Report for DevIncept

Problem Statement and approach :-

1. The problem here was to classify on the basis of text, if the target column belongs to any of the 11 classes (i.e. if it belonged to FinTech or it belonged to Cyber Security etc).
2. I have first tried to visualize the data and checked if there are any null values in the given Data Frame and there were 3 which I just dropped because they were very less in number.
3. I have then performed an EDA(Exploratory Data Analysis) and checked the text column if there were any unnecessary symbols for characters.
4. I have also used word Cloud to see which word repeats the most in the whole dataset where Financial ,technology ,innovation …. were the most used words
5. FinTech has the most number of appearances in the Target column.
6. I have also performed some Feature Engineering about which I will talk at the end.

Model Interpretation :-

The model I have used is Multinomial Naïve Bayes because of its potential to work with text data and multiple class labels. The Naïve Bayes algorithm is based on the principle of Conditional Probability.

Train and Test accuracy :-

The train accuracy is :- 0.9774229074889867
The test accuracy is :- 0.7077736181457829

Limitations of the model :

1. It can be easily seen that the model is overfitting the training data.
2. We could have performed Grid Search for the best value of alpha.
3. I could have used a more powerful model like LSTM

Some Extra points

1. I have created two features namely ‘sentence length’ and the ‘number of repetition of a word in the sentence’.
2. After vectorizing the text data with tf-idf vectorizer I stacked it with the with custom engineered features but since it was a sparse matrix so I had to make the custom features also sparse.
3. With the custom feature the f1 score of some of the classes increased.

Thank You