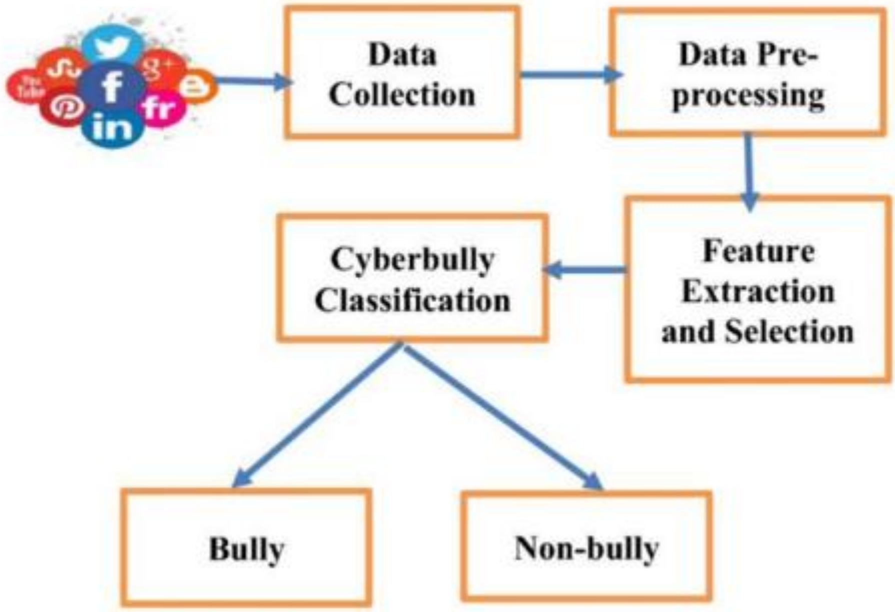
# Table of Contents

## Aim

The aim of the paper is to detect and classify cyberbullying on Twitter using advanced NLP techniques like BERT and LSTM for accurate and context-aware analysis.

Collab Link: https://colab.research.google.com/drive/1wqSfcG2lD1fjDVbUMnPSjIKZXu-QpW3r?usp=sharing

## Paper Work



**Strengths of the Paper**

1. **Relevant Topic**
   The paper addresses an important and socially relevant issue—cyberbullying on social media—which has widespread implications.

2. **Use of Modern Techniques**
   It incorporates state-of-the-art NLP models such as BERT, which effectively captures context in text data better than traditional models.

3. **Hybrid Approach**
   The combination of BERT embeddings with LSTM layers enables the model to understand both the semantics and temporal patterns in text.

4. **Comprehensive Evaluation**
   The paper uses multiple evaluation metrics like accuracy, precision, recall, and F1-score, which provides a well-rounded assessment of model performance.

5. **Data Preprocessing**
   The study emphasizes the importance of preprocessing techniques such as noise removal, tokenization, and normalization, which are critical in text-based applications.

6. **Comparative Analysis**
   By comparing models like LSTM and BERT, the paper provides useful insights into their relative strengths and weaknesses for cyberbullying detection.

7. **Real-World Application**
   The methodology is applicable to real-world social media platforms, providing a framework for safer online environments.

**Weaknesses of the Paper**

1. **Dataset Details Are Limited**
   The paper does not clearly specify dataset size, diversity, and handling of class imbalance, which are essential for evaluating model reliability.

2. **Limited Model Exploration**
   Only a few models are considered. Other architectures like CNNs, attention mechanisms, or ensemble models are not explored.

3. **Scalability Concerns**
   BERT-based models require considerable computational resources, making deployment in real-time or resource-constrained environments challenging.

4. **Ethical Considerations Are Not Addressed**
   The paper does not explore fairness, bias mitigation, or privacy concerns that are critical when deploying such models in social contexts.

5. **Overfitting and Regularization Not Discussed**
   There is no in-depth discussion on how overfitting is prevented or techniques like dropout and regularization that improve model generalization.

6. **Real-Time Implementation Missing**
   The paper lacks strategies for real-time detection, thresholding, or alert mechanisms that could make the system practical for live applications.

**What You Can Implement Further**

1. **Dataset Enhancements**
   - Use more recent datasets and incorporate tweets from multiple languages and regions.
   - Apply data augmentation techniques to balance class distribution.

2. **Model Improvements**
   - Experiment with lighter transformer models such as DistilBERT or RoBERTa.
   - Explore hybrid models combining CNNs, LSTMs, or attention layers for improved performance.

- Implement ensemble methods that combine multiple models for robust predictions.

3. **Explainability and Interpretability**
    - Integrate interpretability tools like SHAP or LIME to explain why certain tweets are classified as bullying.
4. **Real-Time Deployment**
    - Use Twitter's API to fetch tweets in real-time and classify them on-the-fly.
    - Create dashboards for monitoring trends and generating reports.
5. **Ethical and Bias Handling**
    - Apply fairness-aware learning algorithms to reduce bias related to gender, ethnicity, or age.
    - Conduct sensitivity analyses to understand where models might fail.
6. **User Interaction Features**
    - Build interfaces for users to report misclassifications and provide feedback for improving the system.
7. **Monitoring and Alert Systems**
    - Implement threshold-based alerts to flag highly abusive content.
    - Track user behavior over time and generate warnings when patterns of cyberbullying are detected.
8. **Performance Optimization**
    - Use model quantization or distillation to reduce inference time.
    - Optimize for GPU acceleration and use batch processing for faster results.

## What I did

**Implementation Steps**

1. **Data Collection**
    - Gathered tweets labeled as bullying or non-bullying across categories like religion, age, ethnicity, and gender.
2. **Data Preprocessing**
    - Cleaned text by removing URLs, special characters, and stop words.
    - Applied tokenization and normalization techniques.
3. **Text Representation**
    - Used pre-trained BERT to generate contextual embeddings from tweets.
4. **Model Architecture**
    - Combined BERT embeddings with two LSTM layers.

    - Applied dropout and a dense layer with softmax for classification.

5. **Fine-Tuning**
    - Adapted the BERT model to the specific dataset by adjusting learning rates and training parameters.
6. **Training**
    - Split data into training, validation, and test sets.
    - Used sparse categorical cross-entropy loss and Adam optimizer.
    - Evaluated with accuracy, precision, recall, and F1-score.
7. **Evaluation**
    - Compared performance using confusion matrices.
    - Found BERT outperformed LSTM in detecting cyberbullying content.
8. **Results**
    - BERT achieved higher accuracy and better handling of nuanced language.
9. **Applications**
    - Can be used for real-time monitoring on social platforms.
    - Supports safer online environments with future improvements like fairness and explainability.

## Something New – Something Extra

In this extension of the cyberbullying detection project, we integrate real-time data from the **NewsAPI**, a free and publicly available API, to fetch recent news headlines and classify them using the pre-trained BERT-LSTM model. This demonstrates how the model can be adapted to handle live data streams beyond its original dataset.

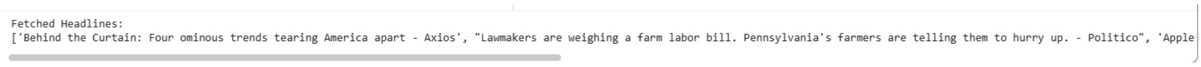
**Steps Involved**

1. **Access NewsAPI**
    - Obtain a free API key from https://newsapi.org/ to access the latest news headlines.
2. **Fetch Real-Time Data**
    - Use the requests library to query the NewsAPI and retrieve news headlines in JSON format.
3. **Preprocess the Text**
    - Clean the headlines by removing URLs, special characters, and unnecessary spaces to prepare them for analysis.
4. **Encode Headlines with BERT**
    - Use the pre-trained BERT tokenizer to convert the cleaned headlines into token IDs and attention masks required by the model.

5. **Run Predictions**
   - Pass the encoded data through the BERT-LSTM model to classify each headline into predefined categories such as bullying or non-bullying.
6. **Display Results**
   - Map the model's predictions back to readable labels and present the results alongside the original headlines.

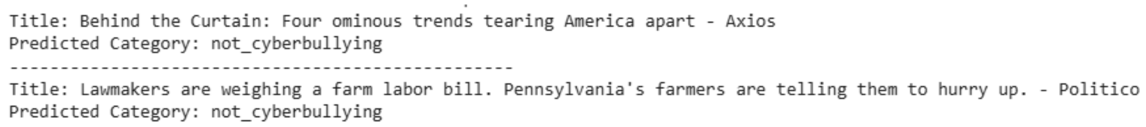
**Outcome**

This setup showcases how advanced NLP models like BERT-LSTM can be extended to classify live, real-world text data. It emphasizes adaptability, practical deployment, and provides a framework for building real-time monitoring systems that can be used in applications like online safety, content moderation, and sentiment tracking.



```
Fetched Headlines:
['Behind the Curtain: Four ominous trends tearing America apart - Axios', "Lawmakers are weighing a farm labor bill. Pennsylvania's farmers are telling them to hurry up. - Politico", 'Apple
```

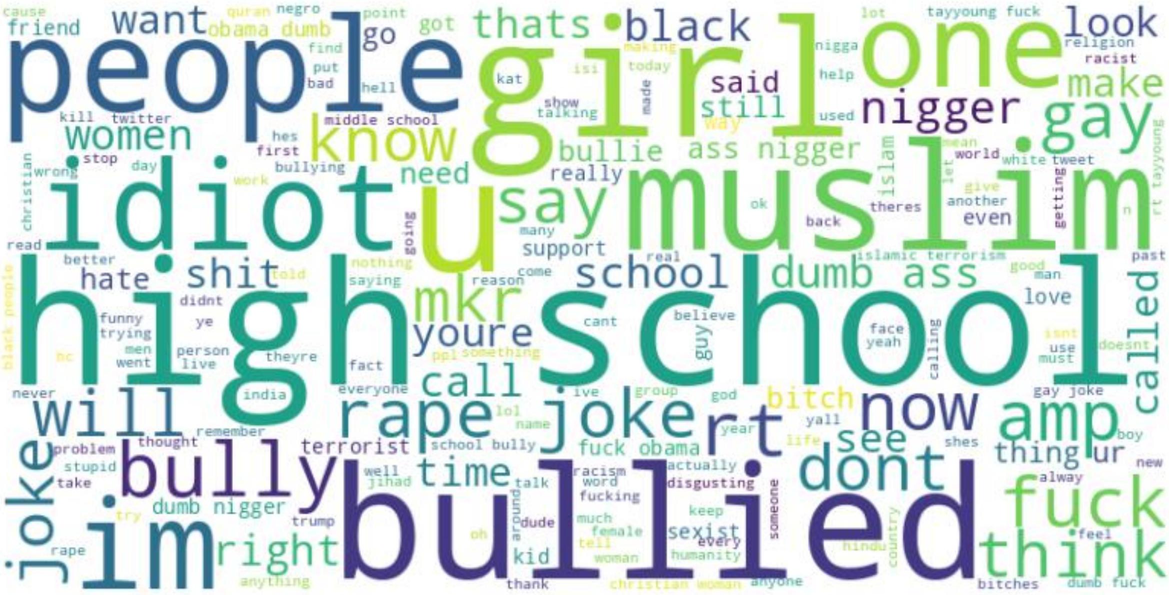
Output



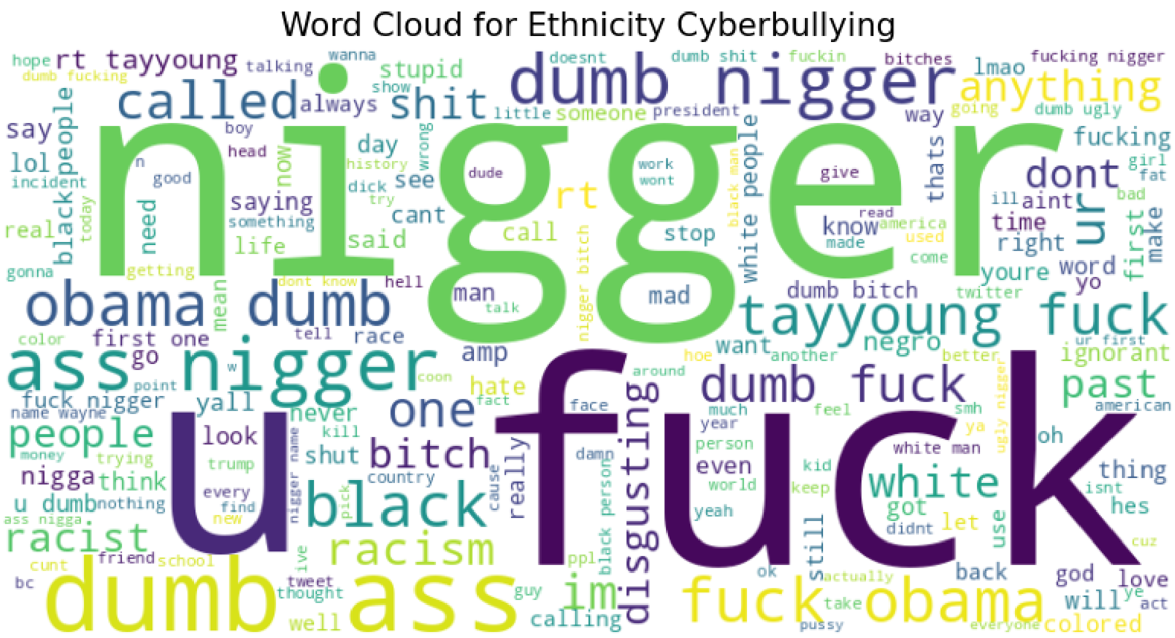
```
Title: Behind the Curtain: Four ominous trends tearing America apart - Axios
Predicted Category: not_cyberbullying
--------------------------------------------------
Title: Lawmakers are weighing a farm labor bill. Pennsylvania's farmers are telling them to hurry up. - Politico
Predicted Category: not_cyberbullying
```

## Output

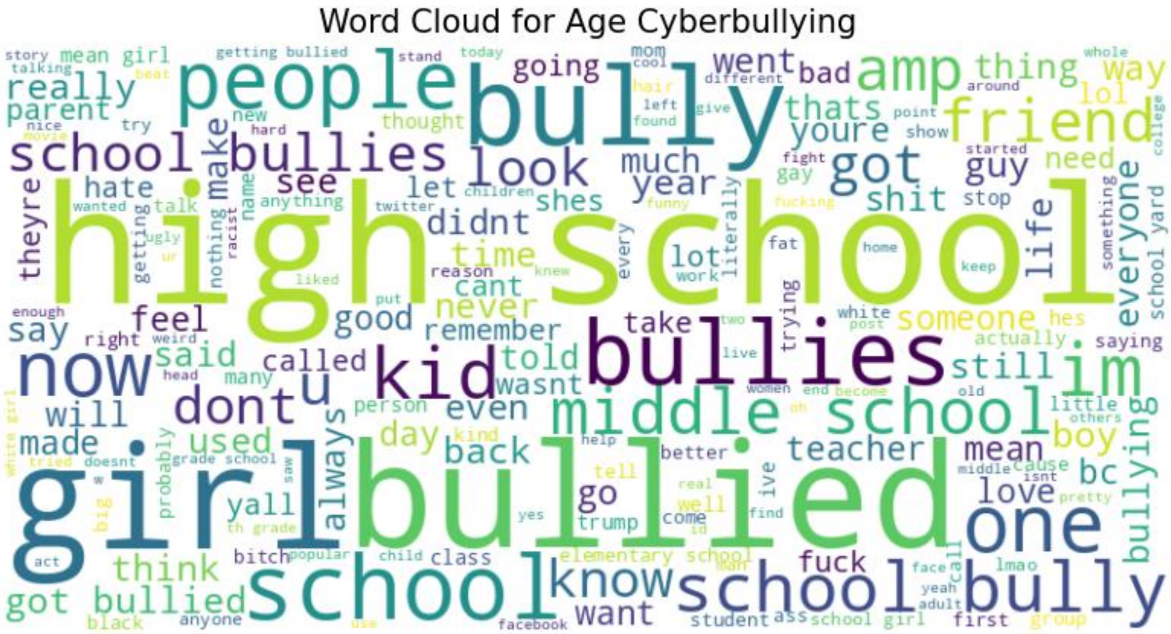
Overall Dataset



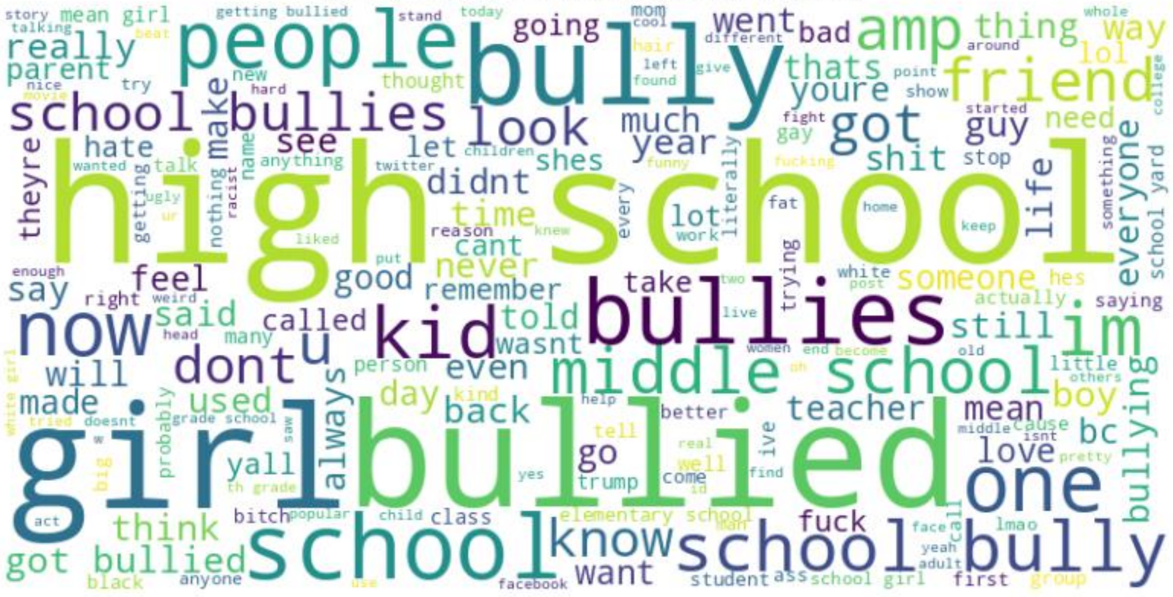
For Ethinicity



For Age

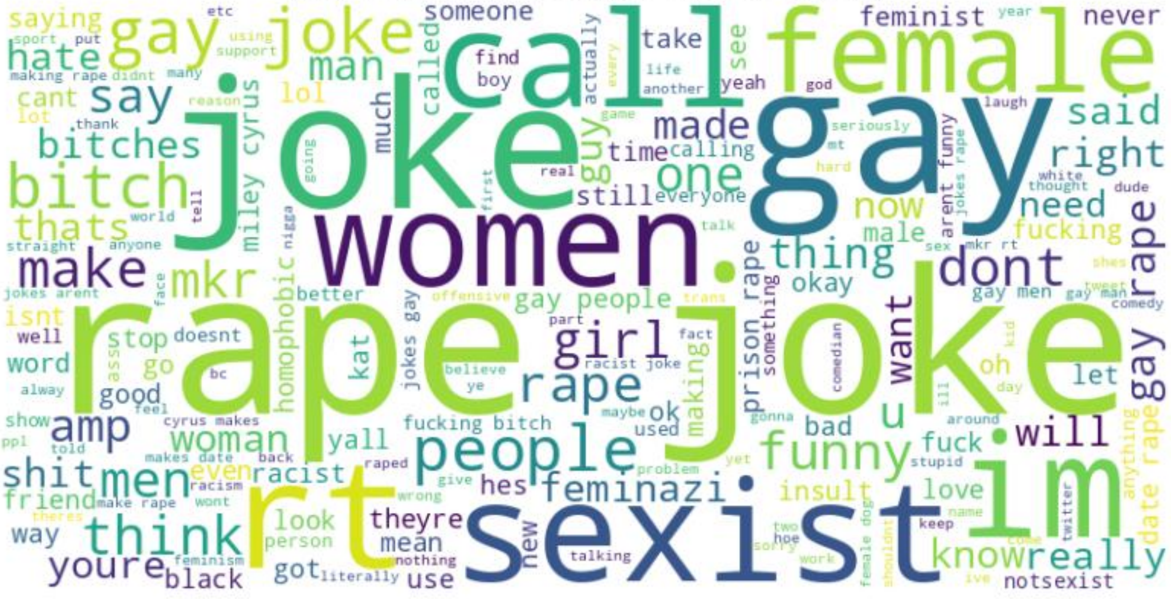


For Religion



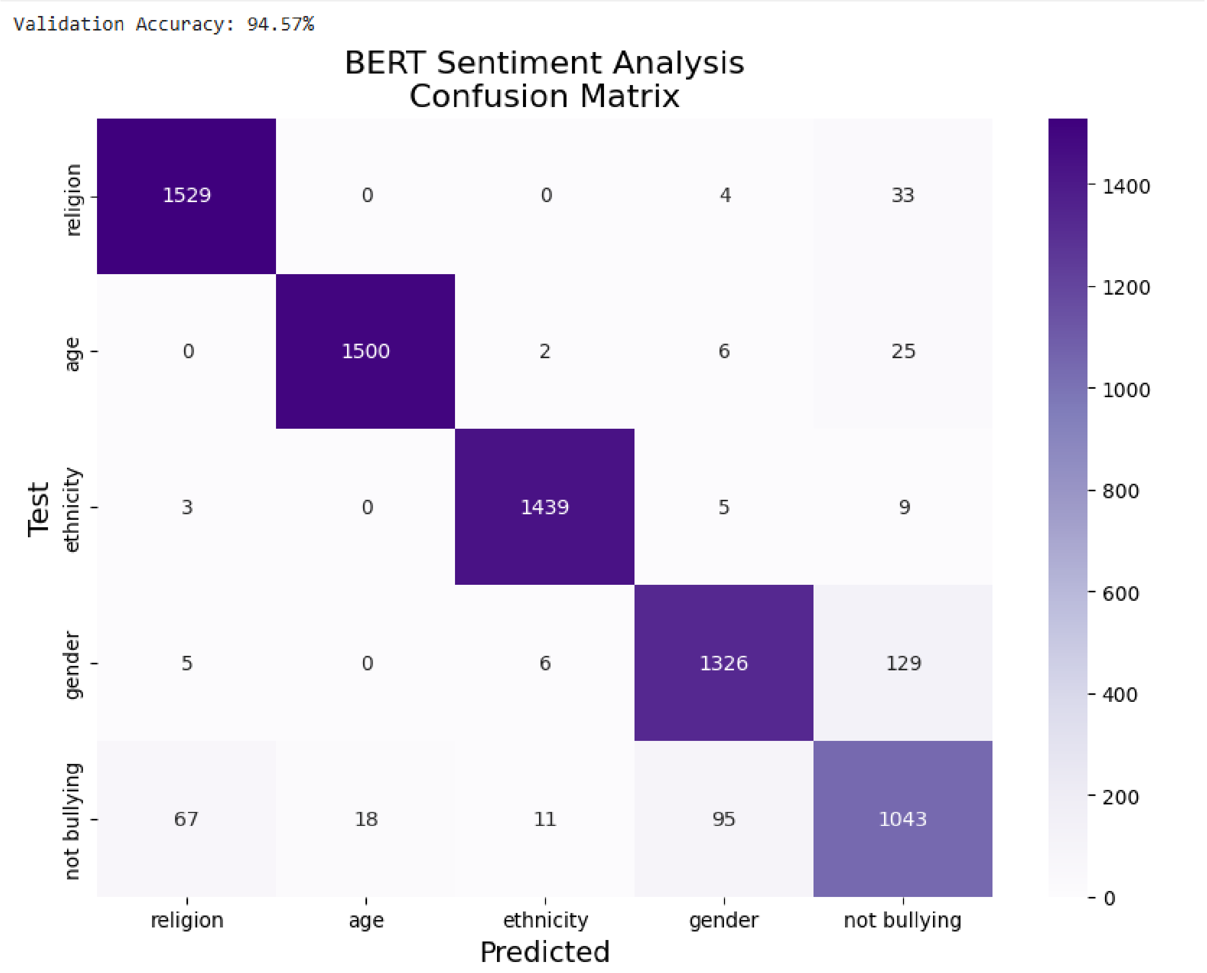
Word Cloud for Age Cyberbullying

For gender



Word Cloud for Gender Cyberbullying

Accuracy & Grpah

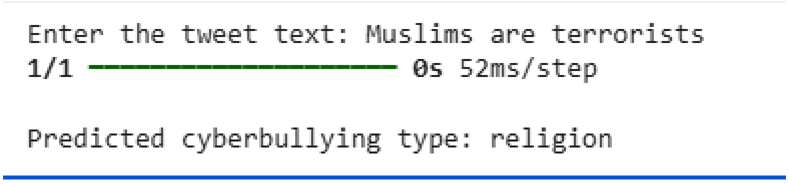


```
Validation Accuracy: 94.57%
```

| | religion | age | ethnicity | gender | not bullying |
|---|---|---|---|---|---|
| religion | 1529 | 0 | 0 | 4 | 33 |
| age | 0 | 1500 | 2 | 6 | 25 |
| ethnicity | 3 | 0 | 1439 | 5 | 9 |
| gender | 5 | 0 | 6 | 1326 | 129 |
| not bullying | 67 | 18 | 11 | 95 | 1043 |

Testing

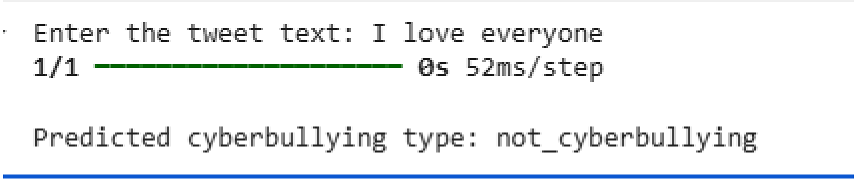
Test1:



```
Enter the tweet text: Muslims are terrorists
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step

Predicted cyberbullying type: religion
```

Test2:



```
Enter the tweet text: I love everyone
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 52ms/step

Predicted cyberbullying type: not_cyberbullying
```

Paper Accuracy: 95%

My accuracy: 94.7%

Almost there : )

## Conclusion

The paper provides a strong foundation for detecting cyberbullying using NLP and deep learning. It successfully leverages BERT and LSTM to enhance accuracy while highlighting the importance of contextual understanding and sequential modeling. However, areas such as dataset diversity, ethical considerations, scalability, and real-time deployment need further exploration. Implementing improved datasets, advanced models, interpretability tools, and fairness-aware techniques will enhance the robustness and usability of the solution.