Malignant Comment Classification

Submitted By

Anita Tarangi

**INTRODUCTION**

* Business Problem Framing

To build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

* Conceptual Background of the Domain Problem

Machine Learning helps to find out the malignant comments and by which we can train a model to control or restrict the abuse.

* Review of Literature

The dataset had no NULL values, cleaned the data, carried out different methods to train the model.

**Analytical Problem Framing**

* Identification of possible problem-solving approaches Logistic Regression, Decision Tree, Random Regressor, XGB Regressor, AdaBoost Classifier, KNeighbours Classifier
* Data Sources and their formats
* There were no null values in the dataset.
* The dataset was not clean
* For some features, there may be values which might not be realistic. You may have to observe them and treat them with a suitable explanation.
* You might come across outliers in some features which you need to handle as per your understanding. Keep in mind that data is expensive, and we cannot lose more than 7-8% of the data.
* Data Preprocessing Done
* **Checking the size of the dataset**
* **Checking the summary statistics of the dataset**
* **Checking the data types and null values**
* **Dropping columns which aren’t important**
* **Converting categorical columns using dummy method.**
* Hardware and Software Requirements and Tools Used
* Python, Pandas, Seaborn, Matplotlib, sklearn, nltk, Model Selection, Logistic Regression, Decision Tree, Random Regressor, XGB Regressor, AdaBoost Classifier, KNeighbours Classifier

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Logistic Regression, Decision Tree, Random Regressor, XGB Regressor, AdaBoost Classifier, KNeighbours Classifier

* Run and evaluate selected models
* Logistic Regression, Decision Tree, Random Regressor, XGB Regressor, AdaBoost Classifier, KNeighbours Classifier.
* Visualizations

Seaborn, Matplotlib, nltk

* Interpretation of the Results
* Carried out Analysis as well as visualization of the Dataset, treated imbalanced dataand then, trained the model using Logistic Regression, Decision Tree, Random Regressor, XGB Regressor, AdaBoost Classifier, KNeighbours Classifier, model predicted with different accuracy for every model, out of all the models, Random Forest, Logistic Regression, have achieved with 96% accuracy rate.

**CONCLUSION**

* Data exploration, cleaning, visualization is the basic steps, inconsistent data was identified and treated. And then, trained the model using Logistic Regression, Decision Tree, Random Regressor, XGB Regressor, AdaBoost Classifier, KNeighbours Classifier, model predicted with different accuracy for every model, out of all the models, Random Forest has achieved with 96% accuracy rate.