# Machine Learning Project Final Evaluation

Team **Nameless**

# Previous Evaluations :

- PreProcessing : NULL ,‘?’  and duplicate values.
- Analysing Data : snsplots
- NLP preprocessing techniques
- CountVectorizer
- LogisticRegression
- GridSearchCV on LogisticRegression
- Model Selection (SVM, AdaBoost, Bagging, DecisionTrees)
- PCA

# TruncatedSVD :

- As PCA didn’t work on our data, we searched for alternatives on internet and found TruncatedSVD.
- We reduced the features using TruncatedSVD and found that the accuracy is being decreased.
- Hence, we decided to avoid TruncatedSVD and make use of all features.

# CountVectorizer Hyperparameters :

- As now we are fixed with using all features, we shifted our focus to tune perfect hyperparameters for CountVectorizer.
- We Studied and tried different combination of hyperparameter options available.
- Discovered ngram_range = (1,4) gives us best results.
- We tried increasing the range to (1,5) and observed that the session is crashing. But we expected to get better results if we are able to compute using this hyperparameter.

# LogisticRegression Hyperparameters :

- As we observed, we are getting best results only for LogisticRegression when compared to other models.
- As we are fixed with the model, we focused more on tuning the best hyperparameter combination.
- Eventually, on a trail and error basis along with some observation of chabginf accuracy by changing parameters we found the best combination to be

  (C=0.77, class_weight= {0:0.25,1:1}, solver = 'liblinear', max_iter=10000, penalty = 'l1')