Cyberbullism Detection

This project is particularly interesting because it addresses the growing issue of cyberbullying by leveraging machine learning techniques to detect and classify harmful content. Machine learning enables the identification of hidden patterns in text data, improving accuracy in distinguishing between safe or hate speech, thus making online spaces safer and more inclusive.

Preprocessing Phase & Classification

- In this project, text data is **preprocessed** through cleaning and stemming to ensure consistency. The processed text is then converted into a numerical format for effective classification by machine learning models.
- In this project, **classification** is performed in two steps:
  - **Level 1: Binary Classification**
    - Determines whether a comment is an attack or not.
    - Helps filter out non-harmful content before further analysis.
  - **Level 2: Multiclass Classification**
    - If a comment is classified as an attack, it is further categorized.
    - Identifies the specific type of attack (e.g., hate speech or offensive language).
    - Ensures a more precise understanding of harmful content.

Interpretability & Explainability

- Objective: provide insight into the classifier’s predictions to improve transparency and trust.
- Evaluating different explanation strategies, considering both effectiveness and computational cost.
- A possible strategy is using **Pattern Mining** to extract and describe class behaviors. Other possibilities include XAI techniques for explainability evaluation, such as **SHAP**, or alternative approaches.

Dataset description

- The dataset consists of **47,692 labeled tweets**, each annotated with a specific category indicating the presence or absence of cyberbullying. It is structured as a two-column CSV file:
  - **tweet_text**: the raw content of the tweet (string format).
  - **cyberbullying_type**: the corresponding label, which can be either *not_cyberbullying* or one of several cyberbullying types.
- This dataset enables both **binary classification** (bullying vs. non-bullying) and **multi-class classification** (recognizing specific forms of abuse such as sexism, racism, etc.).
- *URL: https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data*

Related Works

Numerous recent studies have investigated the application of machine learning techniques for detecting and categorizing cyberbullying on social media platforms. The works selected for review helped with the choice of algorithms and contributed valuable methodological guidance in the development of the proposed approach:

- Title: *A Machine Learning Ensemble Model for the Detection of Cyberbullying*; url: https://arxiv.org/abs/2402.12538
- Title: *AI Powered Anti-Cyber Bullying System using Machine Learning Algorithm of Multinomial Naive Bayes and Optimized Linear Support Vector Machine*; url: https://arxiv.org/abs/2207.11897