Project: *“BullyShield, A Deep Learning Approach to Combat Online Bullying"*

Team Members:

1. Kazi Israrul Karim
2. Eshtiak Alam Shihab
3. Md Shamsur Shafi Nur E Aziz

Introduction:

In the digital age, the prevalence of online bullying demands swift and effective solutions. This project, "BullyShield," harnesses the power of Artificial Intelligence (AI) and deep learning to rapidly detect and counteract online bullying in real-time.

Inspired by successful AI applications in gaming, our project employs advanced natural language processing and deep learning algorithms to analyze text-based communication. "BullyShield"
aims to expedite the identification of bullying patterns, allowing for prompt intervention and the establishment of a safer online environment.

Utilizing cutting-edge deep learning models and Python libraries like scikit-learn, our solution seeks to provide a faster and more accurate means of distinguishing normal communication from harmful behavior. The project's transformative potential extends to its adaptability for use across various online portals, contributing to a more positive online experience.

By demonstrating the speed and efficacy of "BullyShield," we envision its application as a proactive tool for multiple online platforms. This project addresses the pressing issue of online bullying and stands as a scalable solution for fostering respectful and secure digital interactions across diverse online communities.

Problem statement:

The task is to build a machine learning model that can classify online comments as toxic (i.e. disrespectful or harmful/threatening) or non-toxic. The challenge lies in accurately identifying
various forms of cyberbullying, including overt and subtle forms, while minimizing false positives and negative

Description of dataset:

The intended dataset for use contains more than 47000 tweets labeled according to the class of cyberbullying:

- Age
- Ethnicity
- Gender
- Religion
- Other types of cyberbullying
- Not cyber bullying

Methodology:

1. Data Preprocessing:

Conduct extensive data preprocessing to enhance model performance:

- Text Cleaning: Remove noise, HTML tags, and irrelevant characters from the tweets.
- Tokenization: Break down the tweets into individual tokens.
- Stopword Removal: Eliminate common words that may not contribute significantly to identifying cyberbullying.
- Special Character Removal: Standardize the text by removing special characters.
- Remove Rare Words: Eliminate words with low frequencies to reduce noise.
- URL, Emoji, and Rare Word Removal: Enhance the quality of the text data.
- Lemmatization: Reduce words to their base or root form for better representation.

2. Feature Extraction:

- Utilize feature extraction techniques to convert the preprocessed text into numerical representations:
- TF-IDF Vectorization: Convert tokenized and preprocessed text data into TF-IDF weighted vectors.
- Additional Techniques: Explore other feature extraction methods suitable for text data.

3. Model Selection:

Employ a variety of machine learning models, including but not limited to:

- Support Vector Machines (SVM)
- Logistic Regression
- Random Forest
- Naive Bayes

-

4. Model Training:

- Split the dataset into training and validation sets.
- Train each selected model on the preprocessed and feature-extracted data.

5. Hyperparameter Tuning:

- Optimize the hyperparameters of selected models to enhance their performance.

6. Cross-Validation:

- Implement cross-validation to ensure the robustness of the models and mitigate overfitting.

7. Model Evaluation:

Evaluate the performance of each model using a comprehensive set of metrics:

- Accuracy, Precision, Recall, F1-Score: Assess the overall model performance.
- Confusion Matrix: Analyze true positives, true negatives, false positives, and false negatives.
- ROC Curve and AUC: Examine the model's ability to distinguish between toxic and non-toxic comments.

Required technology:

- Python
- Scikit-learn & Pandas
- NLTK
- Matplotlib and Seaborn
- Google Colab & Jupyter Notebooks

Conclusion:

In summary, "BullyShield" employs AI and deep learning to swiftly detect and counteract online bullying. With a diverse dataset and a multi-model approach, it aims to provide a proactive and scalable solution, fostering a positive digital experience across various online platforms.