Project Proposal: Multimodal Hate Speech Detection in Social Media Posts- Group 17

1. Topic and Use Case

The proliferation of hate speech on social media platforms poses significant challenges to maintaining respectful online environments. While traditional hate speech detection methods focus primarily on textual content, many social media posts combine text with images, requiring a multimodal analysis approach. This project aims to develop a classifier capable of detecting hate speech in social media posts by integrating both textual and visual data. By leveraging multimodal data analytics, we seek to enhance the accuracy of hate speech detection and provide comprehensive insights into the nature of such content.

2. Planned Data Sources

We will utilize the MMHS150K dataset, a manually annotated collection of 150,000 tweets, each containing both text and an associated image. This dataset encompasses various hate speech categories, including racism, sexism, homophobia, and religion-based attacks. The dataset is publicly available and has been used in prior research on multimodal hate speech detection. (Link)

3. Preliminary Thoughts on Analysis

- **Data Preprocessing:**
  - *Textual Data:* Perform tokenization, stop-word removal, stemming, and lemmatization to clean and prepare the text for analysis.
  - *Visual Data:* Resize images to a uniform dimension (e.g., 500 pixels on the shortest side) and normalize pixel values to standardize the input for visual models.
- **Feature Extraction:**
  - *Textual Features:* Utilize pre-trained language models such as BERT to extract contextual embeddings from the tweet text.
  - *Visual Features:* Employ convolutional neural networks (CNNs), like ResNet or EfficientNet, to extract features from images.
- **Multimodal Fusion:**
  - Develop a fusion network that integrates textual and visual features to capture the interplay between modalities. This approach aims to improve detection accuracy by considering both text and image content.
- **Model Training:**
  - Train the multimodal classifier using the processed dataset, employing techniques such as cross-validation to ensure robustness.
- **Evaluation:**
  - Assess model performance using metrics like precision, recall, F1-score, and ROC-AUC.
- **Data Visualization:**
  - Create visualizations such as confusion matrices, ROC curves, and attention maps to interpret model performance and highlight areas for improvement.