Approach

Our problem statement was a classification problem that included text as input. So, after **Label Encoding** the classes and doing some Data Visualisation to better understand the data we are working with, we moved to **Cleaning** the text.

Cleaning the text included removing links, hashtags, other special symbols, extra spaces, numbers, etc. We also removed **Stopwords** and applied **Stemming**.

Then to vectorize our data, we chose to use the T**F-IDF vectorizer**. We chose this method over other vectorization techniques since it is more powerful and can determine relevant words (keywords) from the data.

Then we split our data into training and validation set with a 75% training data split and moved on to trying out various classification models to benchmark results.
We started with some probabilistic classification models like Logistic Regression, Naïve Bayes, Decision Trees, etc.
Then we trained a non-linear classification model, SVC and then moved to bagging ensemble models like Random Forest, Voting Classifier between multiple classification models and a boosting ensemble model, AdaBoost.
We also tried an online learning algorithm, Passive Aggressive Classifier.

Since our problem statement included working with text, we also tried an NLP technique. This included One-Hot encoding the cleaned text and then creating the word embeddings which were used to train an LSTM model with a Sigmoid activation function, Binary Cross entropy loss function and an Adam optimiser.