MALIGNANT COMMENTS CLASSIFICATION

Submitted by:

MUHAMMED MARJAN P

ACKNOWLEDGMENT

Online platforms when used by normal people can only be comfortably used by them only when they feel that they can express themselves freely and without any reluctance. If they come across any kind of a malignant or toxic type of a reply which can also be a threat or an insult or any kind of harassment which makes them uncomfortable, they might defer to use the social media platform in future. Thus, it becomes extremely essential for any organization or community to have an automated system which can efficiently identify and keep a track of all such comments and thus take any respective action for it, such as reporting or blocking the same to prevent any such kind of issues in the future.

This is a huge concern as in this world, there are 7.7 billion people, and, out of these 7.7 billion, more than 3.5 billion people use some or the other form of online social media. Which means that every one-in-three people uses social media platform. This problem thus can be eliminated as it falls under the category of Natural Language Processing. In this, we try to recognize the intention of the speaker by building a model that's capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate. Moreover, it is crucial to handle any such kind of nuisance, to make a more user-friendly experience, only after which people can actually enjoy in participating in discussions with regard to online conversation.

INTRODUCTION

- Business Problem Framing

  Given a number of comments, sentences or paragraphs being used as a comment by a user, our task is to identify the comment as whether it is a malignant comment or no. After that, when we have a collection of all the malignant comments, our main task is to classify the tweets or comments into one or more of the following categories – 'Highly malignant', 'Rude', 'Threat', 'Abuse' and 'Loathe'.This problem thus comes under the category of multi-label classification problem.

- Conceptual Background of the Domain Problem

  There is a difference between the traditional and very famous multi-class classification, and the one which we will be using, which is the multi-label classification. In a multi-class classification, each instance is classified into one of three or more classes, whereas, in a multi-label classification, multiple labels (such as –Highly

malignant, Rude, Threat, Abuse and Loathe) are to be predicted for the same instance.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Python is widely used in scientific and numeric computing:
  SciPy is a collection of packages for mathematics, science, and engineering.
  Pandas is a data analysis and modelling library.
  Python is a powerful interactive shell that features easy editing and recording of a work session, and supports visualizations and parallel computing.
  The Software Carpentry Course teaches basic skills for scientific computing, runningbootcamps and providing open-access teaching materials

- Data Sources and their formats

Data selection is defined as the process of determining the appropriate datatype and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity.

	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	0

- Data Preprocessing Done
- The dataset consists of the following fields-
- id: An 8-digit integer value, to get the identity of the person who had written this comment
- comment_text: A multi-line text field which contains the unfiltered comment
- malignant: binary label which contains 0/1 (0 for no and 1 for yes)
- highly_malignant: binary label which contains 0/1
- rude: binary label which contains 0/1
- threat: binary label which contains 0/1
- abuse: binary label which contains 0/1
- loathe: binary label which contains 0/1
- Out of these fields, the comment_text field will be preprocessed and fitted into different classifiers to predict whether it belongs to one or more of the labels/outcome variables (i.e. toxic, severe_toxic, obscene, threat, insult and identity_hate).
- We have a total of 151203 samples of comments and labelled data, which can be loaded from train.csv file.
- Data Inputs- Logic- Output Relationships

The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.

- Hardware and Software Requirements and Tools Used

Python is widely used in scientific and numeric computing: SciPy is a collection of packages for mathematics, science, and engineering. Pandas is a data analysis and modelling library. Python is a powerful interactive shell that features easy editing and recording of a work session, and supports visualizations and parallel computing. The Software Carpentry Course teaches basic skills for scientific computing, runningbootcamps and providing open-access teaching materials

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

I have first gone through the train dataset, firstly I had clean the dataset and train it. Then I gone through test dataset and clean it and predict the outcome.

- Testing of Identified Approaches (Algorithms)
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier
  - xgBoost
  - AdaBoost Classifier
  - KNeighbors Classifier

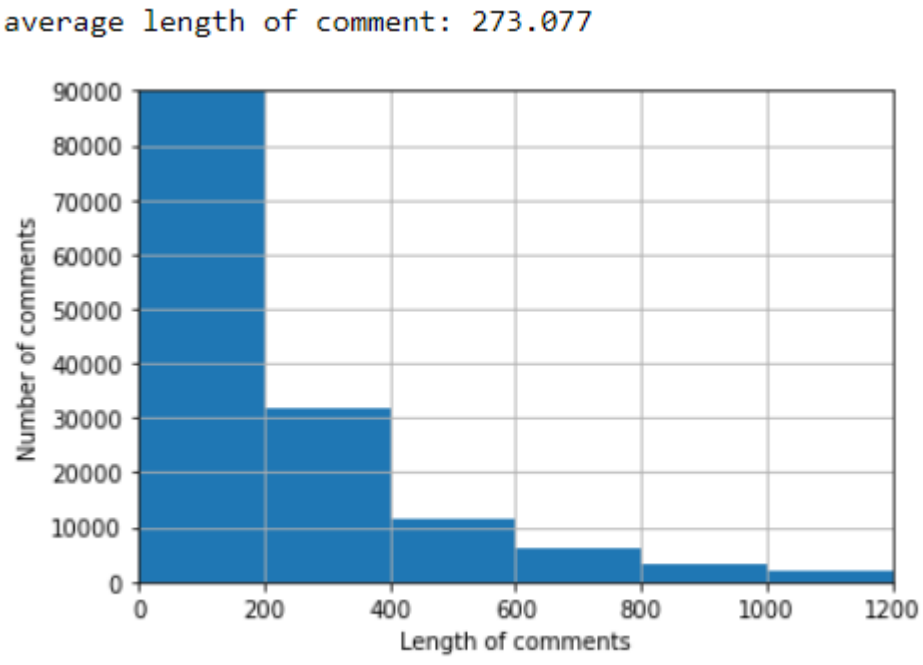
- Run and Evaluate selected models

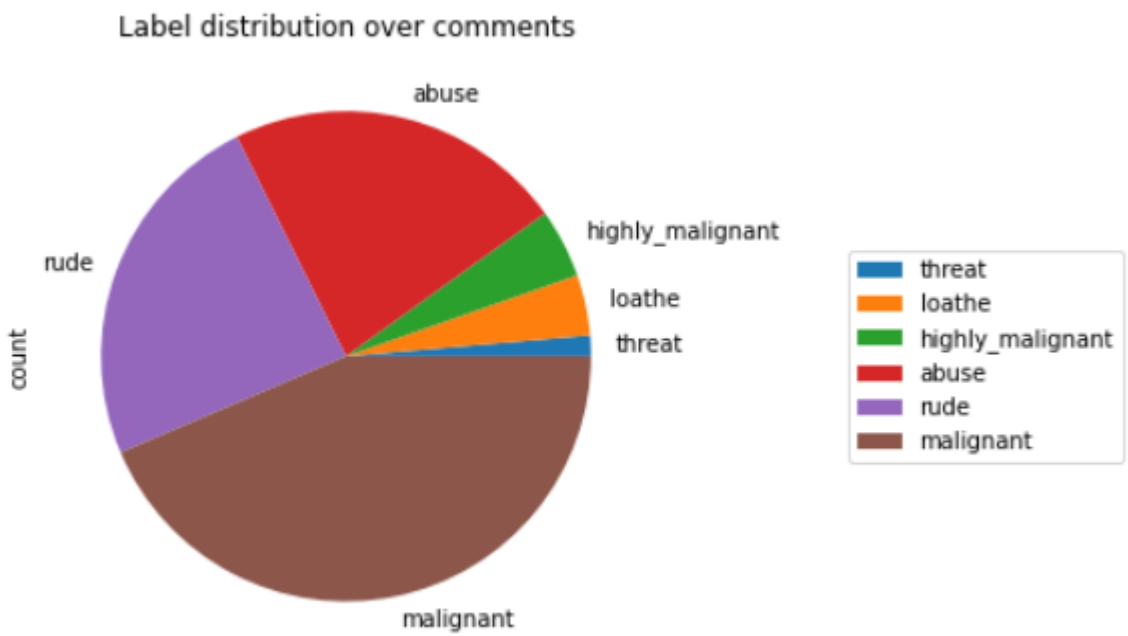
```
print(classification_report(y_test,y_pred_test))
```

```
Training accuracy is 0.9988898736783678
Test accuracy is 0.9549841243315508
cross validation score : 95.70034599895982
[[42403   547]
 [ 1608  3314]]
              precision    recall  f1-score   support

           0       0.96      0.99      0.98     42950
           1       0.86      0.67      0.75      4922

    accuracy                           0.95     47872
   macro avg       0.91      0.83      0.86     47872
weighted avg       0.95      0.95      0.95     47872
```

7]:
```
#Plotting the graph which tells us about the area under curve , more the area under curve more will be the better prediction
# model is performing good :
fpr,tpr,thresholds=roc_curve(y_test,y_pred_test)
roc_auc=auc(fpr,tpr)
plt.plot([0,1],[1,0],'k--')
plt.plot(fpr,tpr,label = 'RF Classifier')
plt.xlabel('False positive rate')
plt.ylabel('True positive rate')
plt.title('RF CLASSIFIER')
plt.show()
```

- Visualizations





CONCLUSION

→ Although we have tried quite several parameters in refining my model, there can exist a better model which gives greater accuracy. → Yes. We were unable to find a clear implementation of the Adaboost.MH decision tree model which we had planned to use. The scikit-multilearn library doesn't even mention of such a model. Also, the research papers were a little vague regarding implementation details.