Sentiment Analysis API Report

MODEL EXPERIMENTED

For Binary Classification of sentiment (Positive/ Negative or 1/0), machine learning algorithms named as LGBM, Random Forest Model, XGBoost Model and ANN was tested. Each experiment was done using hyper-parameter tuning approach which helped in finding best parameters to every model. Validation of test dataset on different splits was also practised using Cross-Validation. Data Generation/Augmentation was done and because of that data set size was increased from 11541 to 16266 which made unbalanced state of data set, balanced. Train test split was of ratio 8:2.

FINAL METRICS OF TEST DATA

Model	Accuracy	TP	TN	FP	FN	F1-Score	Precision	Recall
Random Forest	0.928	1769	1250	98	137	0.914	0.927	0.901
XGBoost	0.938	1780	1271	87	116	0.926	0.936	0.916
LGBM	0.935	1774	1267	93	120	0.922	0.932	0.913
ANN	0.944	1777	1295	90	92	0.934	0.935	0.934

On testing all above models and opting best practicing for tuning the data, Artificial Neural Network model outperformed all other ones. Model resulted in 94.4% accuracy and f1-score of 93.4%.

STEPS USED TO TUNE DATA

Step1: Data cleaning is done as the first and foremost step; removing the '@name', links, emoji characters and punctuation using Regex.

Step2: After doing data analysis, it was seen that data was unbalanced. The proportionality of negative reviews were much more than positive ones. Due to this we won't be able to find a generalize model for our prediction. To have generalized model without reducing the data, text augmentation was opted. Through this each and every row was analyzed and synonym of words were used without affecting the actual meaning of

the line. Generating new data with respect to the previous data will help our model to get more generalized results.

Step3: Now, Dataset is passed for stemming where words will be reduced to their root form.

Step4: Converting each text review into numeric form as a vector, so that it can be passed to the models.

Step5: Random Forest, XGBoost, Light GBM are tuned with the help of RandomizedSearchCV or GridSearchCV where we tuned max_depth, criterion, min_samples_split and n_estimators and on the other hand we have tuned the ANN with the help of Keras Tuner where we find the optimal number of layers and the number of neurons in each layers to get best results.

Step6: To get more generalized result for each model K fold cross validation was used.