Data Collection and Preprocessing Phase

|  |  |
| --- | --- |
| Date | 20 October 2024 |
| Team ID | 739916 |
| Project Title | Toxic Comment Classification for Social Media using NLP |
| Maximum Marks | 6 Marks |

**Preprocessing Template**

The images will be preprocessed by resizing, normalizing, augmenting, denoising, adjusting contrast, detecting edges, converting color space, cropping, batch normalizing, and whitening data. These steps will enhance data quality, promote model generalization, and improve convergence during neural network training, ensuring robust and efficient performance across various computer vision tasks.

|  |  |
| --- | --- |
| **Section** | **Description** |
| Data Overview | The dataset consists of social media comments labeled into categories such as toxic, severe toxic, obscene, threat, insult, identity hate, and neutral. The dataset has 20,000 comments with an approximate class distribution: neutral (60%), toxic (20%), and other categories (5% each). |
| Resizing | Adaptation: Limit comments to a maximum number of words or characters (e.g., 100 characters).  **Example Transformation**: Original: "This comment is very long and unnecessary for our processing steps." Resized: "This comment is very long." |
| Normalization | * Convert all text to lowercase. * Remove unwanted characters such as URLs, HTML tags, and special characters. |
| Data Augmentation | * Generate synthetic data by:   + Replacing words with synonyms.   + Back-translation. * Example:   + Original: "You are terrible." |

|  |  |
| --- | --- |
|  | * Synonym Replacement: "You are awful." * Back-translation (via Spanish): "You are horrible." |
| Denoising | * Remove stopwords (e.g., "is", "and", "the"). * Example: * Original: "This is an offensive comment." * Denoised: "offensive comment" |
| Edge Detection | * Adaptation: Extract key phrases or n-grams from text. * Example: * Original: "I hate you, you are useless." * Key Phrases: ["hate you", "are useless"] |
| Color Space Conversion | * Convert sentences to embeddings (e.g., Word2Vec, GloVe, or BERT embeddings). |
| Image Cropping | * Adaptation: Truncate text to relevant portions, e.g., first 50 words. * Example: * Original: "This is a very lengthy comment that exceeds the limit." * Cropped: "This is a very lengthy comment." |
| Batch Normalization | * Normalize word frequencies in text data (e.g., TF-IDF). * Example: * Comment: "This is toxic toxic toxic." * After Normalization: ["toxic": 3/6, "this": 1/6, "is": 1/6]. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data |  |

|  |  |
| --- | --- |
| Data Description & Null values |  |
| Coreleation between variables |  |
|  |  |

|  |  |
| --- | --- |
| Data Preprocessing |  |
| clean the comment \_text in both the datasets. & training and testing |  |
| Vectorize the data |  |
| Train \_test\_ split& Transform the data& saving word vectorizer |  |

|  |  |
| --- | --- |
| Loading the pickle file |  |