In this document we will see the steps taken into account to solve this particular problem.

1. *Upload the Data set into the cloud storage.*
   - We have used cloud to take advantage of the free GPU provided by Google with Colab (GPU & Python 3)
2. *Load the training & the test data*
   - Loaded the ~ delimited datasets. [train.csv & test.csv]
3. *Clean-up the data*
   - As raw data cannot be fed into the machine learning models, we need to Vectorize the. But before vectorizing the data needs to be cleaned up.
   - To clean the data:
     a. Remove the punctuation
     b. Remove the stopwords which does not add any value in prediction.
     c. Tokenize
     d. Stem/Lemmatize the data
4. *Feature Engineering*
   - Here we tried to extract some features from the available data.
   - We extracted two features :
     1. Length of the text
     2. % of punctuation in the text
   - After evaluation of the extracted features we observe that, they hardly explained any variance towards the dependent variable. So, discarded them as we had enough important features to train with.
5. *Finalize the list of feature for modelling.*
6. *Split the data into training & test sets.*
7. *Machine Learning algorithm implementation*
   - Try multiple classification algorithms
     Like Logistic regression, Naïve Bayes, Random Forest, Gradient boosting classifier etc.
   - Select the one which gives best performance.
   - Use Grid Search CV for Hyperparameter tuning & to avoid overfitting.
8. *Model selection*
   - Generally the model that gives best prediction is selected.
9. *Make Predictions*
   - Make prediction on the unseen data(test.csv)
10. *Report the findings.*