Text Analysis Threat Level Prediction

Submitted By :

ASHUTOSH CHAUDHARY

ACKNOWLEDGMENT

I would like to express my gratitude towards my internship mentor Ms. Swati Mahaseth for helping me in completion of the project.

BUSINESS PROBLEM

From my understanding the problem is about making prediction of threat level of message on the basis of the analysis of what is written in the text.

OBJECTIVE FOR PROBLEM UNDERTAKEN

We have to study the comment_text feature & need to analyse each & every text given in it & on the basis of every text's context we have to predict the level of threat every text is projecting from the content written in it.

ANALYTICAL PROBLEM FRAMING

- **Origin of dataset & data types of every features**

Dataset is provided by the company & we have to import it using various libraries necessary for the project to get completed. Also data types of features are both continuous & categorical.

- **Mathematical/Analytical modelling of the problem**

For visualization we only use two plots most of the times that were countplot & scatterplot & for model building we use Logistic Regression, Decision Tree Classification, Random Forest Classification, Bagging Classification & AdaBoosting Classification models to opt best out of them to work on dataset.

- **Assumptions related to problem statement**

No assumptions were made while working on dataset.

- **Libraries & Tools used**

We used numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickle & warnings libraries for this task.

STEPS TAKEN FOR THE TASK

1. **Importing Libraries for the task**

   Numpy, pandas, matplotlib.pyplot, seaborn, sklearn, pickles & warnings were imported for task to get completed.

2. **Importing Dataset using libraires**

   Imported both datasets using pandas library in jupyter notebook.

3. **Checking Dimension of dataset**

   By checking dimension of train dataset we get to know that it contains 159571 rows & 8 columns & test dataset contains 153164 rows & 2 columns

4. **Checking Description of dataset**

   From description we find mean, min value, max value, etc of every column which contains continuous data in them

5. **Checking for presence of null values in dataset**

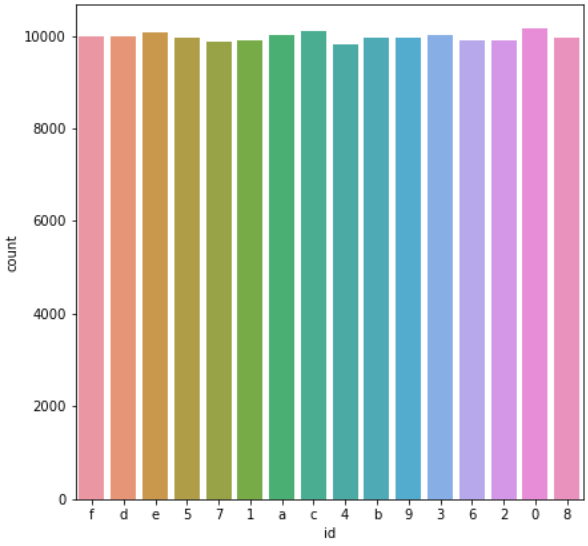
   There was no null value present in either of the datasets

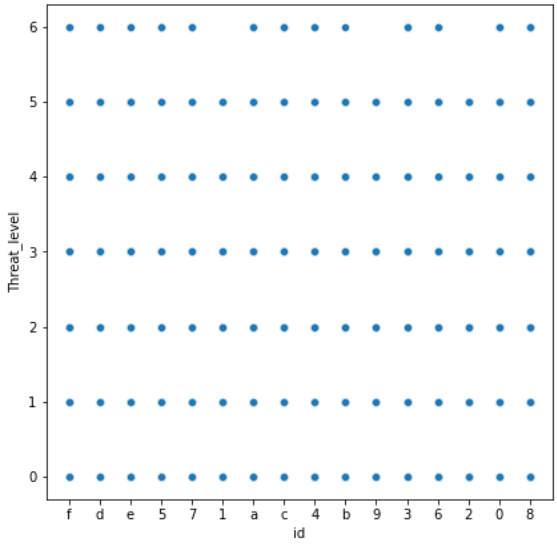
6. Creating Target Variable

By observing both datasets we get to know that there was no target variable present in the test dataset where as train dataset contains 6 target variables. So we combine all 6 of the target variables to create one target variable. And after that we dropped the other 6 variables from the dataset.

7. Performing EDA on Train dataset

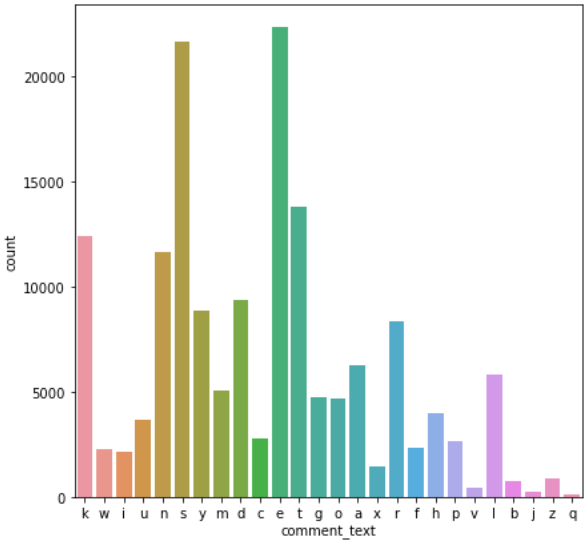
Id

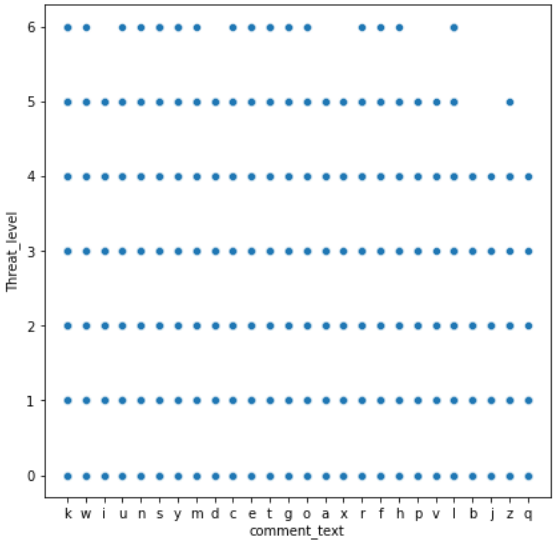




From visualization we conclude most data is present in 2 categories which are C & 0 & every threat_level is present in every category of id except of threat_level 6 which is not present in 3 categories of id which are 1, 9 & 2.

Comment_text





From visualization we concluded that most data is present in e category & every category of threat_level is present in every category of text_comment except of 2 categories of threat_level which are 5 & 6. 5th category is not present in b, j, & q category of comment_text & 6th category is not present I, d, a, x, p, v, b, j, z & q category of comment_text

8. Splitting the Train dataset

We preapared train dataset by changing data types of every feature to continuous datatype & then we scaled the train dataset so we could test every model on it using GridSearchCV to find the best suited model & then apply it on test dataset.

9. Applying GridSearchCV

We chose 5 classification models (Logisitic Regression, Decision Tree Classification, Random Forest Classification, Bagging Classification, AdaBoosting Classification) to check performance of them on the train dataset to find out the best suited model for the dataset. After applying GridSearchCV on every model we get to know that every model is giving the same train accuracy score : 0.9230 & test accuracy score : 0.9207. So we chose model randomly which was Decision Tree Classification model with hyper parameter combination (max_depth : 1, min_samples_leaf : 1, min_samples_split : 2)

10. Preparing Test dataset

After chosing the model we prepare the test dataset, so, we could apply model on it to make prediction

11. Applying Model on dataset

Chose Decision Tree Classification Model & made it ready by tuning hyper parameter. After tuning hyper parameter we applied model on dataset to make prediction. After making prediction we made dataframe of it & concatenate it with test dataset. Also we could not check its performance as there is no target variable present in test dataset.

12. Saving Model

After making prediction we saved the model using pickle library.

CONCLUSION

In the end we conclude that the model is quite accurate as it was giving the train accuracy score of 0.9230 & test accuracy score of 0.9207.