### Objective

* Use NLP and ML to make a model to identify toxic and non-toxic tweets.

**Steps**

1. Load the tweets file using the read\_csv function from the Pandas package.

(downloaded from https://www.kaggle.com/datasets/ashwiniyer176/toxic-tweets-dataset.)

1. Upload the tweets into a list for easy text cleanup and manipulation
2. Clean up the tweets using regular expressions and tokenizer and apply lemmatization
3. Join the tokens back to form strings, which will be required for the vectorizers
4. Assign X and y
5. Perform train\_test\_split using sklearn Task to Perform
6. Use TF-IDF values for the terms as a feature to get into a vector space model
7. Fit and apply the vector space model on the train set and test set
8. Apply ensemble learning to use the given machine learning algorithms.

Decision Trees

• Random forest

• Naive Bayes Model

• K-NN Classifier

• SVM

1. Make predictions for the train and the test sets
2. Model evaluation: accuracy, recall, and f\_1 score
3. Building Confusion Matrix and RoC - AUC curve