Detecting Aggressive Behavior In Conversations

Goal:
Develop a system to automatically detect aggressive behavior in online conversations, such as insults, threats, or derogatory comments. The goal is to identify and classify hostile interactions across platforms like social media, forums, or messaging boards.

Data Sources:
Use publicly available datasets with labeled aggressive and non-aggressive conversations, such as the Offensive Language Dataset or scrape data from social media platforms.
https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset

Technical Aspects:
Use Feature Extraction, NLP, and Text Classification to make the model

Evaluation:
Test the model using standard metrics like accuracy, precision, recall, and F1-score to measure its performance in detecting aggression.

Expected Product:
A trained model capable of identifying aggressive messages and a report on its performance. The model should be able to take in a message or set of messages to output a report on the user(s) aggressiveness.