offensive classes. Other cyberbullying presented the lowest F1-score (0.49), indicating difficulty in separating it from overlapping classes. With a macro average F1-score of 0.80, ANN provides balanced measurements but could be helped by more in-depth context modeling to enhance ambiguous class differentiation.

3.3. Model tuning

The following parameters were used for tuning the model in order to improve performance.

Model	**Tuned Parameters**
Random Forest	n_estimators=200, max_depth=20, min_samples_split=5, min_samples_leaf=2, n_jobs=-1, random_state=42
XGBoost	n_estimators=300, max_depth=6, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8, use_label_encoder=False, eval_metric='mlogloss', random_state=42
Gradient Boosting	n_estimators=200, learning_rate=0.05, max_depth=5, min_samples_split=4, min_samples_leaf=2, random_state=42
ANN (Neural Network)	Dense(512, relu), Dropout(0.4), Dense(256, relu), Dropout(0.3), Dense(output, softmax), optimizer='adam', loss='categorical_crossentropy', epochs=10, batch_size=64, validation_split=0.1

a) Random Forest

```
--- Random Forest (Tuned) ---
                    precision    recall  f1-score   support

               age       0.95      0.98      0.96      1598
         ethnicity       0.98      0.96      0.97      1592
            gender       0.94      0.72      0.82      1595
  not_cyberbullying       0.46      0.66      0.54      1589
other_cyberbullying       0.49      0.39      0.43      1565
          religion       0.94      0.91      0.93      1600

          accuracy                           0.77      9539
         macro avg       0.79      0.77      0.78      9539
      weighted avg       0.79      0.77      0.78      9539
```

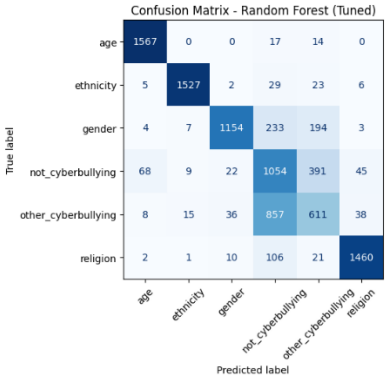


Figure 7: Tuned random forest performance

The tuned Random Forest model scored 77% in total accuracy. It exhibited high F1-scores for ethnicity (0.97), age (0.96), and religion (0.93), with high precision and recall in the identification of the sharply defined bullying types above. For gender class, the F1-score was 0.82, but there was decreased recall to 0.72, indicating some overlap with “not” and “other” cyberbullying. Accuracy for not cyberbullying was better (F1: 0.54) than in the default case, with improved recall (0.66), but other cyberbullying continued as the weakest (F1: 0.43), with ongoing issues with the classification of overlapping or blended abuse postings. With the macro average F1-score of 0.78, the model benefits marginally from the tuning, but there remain issues when it comes to overlapping classes.

b) XGBoost

```
--- XGBoost (Tuned) ---
                    precision    recall  f1-score   support

               age       0.99      0.97      0.98      1598
         ethnicity       0.99      0.98      0.99      1592
            gender       0.89      0.80      0.85      1595
  not_cyberbullying       0.51      0.61      0.55      1589
other_cyberbullying       0.55      0.54      0.54      1565
          religion       0.97      0.93      0.95      1600

          accuracy                           0.81      9539
         macro avg       0.82      0.80      0.81      9539
      weighted avg       0.82      0.81      0.81      9539
```

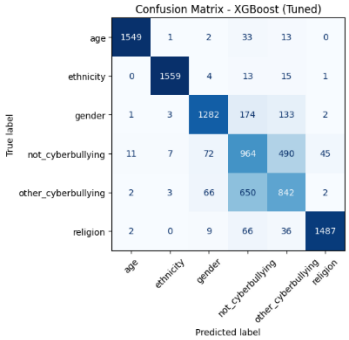


Figure 8: Tuned XGboost performance

The tuned XGBoost performed with a total accuracy of 81%, with high F1-scores for age (0.98), ethnicity (0.99), and religion (0.95), with consistent and accurate results for the clearly defined bullying labels. The gender class also improved to 0.85 in F1-score, with improved precision and recall balance. For not cyberbullying with F1 score of 0.55 and other cyberbullying with F1 of 0.54, the results were just medium but more stable compared to previous models. The macro average F1-score of 0.81 shows balance across the classes generally. This XGBoost version after fine-tuning provides the most accurate results to date, particularly in the context of heterogeneous and overlapping string inputs.

c) Gradient Boosting

```
--- Gradient Boosting (Tuned) ---
                     precision    recall  f1-score   support

                age       0.98      0.97      0.98      1598
          ethnicity       0.99      0.97      0.98      1592
             gender       0.90      0.79      0.84      1595
  not_cyberbullying       0.49      0.65      0.56      1589
other_cyberbullying       0.55      0.49      0.52      1565
           religion       0.97      0.92      0.95      1600

           accuracy                           0.80      9539
          macro avg       0.82      0.80      0.80      9539
       weighted avg       0.82      0.80      0.81      9539
```

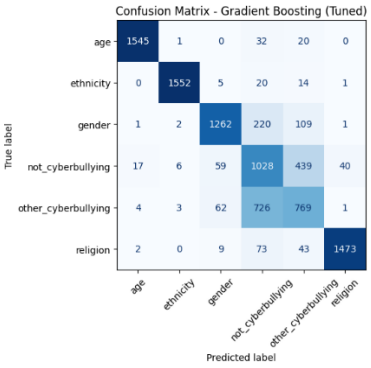


Figure 9: Tuned Gradient Boosting performance

The Gradient Boosting model that was tuned produced a macro accuracy of 80% with great F1-scores for age (0.98), ethnicity (0.98), and religion (0.95) to support consistent classification over strongly defined categories. The gender class received an F1-score of 0.84, with very little misclassification to "not" and "other" cyberbullying. For not cyberbullying, the model performed fairly well with F1 score of 0.56, and there was a good balance between precision and recall. Other cyberbullying fared better to an F1-score of 0.52, though again still hurt by overlap with related classes. With a macro F1-score of 0.80, this tuned model demonstrates consistent performance across all classes and provides a good balance between specificity and generalization.

d) ANN

```
--- ANN (Tuned) ---
                     precision    recall  f1-score   support

                age       0.98      0.97      0.97      1598
          ethnicity       0.99      0.97      0.98      1592
             gender       0.87      0.79      0.83      1595
  not_cyberbullying       0.49      0.44      0.46      1589
other_cyberbullying       0.50      0.63      0.56      1565
           religion       0.96      0.93      0.94      1600

           accuracy                           0.79      9539
          macro avg       0.80      0.79      0.79      9539
       weighted avg       0.80      0.79      0.79      9539
```

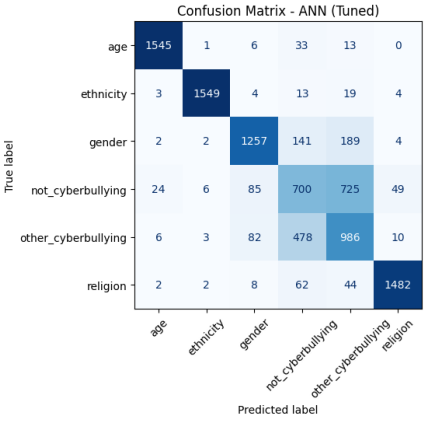


Figure 10: Tuned ANN performance

The tuned ANN model obtained a general accuracy of 79% with high F1-scores for age (0.97), ethnicity (0.98), and religion (0.94), indicating robust performance on well-delineated classes. The gender class held a high F1-score of 0.83, though recall was marginally lower (0.79) owing to misclassifications in some instances. The model demonstrated modest improvement in the case of other cyberbullying with F1 score of 0.56, better than prior ANN results for the class. Yet, not cyberbullying fell to the level of F1-score of 0.46, largely owing to too-frequent misclassification with "other cyberbullying." Despite this, the model attained a macro average F1-score of 0.79, indicating broadly balanced effect across all instances with scope for improvement in differentiation of subtle or overlapping instances.

3.4. Model comparison

Compared to other models, tuned XGBoost showed the best balance between precision and recall across all six categories, particularly maintaining high performance on both clearly defined and overlapping classes. It also demonstrated greater stability and generalization after hyperparameter tuning, making it the most reliable choice for multi-class cyberbullying detection.

4. DISCUSSION

Although the developed cyberbullying classification system showed high-performance results for strongly defined categories like age, ethnicity, and religion-based cyberbullying, some limitations were observed, especially in the subtle and overlapping scenarios. Some of the major areas of concern and the directions of improvement are presented below:

- **Difficulty in Distinguishing Overlapping Classes**

The most challenging task seen was the high misclassification between the "not cyberbullying" and "other cyberbullying" classes. Such classes register ambiguous context-dependent language and hence become more difficult to differentiate using standard ML models and bag-of-words representations. This negatively impacted recall and precision, particularly in the ANN and Random Forest models.

Future Direction: To better capture semantic meanings and intent behind the text, the use of context-aware models like BERT, RoBERTa, or DistilBERT may be useful. Such transformer models comprehend the context by considering the relation of the word with the nearby words.

- **Semantic Lack of Understanding**

Even after preprocessing and feature extraction,traditional models such as XGBoost, Random Forest, and Gradient Boosting depend strongly on word frequency and have little insight into sentence structure or semantics. This prevents them from successfully detecting sarcasm, hidden aggression, or coded speech, all of which are very common in cyberbullying.

Future Direction: The incorporation of word embeddings like Word2Vec, GloVe, or fine-tuned transformer embeddings could add semantic depth. Merging them with the use of sequence models like LSTM or GRU could enhance performance when meaning relies on word order.

- **Model Interpretability and Trust**

Models like ANN and XGBoost, while very capable, are largely black-box models and do not explain why they made a particular prediction. This becomes a problem when such tools are used in the real-world applications for moderation or legal purposes.

Future Direction: Methods such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) may be used to enhance explainability, enabling stakeholders to trust and audit the model's outputs.

- **Domain and Data Generalization**

The dataset used consists entirely of Twitter data and of six preassigned classes. Although useful for preliminary experimentation, it may not capture the range of language and cyberbullying strategies employed across platforms such as Instagram, Reddit, or TikTok.

Future Direction: A larger multi-platform dataset, maybe annotated with the help of human-in-the-loop techniques, might assist in training more general and stronger models. Using cross-domain training and test can also evaluate the scalability of the system.

- **Real-Time Deployment Constraints**

While models worked fine in offline environments, using them in runtime environments (such as social media or chat tools) imposes limitations such as latency, scalability, and resource utilization.

Future Direction: For responsiveness, upcoming implementations may utilize compression techniques such as knowledge distillation and light architectures to run inference on edge devices or live platforms.

In short, while the system performed robust baseline results, particularly in multi-class classification of overt cyberbullying categories, real-world applications will be facilitated by more in-depth linguistic modeling, better explainability, and scalable deployment plans. Such directions create a promising blueprint for the next stages of the project.

5. CONCLUSION

This project was able to successfully create a machine learning-based system to identify and classify different kinds of cyberbullying from social media text data. With heavy data preprocessing, feature engineering, evaluation of the model, and hyperparameter tuning, the system achieved high performance in detecting overt categories such as age, ethnicity, and religion. Although issues persist in differentiating subtle and overlapping classes such as “not cyberbullying” and “other cyberbullying,” the results indicate that machine learning is a practical solution to automated cyberbullying detection. With improvements from context-aware models and improved semantic understanding, this system has the potential to become a very useful tool to increase digital safety and to moderate interaction in the digital world.

References

[1]

N. Yuvaraj *et al.*, “Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking,” *Mathematical Problems in Engineering*, vol. 2021, p. e6644652, Feb. 2021, doi: https://doi.org/10.1155/2021/6644652.

[2]

Z. Ji *et al.*, “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, vol. 55, no. 12, Nov. 2022, doi: https://doi.org/10.1145/3571730.

[3]

K. De Angeli *et al.*, “Class imbalance in out-of-distribution datasets: Improving the robustness of the TextCNN for the classification of rare cancer types,” *Journal of Biomedical Informatics*, vol. 125, p. 103957, Jan. 2022, doi: https://doi.org/10.1016/j.jbi.2021.103957.

[4]

J. Antony Vijay, H. Anwar Basha, and J. Arun Nehru, “A Dynamic Approach for Detecting the Fake News Using Random Forest Classifier and NLP,” *Computational Methods and Data Engineering*, pp. 331–341, Nov. 2020, doi: https://doi.org/10.1007/978-981-15-7907-3_25.

[5]

M. Fayaz, A. Khan, J. U. Rahman, A. Alharbi, M. I. Uddin, and B. Alouffi, “Ensemble Machine Learning Model for Classification of Spam Product Reviews,” *Complexity*, vol. 2020, pp. 1–10, Dec. 2020, doi: https://doi.org/10.1155/2020/8857570.

[6]

Z. Lu *et al.*, “Natural Language Processing and Machine Learning Methods to Characterize Unstructured Patient-Reported Outcomes: Validation Study,” *Journal of Medical Internet Research*, vol. 23, no. 11, p. e26777, Nov. 2021, doi: https://doi.org/10.2196/26777.

[7]

C. J. Harrison and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction to natural language processing,” *BMC Medical Research Methodology*, vol. 21, no. 1, Jul. 2021, doi: https://doi.org/10.1186/s12874-021-01347-1.