Final Reports

Task -1 : IMDB Movie Review Sentiment Analysis

Building a model that can predict the sentiment of IMDb movie reviews.

Entire Process:

1. Data Exploration: Exploring the data which type of data contain and what is it. Checking for null values if any handling by removing or impuring with mean, mode and median based on the dataset. For this dataset no missing values are found. Looking for imbalance, Here no imbalance distributed with same count.

2. Data Preprocessing :

- Removing non-alphabetic characters and punctuations and whitespaces from the reviews text, and as well converting to lower case.
- Applying Word Tokenization and then removing stop words form that through the list comprehension method.
- Applying Lemmatization, to reducing the words to base or dictionary form
- Converting categorical text into numerical through the TF-IDF vectorization

3. Building a model :

- Building different models such as Logistic, Decision Tree, Random Forest and Linear SVC classifiers are used to develop the model.
- Here splitted the data into training and testing.
- Trained the model and then predicted through the different models.

4. Model Evaluation :

- Evaluating the performance of the model using metrics.
- Include Accuracy of the model and classification report
- Here we got 89% accuracy by logistic model we can say the model is predicting good.
- And, by the Decision tree classifier we got 73% accuracy bit lower compare to others but ok and then through the Random forest 85% accuracy and finally by using the Linear SVC we got 89% accuracy.

Conclusion : We can prefer either SVC or logistic regression models for the predictions, here both models are predicting more accurately.

Final Reports

Task – 2 News Article Classification

Build a classification model that can automatically categorize news articles into different predefined categories.

Entire Process:

1. Data Exploration: Exploring the data which type of data contain and what is it. Checking for null values if any handling by removing or impuring with mean, mode and median based on the dataset. For this dataset some missing values are found we can remove those missing value rows, it doesn't affect the model accuracy, it's not a big deal. Looking for imbalance, slightly distributed different from one another.

2. Data Preprocessing :

- Removing non-alphabetic characters and punctuations and whitespaces from the reviews text, and as well converting to lower case.
- Applying Word Tokenization and then removing stop words form that through the list comprehension method.
- Converting categorical text into numerical through the TF-IDF vectorization

3. Building a model :

- Building different models such as Logistic, Random Forest and SVC classifiers are used to develop the model.
- Here splitted the data into training and testing.
- Trained the model and then predicted through the different models.

4. Model Evaluation :

- Evaluating the performance of the model using metrics.
- Include Accuracy of the model and classification report
- Here we got 76% accuracy by logistic model we can say the model is predicting good.
- And, by the Random forest 69% accuracy and by the SVC model we got 78% accuracy.

Conclusion : We can prefer SVC model for the predicitons, here the model is predicting more accurately than other models.