Multi Classification Model

ROOT2AI Technology Private Limited

Data Science,Machine Learning.

Prepared by Mr. Sai Gokul Krishna Reddy

1. Your thoughts on problem or what was your approach to solve the problem:

The problem is the supervised text classification problem, and our goal is to investigate which supervised machine learning methods are best suited to solve it.

This is a Multi Classification problem ,where the dataset consists of just two features one is ‘Text’ and the other one is ‘Target’ column.The classifier makes the assumption that each new complaint is assigned to one and only one category.

2. Model Interpretation:

1. **Importing all necessary Libraries**:

   Firstly I have Imported all the libraries like,pandas,sklearn,tfidfTransformer,

   CountVectorizer,matplotlib,Smoketomek for sampling.

2. **Data Collection:**

   With the help of pandas I have Imported the dataset.Which was given to me.

3. **Data Cleaning:**

   I have checked whether there are any null values or not,Then I found there are 3 null rows.So I have filled them space character instead of removing them.

   There are totally 22701 rows in the dataset,3 rows doesn’t matter at all.

4. **Data Visualization:**

   To check there are any **ImBalanced** data or not. I have used Count plot,Which helped me a lot.There I found that the tuples named “FinTech” have more than

8000 records which are completely Imbalanced dataset.

5. **Converting Text to Numbers:**

   Our Systems can only understand the number format so we need to convert the given text format to numbers.

   To do this I have used the **CountVectorizer(BagOfWords)** technique for Text column and for Target I have used label Encoding and then applied to OverSampling .

6. **Performing OverSampling:**

   Here, I have used **SMOTETomek** which will increase the data of remaining categories and make all the weight the same.Hence it will become a balanced dataset.

7. **Convert to TfIdf:**

   Once The **OverSampling** is done then we can happily convert them into vectors with the help of Tf Idf.(Which will give us the frequency count of the words in a particular sentence).Finally we got our data as a Balanced and in numeric format.

3. Train & test accuracy score:

1. **Model Selection:**

   I have used the **Naive Bayes algorithm,**Which is the best one for Textual data applications.Naive Bayes classifiers are a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, every pair of features being classified is independent of each other. To start with, let us consider a dataset.

2. **Testing:**

   Please change the **'test' list** and give your own text and check.It will definitely match with the **TARGET CLASS.**

3. **Accuracy:**

   Once the training is done,The accuracy was around 54% which was better and the results of the model are pretty much working fine**.**

4. **Confusion Matrix:**

   A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

5. **Hyper parameter Tuning:**

   Once the training is done with Hyperparameter tuning,The accuracy was around 55% which was better and the results of the model are working fine**.**

4. Limitation of the model.

The limitations of the model is accuracy score,And testing with other algorithms are left.But the results are working fine.If you want to test please change the **test list** and fill with your own text.Then execute it.

The results for me are satisfying.So please check this once.

5. You can add your own points as well

This is how I have implemented the multi classification problem.Due to some my personal issues I could only submit this.
I Can Increase the accuracy by performing other algorithms like **Passive Aggressive Classifier Algorithm and HashingVectorizer**