**INTERNSHIP PROGRAM**

**BLEND VIDYA**

**PROJECT REPORT**

**Cyber Bully Tweet Classifier**

**Introduction**

The rise of social media platforms has led to increased instances of cyberbullying, causing psychological distress among users. Detecting and mitigating such behaviour is crucial. This project aims to develop a tool that can automatically classify tweets as cyberbullying or not, and further categorize the type of cyberbullying, enhancing online safety. Cyber Bully Tweet Classifier is a machine learning-based application designed to detect and classify instances of cyberbullying in tweets. As social media platforms become central to communication, the rise of online harassment and bullying has posed significant challenges. This project aims to address this issue by automatically analyzing the textual content of tweets to identify whether they contain cyberbullying, and if so, determine the type.

The classifier supports two modes:

**Binary Classification:** Classifies tweets as Cyberbullying or Not Cyberbullying.

**Multi-Class Classification:** Identifies specific types of cyberbullying such as sexual, religion, age, ethnicity, or other.

By leveraging Natural Language Processing (NLP) and machine learning techniques, the system provides a tool that can assist moderators, researchers, and developers in real-time detection and mitigation of harmful online behaviour.

**Abstract**

This project presents a machine learning-based approach to detect cyberbullying in tweets. Utilizing Natural Language Processing (NLP) techniques, the system preprocesses tweet data, transforms it using the TF-IDF vectorization, and employs Logistic Regression for classification. The model supports both binary (cyberbullying vs non-cyberbullying) and multi-class (specific types of cyberbullying) classifications. A user-friendly interface is developed using Streamlit, allowing users to input single tweets or upload CSV files for batch predictions.

**Problem Statement**

Manual monitoring of social media for cyberbullying is inefficient and impractical given the volume of data. There is a need for an automated system that can accurately detect and classify cyberbullying content in real-time.

**Objectives**

* Develop a preprocessing pipeline to clean and prepare tweet data.
* Implement TF-IDF vectorization for feature extraction.
* Train a Logistic Regression model for both binary and multi-class classification.
* Evaluate the model's performance using appropriate metrics.
* Deploy the model using Streamlit for user interaction.

**Literature Review:**

Previous studies have explored various machine learning techniques for cyberbullying detection. Logistic Regression has been noted for its efficiency in binary classification tasks, while other models like Random Forest and SVM have shown promise in multi-class scenarios. The choice of model often depends on the specific dataset and classification requirements.

**System Architecture:**

Modules:

* **Data Preprocessing:** Cleaning and preparing tweet data.
* **Feature Extraction:** Transforming text data using TF-IDF.
* **Model-Training:** Training Logistic Regression models for classification.
* **Evaluation:** Assessing model performance using metrics like accuracy and F1-score.
* **Deployment:** Developing a Streamlit app for user interaction.

**Methodology:**

* **Data-Preprocessing:** The clean\_text function converts text to lowercase, removes URLs, mentions, hashtags, and punctuation, and strips whitespaces.
* **Feature Extraction:** TF-IDF Vectorization is used to convert text data into numerical features.
* **Model Training:** The train\_model function loads and preprocesses data, splits it, trains a Logistic Regression model with balanced class weights, and saves it using joblib.
* **Evaluation:** The evaluate\_model function loads the test data, makes predictions, calculates accuracy and classification report, and identifies any unpredicted classes.
* **Deployment with Streamlit:** The app.py script provides a UI to classify a tweet, upload a CSV file for predictions, and evaluate the model with metrics and visualizations.

**Code Explanation:**

* **Preprocess.py:** Contains functions for cleaning text and loading data using TF-IDF vectorization.
* **train.py:** Trains Logistic Regression models for binary and multi-class classifications.
* **evaluate.py:** Evaluates the trained models and generates performance metrics.
* **app.py:** Develops a Streamlit web application for classification, batch prediction, and evaluation.

**Results:**

The Logistic Regression model demonstrated satisfactory performance in both binary and multi-class classifications. The Streamlit app allows users to interact with the model seamlessly, providing real-time predictions and evaluations.

**Ethical Considerations:**

The development and deployment of AI models for cyberbullying detection require careful ethical reflection:

**Bias and Fairness:**

The model is trained on labelled data, which may include human biases. This could lead to unfair labelling of certain dialects, cultures, or minority groups.

To mitigate this:

* We use class balancing techniques.
* We provide both binary and multi-class options for better interpretability.

**Interpretability and Confidence:**

The model predictions are presented with confidence levels, and users are warned that predictions are advisory and not to be used for punitive measures without human judgment.

**Data Privacy:**

User-uploaded data and input tweets are processed in-memory and not stored, ensuring privacy and compliance with data protection principles.

**Misuse Prevention:**

The application includes disclaimers to prevent misuse. It should only be used for research or awareness purposes, not as a standalone tool for moderation or enforcement.

**Conclusion:**

The project successfully developed a system capable of detecting and classifying cyberbullying in tweets. By leveraging NLP techniques and Logistic Regression, the model offers a practical solution for real-time monitoring of social media content.

**Future Scope:**

* Incorporate more advanced models like SVM or neural networks for improved accuracy.
* Expand the dataset to include more diverse examples.
* Integrate the system with social media platforms for live monitoring.
* Enhance the user interface with additional features and visualizations.

**References:**

1. Cyberbullying Detection on Twitter Data Using Machine Learning: <https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1021&context=data-science-mining>
2. Cyber Bullying Detection in Twitter using Machine Learning Algorithms: <https://ijaem.net/issue_dcp/Cyber%20Bullying%20Detection%20in%20Twitter%20using%20MachineLearning%20Algorithms.pdf>
3. A Machine Learning-Based Cyberbullying Detection System for Twitter: <https://ijrpr.com/uploads/V5ISSUE12/IJRPR36641.pdf>

**Overview:**

Thank you for your continued support and mentorship throughout my internship. I remain committed to delivering high-quality work and exceeding expectations.

Best regards,

Palamangalam Prabhas, Intern, Blend Vidya