Malignant Comments Classification Project

Submitted by

Rohan V Borade

ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot. I am also grateful to Mr Shubham Yadav for his constant guidance and support.

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.

Therefore, 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 cyberbullying.

CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM

Internet is one of the important inventions and large number of peoples are its users. The usage of the internet differs from person to person. There are different types of social media, which are preferred by user for entertainment purpose. The user can post or spread the news through these online platforms. These platforms do not verify the users or their posts. So, some of the users for fun spread online hate. There has been a remarkable increase in the cyberbullying

and trolls on various social media platforms. Many celebrities and influencers are facing backlashes from the people and have to come across the hateful and offensive comments. This can take a toll on any one affect them mentally leading to depression, metal illness, self-hatred, suicidal thoughts.

REVIEW OF LITERATURE

Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet aggression and hate of 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 insult to the third parties such as celebrities will be tagged as inoffensive but “You are an idiot” is clearly offensive.

Our goal is to build a prototype of online hate and abuse comments classifier. Which can be used to classify the hate and offensive comments so that it can controlled and restricted from spreading hatred and cyberbullying.

There is a difference between traditional and multi classification, and the one which will be using, which is the multi label classification. In multi label classification, each instance is classified into two or three or more classes, whereas in multi label classification, multiple labels (such as – toxic, obscene, threat, insult or identity hate) are to be predicted for the same instance.

Multiple ways are there to approach this classification problem. It can be done – using multi label method which belongs to problem transformation category: Label Power Set (LP), Binary Relevance (BR), BR+, Classifier Chain.

Base and adapted algorithms like (Decision Tree), Navie Bayes, K-nearest Neighbor (KNN), SMO (Support Vector Machines).

Further, out of the total dataset used for experimenting these algorithms, 70% was used for training and 30% was used for testing data set. Each testing datasets was labelled and thus for each algorithm using the predictions and labels, calculation of hamming loss, accuracy and log loss was done. The final results have been complied on the basis of values obtained by algorithmics models in hamming loss combined.

Motivation for the Problem Undertaken

The project first was provided to me by FlipRobo as a part of internship program. The exposure to real world data and the opportunity to deploy my skillset in solving real world problems has been the primary objective. However, the motivation for taking this project was that it is relatively a new field of research. Here we have many options but less concrete solutions The main motivation was to classify the news in order to bring awareness and reduce unwanted chaos and make good model which will help us to know of miscreants.

Analytical Problem Framing

Anyone can be victim of hate or cyberbullying. The social media has become dangerous place to dwell in. The use abusive language, aggression, cyberbullying, hatefulness, insults, personal attacks, provocation, racism, sexism, threats or toxicity have significantly high negative impact on individual. We can use Machine Learning and NLP Technologies to deal with such toxic comments.

We were provided with two different datasets. One for training and another to test the efficiency of the model created using training dataset. The training dataset provided here has a shape of 159571 row and 8 columns. As it is a multi-class problem it has 6 dependent / target column. Here the target columns or dependent variables named as "Malignant, Highly Malignant, rude, threat, abuse, loathe" have two distinct values 0 and 1. where has 1 is represents as yes and 0 represents no for each class. As the target columns are giving binary output and all the independent variables has text so it is clear that it is a supervised machine learning

problem where we can use the techniques of NLP and classification-based algorithms of Machine Learning

Here we will use NLP techniques like word tokenization, Lemmatization, stemming and tf-idf vectorizer then

those processed data will be used to create best model using various classification based supervised machine

learning algorithms like Logistic Regression, Passive Aggressive Classifier, Multinomial NB, Complement NB with

help of One Vs Rest Classifier Which is helpful to deal with multilabel classification problems. The passive Aggressive Classifier belongs to the family of online Machine Learning algorithms and it is very much helpful in processing larger scale data. It remains passive for a correct classification and turns aggressive in case of misclassification. It aims is to make updates that corrects the loss causing very little change in the weight vector.

Data Source and Formats

The data was provided by FlipRobo is CSV format. After loading the training dataset into Jupyter Notebook using pandas and using df.head() it can be seen that there are eight columns named as "id, comment text, Malignant, highly malignant, rude, threat, abuse, loathe".

Similarly, the test file can be load using pandas and first five rows of the dataset can be seen using df.head() method. As mentioned earlier the shape of the training dataset is (159571,8) and the shape of test dataset is (153164,2). The shape of the datasets in form of a tuple can be accessed using df.shape(). The column names of the datasets in form of a list can be seen using df.columns.values()

The datasets have no duplicated values or null values. Both the dataset has no trace of any null or duplicated values.

The number of duplicated values of a dataset can be seen using df.duplicated(). sum () and the null values can be seen using df.isnull(). sum (). The null values can also be visualized

with help of seaborn and matplotlib library. Visualization gives a better idea.

Data Pre-Processing Done

After the dataset is loaded and the shape, null values and duplicated values were checked then the data-set is further treated where the unwanted column "id" is removed from the training dataset as we will work on the columns like "comment_text, Malignant, highly_malignant, rude, threat, abuse, loathe".

So, a copy of the training dataset was made using df.copy() and the column was dropped from the new dataset using df.drop().

Similarly, the 'id' column is also dropped from the test dataset. After removing the unwanted column, a new column named 'normal', was created in the training dataset which represents the statements not falling under malignant, highly malignant, rude, threat, abuse, loathe category.

After the new column 'normal' was added and unwanted column 'id', was dropped a new column named 'raw length of the 'comment_text' column is added to the dataset. it will help to know the length of the 'comment_text' columns before pre-processing and later a new column will be created to compare the length of string before and after pre-processing

```
[ ] # Adding all the outputs. If the sum is 0 indicating non-offensive comments and sum > 0 indicating malignant comments
    Target = []
    for i,j,k,l,m,n in zip(malignant,highly_malignant,rude,threat,abuse,loathe):
        Target.append(i+j+k+l+m+n)
```

```
[ ] from collections import Counter
    print(dict(Counter(Target)))

    {0: 143346, 4: 1760, 1: 6360, 3: 4209, 2: 3480, 5: 385, 6: 31}
```

Here we can see that the dataset contains mostly outputs with 0 i.e. non-malignant comments and if Target > 1 indicating that the comment falls in more than one category such as rude, malignant, loathe, abuse etc

```
#Changing Target with value more than 1 into 1 so that we have a binary classification problem:- malignant or non-malignant
Target2= []
for i in Target:
    if i == 0:
        Target2.append(0)
    else:
        Target2.append(1)
```

```
print(dict(Counter(Target2)))

{0: 143346, 1: 16225}
```

Here Label "1" are offensive comments and "0" are non-offensive

Here Label 1 are offensive comments and 0 are non-offensive to create it into binary classification we have converted 6 dependent columns into label 1 because in the dataset mostly output with 0 I.e., which is considered has Non-Malignant comments, if Target is greater than 1 indicating that the comments fall in more than one category such as rude, malignant, loathe, abuse etc. Wrote a function for word clouds to take care of the Captial letters using token.lower(), also stop word are defined in the function.

Then I have changed Target with value more than 1 into 1 so that we have a binary classification problem Malignant or Non-Malignant using list and running the if loop I equal to equal to zero append it in the Target2 and else append one in the Target2 list, at last step replaced empty Target column with list Target2 now it contains label 1 as offensive comments and 0 as Non offensive comments. Plotting Word

Cloud using the function defined for a sample of non-malignant comments with Target = 0, using

Word_Cloud(list(train_df[train_df == 0].sample (n = 15000).comment_text)). Taking only 15000 samples for the plot since Target = 0 has 143346 records which will take huge processing time for the plot.

Plotting the Word_Cloud(list(train_df[train_df.Target ! = 0].comment_text))

Then splitting train data into training and validation dataset. After splitting data then using Tf-idf to convert text data into numerical format for machine learning models.

State the set of assumptions (if any) related to the problem under consideration

There was no presumption for this problem taken by me.

Hardware and Software Requirements and Tools Used

Hardware Laptop Hp Notebook

OS Window 10 Home basic

Processor Intel Core i3-4005U processor

RAM 4 GB

Language Python 3.8

IDE Jupyter Notebook

Packages Pandas, NLTK, Sklearn, Matplotlib, Seaborn

Identification of possible problem-solving approaches (methods)

After Tf-idf implementation array conversion we have x and y for modelling. Then x and y were split for training and testing using train_test_split in a ratio of 80:20 respectively. After split the shape of x_train and x_test found be

(111699,10000) and (47872,1000) and y_train and y_test found to be (111699,7) and (47872,7) respectively.

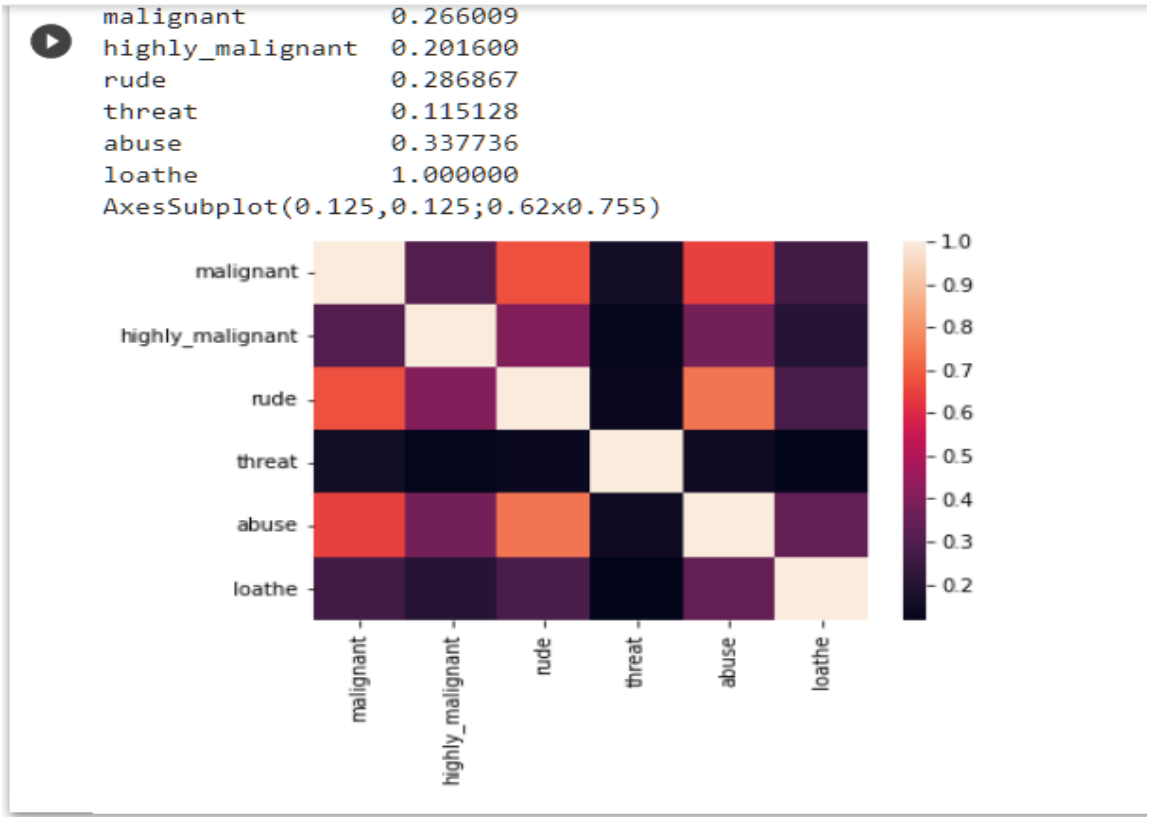
Testing of Identified Approaches (Algorithms)

As it is multi-label classification problem, we will use Classifier from sklearn Library.

1. Logistic Regression ()
2. Ada Boost Classifier ()
3. Random Forest Classifier ()

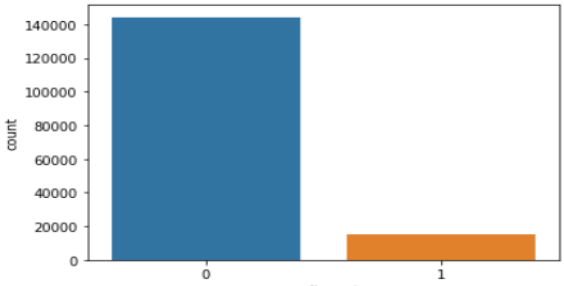
Visualizations

Visualization plays a crucial role in EDA as well as during modelling. It gives a better idea about the things going on beautifully. Below are the few visualizations used during this project to understand the dataset and performance of the algorithms.



```
[18]  col = ['malignant','highly_malignant','rude','threat','abuse','loathe']
      for i in col:
        print(i)
        print(train_df[i].value_counts())
        sns.countplot(train_df[i])
        plt.show()
```

```
0    144277
1     15294
Name: malignant, dtype: int64
```




```
[27] #Plotting Word Cloud using the function defined for a sample of non-malignant comments with Target = 0
     Word_Cloud(list(train_df[train_df.Target == 0].sample(n=15000).comment_text))
     #Taking only 15000 samples for the plot since Target=0 has 143346 records which will take huge processing time for the plot
```

```
[31] # Plotting Word Cloud for all types of Offensive comments
     for cols in Malignant_Columns:
         print("Comment_Type :",str(cols).upper(),"    Count :",len(train_df[train_df[cols] != 0]))
         Word_Cloud(list(train_df[train_df[cols] != 0].comment_text))
         print("_____________________________________________________________________________________")
         print()
```

Interpretation of the Results

Basing on the result obtained 'Random Forest Classifier' have performed well and has given better result as compared to other models so it has been selected as final model and it will be saved using joblib library.

```
[ ] joblib.dump(prediction3,"Model")
```

CONCLUSION

Key Findings and Conclusions of the Study

From the above analysis the below mentioned results were achieved which depicts the chances and conditions of a comment being a hateful comment or a normal comment,

With the increasing popularity of social media, more and more people consume feeds from social media and due differences they spread hate comments to instead of love

and harmony. It has strong negative impacts on individual user and broader society.

Learning Outcomes of the Study in respect of Data Science

It is possible to classify the comments content into the required categories of authentic and however, using this kind of project an awareness can be created to know what is fake and authentic.

Limitations of this work and Scope for Future Work

Every effort has been put on it for perfection but nothing is perfect and this project is of no exception. There are certain areas which can be enhanced. Comment detection is an emerging research area with few public datasets. So, a lot of works need to be done on this field.