MALIGNANT COMMENTS CLASSIFIER

Submitted by:

AJITESH KUMAR

ACKNOWLEDGMENT

I would like to thank FlipRobo Technologies for giving me the opportunity to work on this project. I am very grateful to DataTrained team for providing me the knowledge which helped me a lot to work on this project.

Reference sources are:

1. Google

2. YouTube

3. TowardsDataScience

4. Stackoverflow

5. DataTrained Notes

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, cyberbully, 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 cyberbully 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

  In the past few years its seen that the cases related to social media hatred have increased exponentially. The social media is turning into a dark venomous pit for people now a days. Online hate is the

result of difference in opinion, race, religion, occupation, nationality etc.

In social media the people spreading or involved in such kind of activities uses filthy languages, aggression, images etc. to offend and gravely hurt the person on the other side. This is one of the major concerns now.

The result of such activities can be dangerous. It gives mental trauma to the victims making their lives miserable. People who are not well aware of mental health online hate or cyberbully become life threatening for them. Such cases are also at rise. It is also taking its toll on religions. Each and every day we can see an incident of fighting between people of different communities or religions due to offensive social media posts.

Online hate, described as abusive language, aggression, cyberbully, hatefulness, insults, personal attacks, provocation, racism, sexism, threats, or toxicity has been identified as a major threat on online social media platforms. These kinds of activities must be checked for a better future

- Review of Literature

  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 cyberbully.

- Motivation for the Problem Undertaken

  This project is provided to us by FlipRobo Technologies. The exposure to real world data and the opportunity to deploy our skill-set in solving a real time problem 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 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 cyberbully.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- Data Sources and their formats

  The data-set is provided by FlipRobo Technologies as part of the ongoing Data Science internship program. We have been provided with two CSV files namely Train dataset (To train the model) and Test Dataset (Use to test/predict the results).

```
df_train = pd.read_csv('train.csv')

# looking for the dataset

df_train.head()
```

	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

```
# Checking for the shape of the train dataset

df_train.shape
```

```
(159571, 8)
```

- The train dataset contains 1,59,571 rows and 8 columns including the target columns.

```
df_test = pd.read_csv('test.csv')

# looking for the dataset
df_test.head()
```

	id	comment_text
0	00001cee341fdb12	Yo bitch Ja Rule is more succesful then you'll...
1	0000247867823ef7	== From RfC == \n\n The title is fine as it is...
2	00013b17ad220c46	" \n\n == Sources == \n\n * Zawe Ashton on Lap...
3	00017563c3f7919a	:If you have a look back at the source, the in...
4	00017695ad8997eb	I don't anonymously edit articles at all.

```
# Checking for the shape of the test dataset

df_test.shape
```

```
(153164, 2)
```

- The test dataset contains 1,53,164 rows and 2 columns.

- Data Preprocessing Done

```
# Converting all the comments to lower case

df_train['comment_text']=df_train['comment_text'].str.lower()

df_train.head()
```

	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe	length
0	0000997932d777bf	explanation\nwhy the edits made under my usern...	0	0	0	0	0	0	264
1	000103f0d9cfb60f	d'aww! he matches this background colour i'm s...	0	0	0	0	0	0	112
2	000113f07ec002fd	hey man, i'm really not trying to edit war. it...	0	0	0	0	0	0	233
3	0001b41b1c6bb37e	"\nmore\ni can't make any real suggestions on ...	0	0	0	0	0	0	622
4	0001d958c54c6e35	you, sir, are my hero. any chance you remember...	0	0	0	0	0	0	67

```
#Replacing email address with 'email'

df_train['comment_text']=df_train['comment_text'].str.replace(r'^.+@[^\.].*\[a-z]{2,}$','emailaddress')

#Replacing URLs with 'webaddress'
df_train['comment_text']=df_train['comment_text'].str.replace(r'^http\://[a-zA-Z0-9\-\.]+\.[a-zA-Z]{2,3}(/\s*)?$','webaddress')

#Replacing money symbol with 'moneysymb'(£ can type with ALT key+156)
df_train['comment_text']=df_train['comment_text'].str.replace(r'£|\$','dollers')

#Replacing 10 digit phone number(format include paranthesis, space, no spaces,dashes) with 'phone number'
df_train['comment_text']=df_train['comment_text'].str.replace(r'^\(?[\d]{3}\)?[\s-]?[\d]{3}[\s-]?[\d]{4}$','phonenumber')

#Replacing whitespace between terms with a single space
df_train['comment_text']=df_train['comment_text'].str.replace(r'\s+',' ')

#Replacing number with 'numbr'
df_train['comment_text']=df_train['comment_text'].str.replace(r'^\d+(\.\d+)?','numbr')

#Removing punctuation
df_train['comment_text']=df_train['comment_text'].str.replace(r'[^\w\d\s]',' ')

#Removing leading and trailing whitespace
df_train['comment_text']=df_train['comment_text'].str.replace(r'^\s+|\s+?$',' ')

df_train.head()
```

	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe	length
0	0000997932d777bf	explanation why the edits made under my userna...	0	0	0	0	0	0	264
1	000103f0d9cfb60f	d aww he matches this background colour i m s...	0	0	0	0	0	0	112
2	000113f07ec002fd	hey man i m really not trying to edit war it...	0	0	0	0	0	0	233
3	0001b41b1c6bb37e	more i can t make any real suggestions on imp...	0	0	0	0	0	0	622
4	0001d958c54c6e35	you sir are my hero any chance you remember...	0	0	0	0	0	0	67

```
# Removing the stopwords

stop_words = set(stopwords.words('english') + ['u', 'ü', 'ur', '4', '2', 'im', 'dont', 'doin', 'ure'])

df_train['comment_text']=df_train['comment_text'].apply(lambda x:' '.join(term for term in x.split() if term not in stop_words))
```

```
#Checking for the length of the comments after removing the stopwords

df_train['clean_length']=df_train.comment_text.str.len()
df_train.head()
```

	id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe	length	clean_length
0	0000997932d777bf	explanation edits made username hardcore metal...	0	0	0	0	0	0	264	171
1	000103f0d9cfb60f	aww matches background colour seemingly stuck ...	0	0	0	0	0	0	112	83
2	000113f07ec002fd	hey man really trying edit war guy constantly ...	0	0	0	0	0	0	233	141
3	0001b41b1c6bb37e	make real suggestions improvement wondered sec...	0	0	0	0	0	0	622	374
4	0001d958c54c6e35	sir hero chance remember page	0	0	0	0	0	0	67	29

- ## Hardware and Software Requirements and Tools Used

We have used the following Software & Libraries

1. Jupyter Notebook
2. Python 3
3. Pandas
4. Numpy
5. Matplotlib

6. Seaborn

7. NLTK

8. SkLearn

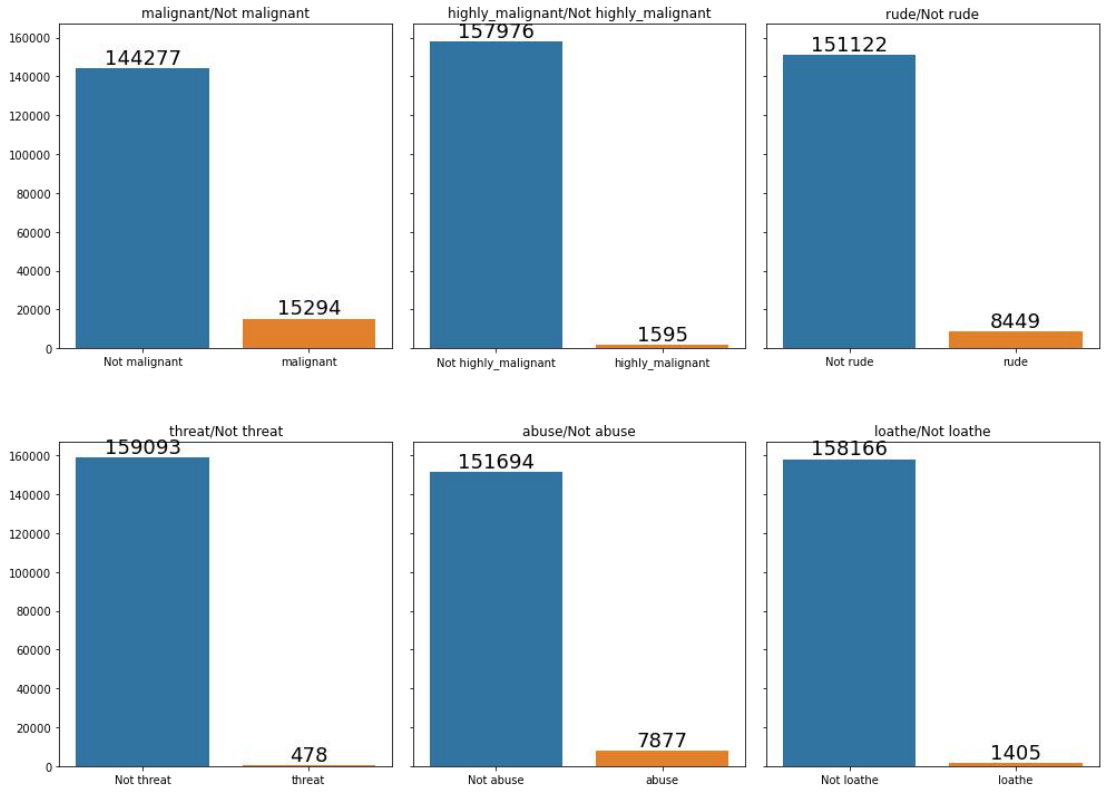
Model/s Development and Evaluation

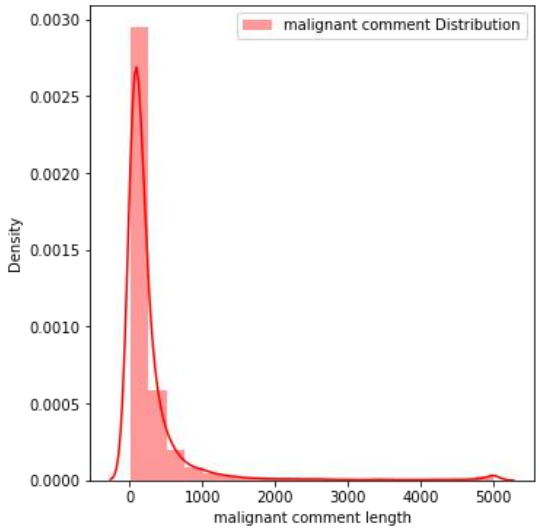
- Key Metrics for success in solving problem under consideration

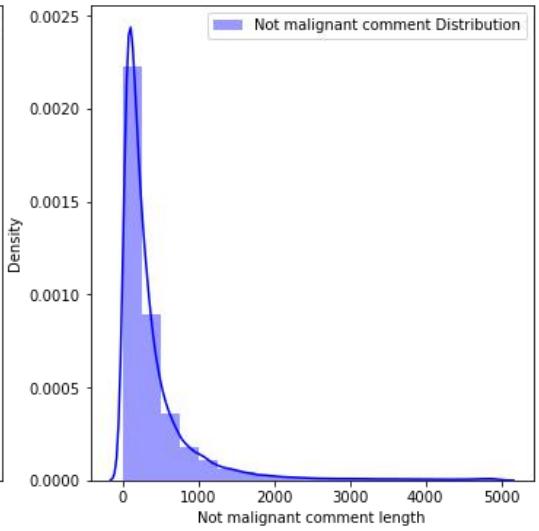
	Model	Accuracy_score	Cross_val_score	Difference	Roc_auc_curve
0	KNeighborsClassifier	91.740406	91.779834	-0.039429	60.806311
1	LogisticRegression	95.608252	95.564357	0.043895	80.410549
2	DecisionTreeClassifier	94.051588	94.183153	-0.131565	82.946990
3	RandomForestClassifier	95.668413	95.669639	-0.001226	84.000003

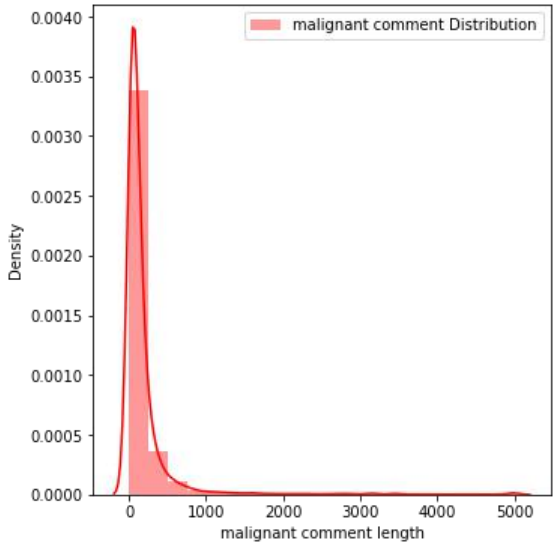
  - From the above table, we found that the minimum difference between the accuracy score and cross validation score is for RandomForestClassifier. So, the best fit model for our project is RandomForestClassifier.

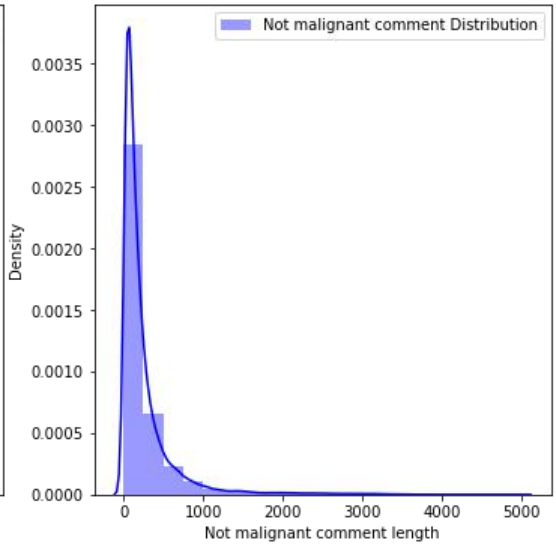
- Visualizations





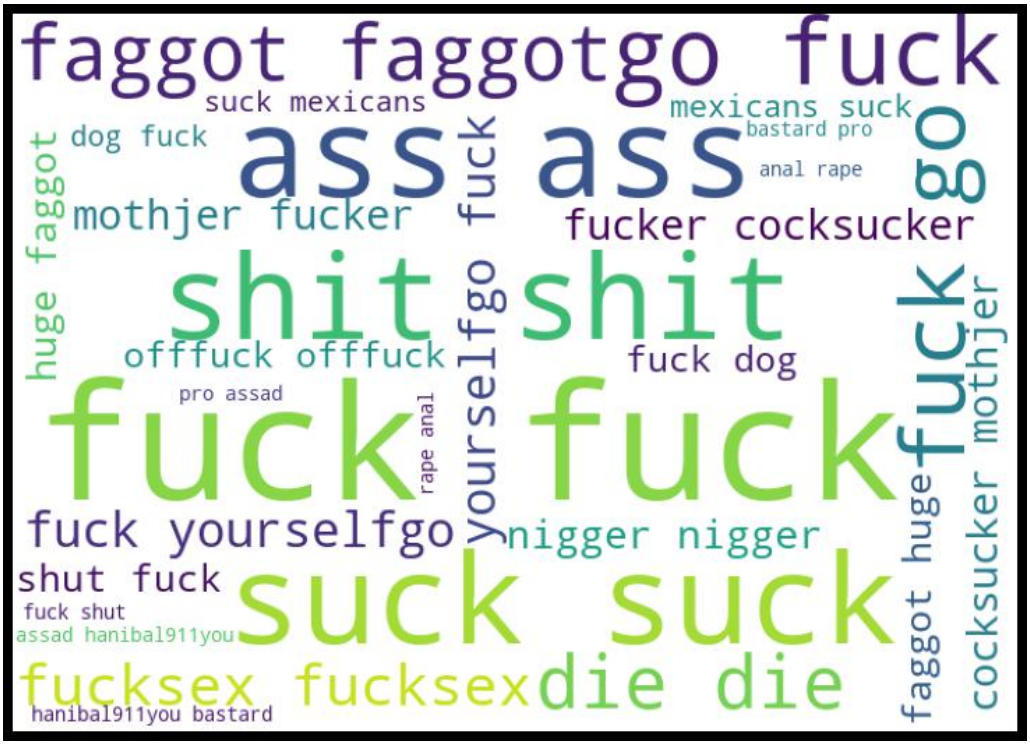




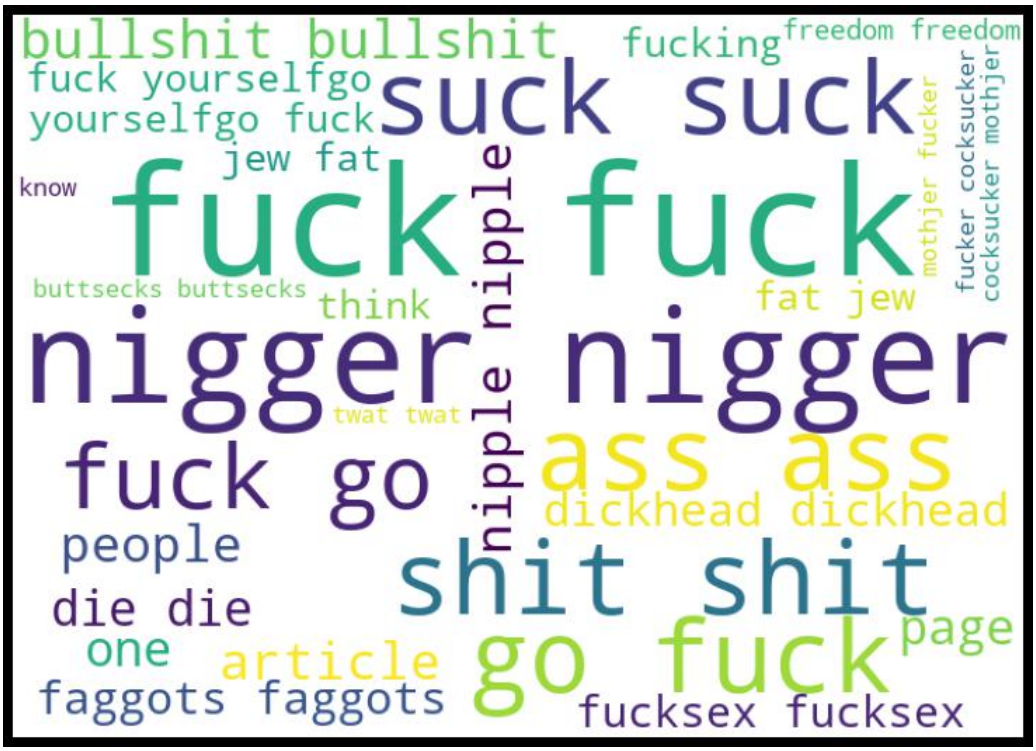


Word Cloud of Malignant Comments

Word Cloud of Highly-Malignant Comments



Word Cloud of Rude Comments

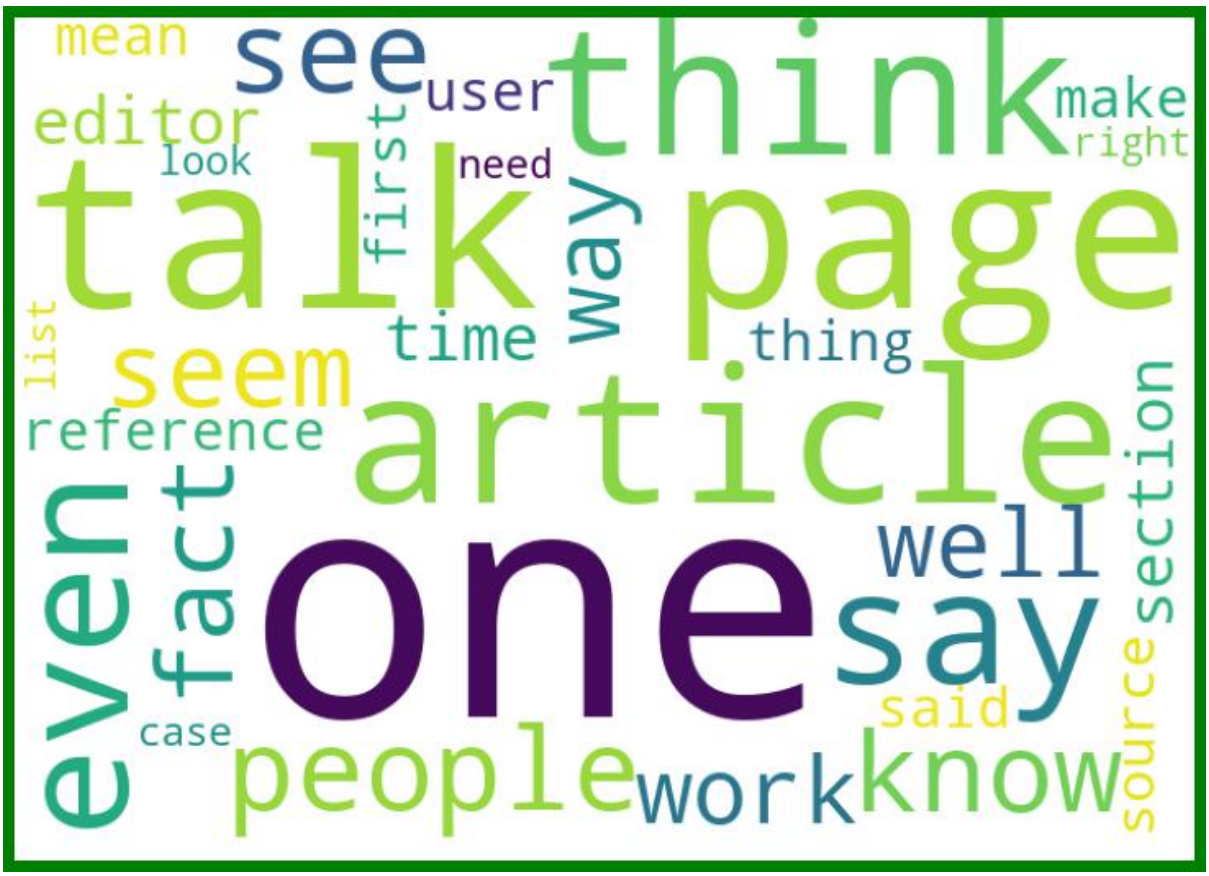


Word Cloud of Abuse Comments

Word Cloud of Threat Comments

Word Cloud of Loathe Comments

Word Cloud of Not Malignant Comments



Looking at the Correlation Table

- Interpretation of the Results

  Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

CONCLUSION

- Key Findings and Conclusions of the Study

  - 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.

- 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 users and broader society.

- Learning Outcomes of the Study in respect of Data Science

  This project helped us to know the concepts of NLP and its application in the field of Data Science. This also helped me to sharpen my knowledge in the different classification models.

THANK YOU