DS-GA-1003: Fake review detection

Machine Learning project proposal

Group:

Aajan Quail	aqd215	
Abha Sahay	as13492	
Christine Shen	zs1534	
Meenakshi Jhalani	mgj265	
Viraj Thakkar	vt943	(Member responsible for submission)

Preprocessing:

For data pre-processing, we will tokenize the words, removing whitespaces and use the top 2000 words as the vocabulary and then employ TF-IDF to create feature vectors.

Model Selection:

After that, we implement logistic regression from scratch by using the hypothesis as the sigmoid function and calculating the log-loss cost function to update our parameters by using gradient descent. Then we modify the weights for the training examples in the loss function in such a way that the minority class gets penalized more as compared to the majority class in case of misclassification. Then, we implement decision trees by creating functions to calculate gini index and splits to make predictions.

Baseline models such as logistic regression tend to overfit due to their inability to capture non-linearity. Hence we decided to try some ensemble models that will provide robust and accurate results while capturing non linearity in the data. We will use models such as Light GBM and XG-Boost that have improvements over regular sequential gradient boosted machines such as histogram based splitting and Gradient based one sided sampling that lead to faster convergence. In order to mitigate any overfitting, we will try random forest to see if we can achieve a better performance by combining multiple high variance uncorrelated decision tree models to achieve a final ensemble model with high accuracy and less variance.

We will be using Random Search and Grid Search parameter optimization techniques for hyperparameter tuning and evaluation. Using Random Search, random combinations of hyperparameters are evaluated to narrow down to a range of hyperparameter values. On these ranges of values, we will use Grid Search on all possible combinations of hyperparameters and we will choose the combination with the best cross-validation score.

Suggested Experiments:

In addition to modifying our algorithms to be better fitted to handle imbalanced data, we make use of Synthetic Minority Over-sampling Technique (SMOTE), a sampling technique, to artificially create more balanced data. SMOTE will only be applied to training data to optimize our model. Since we have already chosen algorithms appropriate for imbalanced data, we will compare the outcomes of the model using both treated and untreated training data and choose the best configuration.

Evaluation:

Finally we evaluate our model based on metrics such as auROC and AP(average precision score) which are a good measure of performances for imbalanced classes.