ROOT2AI Internship Task Report

1.Problem and my approach to solve

Text statements were given as training data, each text had corresponding Target Text. Data was unbalanced i.e. Training examples of some Targets(like ‘Blockchain’) were much more in number than others.

So, I made two different types of training data

(I) Balanced data: Training examples of all Targets were equal in number.
(II) Unbalanced data: data as it is.

Then working on each data individually, I took some significant words(e.g., words excluding punctuations, pronouns, stop words, etc.) from each Text and trained my models on those selected words. And analyzed the accuracies.

2.Model Interpretation

I used three Models(Random Forest Classifier, Support Vector Machines, Logistic Regressions) to train on and compared their accuracies in between and also corresponding to each type of data(Balanced and Unbalanced).

3.Train & Test Accuracy scores and classification report

1) Balanced Data

Balanced Data	Random Forest Classifier	Support Vector Machines	Logistic Regressions
Train Accuracy	0.994	0.984	0.994
Test Accuracy	0.500	0.522	0.434

2) Unbalanced Data

Unbalanced Data	Random Forest Classifier	Support Vector Machines	Logistic Regressions
Train Accuracy	0.987	0.875	0.979
Test Accuracy	0.623	0.635	0.496

Best Classifier along with both the data categories(Balanced and Unbalanced) as we can interpret from above table is ‘Support Vector Machines’. Hence, we’d finalize our model for training as ‘Support Vector Machines(SVM)’.

4.Limitations of the Model

- Since our data has moderate noise i.e., Target classes are overlapping hence SVM couldn’t give nicer accuracy.

- there is no probabilistic explanation for the classification, since SVM works by putting data points, above and below the classifying hyperplane.
- SVM underperforms when the number of features for each data point exceeds the number of training data samples.