Malignant Comments Classifier Project

Submitted by:

Pavankumar Bolli

ACKNOWLEDGMENT

I would like to express my special thanks to flip robo technlogies for given me such a wonderfull of apportunity to explore and insite work on a data science projects.

And and also thanks to Datatraiend institute's mentors to motivate and understand the concepts of Data science. I have got lot of help from their resourses and concept to complete these project.

Link that helps me to complete this project

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

INTRODUCTION

- Business Problem Framing

The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

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. This means that insults to third parties such as celebrities will be tagged as unoffensive, but “u are an idiot” is clearly offensive.

- Conceptual Background of the Domain Problem

  D The Data is provided from the company which consist of 8 columns i.e Id Comment-text, malignant, heighly_malignant, rude, threat, abuse and loathe. Id is a unique number to indentify the comment text id and Comment_text is a comment which is a text format people wrote a comments in there respective languages. And rest of columns are in binary formate 1/0 1 represents rudeness or threat typeis present 0 represent not present in all the binary columns. The shape of the data set is 159571 observations and 8 columns.

- Motivation for the Problem Undertaken

  Describe your objective behind to make this project, this domain and what is the motivation behind.

  Motivation of this project is to determine how the comments are offensive, abusing, hatefulness, aggression and threat. So this can help the people to report or bolck those people and take an action on that people.

Analytical Problem Framing

- Data Sources and their formats

  In this project i've data almost 1 lack 59 thousand rows. Due to large data set i have used Vaex package from this data is divided into chunks and it will create a new data set file with an extension hdf5. Beacuse of this useage of RAM will decrease and we can load entire dataset and work on it quickly. Vaex is nothing but a package like a pandas which is better than pandas beacause of memory management technique.

- Data Preprocessing Done

  In this project each data is in text formate machine can understand only numbers. So for this i have used tool kit of NLP (Natual Language Processing) that is NLTK. In the part of data cleaning i have removed stopwords, punctuation, emogis, numbers extra spaces, words length have less than 2, and removed email addresses etc. Because this type of data is not use full for the prediction rating this type of words are increasing the length of the text. After cleaning the data i have stem each words to give the base words and remove plural's and tenses. For stemming the words i have used Lemmatization package which is in NLTK. After all these steps i have converted text to numbers by using TFIDF and Count Vectorizer

- ## Data Inputs- Logic- Output Relationships

  Comments are in text formate and in any certain languge’s. Based on the words present in teh comment model will predict that comment is malignant, highly malignant, threat, rude etc. Suppose comments have word’s like “kill” “muder” “hack” so model will predict hou much abuse or malignant have in the text.

- ## Hardware and Software Requirements and Tools Used

  Tools and libraries that i have used to solve project:

  - Software: Anaconda, Jupyter Notebook, Python3, Google Colab.
  - Libraries: Numpy, Pandas, Matplotlib, Seaborn, Sklearn.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  The data set contain lable data and target data lable has cleaned by some EDA steps. Target data is in the form of classified i.e (How much malignant). By seeing the our target column we can conclude that this problem comes under supervised and Classification Problem. Target column having in factor . For more testing SVC, Decision Tress Classifier, KNN algorithms have used. Rather testing on one by one algorithm and choosing best alogithm and hyper tuning that best model. I have hyper tunned all the algorithm and added into pipeline so i will get best algorithm along with the best parameters.

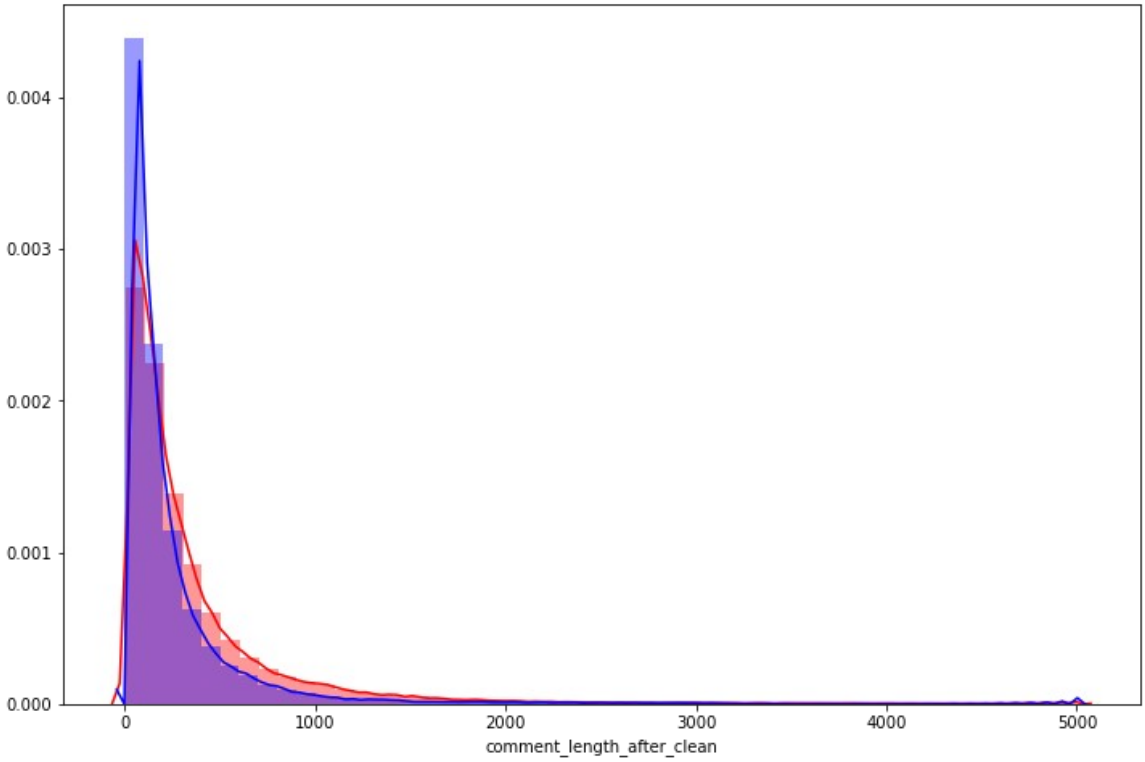
- Testing of Identified Approaches (Algorithms)

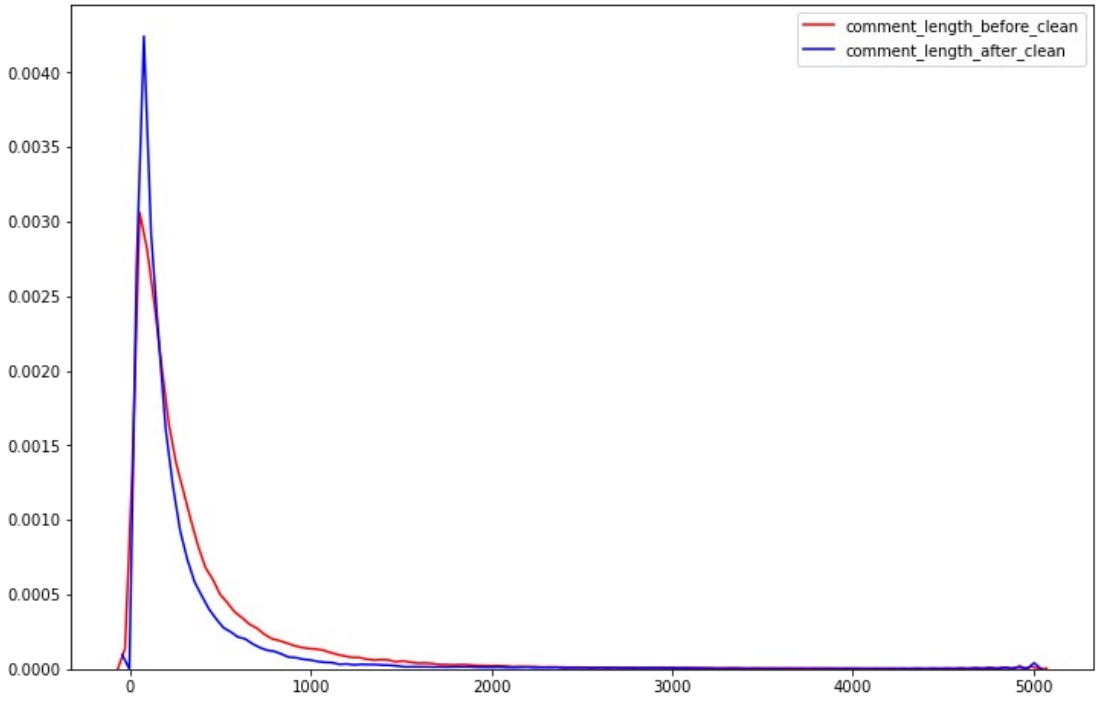
  For testing how good model is generated. I have splitted data set into two parts (70% 30%) and each part is splited into further two part that are lable data and target data. First 70% splited data is used for training the model and another rest of 30% used for Prediction testing the model how good model is predicted in terms of percentage.

- Run and Evaluate selected models

  I have runn two model that is multinomina and random forest. Multinominal works better when the features re independent to each other and random forest is used beacuse of it is one of ensemble technique this model can also work better when the data set have imbalace.

- Visualizations




comment_length_before_clean
comment_length_after_clean
0.0040
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
1000
2000
3000
4000
5000

CONCLUSION

- Learning Outcomes of the Study in respect of Data Science

  I have learn some inbuild funtion of python at the time of data cleaning. Object columns to Integer are done by using python map function over on lambda function. The main challenging is that to run all model once and get the best model name with best parameters also with a AUC ROC curve plot. This is

done by using pipeline. This will helps me to get out from overfitting and underfitting.