# NLP ASSIGNMENT – 1

Kanishk Ramagiri-1800179C202

Likith Reddy - 1800272C202

Shashank Sai Palani - 1800183C202

Vishwak Subbareddy - 1800214C202

Anirudh Jilla - 1800137C202

* The dataset provided contains a column called Comment\_text which contains information with punctuations, spacing & numeric values. These attributes are not useful for the objective, so we cleaned these attributes from the corpus using regular expressions.
* We created separate categorical variables for each of the comment types such as toxic, severely toxic, obscene ,threat, insult , identity hate.
* We used the standard vectors pre-defined in the Glove Vectorizer by importing the file glove.6B.100d.txt of the Glove Vectorizer.
* We also got rid of stop words from the corpus since they are useless information. Doing this we were able to achieve smaller vector size which is helpful for improving the efficiency. We have used 2 kinds of vectorization methods Count Vectorizer and TFIDF Vectorizer. We have used the stop words from both the Count Vectorizer and the TFIDF Vectorizer to capture the useful words. But the Count Vectorizer has failed to differentiate between words that qualify as toxic comments and words that are important and have comparatively lesser frequency.
* The vocabulary of the vectorizer has been reduced from 1,59,571 to 21,547. Hence the dimensionality has been reduced.
* We have used classification algorithms such as Decision Tree, Random Forest both of them returned 69% and 90% accuracies respectively.
* The number estimators used in Random Forest is 50, it yields an accuracy of 90% that is better than the Decision Tree. The 50 multiple trees in Random Forest has allowed it to yield a better accuracy rate than a single Decision Tree.
* The n estimators have been decreased to 25 in Random Forest and the maximum depth in Decision Tree and Regression is 10. We observe that the accuracy has increased correspondingly to 84% and 91% respectively. Hence we can infer that the above model overfits the data.
* The accuracy rates in case of Random Forest algorithm shows lower variability with the change in n estimators value; the accuracy rate lies in the same range. But in the case of the Decision tree algorithm the accuracy shows a large difference with change in maximum depth values. However, the exact reasoning for such behaviour cannot be determined at this point yet we can identify the factor determining this behaviour, it appears to be the use of a single Decision Tree that yields a higher scale of variability compared to multiple Decision trees of the Random Forest algorithm.