MALIGNANT COMMENT CLASSIFIER

Submitted by:

SANDES P

ACKNOWLEDGMENT

I used informative tutorials to follow some steps in the task from websites like geeksforgeeks, stackoverflow, etc.
I researched the topic on
https://www.brainerddispatch.com
https://cacm.acm.org/

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.

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

With the rise of technology, social media is a great place for seeing and communicating with each other. But like every coin has two sides, it has a dark side too. Since there is no monitoring between the conversations, anyone can pass any comments. Especially when there is a chance to create anonymous/fake accounts. We can see various degrees of hatred in social media comments. If we build a model that predicts whether a comment is malignant or not, we can decide to put it on the platform or not. For this, we have to use NLP techniques to find harmful words and train the model with these words

· Review of Literature

Malicious online comments have emerged as an unwelcome social issue worldwide. In the U.S., a 12-year-old girl committed suicide after being targeted for cyberbullying in 2013. In Singapore, 59.4% of students underwent at least some kind of cyberbullying, and 28.5% were the targets of nasty online comments in 2013. In Australia, Charlotte Dawson, who at one time hosted the "Next Top Model" TV program, committed suicide in 2012 after being targeted with malicious online comments. In Korea, where the damage caused by malicious comments is severe, more than 20% of

> Internet users, from teenagers to adults in their 50s, posted malicious comments in 2011.

Recognizing the harm due to malicious comments, many concerned people have proposed anti-cyberbullying efforts to prevent it. In Europe, one such campaign was called The Big March, the world's first virtual global effort to establish a child's right to be safe from cyberbullying. The key motivation behind these campaigns is not just to stop the posting of malicious comments but also to motivate people to instead post benevolent comments online. Research in social networking has found benevolent comments online are not alone but coexist in cyberspace with many impulsive and illogical arguments, personal attacks, and slander. Such comments are not made in isolation but as part of attacks that amount to cyberbullying.

Both cyberbullying and malicious comments are increasingly viewed as a social problem due to their role in suicides and other real-world crimes. However, the online environment generally lacks a system of barriers to prevent privacy invasion, personal attacks, and cyberbullying, and the barriers that do exist are weak. Social violence as an online phenomenon is increasingly pervasive, a phenomenon manifesting itself through social divisiveness.

Research is needed to find ways to use otherwise socially divisive factors to promote social integration. However, most previous approaches to online comments have focused on analyzing them in terms of conceptual definition, current status, and cyberbullying that involves the writing of malicious comments. Still lacking is an understanding of why people post malicious comments in the first

place or even why they likewise post benevolent comments that promote social integration. Unlike previous studies that focused on cyberbullying itself as a socially divisive phenomenon, this study, which we conducted in Korea in 2014, involved in-depth interviews with social media users in regard to both malicious and benevolent comments. To combat the impropriety represented by the culture of malicious comments and attacks, our study sought to highlight the problem of malicious comments based on the reasons people post comments. Here, we outline an approach toward shaping a healthier online environment with fewer malicious comments and more benevolent ones that promote social integration.

It has become evident that human behavior is changing our emotions are getting attached to the likes, comments, and tags we receive on social media. We get both good and bad comments but seeing hateful words, slurs, and harmful ideas on digital platforms on a daily basis makes it look normal when it shouldn’t be. The impact of toxic comments is much more catastrophic than we think. It not only hurts one’s self-esteem or deters people from having meaningful discussions, but also provokes people to such sinister acts as recent capital riots at US Congress and attacks on farmers for protesting in India. Therefore, having a solid toxicity flagging system in place is important if we want to maintain a civilized environment on social media platforms to effectively facilitate conversations.

· Motivation for the Problem Undertaken

The motivation to make this project is the mental suffering of people who facing cyberbullying or hateful comments. If we want to end this, we have to make a system that detects bullying/hatred comments before posting, so that people will not see those malignant comments.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Here we tried to use different multilabel classification models such as BinaryRelevance, ClassifierChain, LabelPowerset, MLkNN to find the best model.

- Data Sources and their formats

  The data required to build this model is in CSV format. There are a training dataset and a test dataset. The training dataset consist of 159571 rows and 8 columns. The test dataset consists of rows.

  Description of columns:

abuse	Binary column with labels with abusive behaviour.
loathe	Label to comments that are full of loathe and hatred.

There may be some comments which have multiple labels on them, i.e. some comments may be both malignant and loathe.

You are provided with a large number of comments which have been labeled by human raters for malignant behaviour.

The types are:

Malignant

Highly_malignant

Rude

Threat

Abuse

Loathe

- Data Preprocessing Done

  Since the data is scraped by ourselves, there is not much cleaning was required.

  - There were no missing values
  - Removed the stopwords from the message
  - Lemmetized the message
  - Crerated three new columns which
    - Character length of the message
    - Word length of the message
    - The intensity of badness(sum of the column values loathe, abuse, threat, rude, malignant, highly malignant

```
data['c_af_length'] = data['comment_text'].str.len()

data['w_af_length'] =
data['comment_text'].apply(lambda x:len(x.split()) )
```

```
data['bad_intensity'] =data[["malignant",
"highly_malignant", "rude", "threat", "abuse",
"loathe"]].sum(axis =1)
```

- Removed the id column which is unique
- Transformed the text using TfidVectorizer

· Data Inputs- Logic- Output Relationships

- Most of the messages are not hateful
- The training dataset is imbalanced
- High bad comments contain less number of words

· Hardware and Software Requirements and Tools Used

- 8GB RAM
- i5 7th gen processor

Software requirements

- Python
- Jupyter notebook
- Google Colab

Libraries

- Pandas
- Numpy
- Matplotlib
- Seaborn
- drive
- nltk
- Sklearn
- sk-multilearn
- joblib

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
    - EDA for finding the relationship
    - Vectorize
- Testing of Identified Approaches (Algorithms)
    - BinaryRelevance
    - ClassifierChain
    - LabelPowerset
    - MLkNN

- Run and Evaluate selected models

[73]
```
from skmultilearn.problem_transform import BinaryRelevance
from sklearn.naive_bayes import GaussianNB
```

[74]
```
# initialize binary relevance multi-label classifier
# with a gaussian naive bayes base classifier
br = BinaryRelevance(GaussianNB())

# train
br.fit(X_train, y_train)

# predict
predictions = br.predict(X_test)
```

[75]
```
predictions.toarray()
```

```
array([[1, 1, 1, 0, 1, 1],
       [1, 1, 1, 1, 1, 1],
       [0, 0, 0, 0, 0, 0],
       ...,
       [0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0],
       [0, 0, 0, 0, 0, 0]], dtype=int64)
```

[76]
```
from sklearn.metrics import accuracy_score
accuracy_score(y_test,predictions)
```

```
0.4964697526737968
```

[77]
```
# using classifier chains
from skmultilearn.problem_transform import ClassifierChain
from sklearn.naive_bayes import GaussianNB

# initialize classifier chains multi-label classifier
# with a gaussian naive bayes base classifier
gn = ClassifierChain(GaussianNB())

# train
gn.fit(X_train, y_train)

# predict
predictions2 = gn.predict(X_test)

accuracy_score(y_test,predictions2)
```

```
0.5180690173796791
```

[78]
```
# using Label Powerset
from skmultilearn.problem_transform import LabelPowerset
from sklearn.naive_bayes import GaussianNB

# initialize Label Powerset multi-label classifier
# with a gaussian naive bayes base classifier
lbl = LabelPowerset(GaussianNB())

# train
lbl.fit(X_train, y_train)

# predict
predictions3 = lbl.predict(X_test)

accuracy_score(y_test,predictions3)
```

```
0.37034174465240643
```

[79]
```
from skmultilearn.adapt import MLkNN

ml = MLkNN(k=20)

# train
ml.fit(X_train, y_train)

# predict
predictions4 = ml.predict(X_test)

accuracy_score(y_test,predictions4)
```

```
0.9022810828877005
```

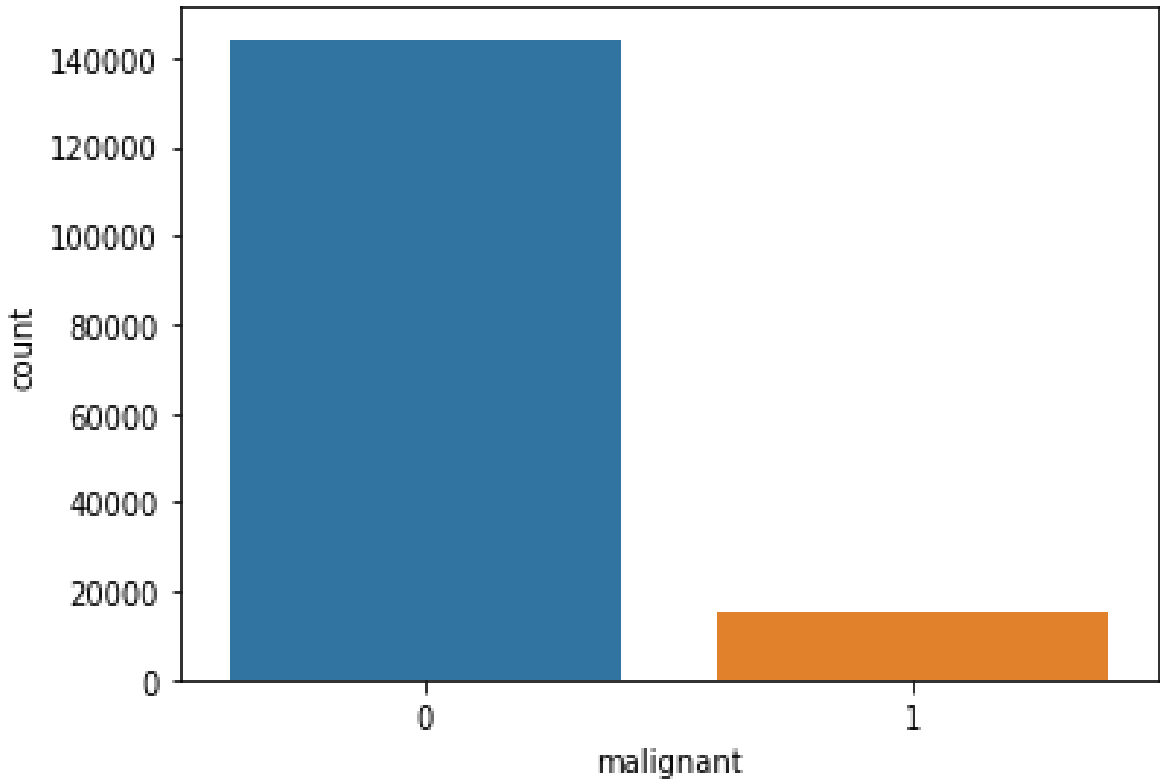
We have used BinaryRelevance, ClassifierChain, LabelPowerset, and MLkNN.

- Key Metrics for success in solving problem under consideration

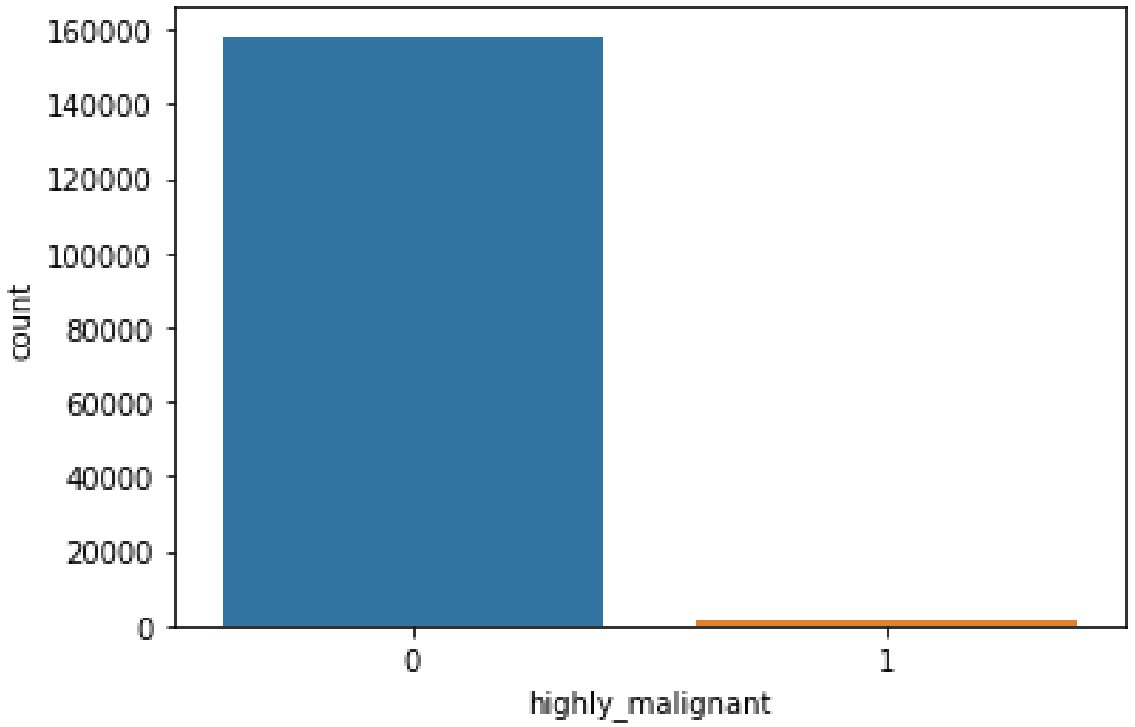
Here we used accuracy score since this is a classification problem

- Visualizations

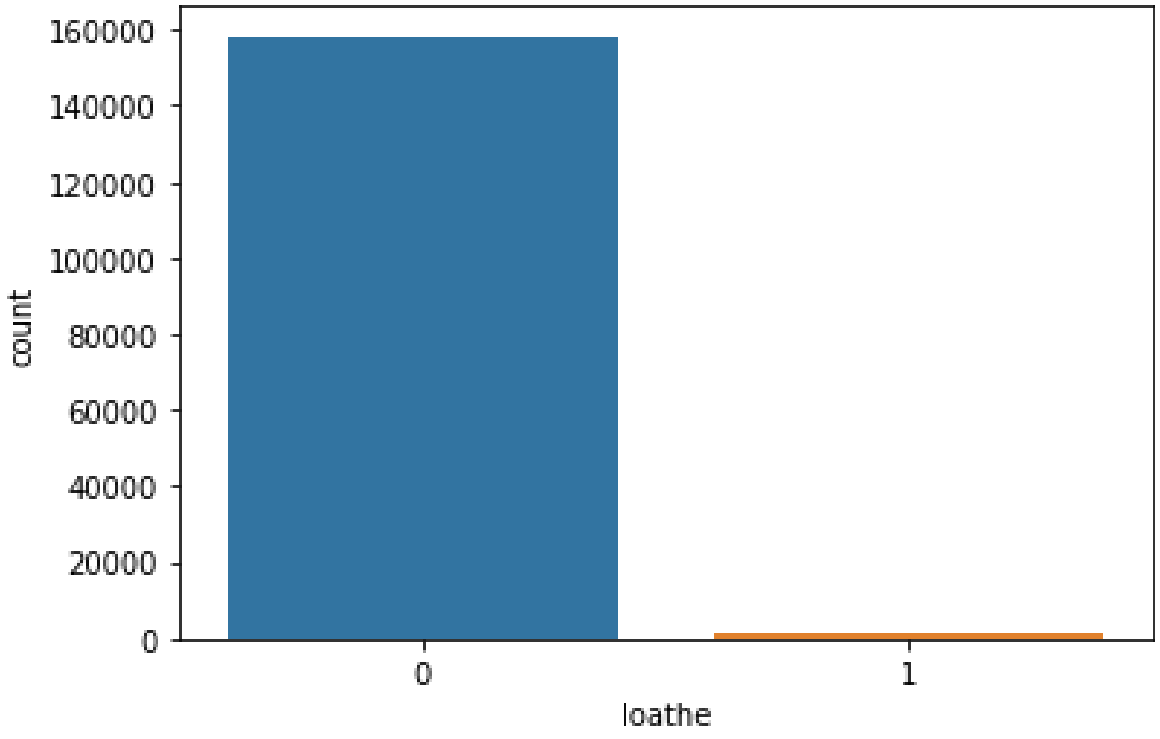
Malignant



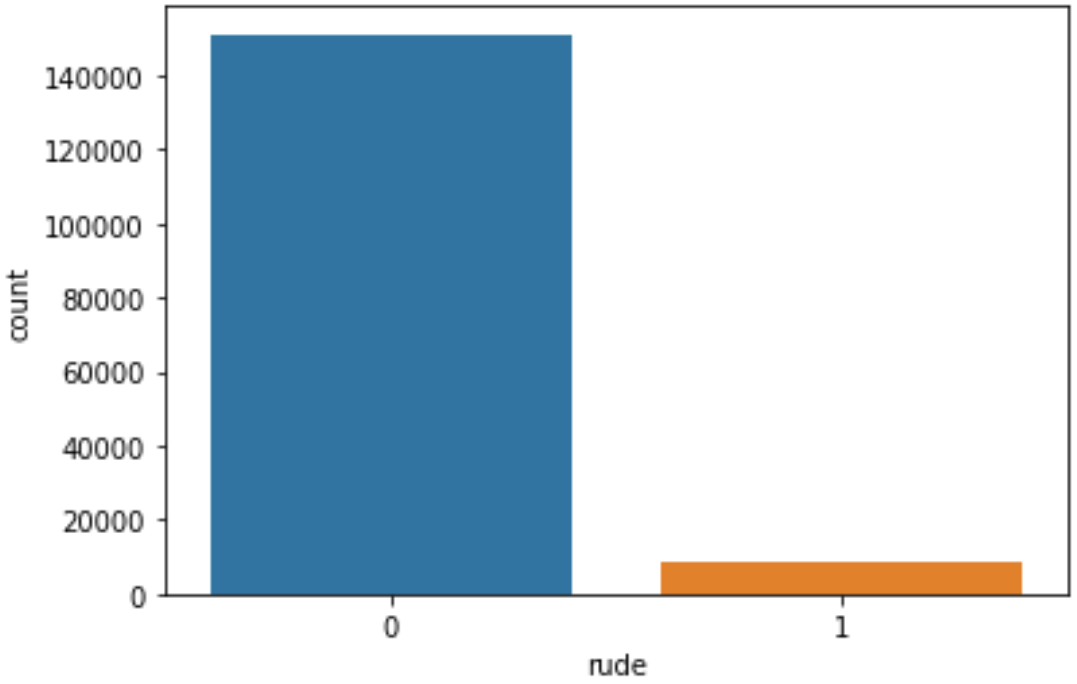
Highly malignant



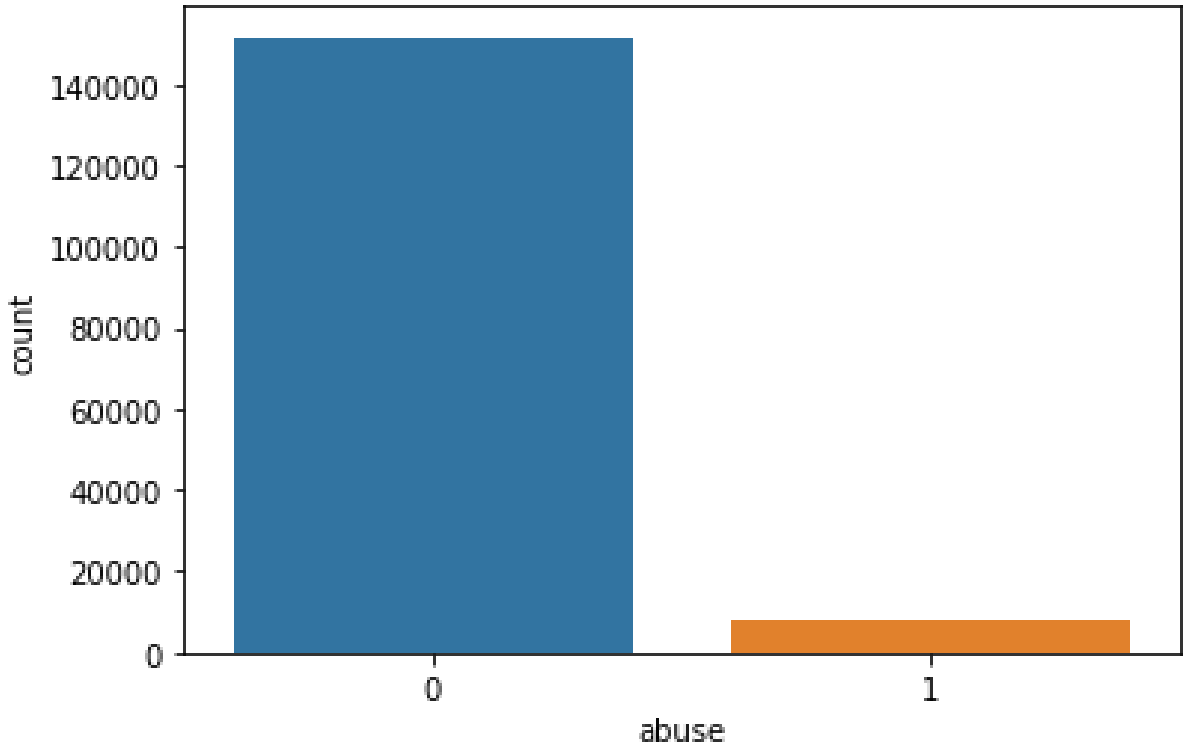
Loathe



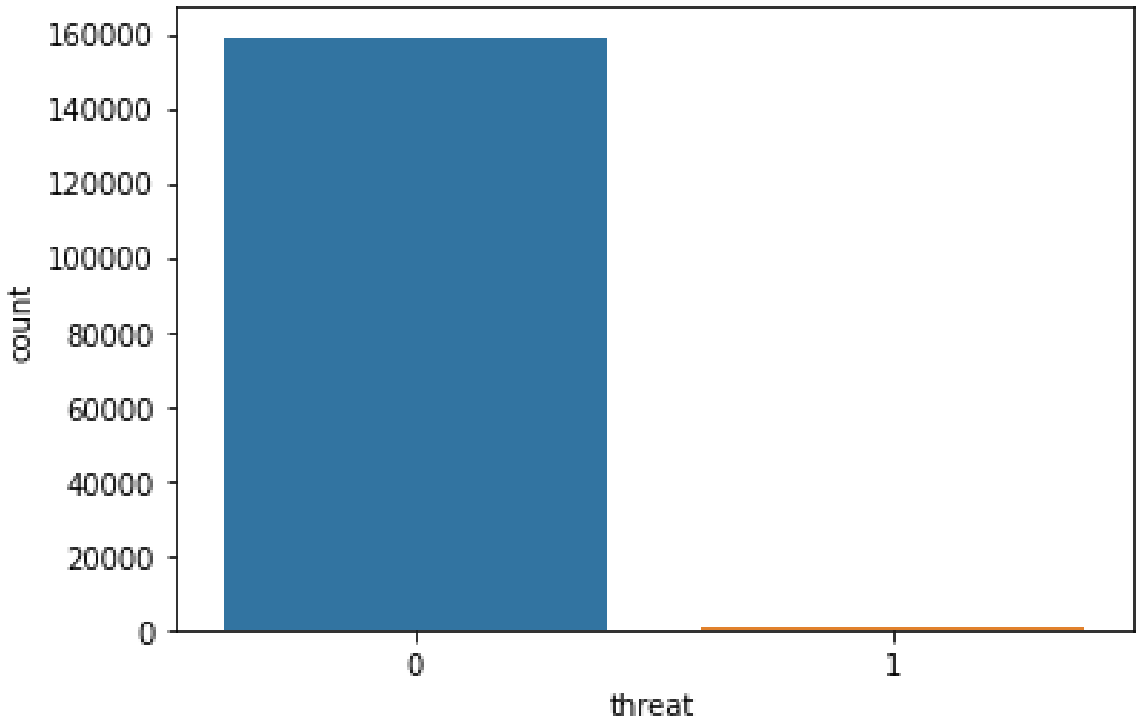
Rude


140000
120000
100000
80000
60000
40000
20000
0
count
0
1
rude

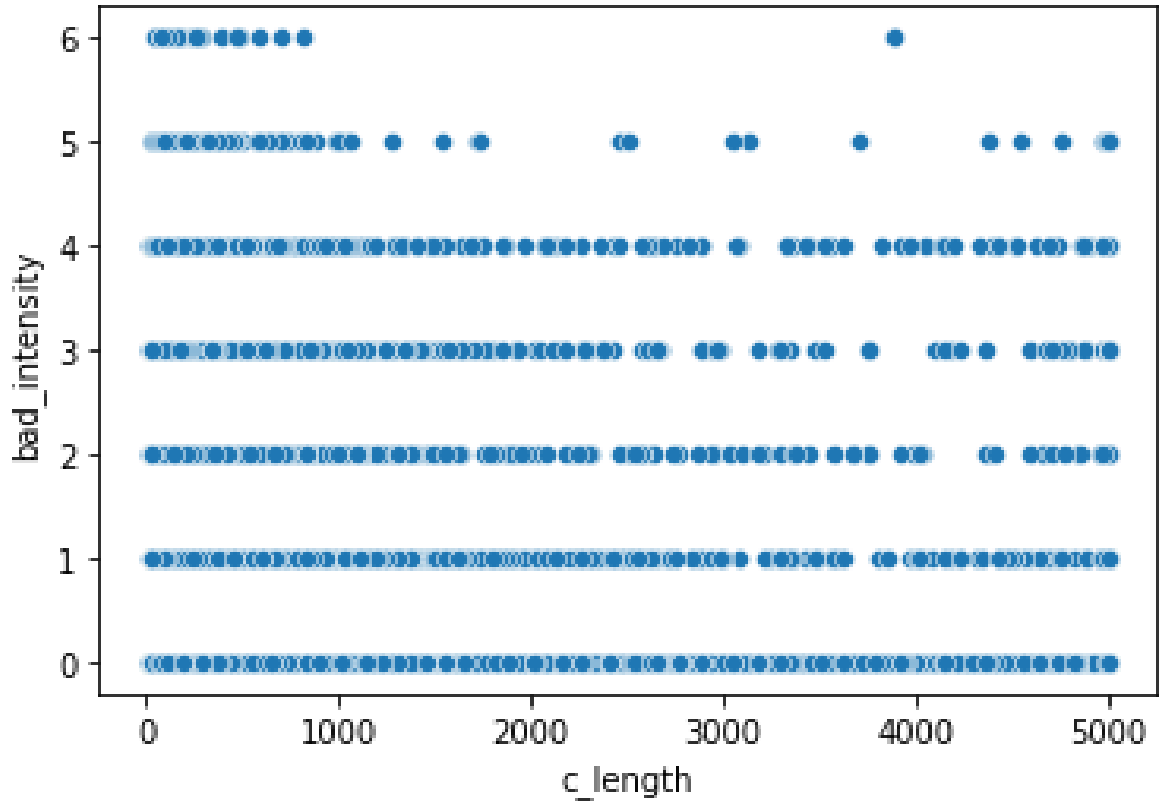
Abuse


140000
120000
100000
80000
60000
40000
20000
0
count
0
1
abuse

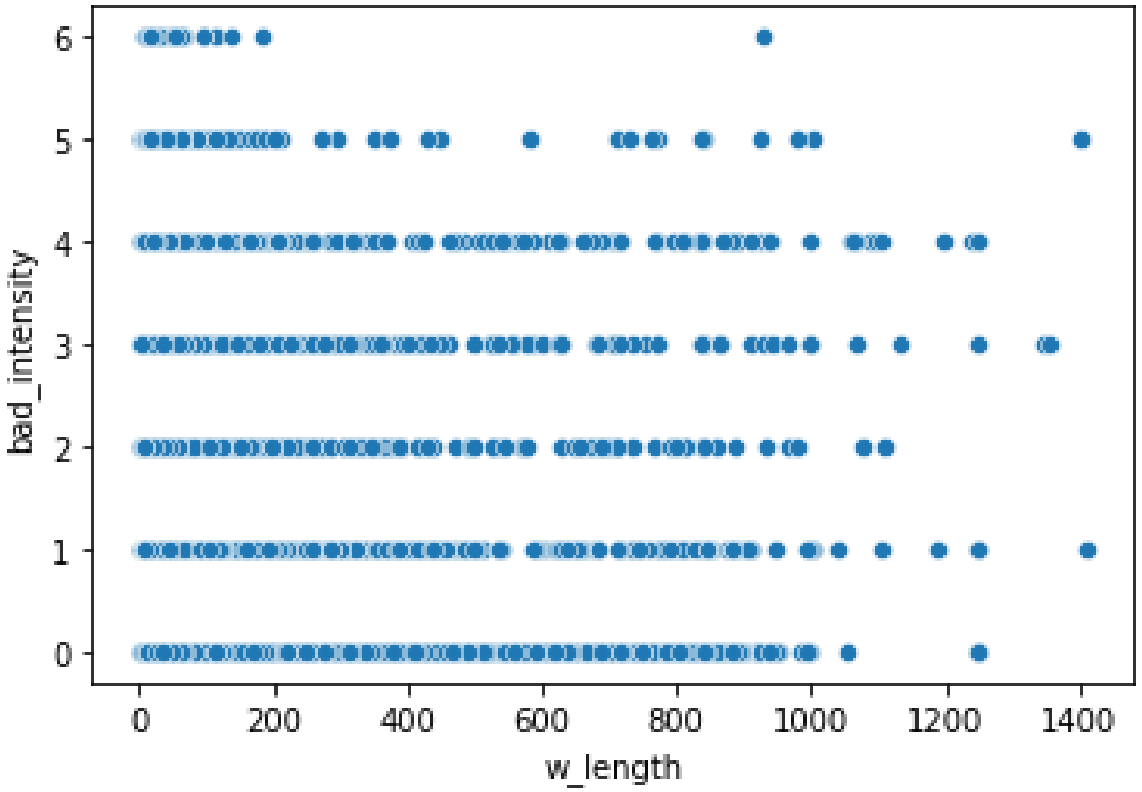
Threat



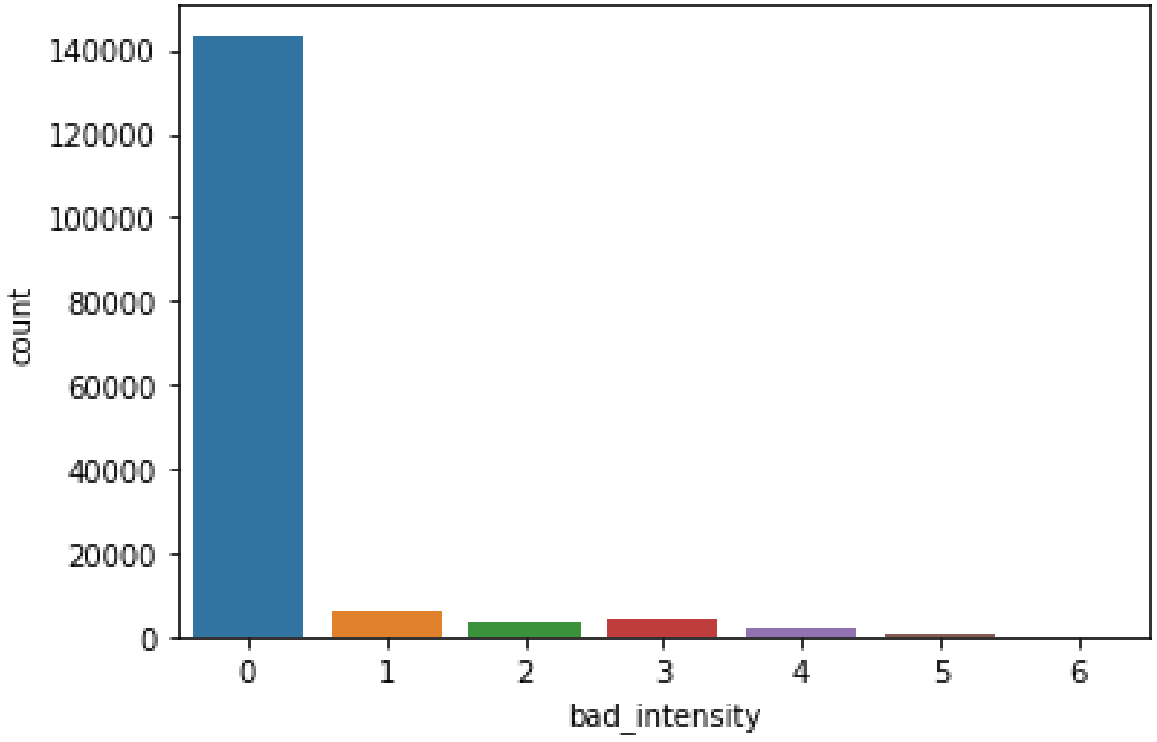
Bad intensity vs char length of the message



Bad intensity vs word length of the message



Count of bad intensity



- Interpretation of the Results

    - Most of the messages are not hateful
    - The training dataset is imbalanced
    - High bad comments contain less number of words
    - Output accuracy is lesser in test dataset

CONCLUSION

- Key Findings and Conclusions of the Study

    All comments are vectorized and trained to find out the bad words and find out the toxicity of the message.

- Learning Outcomes of the Study in respect of Data Science

    Got the opportunity to work with different Ml algorithms like BinaryRelevance, MLkNN , etc

- Limitations of this work and Scope for Future Work

    I had to cut off the feature number to 3000 because of lesser resources