PROJECT REPORT

ON

MALIGNANTS

PROJE CT

SUBMITTED BY:-

HITESH KUMAR SHARMA

ACKNOWLEDGMENT

Reference :-

I got references about this project from Flip Robo Technologies pvt.ltd .These provide an AI and software development company catering the new age demands of the industry and this company also tie up to Datatrained institute which also provide data science and other courses and this company provide 6 month internship for PG program of data science.

This company website – https://www.fliprobo.com

Guided to Datatrined Academy :-

This academy provide many practice project , evaluation project and understanding concept and detailed discussion .I have learned all insight of the data and use methods and technique.

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, cyber bullying, 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 cyber bullying 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.
- I have used train and test both data to merge and doing all preprocessing steps and in this project basically malignants classifier problem.
- I have used natural language processing and apply to many methods and technique to solve this problem and identify easily to understand and build the best model then predicting the new textual data from comment then identifying malignant class.

Conceptual the Domain Problem:-

This data has classification problem for malignant column which has two class to identify for malignant class to comment data.

In this type classification problem to solve NLP (natural language processing) task and different methods and technique used for NLP task to solve and this data clean and process to easily readable and understandable.

More column and rows are used to this shape (312735, 2) and in this data output column is malignant in this two class is present.

I will create the best model using the input/output variables in this dataset then by using the best model, I will be able to perform better prediction in the new input testing data.

Review of Literature:-

I have used input columns like that one is comment_text and output label column malignants given.

I have used many attribute /methods, visualization methods, some pipeline, some important technique for NLP task also use in this session.

Motivation for the problem undertaken:-

- ❖ Firstly collect the both data train and test
- ❖ Clean the review text data using all NLP pre – processing
- ❖ Using machine learning models
- ❖ All text data to convert vector form using tfidfvectorizer
- ❖ These models checking cross val score and choose the best model
- ❖ Clean the review text data using different method
- ❖ Using word cloud library and sentiment intensity analyzer for NLP task
- ❖ Using basic EDA process and cleaning methods for easily to understand the textual content.
- ❖ After using pre – processing pipeline technique.

- ❖ After that building machine learning models
- ❖ Then choosing best model and predict the class for new test data.

Analytical Problem Framing

Mathematical/Analytical Modeling of the Problem:-

In this project , I have used mathematical analysis for some attribute like that shape, columns, attribute count, datatypes, columns name using info() methods, groupby() ,value_counts() ,unique() ,nunique() ,rename() methods dtypes and methods like that describe(),info(), check the null values etc. and another important technique.

Using Input /output variables in this dataset and choosing the best model, we will be able to perform better prediction in the new input data.

Data Sources and Their Format:-

mg

	comment_text	malignant
0	Explanation\nWhy the edits made under my usern...	0.0
1	D'aww! He matches this background colour I'm s...	0.0
2	Hey man, I'm really not trying to edit war. It...	0.0
3	"\nMore\nI can't make any real suggestions on ...	0.0
4	You, sir, are my hero. Any chance you remember...	0.0
...	...	...
153159	. \n i totally agree, this stuff is nothing bu...	NaN
153160	== Throw from out field to home plate. == \n\n...	NaN
153161	" \n\n == Okinotorishima categories == \n\n I ...	NaN
153162	" \n\n == ""One of the founding nations of the...	NaN
153163	" \n :::Stop already. Your bullshit is not wel...	NaN

312735 rows × 2 columns

Data Pre- Processing Done:-

In this, I have used multiple methods, attributes so that we can understand the data well such as –

Shape – check that how many row and columns are there in this dataset. This dataset input columns 1 and output column is 1 and 312735 rows use.

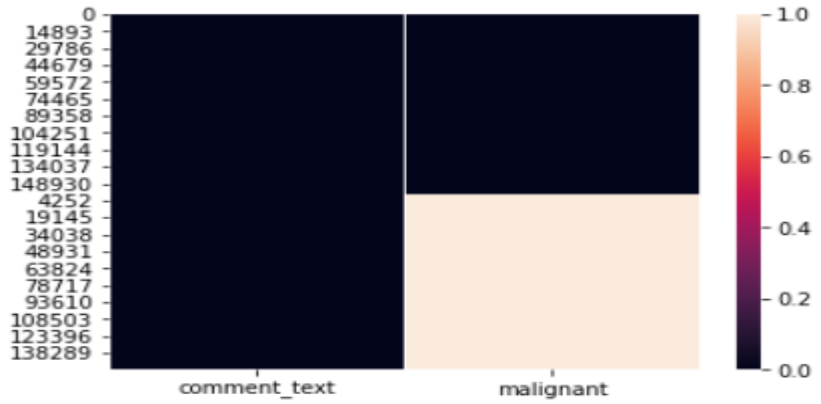
Columns – checking the which columns are used in this dataset. In this dataset 2 columns are used.

Info() – what does this method describe that how many count values, particular data types, null values , columns name in column wise.

Is null().sum() – check the null/missing values in dataset and there is many null value present in this dataset.

```
sns.heatmap(mg.isnull())
```

```
<AxesSubplot:>
```



Head() – This method use all dataset only showing that top five rows and columns.

Dtypes – This is attribute and using for every columns showing data type.

Index – this is attribute and show that range for all rows.

Value_counts() – This function return object containing counts of unique value.

Unique() – this method return only unique value every column present in type of data showing unique values.

Nunique() – this method return unique value every columns.

Rename() – this method quite useful when we need to rename some selected columns.

I have used in dataset some methods/attribute for clean the data and understand that these data how to work. I have used many methods and technique for NLP that means cleaning text /document format data.

Stopword - In natural language processing, useless words (data),are referred to as stop words. **stopword** are words that are so common they are basically ignored by typical tokenizers. By default, **NLTK** (Natural Language Toolkit) includes a list of 40 **stop words**, including: **“a”, “an”, “the”, “of”, “in”**, etc. The **stopwords in nltk** are the most common words in data. I have removed all stopwords from text data.

Regex () - A regular expression is **a powerful tool for matching text, based on a pre-defined pattern.** It

can detect the presence or absence of a text by matching with a particular pattern, and also can split a pattern into one or more sub-patterns. This methods use for substitution to many punctuation marks, white space, emoji expression, phone number, email/web address, any type of digit etc.

Data Input – Logic –output Relationship:-

Input output relationship to the show best methods using word cloud for visualize the maximum number of words used in any textual content. , this is best visualization methods and better way to understand the data how to find that insight from it pattern. It makes it easy to understand the subject and topics discussed in the text by just running this code.

Many times you might have seen a cloud filled with lots of words in different sizes, which represent the frequency or the importance of each word. This is called Tag Cloud or WordCloud.

```
plt.figure(figsize=(14,10))
wc=WordCloud(max_words=500,width=1000,height=800,min_word_length=8,
             background_color='white').generate(" ".join(mg[mg.malignant == 1].comment_text))
plt.axis('off')
plt.imshow(wc,interpolation="bilinear")
```

Hardware and Software Requirements and Tool used:-

I have used software and tool like that in python numpy library use statistical calculation , matplot library, seaborn library for visualizing plot through word cloud, count , scikit learn library from many models building approach, some technique like that vader to sentiment intensity analyzer, tfidf vectorizer ,cross val score, grid search cv, and many library and those inbuilt package used in this data insight information comes out and easily understand and better prediction can be take on future output data so that business problem can be

easily handle and can be growing to the industry and organization.

Model Development and Evaluation

Identification of Possible Problem solving approaches (Methods):-

Cross val score – cross val score is technique cross validation and this technique is a statistical method used to estimate the skill of the machine learning and K-folds is parameter use in this technique to estimate the skill of the model of new data.

This method resampling procedure used to evaluate model. This technique used after checking the different models so that I can make the best model choose for this dataset.

Sentiment Intensity Analyzer - Using sentimental intensity analyzer for especially to sentiment expressed in social media. VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as **sentiment scores**. The

sentiment score of a text can be obtained by summing up the intensity of each word in the text. For example- Words like ‘love’, ‘enjoy’, ‘happy’, ‘like’ all convey a positive sentiment.

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sia=SentimentIntensityAnalyzer()
```

```
empty=[]
for l in mg["comment_text"]:
    vs=sia.polarity_scores(l)
    empty.append(vs)

mg_senti=pd.DataFrame(empty)
mg_senti.head()
```

	neg	neu	pos	compound
0	0.000	0.745	0.255	0.6369
1	0.110	0.442	0.448	0.4767
2	0.158	0.721	0.121	-0.2415
3	0.082	0.801	0.117	0.2500
4	0.000	0.349	0.651	0.6808

TfIdfVectorizer - Using TfIdfVectorizer technique for convert the collection of row documents to matrix of Tf – Idf features.

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
tdi=TfidfVectorizer()
pt=tdi.fit_transform(mg["comment_text"])
```

```
tdi.vocabulary_
```

```
{'explanation': 87244,
 'edits': 77922,
 'made': 157633,
 'username': 282607,
 'hardcore': 112179,
 'metallica': 166178,
 'fan': 89186,
 'reverted': 226762,
 'vandalisms': 284168,
 'closure': 48917,
 'gas': 100471,
 'voted': 288951,
 'new': 180375,
 'york': 301304,
 'dolls': 72599,
 'fac': 88118,
 'please': 206298,
 'remove': 224439,
 'template': 264891,
 'talk': 262377,
```

Testing of identified approach (Algorithms):-

Many algorithm and model used for the training and testing to the data.

This project multiple models used for the better prediction which is as follows –

Logistic Regression, k – neighbors classifier, Decision Tree classifier.

Run and Evaluate Selected Models:-

In this project I have used different models which are written above but before using these models, I have to use best random state and I will use a Logistic Regression model for the best random state.

Me on using the logistic model then I found that best random state value 10 and accuracy score 95.62 %.Now I will use the random state value 10 in all models.

I this project using these model , I have to from scikit library import train test split method so that divide training and testing data and I can use those model.

Decision Tree Classifier –

```
from sklearn.tree import DecisionTreeClassifier
dt=DecisionTreeClassifier()
dt.fit(x_train,y_train)
preddt=dt.predict(x_test)
print(accuracy_score(y_test,preddt))
print(confusion_matrix(y_test,preddt))
print(classification_report(y_test,preddt))
```

```
0.9458765040106952
[[42167  1209]
 [ 1382  3114]]
              precision    recall  f1-score   support

           0       0.97      0.97      0.97     43376
           1       0.72      0.69      0.71      4496

    accuracy                           0.95     47872
   macro avg       0.84      0.83      0.84     47872
weighted avg       0.94      0.95      0.95     47872
```

k – neighbors classifier-

```
from sklearn.neighbors import KNeighborsClassifier
knn=KNeighborsClassifier()
knn.fit(x_train,y_train)
predknn=knn.predict(x_test)
print(accuracy_score(y_test,predknn))
print(confusion_matrix(y_test,predknn))
print(classification_report(y_test,predknn))
```

```
0.8911472259358288
[[41904  1472]
 [ 3739   757]]
              precision    recall  f1-score   support

           0       0.92      0.97      0.94     43376
           1       0.34      0.17      0.23      4496

    accuracy                           0.89     47872
   macro avg       0.63      0.57      0.58     47872
weighted avg       0.86      0.89      0.87     47872
```

Logistic Regression-

```
from sklearn.linear_model import LogisticRegression
lg=LogisticRegression()
lg.fit(x_train,y_train)
predlg=lg.predict(x_test)
print(accuracy_score(y_test,predlg))
print(confusion_matrix(y_test,predlg))
print(classification_report(y_test,predlg))
```

```
0.9562583556149733
[[43203   173]
 [ 1921  2575]]
              precision    recall  f1-score   support

           0       0.96      1.00      0.98     43376
           1       0.94      0.57      0.71      4496

    accuracy                           0.96     47872
   macro avg       0.95      0.78      0.84     47872
weighted avg       0.96      0.96      0.95     47872
```

```
from sklearn.model_selection import cross_val_score
scr= cross_val_score(lg,x,y)
print("cross validation score ",scr.mean())
```

```
cross validation score  0.9543716924256657
```

```
from sklearn.model_selection import cross_val_score
scr= cross_val_score(knn,x,y)
print("cross validation score ",scr.mean())
```

```
cross validation score  0.8981274155335862
```

```
from sklearn.model_selection import cross_val_score
scr= cross_val_score(dt,x,y)
print("cross validation score ",scr.mean())
```

```
cross validation score  0.9437432630587936
```

```
minimum difference is accuracy and cross validation score is DecisionTreeClassifier() so this is our best model.
```

Interpretation of the Results:-

Now further to choose best model , I need to see of these models minimum difference accuracy score and cross val score and the model that will have the least difference , The same model would be best for this dataset. Minimum difference is accuracy and cross val score decision tree classifier() model so this is our best model.

To be using the best model predict the output data and better prediction of future new data and taking the decision for company /industry and these prediction use better decision future client/customer for malignant class .

I have using model apply that on testing data and predict malignant class accordingly.

```
prediction=loaded.predict(prt)
```

```
prediction
```

```
array([0, 0, 0, ..., 0, 0, 1])
```

THANK YOU ??