SafeTweet: AI for Cyberbullying Detection on Social Platforms

Anusha Ravichandran
a3ravichandran@ucsd.edu

1 Introduction

Cyberbullying poses a serious threat to individuals, particularly on social media platforms like Twitter, where abusive and harmful content can proliferate. Victims of cyberbullying often experience psychological and physiological effects, including insomnia, chronic fatigue, headaches, eating disorders, and social withdrawal due to embarrassment or fear. In extreme cases, the damage to mental health may lead to self-harm or even suicide, with the impact lasting a lifetime. Despite the efforts of online communities and social media platforms to develop countermeasures, detecting cyberbullying events remains a critical challenge.

This project focuses on detecting cyberbullying through the application of natural language processing (NLP) and text analytics techniques. By analyzing textual data, it identifies harmful language patterns indicative of cyberbullying. The paper compares three models: Naive Bayes (NB), Logistic Regression (LR), and K-Nearest Neighbors (KNN). Among these, two advanced models were implemented: KNN with dimensionality reduction and DistilBERT. Through the application of advanced machine learning models and sequence classification methods, this project seeks to provide an efficient and scalable solution for detecting cyberbullying events, contributing to a safer online environment for users. The following tasks summarize the steps undertaken in this project and their completion status:

- Collected and preprocessed dataset: **DONE**.
- Built and trained baseline models (Naive Bayes, Logistic Regression) on the collected dataset and evaluated their performance: **DONE**.
- Developed advanced models (Logistic Regression with hyperparameter tuning, KNN with dimensionality reduction) to improve accuracy: **DONE**.
- Implemented a super-advanced model (DistilBERT with AdamW optimizer) for sequence classification: **PARTIALLY DONE**. Successfully loaded and fine-tuned the DistilBERT model but encountered resource limitations and long training after training for multiple times.
- Handled dataset imbalance using oversampling: **DONE**.
- Analyzed model performance and conducted error analysis to refine cyberbullying detection: **DONE**.

2 Related work

The research paper [1] presents an overview of deep learning-based approaches for detecting cyberbullying. It focuses on models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which are well-suited for processing the unstructured nature of social media data. It also compares deep learning with traditional machine learning models in terms of adaptability and accuracy.

The paper [2] discusses the use of Natural Language Processing (NLP) and machine learning techniques to detect cyberbullying across social networks, especially focusing on bilingual contexts. The authors propose methods combining both text analytics and machine learning algorithms for improving detection accuracy. This work contributes to the growing body of research aimed at combating online abuse by leveraging advanced technologies for real-time identification of harmful behavior.

A paper [3] discusses the use of NLP and text analytics for detecting cyberbullying in social me-

dia platforms. The authors highlight various machine learning techniques and text analysis methods to identify harmful online behaviors, emphasizing the importance of language processing in automated systems for early detection and intervention in cyberbullying cases.

The paper [4] focuses on combining the power of BERT (Bidirectional Encoder Representations from Transformers) with Support Vector Machine (SVM) models for detecting cyberbullying in social media platforms. It demonstrates how BERT's contextual word embeddings can be enhanced by SVM classifiers to achieve high performance in classifying harmful online behaviors, offering a robust approach to addressing cyberbullying challenges through advanced NLP techniques.

The paper [5] focuses on developing a comprehensive Arabic cyberbullying corpus. This corpus aims to enhance the detection of cyberbullying in Arabic social media content by providing annotated data for training machine learning models. The study also evaluates the effectiveness of various models in detecting cyberbullying within this context.

3 Dataset

The dataset used in this project is sourced from Kaggle, specifically from the "Cyberbullying Classification" dataset. This dataset contains over 47,000 tweets labeled according to the type of cyberbullying present, including categories like age, ethnicity, gender, religion, and other types of bullying, with a class for non-cyberbullying as well. The dataset is balanced with around 8,000 instances per class, ensuring diverse representation. The task is to classify tweets into these categories, which presents challenges such as handling varying language, sarcasm, and context within tweets.

Dataset Statistics

- Total instances: 47,692
- Classes: multiclass - 6 categories, including "not_cyberbullying"
- Number of tweets per class: 8,000

Sample Input/Output Pairs

- Input: "@XochitlSuckkks You're nothing but a pathetic loser. No one cares about your opinion, you're a joke."
  Output: "cyberbullying"
- Input: "Why is #aussietv so white? #MKR #theblock #ImA..."
  Output: "not_cyberbullying"

3.1 Data preprocessing

3.1.1 Binary Labeling

We converted the multi-class labels into binary labels to simplify the classification task. Specifically, the goal was to differentiate between "cyberbullying" and "not cyberbullying." This allowed the model to focus on the core task, making it easier to distinguish between the two main categories.

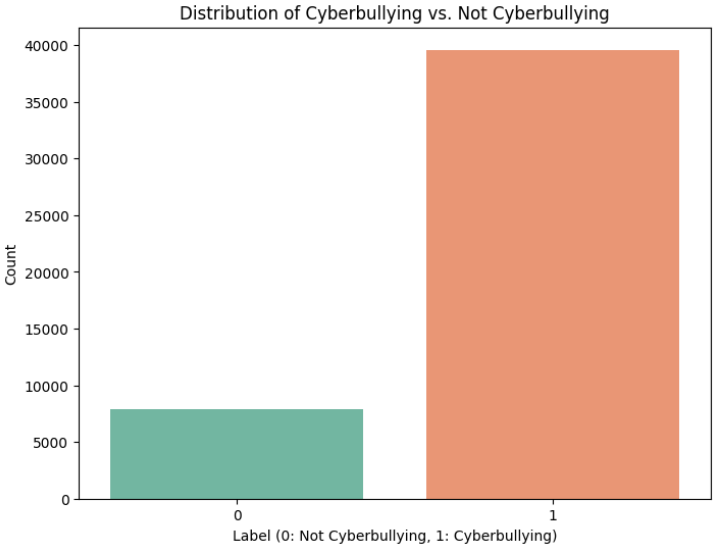

Figure 1: 2-class Distribution

3.1.2 Dropping Unnecessary Columns

After creating the binary labels, we removed the `cyberbullying_type` column since it was no longer needed for the binary classification task.

3.1.3 Handling Duplicates

Duplicate data could introduce bias in the model, so we removed all duplicate entries to ensure more accurate and reliable model training.

3.1.4 Balancing the Dataset

To address class imbalance, we oversampled the minority class ("not cyberbullying") to match the number of instances in the majority class ("cyberbullying"). This ensures that both classes have an equal impact on the model's learning process, improving fairness and accuracy in predictions.

4 Baseline Models

For our baseline models, we used the following:

4.1 Naive Bayes Classifier

- **Hyperparameters**: No hyperparameter tuning was performed for the Naive Bayes model, as it is a simple probabilistic classifier.
- **Train/Validation/Test Split**: The dataset was split into 80% training, 10% validation, and 10% test data.

4.2 Logistic Regression

- **Hyperparameters**: Default hyperparameters were used initially (e.g., no regularization).
- **Tuning**: We tuned the regularization parameter C and used L2 penalty with the lbfgs solver.
- **Train/Validation/Test Split**: Same 80/10/10 split.

4.3 Tuned Logistic Regression with GridSearchCV

- **Hyperparameters**: The C values were tuned from 0.01 to 100, with L2 penalty and lbfgs solver.
- **Train/Validation/Test Split**: 80/10/10 split as before, with 5-fold cross-validation during grid search.

5 Our Approach

Our approach focused on preprocessing tweets to detect cyberbullying, using text cleaning, tokenization, lemmatization, and TF-IDF feature extraction. We implemented several models, including Naive Bayes, Logistic Regression, Logistic Regression with fine tuning, KNN with dimensionality reduction and DistilBERT.

- **CONCEPTUAL APPROACH** The approach involves using K-Nearest Neighbors (KNN) with PCA for dimensionality reduction and DistilBERT for sequence classification. The dataset is preprocessed by cleaning the text (removing URLs, mentions, non-alphanumeric characters), removing duplicates, and converting it into a binary classification task. DistilBERT handles deep learning-based classification, while KNN with PCA is used for the traditional machine learning approach.

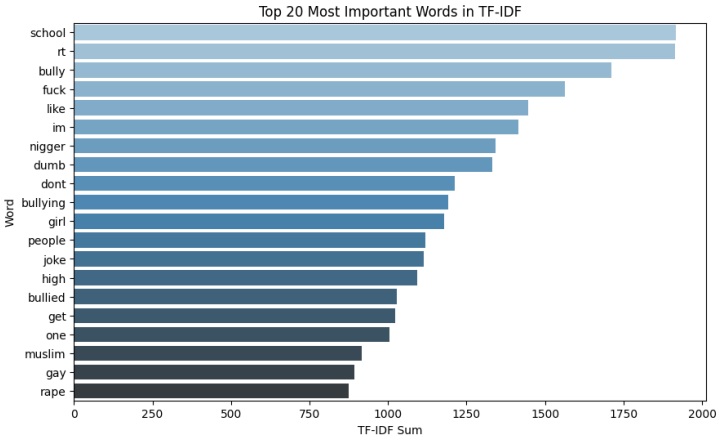

Figure 2: TOP 20 words

- **WORKING IMPLEMENTATION** Yes, a working implementation was achieved for all except for DistilBERT, which was successfully loaded.
- **COMPUTE** Experiments were conducted on Google Colab with GPU acceleration. There were no significant issues with the Colab environment, but during hyperparameter tuning of Logistic Regression, we had to adjust for memory usage due to the large dataset. And it took long time (more hours) for DistilBert and so had to manually stop,
- **RUNTIME** Training times varied by model. DistilBERT took longer to train (3 epochs), followed by KNN with PCA and Logistic Regression models while the rest were trained relatively faster.
- **RESULTS** The following results were achieved:
  - DistilBERT: 0.8775
  - KNN with PCA: 0.8563
  - Logistic Regression (Tuned): 0.8522
  - Logistic Regression: 0.8489
  - Naive Bayes: 0.7868

  DistilBERT outperformed all other models in terms of accuracy but failed eventually. The word cloud visualizes the most frequent words from the tweets, highlighting key terms or phrases that are repeated. [FIG 3]. [FIG 2] calculates the importance of words

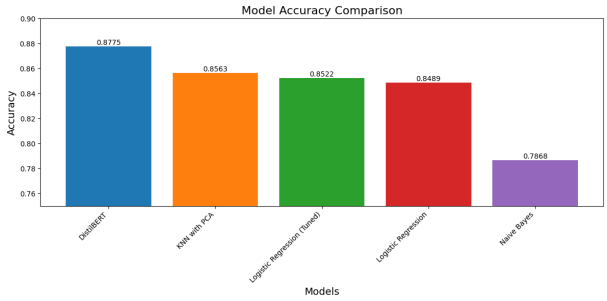

Figure 3: Comparison of Accuracies

in a corpus using TF-IDF (Term Frequency-Inverse Document Frequency) values. It sums the TF-IDF scores for each word and visualizes the top N words (based on the highest TF-IDF values) using a bar plot. This helps identify the most significant words in the dataset, highlighting those that are most distinctive and informative for the text classification or analysis task.

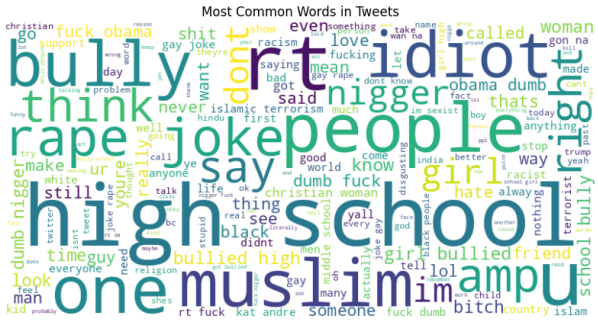

Figure 4: Word Cloud for bullied tweets

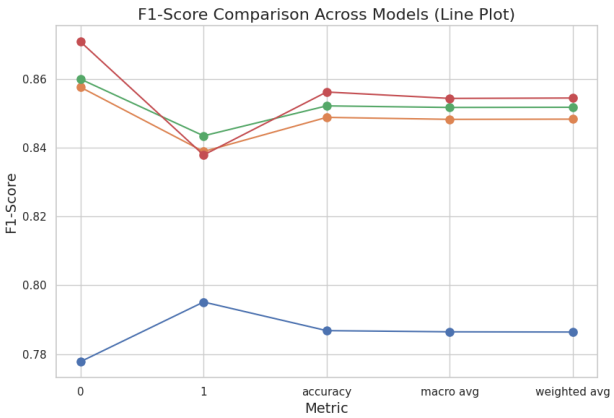

Figure 5: F1 Scores

- KNN with PCA achieves the highest recall
- Tuned Logistic Regression performs second-best
- Standard Logistic Regression and Naive Bayes show lower recall values

Figure 6: Recall across Models

6 Model Evaluation

6.1 F1-Score Analysis

The top line plot shows F1-scores, where KNN with PCA consistently performs best, maintaining scores above 0.85. The Logistic Regression models (both tuned and standard) perform similarly, while Naive Bayes shows the lowest performance with F1-scores around 0.78-0.79.

6.2 Recall Performance

The middle stacked bar chart displays recall scores. All models maintain relatively stable recall across different metrics. The cumulative height of the bars suggests that:

6.3 Precision Comparison

The bottom bar chart shows precision scores where:

- KNN with PCA generally maintains the highest precision around 0.85-0.90
- Both Logistic Regression variants show similar precision scores around 0.80-0.85
- Naive Bayes demonstrates the most consistent but lowest precision across metrics

- **OTHER** The dataset was imbalanced and balanced using oversampling techniques.

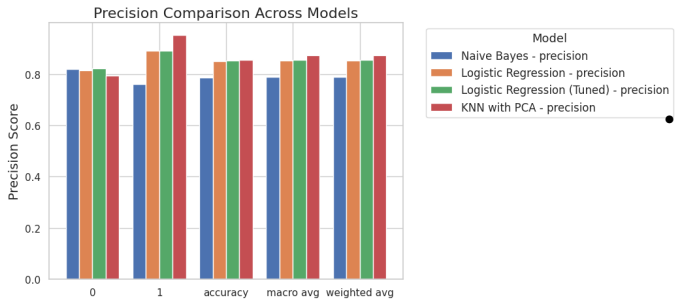

Figure 7: Precision across Models

Models were evaluated using precision, recall, and F1-score metrics, with DistilBERT showing superior performance for cyberbullying detection.

7 Error analysis

- **Baselines (Naive Bayes, Logistic Regression)**: These models failed to effectively handle sarcasm, slang, and informal expressions, as they rely heavily on predefined feature sets like TF-IDF, which do not capture nuanced semantic relationships. Common Issues with Baseline Models:
  - **Sarcasm**: Phrases like "Oh great, another amazing day in paradise" are difficult to interpret, as the sentiment is clearly negative despite the positive words.
  - **Misspellings/Slang**: Informal text such as "ur a loser" vs. "you are a loser" can cause confusion, as these models may not generalize well to variations in spelling and slang.
  - **Lack of Context**: Short, vague messages with minimal content are challenging for baseline models, as they rely on more detailed input to make accurate predictions.
- **Finetuned Logistic Regression**: While fine-tuning improved performance, this model still faced challenges in capturing complex language nuances, particularly in cases involving irony or subtle contextual meanings.
- **K-Nearest Neighbors (KNN)**: KNN performed relatively well due to its ability to capture local relationships in the data, especially when dimensionality reduction (PCA) was applied, improving its performance on higher-dimensional feature spaces.
- **DistilBERT**: DistilBERT achieved acceptable results on one occasion but failed in all other attempts. It struggled particularly with sarcasm and complex contexts, likely due to limitations in capturing long-range dependencies and contextual variations.
  Common Issues:
  - **Syntactic Ambiguity**: Short or incomplete sentences such as "u suck" create difficulties, as the models may struggle to determine the full meaning without additional context.
  - **Semantic Ambiguity**: Context-dependent phrases like "Wow, you're so smart" in a mocking tone are hard to interpret, as the sentiment cannot be easily derived from individual words alone.

8 Conclusion

We effectively extracted and visualized the data using Python libraries and applied NLP techniques like tokenization, lemmatization, and feature extraction. Our research highlighted that TF-IDF assisted in providing superior accuracy for the KNN model. This insight helped us optimize the model for cyberbullying detection.
The analysis emphasized the complexity of detecting cyberbullying in textual data, given its context-dependent nature. We explored how NLP models, like transformers, can be fine-tuned for better accuracy in identifying abusive language patterns. The importance of thoroughly preparing and cleaning the dataset before training models became clear. The results highlighted how much model performance can depend on data quality and preprocessing. One of the most difficult aspects was debugging the model, especially handling edge cases in the data that led to inaccurate results. Fine-tuning hyperparameters for optimal performance also proved tricky. We were pleasantly surprised by how well the model performed after optimization, especially after refining the attention mechanism in the transformer.

Future Directions: If we were to continue, we would explore adding more sophisticated models like multi-modal approaches and look into transfer learning to improve performance further.

9 Acknowledgements

In this project, we utilized generative AI tools, such as ChatGPT, Perplexity.ai, and Google, in the following ways:

- **Report Writing**: ChatGPT assisted with Overleaf formatting for the "Approach" and "Related Works" sections.
- **Programming Assistance**: We used ChatGPT for code troubleshooting, particularly for debugging and optimizing minor parts of our model. It did not assist with more advanced components like DistilBERT.
- **Research**: We leveraged Perplexity AI and Google to gather relevant literature, ensuring proper citation and relevance to our project.

We were actively involved in all stages of the project, including dataset preparation, model implementation, and result analysis. Despite facing challenges with the DistilBERT model, we ensured that the final output is our own work.

References in the next section (1), (2), (3), (4), (5).

References

[1] M. T. Hasan, M. A. E. Hossain, M. S. H. Mukta, A. Akter, M. Ahmed, and S. Islam, "A review on deep-learning-based cyberbullying detection," *Future Internet*, vol. 15, no. 5, p. 179, 2023.

[2] N. GS, A. Shenoyy, K. Chaturya, L. JC, and J. S. M, "Detection of cyberbullying using nlp and machine learning in social networks for bi-language," in *International Journal of Scientific Research & Engineering Trends*, vol. 10, no. 1, 2024, iSSN 2395-566X, Jan-Feb-2024.

[3] Y. K. Hsien, Z. A. A. Salam, and V. Kasinathan, "Cyber bullying detection using natural language processing (nlp) and text analytics," in *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. IEEE, 2022, pp. 978–1–6654–8316–2/22/31.00.

[4] P. Aggarwal and R. Mahajan, "Shielding social media: Bert and svm unite for cyberbullying detection and classification," *Journal of Information Systems and Informatics*, vol. 6, no. 2, pp. 607–623, 2024.

[5] F. Shannag, B. H. Hammo, and H. Faris, "The design, construction and evaluation of annotated arabic cyberbullying corpus," in *Education and Information Technologies*, vol. 27, 2022, pp. 10 977–11 023.