In this task, multiple methods, which have been learnt so far, were implemented to find the optimal one.

It uses multiple kinds of dataset as well. Raw text dataset can only be used in baseline methods.

In the data processing, some rows with blank comments or nearly blank comments are deleted since these rows may potentially affect the performance of the classifiers. The datasets after deletion of rows are renamed as modified + original dataset name.

**Baseline Method:**

The first 2 baseline methods are zero-R and random baseline models respectively. These 2 methods use raw-text, tf-idf and embedding dataset.

And it also uses the modified datasets to test the performance of these 2 baseline methods.

**Naïve Bayes Model:**

Only tf-idf dataset and embedding dataset can be used in naïve bayes.

For tf-idf dataset, 3 types of naïve bayes models are used. Gaussian, Bernoulli and multinomial.

For embedding dataset, multinomial cannot be used anymore.

Corresponding modified datasets are used as well.

**K-nearest Neighbour:**

KNN with majority and KNN with inverse distancing are used.

Tf-idf dataset, embedding dataset, modified tf-idf and modified embedding are used to test the performance of KNN with 2 aspects.

**Logistic Regression:**

The maximum iteration is set as 10000.

Tf-idf dataset, embedding dataset, modified tf-idf and modified embedding are used to test the performance of logistic regression.