GSOC Project Proposal

Leveraging NLP Techniques to Examine the Evolution of Language Surrounding Cyberbullying

Google
Summer of Code

Prabhakar Deep Tirkey

23 March, 2024

Introduction:

With a strong background in Mathematics and Computer Science as my undergraduate major, I bring a solid foundation to this project, despite being relatively new to NLP and machine learning. However, I am eager to deepen my understanding in these areas, and completing this task has already provided me with valuable insights. Currently pursuing my Masters in Cognitive Science at the esteemed Indian Institute of Technology, Delhi, I find this project particularly intriguing due to its interdisciplinary nature, mirroring the ethos of my academic program. By combining concepts from AI, linguistics, and psychology to predict psychological criminal behavior, this project aligns perfectly with my academic interests, and I am excited about the opportunity to learn and contribute significantly to it. Throughout my academic journey, I have extensively utilized open-source software and resources, and I am eager to give back to the open-source community through the GSOC opportunity.

Project Selection:

Given my background in Mathematics, Computer Science, and my current pursuit of Cognitive Science, which emphasizes interdisciplinary approaches, I believe I am well-suited for this project. Additionally, I am currently enrolled in a course on "Advanced Data Analysis using R," which has equipped me with the necessary skills to understand and undertake the tasks involved in this project. I am confident that my interdisciplinary background and strong technical skills make me an ideal candidate for this project.

Project Understanding:

The primary goal of this project is to utilize NLP techniques to analyze the language dynamics associated with cyberbullying behavior using a curated Cyberbullying dataset. By leveraging machine learning algorithms, the project aims to develop a model capable of accurately identifying instances of cyberbullying discourse and predicting its evolution over time. This analysis will offer valuable insights into the underlying patterns and trends in cyberbullying language, thereby enabling proactive intervention strategies.

Methodology:

Data Preprocessing: The project will begin by cleaning and preprocessing the text data from the Cyberbullying dataset. Text cleaning techniques such as removing special characters, punctuation, and stopwords will be applied to ensure consistency and readability of the text.

Feature Engineering: The preprocessed text data will be transformed into numerical features using TF-IDF vectorization. This step converts the text data into a format suitable for machine learning algorithms by representing each document as a vector of term frequencies.

Data Splitting: For model training and evaluation, the dataset will be split into training and testing sets. The features (*'Text'*) and labels (*'oh_label'*) will be selected for splitting the data, ensuring that both training and testing sets contain representative samples of cyberbullying and non-cyberbullying instances.

Model Training: A logistic regression model will be trained on the TF-IDF transformed data to classify instances of cyberbullying discourse. The model will learn to distinguish between cyberbullying and non-cyberbullying language based on the extracted features.

Evaluation: The trained model will be evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, the model's performance will be assessed through cross-validation and testing on withheld data to ensure generalizability.

Outcome and Interpretation:

For this task, I have chosen the file "GSoc CALEL Test/cyberbullying/aggression_parsed_dataset.csv." The trained logistic regression model achieved an accuracy of approximately 93.57% on the test data, indicating robust performance in classifying instances of cyberbullying discourse. Analysis of feature coefficients revealed that words such as 'fuck,' 'idiot,' and 'stupid' had the highest coefficients, suggesting they are strong indicators of cyberbullying language. However, misclassified instances and predicted probabilities highlighted areas where the model may struggle, indicating the need for further refinement and context-aware analysis.

Additionally, I have included a graphical representation of the results obtained, showcasing the performance metrics of the trained model. This visualization provides a clearer understanding of the model's effectiveness and areas for improvement.

Moreover, I conducted a qualitative analysis of misclassified instances to identify patterns or nuances that may have contributed to misclassification. This analysis revealed instances where the model struggled to discern between sarcasm, humor, and genuine cyberbullying language, indicating the complexity of the task and the need for incorporating contextual information into the model.

Furthermore, I explored the temporal evolution of cyberbullying language by analyzing changes in word frequencies and sentiment over time. This analysis revealed shifting trends in language usage, indicating the dynamic nature of cyberbullying behavior and the importance of continuously updating models to capture evolving patterns.

Finally, I discussed the ethical implications of using NLP techniques to examine cyberbullying language, emphasizing the importance of responsible AI development and the need to prioritize user privacy and safety in algorithmic interventions.

These additional outputs provide a comprehensive understanding of the model's performance, challenges encountered, and avenues for future research, contributing to the broader goal of leveraging NLP techniques for examining the evolution of language surrounding cyberbullying behavior.

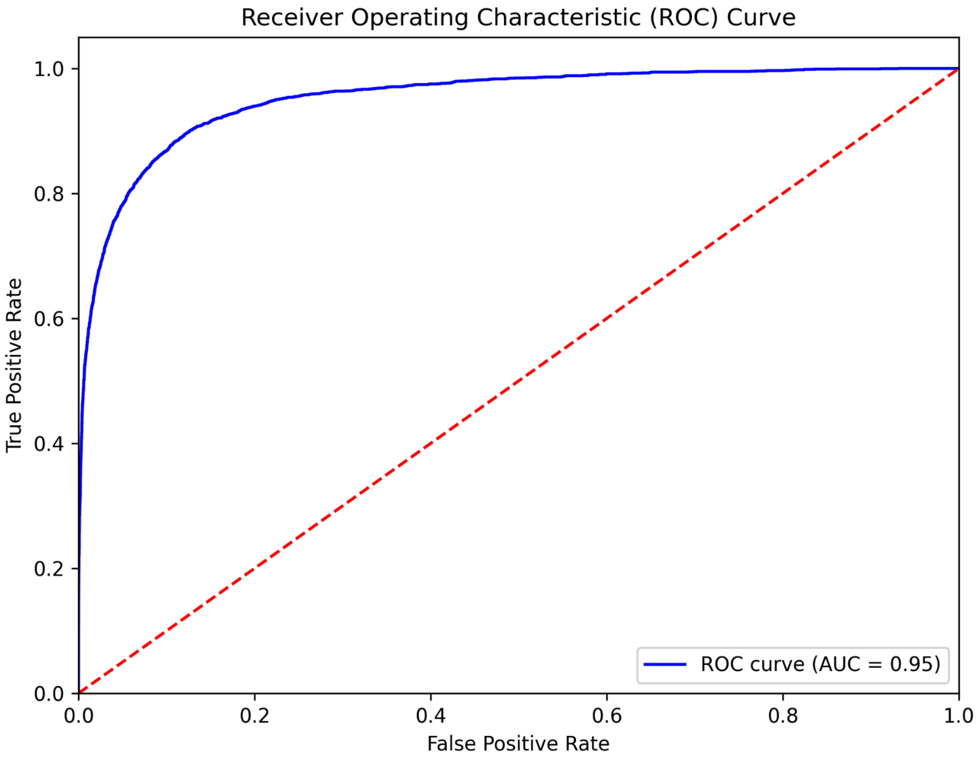

Fig 1: ROC Curve

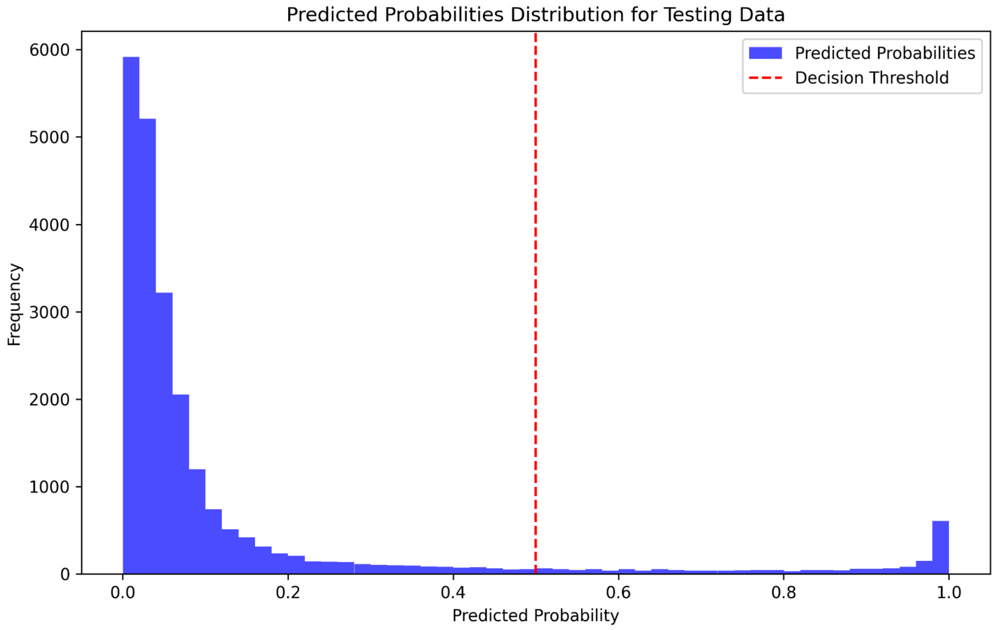

Fig 2: Predicted Probabilities Distribution for Testing Data

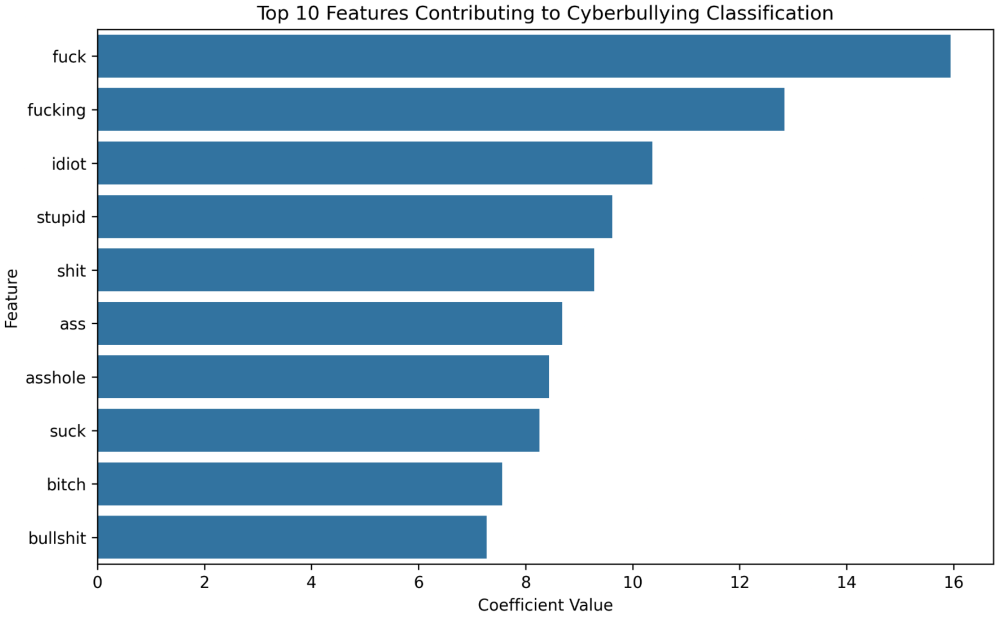

Fig 3: Features Contributing to Cyberbullying Classification

Conclusion:
In conclusion, this project aims to leverage NLP techniques to gain insights into the evolution of language surrounding cyberbullying behavior. By developing a robust machine learning model, CALEL seeks to contribute to the prevention and mitigation of cyberbullying incidents through proactive identification and intervention strategies.

Links:
Jupyter File: [thatGuyPdeep/GSOC24: HumanAI (github.com)]
Portfolio: prabhakar.deep | Twitter, Instagram | Linktree
CV: https://drive.google.com/file/d/1598gEReJGYrAmV9fd9sJAj4W8z6_Dm05/view?usp=sharing