# Week 5 Lab Report: Cyberbullying Detection Using NLP

## 1. Objective

## The main objective of this lab assignment was to come up with an efficient machine learning algorithm that helps determine the cases of cyberbullying using the social media post, especially tweet. This was a work that needed the application of improved natural language processing (NLP) methods, broad feature engineering and the usage of transformer-based models to correctly group text data into cyberbullying or non-cyberbullying. The goal was not just to attain great accuracy in classification but also learn valuable linguistic patterns that are typical of cyberbullying content and so help create a safer online space.

## \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## 2. Cleaning Data, Basic EDA

## Downloaded a dataset of tweets labelled in a number of ways as cyberbullying or non-cyberbullying tweets. This dataset was used as the basis of the classification model training and evaluation.

## Normed categorical categories into one variable family in binary form to enable processing by machine learning:

## 0 coded non-cyberbullying tweets

## Cyberbullying tweets were indicated by 1

## Defined data quality by dropping rows containing missing/null text data that would interfere with data analysis.

## Have conducted an exploratory data analysis (EDA), i.e. the analysis of sample data points, this aspect has served to confirm preprocessing procedures as well as to gain insights into the distribution and the nature of the tweet content, including keywords, sentiment patterns, and linguistic properties.

3. Feature Enhancement

Retrieved sentiment polarity scores of each of the tweets using TextBlob that measures the emotional tone as well as an intensity of the given text. The sentiments analysis offered attractive features that could be linked with the cyberbullying behavior.

The extracted n-gram feature was the Unigram and bigram features using CountVectorizer. These functions supported visualization and pattern identification that showed frequent phrases and linguistic patterns which are common in cyberbullying.

Although these feature improvements enriched the dataset, they were mostly beneficial in exploratory analysis and visualization since the core classification was based on transformer-based models that take unprocessed text as input.

**4. Model Preparation**

Selected DistilBERT that is a lightweight, but powerful transformer model optimised to work with textual data in binary classification tasks.

Detokenized the textual information to transform the raw tweets into input to transformer the model with successful adaptability to the BERT architecture.

Managed the problem of class imbalance by determining class weights, which were introduced in the algorithm training by use of a custom weighted trainer. This method contributed to the increase in the performance of the model in detecting examples of the minority class and reducing bias.

## **5. Training and Evaluation**

## Training was done in a series of epochs where the model consisted of learning some repeat in the data. To overcome the problem of class imbalance, the training was supplemented by weighted loss functions that place necessary focus on minority classes. This strategy improved the capability of the model to deal with multiple data distributions and exhibit reliability in preconditioning in real-life situations.

## The performance of the model was evaluated with the help of some significant metrics. Accuracy was a good overall average of the number of predictions achieved, whereas the F1-score was a balanced approach including both the precision and recall. The combined use of these metrics provided a complete picture of the efficacy of the given model, particularly in regards to imbalanced datasets whereby accuracy can be deceptive on its own.

## In a further bid to obtain a better grasp of the way the model behaves, several visualization techniques were used. Confusion matrices were broken down to determine true positives, false positives, false negatives, and the true negatives giving an idea of areas where the model performed exceptionally or not. ROC curve was generated to study the trade-off between the true positive rate and false positive rate at various thresholds in the process of choosing optimum decision boundaries. Further to this, precision-recall curve was examined to understand the capability of the model on preserving high levels of precision at various categories of recalls as this is critical in the case of class imbalance.

## Moreover, misclassified tweets were turned into a word cloud. This visual illustration was useful in locating shared features and patterns of language that explain errors in classification. It is important to identify the patterns to make any subsequent improvements, e.g., work on the feature extraction methods or expansions to the training set to extract more complex language constructs.

## 6. Outcome

## The implemented combination of transformer models with feature engineering methods and weighted training strategies on classes was especially successful in identifying cyberbullying tweets. This combined mechanism took advantage of the strengths of deep learning models, e.g. transformers, with traditional NLP features, leading to stronger classification and robustness. The project was an important practical experience regarding most elements of NLP, including data normalization, training, testing, and reporting of the results as well as understanding the data. Such abilities are necessary to create safer social media networks that can allow content filtration, and moderation of interactions on the internet in a responsible manner. The knowledge developed in the course of this work would help in continuing to develop healthier online environments by identifying harmful content more effectively and more efficiently so that mitigation can be done to combat these harmful contents.