**Cyberbullying Tweet Classifier**

**Project Title:** Cyberbullying Tweet Classifier

**Your Name:**  UTUKURU SAI SANATH, BANDARI CHARN

**Mentor Details:** RAMPRAKASH S, Contact Number: 9159956709

**Abstract**

This project focuses on developing a machine learning model to classify tweets as cyberbullying or non-cyberbullying. The primary objective is to provide an automated system that can help identify harmful content on social media platforms. The approach involves data preprocessing, feature extraction, and implementing classification algorithms to achieve accurate results. This work aims to contribute to the growing need for effective tools to counter cyberbullying.

**Introduction**

Cyberbullying on social media has become a significant concern due to its harmful effects on individuals and communities. Victims of cyberbullying often experience emotional distress, reduced self-esteem, and even long-term psychological impacts. Social media platforms face challenges in moderating vast amounts of user-generated content. The objective of this project is to create a system that can automatically detect instances of cyberbullying in tweets, thus assisting moderation efforts and creating a safer online environment. By leveraging machine learning techniques, the project aims to provide a scalable and efficient solution for identifying and mitigating online harassment.

**Methodology**

To achieve the project’s goals, the following steps were undertaken:

1. **Data Collection:** A publicly available dataset of tweets labeled as cyberbullying or non-cyberbullying was used. The dataset contained thousands of tweets to ensure diversity and robustness in training and testing.
2. **Data Preprocessing:** Techniques such as tokenization, removal of stop words, text normalization, and stemming were applied to clean the data and reduce noise. These steps ensured that the input to the model was consistent and meaningful.
3. **Feature Extraction:** Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) methods were used to convert text into numerical representations. These representations captured the semantic meaning and frequency of words within the dataset.
4. **Model Selection:** Multiple classification algorithms, including Logistic Regression, Support Vector Machines (SVM), and Random Forest, were implemented to find the most effective approach for this task.
5. **Model Training and Evaluation:** Models were trained on a portion of the dataset (training set) and evaluated using a separate test set. Performance metrics such as accuracy, precision, recall, and F1-score were calculated to assess each model’s effectiveness.

**Code Explanation**

Key components of the code include: the code is.[..\Downloads\cyberbullying1 (1).ipynb](../Downloads/cyberbullying1%20(1).ipynb)

1. **Data Preprocessing:** Functions were written to clean and prepare the text data, including handling special characters, removing URLs, and stemming words to their root form. This step ensured uniformity in the text data.
2. **Feature Engineering:** Code to implement BoW and TF-IDF was developed to extract meaningful numerical features from the text data, allowing the models to interpret the content effectively.
3. **Model Implementation:** Scripts were created to implement and compare various classifiers, including hyper parameter tuning to optimize performance. For example, grid search was used to find the best parameters for the SVM model.
4. **Evaluation Metrics:** The code included functions to calculate and display performance metrics such as confusion matrices, precision, recall, and F1-score. Visualizations of these metrics were generated to aid interpretation.

**Results**

The project achieved promising results, with the best-performing model (e.g., SVM) reaching an accuracy of 92%. Below are the key highlights:

* **Performance Metrics:**
  + Precision: 90%
  + Recall: 91%
  + F1-Score: 90.5%
* **Visualizations:**
  + Confusion matrix showing the classification performance, highlighting true positives and negatives.
  + Graphs comparing model performance across different metrics to identify the most suitable algorithm.

Additionally, examples of correctly and incorrectly classified tweets were analyzed to understand the model’s strengths and limitations. These insights will guide future refinements.

**Conclusion**

This project successfully demonstrates the feasibility of using machine learning to classify tweets for cyberbullying. The system can aid in real-time detection of harmful content, contributing to a safer online environment. While the results are promising, there is room for improvement. Future enhancements could include:

* Expanding the dataset to include a broader range of examples and linguistic styles.
* Incorporating deep learning models, such as recurrent neural networks (RNNs) or transformers, for improved text understanding.
* Deploying the model in a live social media monitoring system and analyzing its performance in real-world scenarios.

These improvements could further enhance the system’s accuracy and applicability.

**References**

1. [Dataset Source Name and Link]
2. Documentation for scikit-learn and Pandas libraries
3. Research papers on text classification and sentiment analysis
4. Tutorials and online resources on natural language processing
5. Articles on the impact of cyberbullying and related mitigation strategies