Documentation: MCRS_by_Databoost

Name: Dataset_Toxicity
Number of Rows: 24,000
Number of Columns: 4

Column Descriptions:

1. **Text**:
    - **Type**: string
    - **Description**: This column contains all tweets or user comments collected from various sources.
2. **Degree_of_toxicity**:
    - **Type**: float64
    - **Description**: This column measures the degree of toxicity of the text. It helps identify whether the text is toxic.
    - **Value**:
        - If toxicity is greater than or equal to 0.1, the text is considered toxic.
        - Otherwise, it is labeled as non-toxic.
3. **Contains_offensive_word**:
    - **Type**: int8
    - **Description**: This column is generated when one or more offensive words are found in the text.
    - **Value**:
        - `1` if the text contains offensive words.
        - `0` if no offensive words are present.
4. **Label**:
    - **Type**: int8
    - **Description**: This column is the most important as it determines whether the text is toxic or not.
    - **Value**:
        - `1` = toxic
        - `0` = non-toxic

General Description:
This dataset is designed to analyze, classify, and model content published on social media. It includes texts extracted from various platforms (tweets, posts, comments, etc.) accompanied by metadata and labels describing the nature or characteristics of the content (e.g., toxicity). It is suitable for machine learning tasks such as toxic content detection, sentiment analysis, or identifying recurring themes in discussions.

Dataset Source:
The data comes from multiple public datasets merged into one:

- Jigsaw Toxic Comment Classification Challenge
- TweetEval
- Hate Speech Offensive

Purpose of Dataset Creation:

- **Offensive Content Detection**: Identifying messages containing offensive language to improve moderation on social media.
- **Measuring Toxicity Levels**: Quantifying toxicity to prioritize interventions or adjust automated responses.
- **Enhancing AI Models**: Training algorithms to differentiate between various levels of toxicity in online interactions.
- **Promoting a Safe Environment**: Helping platforms create safer and more respectful spaces by detecting and reducing harmful behavior.

Transformations Applied:

- **Data Cleaning**: Removed special characters, quotes, unnecessary symbols, and extra spaces from the "Text" column.
- **Label Standardization**: Normalized values in the "Label" column to ensure consistency (e.g., 0 for "non-toxic" and 1 for "toxic").
- **Filtering and Balancing**: Selected a balanced subset of data between classes to ensure fair learning for models.

Dataset Use Cases:

1. **Toxic Content Detection**: Training machine learning models to automatically detect toxic or offensive messages on social platforms.
2. **Automated Moderation**: Supporting moderation systems by filtering harmful or inappropriate content in online communities, forums, or messaging platforms.
3. **Linguistic Analysis**: Studying linguistic trends in offensive or toxic messages to inform policies or actions against harmful online behavior.
4. **Enhancing User Experience**: Helping companies improve platform quality by removing or reducing negative interactions for users.

AI Models Suitable for This Dataset:

- **BERT (Bidirectional Encoder Representations from Transformers)**: A powerful natural language processing (NLP) model for analyzing text context and classifying message toxicity.
- **RoBERTa (Robustly Optimized BERT)**: A BERT variant optimized for tasks like detecting offensive content.
- **FastText**: A lightweight and fast model suitable for text classification based on simple yet effective word and phrase vector representations.

Strengths of the Dataset:

- **Granularity of Information**: Columns like `Contains_offensive_words` and `Degree_of_toxicity` provide a detailed analysis of the content, enabling fine-grained classification.
- **Versatility**: Usable for various modeling tasks such as toxicity detection, sentiment analysis, or offensive speech identification.
- **Content Diversity**: Derived from multiple social networks, making it representative of various contexts, languages, and toxicity levels, which enhances model robustness.
- **Label Accuracy**: Clearly defined labels (`toxic` or `non-toxic`) ensure precise training for supervised AI models.