FLIP ROBO

Malignant Comments Classification Project

Submitted by:

LALIT JOSHI

ACKNOWLEDGMENT

INTRODUCTION

- ## Business Problem Framing
    1. Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users.
    2. Our goal is 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 cyber bullying.

- ## Conceptual Background of the Domain Problem
    1. Initially we will check and classify the problem (Regression or Classification).
    2. We will also check whether our dataset is balanced or imbalanced. If it is an imbalanced one, we will apply sampling techniques to balance the dataset
    3. Then we will do model building and check its accuracy. Our main aim is to build a model with good accuracy.

- ## Review of Literature
    Below is the summary of the research done to build this model.

    1. Firstly, I imported all the necessary libraries required to build the project/model.
    2. Then, I created the data frame for both train and test data.
    3. I checked the basic details of both the datasets like: checking information, looked for the null values, description of the dataset.
    4. Once these steps are done I proceed further with data visualization by using heatmap for the correlation among the data, count plots to check the value counts, converting text to vectors using TD-IDF and then splitting the data for testing and training the model.

- # Motivation for the Problem Undertaken

  The objective behind to make this project is 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 cyber bullying.

Analytical Problem Framing

- # Mathematical/ Analytical Modeling of the Problem

  If you look at data science, we are actually using mathematical models to model business circumstances, environment etc and through these models, we can get more insights such as the outcomes of our decision undertaken, what should we do next or how shall we do it to improve the odds. So mathematical models are important, selecting the right one to answer the business question can bring tremendous value to the organization.

  I have used 5 machine learning algorithms for model building, out of which Random Forest Classifier is giving us the best accuracy which is:

  ```
  Training accuracy is 0.9987018684142204
  Test accuracy is 0.9558823529411765
  ```

- # Data Sources and their formats

  Data Source: The read_csv function of the panda's library is used to read the content of a CSV file into the python environment as a pandas Data Frame. The function can read the files from the OS by using proper path to the file.

  Data description: Pandas describe () is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values.

- # Data Preprocessing Done

  I have checked the correlation between the target variables.

  ♣ I have studied the statistical summary of the train and test dataset.

  ♣ I have done some visualization using count plots.

♣ Then I have converted the comments in train data into lowercase.

♣ Then I have removed punctuations, stop words to get a clean length.

♣ Then I have converted text into vectors using TF-IDF so that the data gets ready for model building.

♣ I have splitted the independent and dependant variables into x and y.

- ## Hardware and Software Requirements and Tools Used

Hardware Requirements:

Processor: Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz (4 CPUs), ~2.7GHz

RAM: 8192MB

Software Requirements:

Python: Programming language

Tools used:

- Jupyter notebook: NumPy
- Pandas
- Matplotlib
- Seaborn
- Scikit-learn
- SciPy. Stats

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)
    1. To check the correlation among the data, I have used heatmap to visualize it.
    2. To get a clear view of the columns visually, I have used count plots.
    3. For training and testing the data, I have imported train_test_split library from scikit-learn.

4. For model building, I have used 5 algorithms, out of which Random Forest Classifier gives the best accuracy.

- Testing of Identified Approaches (Algorithms)
  1. Logistics Regression.
  2. Decision Tree Classifiers
  3. Random Forest Classifiers
  4. AdaBoost Classifiers
 5. KNeighbors Classifiers

- Run and Evaluate selected models

  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

- Key Metrics for success in solving problem under consideration

  The key metrics used are accuracy_score, confusion_matrix and classification_report.

- Visualizations
  1. **PIE CHART**: Showing the coverage of all the malignant words in the dataset.

Label distribution over comments
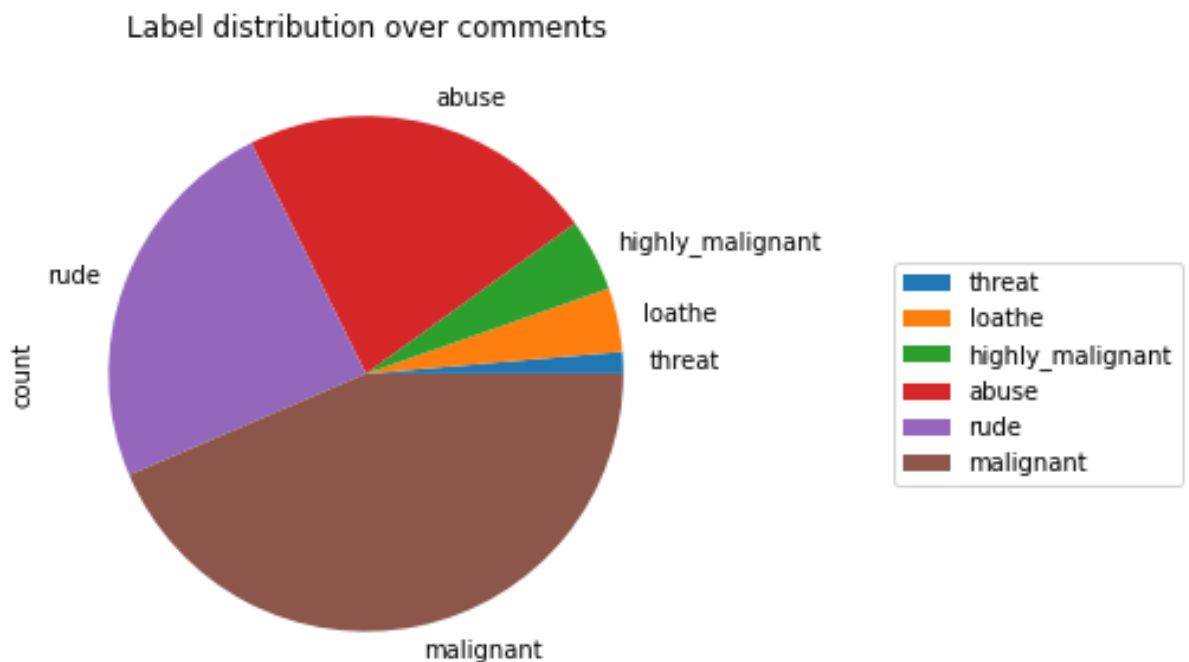
- Interpretation of the Results

    2. **VISUALIZATION**: I have seen how my data looks like using heatmap, count- plots, pie- plot etc.
    3. **PRE-PROCESSING**: I have cleaned my data using some NLP methods and other methods too.

4. **MODELLING:** I have designed our model using algorithms like Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier and KNeighbors Classifier. The accuracy, f1 score, confusion_matrix, classification_report is achieved for each model

CONCLUSION

- Key Findings and Conclusions of the Study

    The key findings are we have to study the data very clearly so that we are able to decide which data are relevant for our findings. The conclusion of our study is we have to achieve a model with good accuracy and f1-score.

- Learning Outcomes of the Study in respect of Data Science

    We will develop relevant programming abilities. We will demonstrate proficiency with statistical analysis of data. We will develop the ability to build and assess data-based models. We will execute statistical analyses with professional statistical software. The best algorithm for this project according to my processing is Random Forest Classifier because the accuracy and f1 score that I have achieved is quite satisfactory than the other model.