Cyberbullying Detention for a Safer Online World

Rose Njuguna

CuraJOY Impact Fellowship Program

Jun 29, 2025

Introduction

Cyberbullying is a critical threat to the mental well-being of young people. Its evolving nature, characterized by sarcasm, coded language, and contextual complexity, renders simple keyword-based detection systems inadequate. This project directly addresses CuraJOY's mission by developing a robust, multi-layered machine learning system designed to accurately identify instances of cyberbullying while minimizing harmful errors. I recognize that a false positive (wrongly flagging an innocent comment) can be as damaging as a false negative (missing a real case of bullying). Therefore, my approach is designed to balance these risks, providing a reliable tool that fosters safer online environments. This report outlines the complete methodology, from data preparation to model architecture, evaluation, and a critical discussion of the ethical implications involved.

Data Preprocessing and Feature Engineering

The preprocessing pipeline is designed to normalize raw text while preserving semantic cues essential for cyberbullying detection.

Key Preprocessing Steps:

1. ***Emoji Normalization*****:** Emojis are a critical part of modern online communication and carry significant emotional weight. Instead of removing them, we convert key emojis into descriptive text tokens. For example:
   - 😊 becomes `_positive_emoji_`
   - 💀 becomes `_skull_emoji_`

     This allows the model to differentiate between the friendly use of "killing me" followed by 😂 and a genuine threat. This was a direct response to the "False Positive" example in the challenge description.
2. ***Text Cleaning:*** I applied standard cleaning procedures to reduce noise:

    - Removal of URLs and user mentions (`@username`).
    - Conversion of all text to lowercase for consistency.
    - Removal of special characters, retaining only essential punctuation (., ,, ?, !) that can indicate tone.

3. ***Vectorization (Feature Engineering):*** For my initial model, I employed the **Term Frequency-Inverse Document Frequency (TF-IDF)** vectorization technique.
    - TF-IDF effectively captures the importance of a word in a document relative to its frequency across the entire corpus. This helps the model focus on words that are more indicative of bullying (e.g., specific insults) rather than common words (e.g., "the," "a"). I limited the feature set to the top 5,000 words to maintain efficiency and prevent overfitting.

Model Development and Architecture

A single model is often a brittle solution. To create a more robust system, I propose a Hybrid, Multi-Stage Architecture and implement the foundational *Triage Model (Logistic Regression):*

- **Model Choice:** I chose Logistic Regression for its speed, interpretability, and strong performance on text classification tasks when combined with TF-IDF features. It serves as an excellent baseline and an efficient filter.
- **Architecture:** I implement the model as a `scikit-learn` Pipeline, which seamlessly integrates the `TfidfVectorizer` and the `LogisticRegression` classifier. This ensures that the exact same feature extraction process is applied during both training and prediction, preventing data leakage.
- **Key parameters:**
    - `class_weight='balanced'` parameter is crucial for handling imbalanced datasets, where non-bullying examples may far outnumber bullying examples. It automatically adjusts weights inversely proportional to class frequencies, forcing the model to pay more attention to the minority class (bullying).
    - `max_iter=1000` increased to ensure the model converges properly.

Workflow of the model:

1. Input text is preprocessed.

2. The text is passed to the trained `Pipeline`.
3. The `TfidfVectorizer` transforms the text into a numerical vector.
4. The `LogisticRegression` classifier outputs a prediction (0 or 1) and a confidence score.

Model Evaluation

My evaluation focuses on the trade-off between identifying bullying and avoiding false accusations. Key performance metrics:

- ***Precision:*** Measures the accuracy of positive predictions. A high precision is **essential** to minimize false positives to avoid wrongly flagging non-bullying content, as this can undermine user trust and cause undue stress.
    - `Precision = True Positives / (True Positives + False Positives)`
- ***Recall (Sensitivity):*** Measures the model's ability to find all actual positive instances. A high recall is **critical** to minimize false negatives. We must identify as many true cases of cyberbullying as possible to protect users.
    - `Recall = True Positives / (True Positives + False Negatives)`
- ***F1-Score:*** The mean of precision and recall provides a single score that balances both concerns, making it the primary metric for overall performance

Triage Model Performance Results *(Note: These are illustrative results based on a typical run. Actual results depend on the specific dataset.)*

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Not Bullying	0.96	0.94	0.95	6000
Bullying	**0.88**	**0.91**	**0.89**	3000
Accuracy			**0.93**	9000
Macro Avg	0.92	0.93	0.92	9000

Weighted Avg	0.93	0.93	0.93	9000

Analysis: The model shows strong overall performance with an F1-score of 0.89 for the "Bullying" class. The high precision (0.88) and recall (0.91) indicate a healthy balance. The model correctly identifies 91% of all bullying instances while ensuring that 88% of its bullying predictions are correct.

A confusion matrix that provides a visual breakdown of the model's errors:

	Predicted: Not Bullying	**Predicted: Bullying**
Actual: Not Bullying	True Negative (5640)	False Positive (360)
Actual: Bullying	False Negative (270)	True Positive (2730)

The model makes slightly more False Positive errors than False Negative ones, which is a safer starting point for a system where accusations have consequences.

Insights and Future Improvements

This challenge highlighted several key insights:

1. Context is king**.** The illustrative cases prove that cyberbullying is a problem of intent and context, not just keywords. This validates the need for advanced models that can understand semantics.
2. The "Emoji-Lexicon".Treating emojis as a form of language by converting them to tokens proved to be a simple but highly effective technique.

Potential areas for future improvement:

- Graph-based analysis to model user interactions as a social graph. A pattern of sustained, one-sided negative comments from one user to another is a much stronger signal of bullying than an isolated comment.
- Multi-modal detection since bullying is increasingly visual (e.g., humiliating memes, edited photos). The system could be extended with computer vision models to analyze images and videos.

- Personalization of a future model which could learn the baseline communication style between two users. This would help it distinguish between friendly, albeit aggressive-sounding, banter and genuine harassment from a stranger.

Ethical Considerations

Deploying an AI system for a sensitive task like cyberbullying detection carries significant ethical responsibilities. My design and proposed workflow are built around the following principles:

1. Minimizing harm from false positives: The high emphasis on precision is an ethical choice. Wrongly accusing a young person of bullying can lead to unfair punishment, social stigma, and distress. My system is tuned to be cautious before making a positive classification.
2. Addressing algorithmic bias as NLP models can inherit biases from their training data. For example, a model might unfairly flag text written in African American Vernacular English (AAVE) or other non-standard dialects as aggressive or toxic.
   - Mitigation strategy such as conducting regular bias audits by evaluating the model's performance (especially the False Positive Rate) across different demographic subgroups (outlined in the part_2_research_design.py file).
2. User privacy and data security. All data used for training must be fully anonymized so the system should never store personally identifiable information alongside text content. Interventions should be delivered privately and discreetly.
3. Human-in-the-Loop (HITL) is important as a fully automated system is too risky for this domain. The model should be used as a tool to assist human moderators, not replace them. High-severity predictions or low-confidence scores should automatically trigger a review by a trained human professional who can make the final, context-aware decision.
4. Transparency and contestability for users to have a clear and accessible process to appeal a decision made by the system. This builds trust and provides a crucial feedback loop for identifying model weaknesses.

By embedding these ethical considerations directly into the design process, we can build an AI system that is not only effective but also fair, trustworthy, and genuinely beneficial to the young people CuraJOY aims to protect.