New Category Classification

We are given the dataset of new of various categories. There are 25k records present in the data. We combined 'headline', and 'short\_description' as 'news\_text' and retained the 'category' data column with it. Necessary text preprocessing is applied on text and vectorized using tf-idf weighting scheme. three models include SVM, Logistic Regression, and Naïve Bayes Classifier used on this data with a 75:25 ratio.

The performance results for the SVM model are as follows:

he Support Vector Classifier (SVC) model's performance is evaluated based on several metrics. The accuracy of the model is 0.4768, indicating that it correctly classifies approximately 47.68% of the data samples. The F1 score, which measures the trade-off between precision and recall, varies for different classes, with values ranging from 0.0 to 0.5614. The recall, which reflects the model's ability to correctly identify positive samples, also varies across classes, with values ranging from 0.1125 to 0.7311. Overall, the SVC model's performance is moderate, with varying levels of accuracy, F1 score, and recall for different classes, suggesting the need for further refinement and optimization to achieve better classification results..

The performance results for the Logistic Regression model are as follows:

The accuracy is 0.49456, F1 scores range from 0.08989 to 0.75806, with an average of 0.43445. Recall scores range from 0.04706 to 0.79774, with an average of 0.40809. Precision scores range from 0.0 to 1.0, with an average of 0.49883. The cross-validation accuracy is 0.49604. The confusion matrix shows the number of correctly and incorrectly predicted instances for each class. For example, class 1 has 81 correct predictions, 10 false positives, and 31 false negatives, among others. Overall, the Logistic Regression model's performance is moderate, with room for improvement in terms of accuracy, F1 scores, recall, and precision for some classes.

The performance results of Naïve Bayes Classifier, we can see the following:

The accuracy of the classifier was found to be 0.40944. The F1 score, which combines precision and recall, was calculated for multiple categories and ranged from 0.0 to 0.90056497. The recall, representing the true positive rate, varied from 0.0 to 0.90056497, while the precision, representing the positive predictive value, ranged from 0.0 to 1.0. Overall, the results suggest that the Naïve Bayes Classifier performed moderately well with room for improvement, particularly in categories with lower scores. Further analysis and optimization may be needed to enhance the performance of the classifier for specific use cases.