MALIGNANT COMMENTS CLASSIFICATION

Submitted by:

Aneesha B Soman

ACKNOWLEDGMENT

The success and final outcome of this project required a lot of guidance and assistance from Keshav Bansal Sir and I am Extremely fortunate to have got this all along the completion of my project work

Data was obtained from FlipRobo Company.

Various references were used like Anlyticsvidhya, Medium,Data trained Reference materials and Github which helped me in completion of the project

Research papers referred to: Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach

INTRODUCTION

- **Business Problem Framing**

  Our goal of the buisness to understand a comment posted in social media can be harassing.

  The goal of the project 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**

  The spread of aggressive tweets, status and comments on social network are increasing gradually. People are using social media networks as a virtual platform to troll, objurgate, blaspheme and revile one another. These activities are spreading animosity in race-to-race, religion to religion etc. So, these comments should be identified and blocked on social networks. This work focuses on extracting comments from social networks and analyzes those comments whether they convey any blaspheme or revile in meaning.
  Describe the domain related concepts that you think will be useful for better understanding of the project.

- **Review of Literature**

  Chatzakok et al had predicted cyber bullying and user's aggressive behavior on social media networks. They had collected their experiment data from twitter's streaming API and built their classifier using random forest model. Three feature extraction methods were applied; user-based, text-based and network-based. 10 times repeated 10-fold cross validations were applied. They ended up their works with 90% and 81% accuracy using 3 class and 4 class classifications respectively.

  Chavan and Shylaja , were also tried to identify and predict Cyber aggressive Comments but used binary class classification approach only. They had collected data from Kaggle website. They applied capturing pronouns, skip-grams, counting words, n-gram as feature type and Chi-Square for feature selection process. They used Support Vector Machine (SVM) and Logistic Regression as machine learning classifiers. Finally, they got 77.65% classification accuracy for SVM and 73.76% for Logistic Regression.

  Davidson et al., undertook experiments for predicting hate speech and the problems of offensive language use. They had collected their experiment data from Twitter's API. The number of classes was three in their experiment dataset. They had used TF-IDF, N-gram and count indicators as feature type and applied logistic regression with L1 and L2 regularization for feature selection process and Linear SVM was used as machine learning classifier to build their classification model. Finally, they ended up their experiments with 90% classification accuracy .

Papegnies et al, tried to find out the impact of content features for detecting abusive comments on online media.

Reynolds et al, applied decision tree and instance based learning to detect cyber bullying. They had used questionnaire data that were collected from Formspring.me website. Their proposed model predicted cyber bullying with 78.5% accuracy.

Sherly and Jetha , also worked for detecting Cyber bullying. They used Twitter's API dataset and only worked for binary class classification. They had used noun phrases as feature selecting technique and used supervised feature selection applying ranking method. Extreme Learning Machine (ELM) was applied as classifier to detect cyber bullying. Finally, they got 93% classification accuracy in binary class classification.

Nandhini and Sheeba , also worked for detecting cyber bullying by applying machine learning and information retrieval algorithms. They had used Levenshtein algorithm and Naïve Bayes classifier.

Vijayarani et al, made their experiments to present an overview of different processing techniques for text mining.

Xiang et al, also worked for detecting offensive tweets on social media. Their motive was to work with a large scale twitter corpus by applying topical feature discovery.

Zheng et al undertook a study to show how to select useful features from an imbalance dataset for text categorization. All these works tried to classify cyber bullying using binary or multi class classification approaches

- **Motivation for the Problem Undertaken**

1. Hate comments online has been linked to a **global increase in violence toward minorities,** including mass shootings, lynchings, and ethnic cleansing.Policies used to curb hate speech risk limiting free speech and are inconsistently enforced.

Countries such as the United States grant social media companies broad powers in managing their content and enforcing hate speech rules. Others, including Germany, can force companies to remove posts within certain time periods.

Social scientists and others have observed how social media posts, and other online speech, can inspire acts of violence:

- In Germany a correlation was found between anti-refugee Facebook posts by the far-right Alternative for Germany party and attacks on refugees. Scholars Karsten Muller and Carlo Schwarz observed that upticks in attacks, such as arson and assault, followed spikes in hate-mongering posts.

- In the United States, perpetrators of recent white supremacist attacks have circulated among racist communities online, and also embraced social media to publicize their acts. Prosecutors said the Charleston church shooter, who killed nine black clergy and worshippers in June 2015, engaged in a "self-learning process" online that led him to believe that the goal of white supremacy required violent action.

- The 2018 Pittsburgh synagogue shooter was a participant in the social media network Gab, whose lax rules have attracted extremists banned by larger platforms. There, he espoused the conspiracy that Jews sought to bring immigrants into the United States, and render whites a minority, before killing eleven worshippers at a refugee-themed Shabbat service. This "great replacement" trope, which was heard at the white supremacist rally in Charlottesville, Virginia, a year prior and originates with the French far right, expresses demographic anxieties about nonwhite immigration and birth rates.

- The great replacement trope was in turn espoused by the perpetrator of the 2019 New Zealand mosque shootings, who killed forty-nine Muslims at prayer and sought to broadcast the attack on YouTube.

- In Myanmar, military leaders and Buddhist nationalists used social media to slur and demonize the Rohingya Muslim minority ahead of and during a campaign of ethnic cleansing. Though Rohingya comprised perhaps 2 percent of the population, ethnonationalists claimed that Rohingya would soon supplant the Buddhist majority. The UN fact-finding mission said, "Facebook has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the Internet [PDF]."

- In India, lynch mobs and other types of communal violence, in many cases originating with rumors on WhatsApp groups, have been on the rise since the Hindu-nationalist Bharatiya Janata Party (BJP) came to power in 2014.

- Sri Lanka has similarly seen vigilantism inspired by rumors spread online, targeting the Tamil Muslim minority. During a spate of violence in March 2018, the government blocked access to Facebook and WhatsApp, as well as the messaging app Viber, for a week, saying that Facebook had not been sufficiently responsive during the emergency.

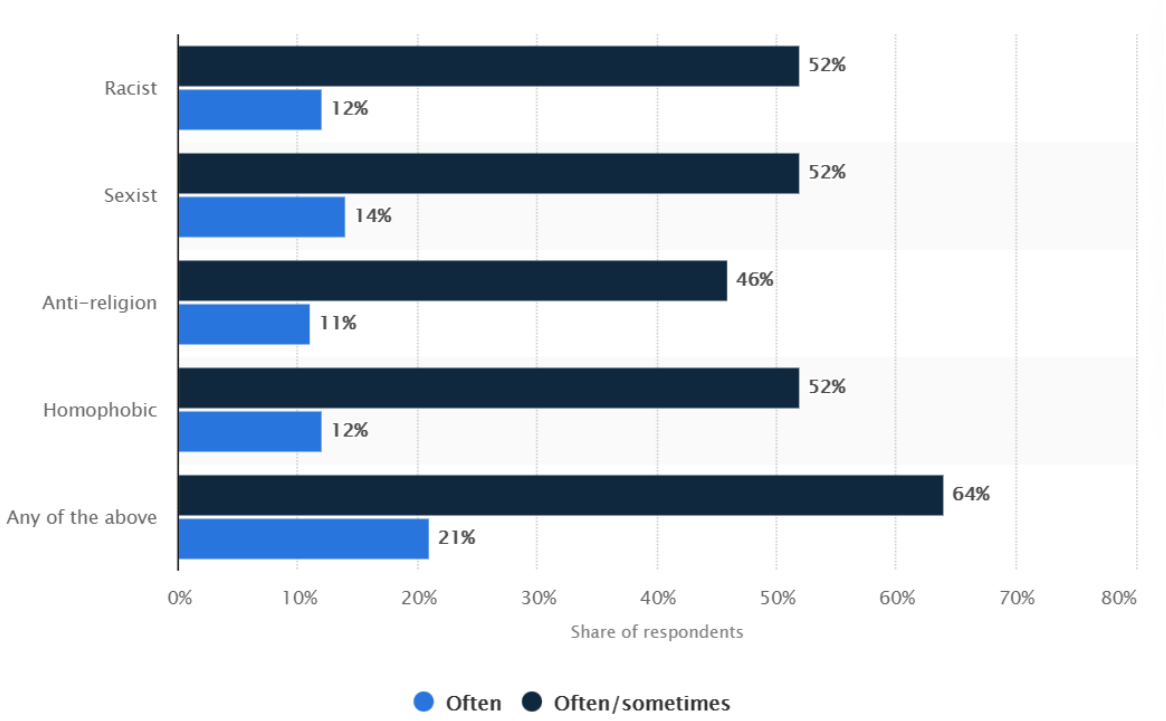
2.Pchychological impact on young people

Young people are connected more than ever before and while this can be a huge benefit in linking them with friends, communities, loved ones and knowledge, it can, of course, be problematic in that they are exposed to an almost constant stream of information which they may not have the critical skills to filter and navigate.

Many young people have a clear digital identity which very often reflects the core of who they are. For example, they may not be 'out' as LGBTQ+ offline but are in their online life. If this is attacked, it hits very hard at a unique part of themselves that they should rightfully be proud of. To be exposed to any form of hate speech that attacks their community or identity is painful and can sadly lead to some not wanting to 'reveal' that part of themselves.

Hate comments statistics

Percentage of teenagers in the United States who have encountered hate speech on social media platforms as of April 2018, by type



Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

a.**Statistical models used:**

- Binary Relevance
- pipeline for applying logistic regression and one vs rest classifier
- Classifier Chains
- Label Powerset

b. **b.Analytical models:**

Descriptive Analytics	**Diagnostic Analytics**	**Predictive Analytics**	**Prescriptive Analytics**
Data visualization done through between the comments and the labels through bar plot	The reason for change is understood through text analytics-vectorised	Prediction is done through various Maultilabel classification techniques	Prescritive analysis is done through the Model created and the test outputs are created through it

- Data Sources and their formats

To undertake the experiments we have collected social media text data in raw format

There is are 2 features id and comment_text

There are 6 label data

malignant	It is a column with binary values depicting which comments are malignant in nature.
highly_malignant	Binary column with labels for highly malignant text.
rude	Binary column with labels for comments that are rude in nature.
threat	Binary column with labels for threatening context in the comments.
abuse	Binary column with labels with abusive behaviour.
loathe	Label to comments that are full of loathe and hatred.

Format for data:csv

Train data

id	comment_text	malignant	highly_malignant	rude	threat	abuse	loathe
0000997932d777bf	Explanation	0	0	0	0	0	0
000103f0d9cfb60f	D'aww! He matches	0	0	0	0	0	0
000113f07ec002fd	Hey man, I'm really r	0	0	0	0	0	0
0001b41b1c6bb37e	"	0	0	0	0	0	0
0001d958c54c6e35	You, sir, are my hero	0	0	0	0	0	0
00025465d4725e87	"	0	0	0	0	0	0
0002bcb3da6cb337	COCKSUCKER BEFOR	1	1	1	0	1	0
00031b1e95af7921	Your vandalism to th	0	0	0	0	0	0
00037261f536c51d	Sorry if the word 'no	0	0	0	0	0	0
00040093b2687caa	alignment on this sul	0	0	0	0	0	0
0005300084f90edc	"	0	0	0	0	0	0
00054a5e18b50dd4	bbq	0	0	0	0	0	0
0005c987bdfc9d4b	Hey... what is it..	1	0	0	0	0	0
0006f16e4e9f292e	Before you start	0	0	0	0	0	0
00070ef96486d6f9	Oh, and the girl abov	0	0	0	0	0	0
00078f8ce7eb276d	"	0	0	0	0	0	0
0007e25b2121310b	Bye!	1	0	0	0	0	0
000897889268bc93	REDIRECT Talk:Voyda	0	0	0	0	0	0

Test data:

id	comment_text
00001cee3	Yo bitch Ja Rule is more succesful then you'll ever be whats up with you and hating you sad mofuckas...i should bitch slap ur pethedic white faces and get yo
00002478(	== From
00013b17	"
00017563(	:If you have a look back at the source, the information I updated was the correct form. I can only guess the source hadn't updated. I shall update the inform
00017695	I don't anonymously edit articles at all.
0001ea87	Thank you for understanding. I think very highly of you and would not revert without discussion.
00024115(	Please do not add nonsense to Wikipedia. Such edits are considered vandalism and quickly undone. If you would like to experiment, please use the sandbo
000247e8	:Dear god this site is horrible.
00025358(	"
00026d10	==
0002eadc	I think its crap that the link to roggenbier is to this article. Somebody that knows how to do things should change it.
0002f87b1	"::::
0003806b	, 25
0003e1cc	"
00059ace	"
00063427	==Curren
000663aff	this other one from 1897
000689dd	==
00083476	:: Wallamoose was changing the cited material to say things the original source did not say. In response to his objections, I modified the article as we went a

- Data Preprocessing Done

- **Data Inputs- Logic- Output Relationships**

  If the input has a particular word or relational words then, the value of the labels changes into 1

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

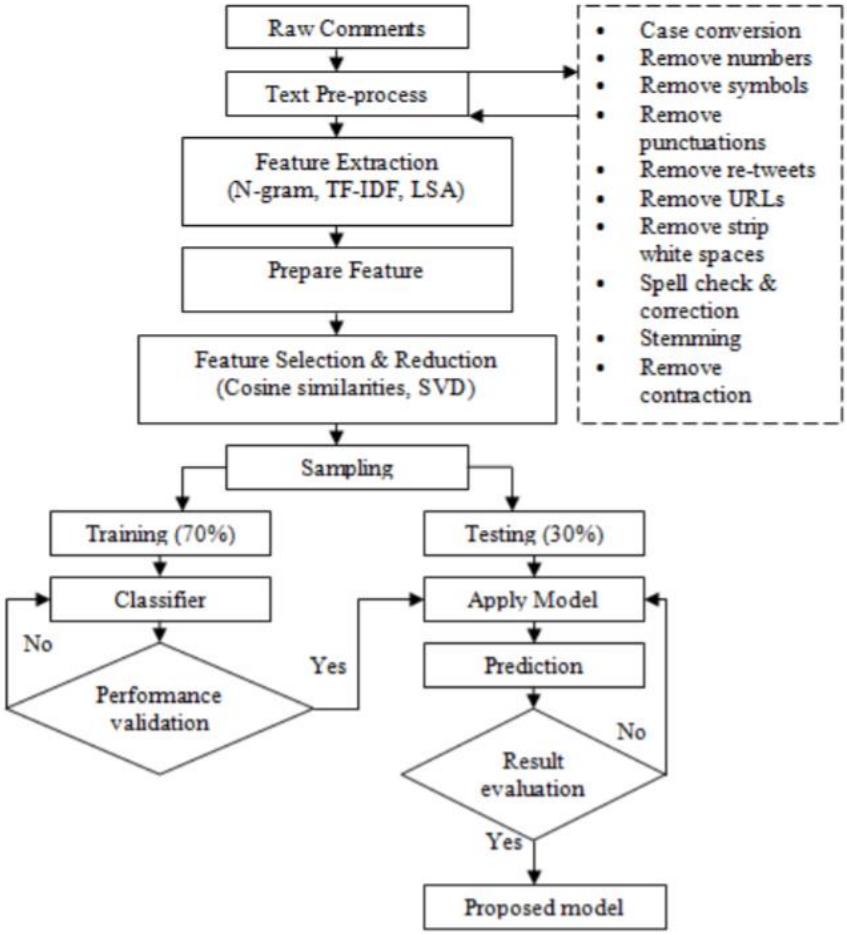
  a.All stop words are present inside the stopwords library

  b.The offensive words are always explicit and sarcasms are not considered

- **<u>Hardware and Software Requirements and Tools Used</u>**
  - Hardware:Windows 10
  - Software:Jupyter notebook
  - Libraries: seaborn, matplotlib, statsmodels, numpy, pandas, sklearn, statsmodels, scipy, pickle, skmultilearn,scipy, nltk, os,re

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)



The problem at hand is a multilabel classification problem.

Hence the classifiers that can be considered are:

1. Binary Relevance
2. pipeline for applying logistic regression and one vs rest classifier
3. Classifier Chains
4. Label Powerset

- **Testing of Identified Approaches (Algorithms)**

Binary Relevance

```
from skmultilearn.problem_transform import BinaryRelevance
from sklearn.svm import SVC

# initialize Binary Relevance multi-label classifier
# with an SVM classifier
# SVM in scikit only supports the X matrix in sparse representation

classifier = BinaryRelevance(
    classifier = SVC(),
    require_dense = [False, True]
)

# train
classifier.fit(x_train, y_train)

# predict
predictions = classifier.predict(x_test)
```

Classifier Chain

```
# initialize classifier chains multi-label classifier
classifier = ClassifierChain(LogisticRegression())
# Training logistic regression model on train data
classifier.fit(x_train, y_train)
# predict
predictions = classifier.predict(x_test)
# accuracy
print("Accuracy = ",accuracy_score(y_test,predictions))
print("\n")
```

LabelPowerset

```
# initialize label powerset multi-label classifier
classifier = LabelPowerset(LogisticRegression())

# train
classifier.fit(x_train, y_train)

# predict
predictions = classifier.predict(x_test)

# accuracy
print("Accuracy = ",accuracy_score(y_test,predictions))
print("\n")
```

Adapted algorithm

```
#adapted algorithm
from skmultilearn.adapt import MLkNN

classifier=MLkNN(k=50)

classifier.fit(x_train,y_train)

predictions=classifier.predict(x_test)
```

- # Run and Evaluate selected models

Hyperparameter tuning on BinaryRelevance

```
from skmultilearn.problem_transform import BinaryRelevance
from sklearn.model_selection import GridSearchCV
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC

parameters = [
    {
        'classifier': [MultinomialNB()],
        'classifier__alpha': [0.7, 1.0],
    },
    {
        'classifier': [SVC()],
        'classifier__kernel': ['rbf', 'linear'],
    },
]

clf = GridSearchCV(BinaryRelevance(), parameters, scoring='accuracy')
clf.fit(x_train, y_train)

print (clf.best_params_, clf.best_score_)
```

```
mod1=BinaryRelevance(classifier=MultinomialNB(),classifier__alpha= [0.7, 1.0])
```

```
mod1.fit(x_train,y_train)
pred=mod1.predict(x_test)
print("BinaryRelevance")
print("accuracy score",accuracy_score(y_test,pred)*100)
print("log loss score:",log_loss(y_test,pred)*100)
print('F1 Score:', f1_score(y_test, y_pred))
print('Recall:',recall_score(y_test, y_pred))
print('Precision:', precision_score(y_test, y_pred))
```

Hyperparameter tuning on ClassifierChain

```
from sklearn.datasets import make_multilabel_classification
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.multioutput import ClassifierChain

#Decision Tree CLassifier
paramters={'base_estimator':['estimator'],
           'random_state':[2,3,4,5,6]
}

GCV=GridSearchCV(ClassifierChain(),paramters,cv=5)
GCV.fit(x_train,y_train)

GCV.best_params_
```

```
mod2=ClassifierChain(base_estimator:'estimator',random_state=4)
```

```
mod2.fit(x_train,y_train)
pred=mod2.predict(x_test)
print("ClassifierChain")
print("accuracy score",accuracy_score(y_test,pred)*100)
print("log loss score:",log_loss(y_test,pred)*100)
print('F1 Score:', f1_score(y_test, y_pred))
print('Recall:',recall_score(y_test, y_pred))
print('Precision:', precision_score(y_test, y_pred))
```

Hyperparameter tuning on Label Powerset

```
from skmultilearn.problem_transform import LabelPowerset
from sklearn.model_selection import GridSearchCV
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier

parameters = [
    {
        'classifier': [MultinomialNB()],
        'classifier__alpha': [0.7, 1.0],
    },
    {
        'classifier': [RandomForestClassifier()],
        'classifier__criterion': ['gini', 'entropy'],
        'classifier__n_estimators': [10, 20, 50],
    },
]
```

```
clf = GridSearchCV(LabelPowerset(), parameters, scoring='accuracy')
clf.fit(x_train,y_train)

print (clf.best_params_, clf.best_score_)
```

```
mod3=ClassifierChain(classifier: MultinomialNB(),classifier__alpha: [0.7, 1.0],)
```

```
mod2.fit(x_train,y_train)
pred=mod2.predict(x_test)
print("Label Powerset")
print("accuracy score",accuracy_score(y_test,pred)*100)
print("log loss score:",log_loss(y_test,pred)*100)
print('F1 Score:', f1_score(y_test, y_pred))
print('Recall:',recall_score(y_test, y_pred))
print('Precision:', precision_score(y_test, y_pred))
```

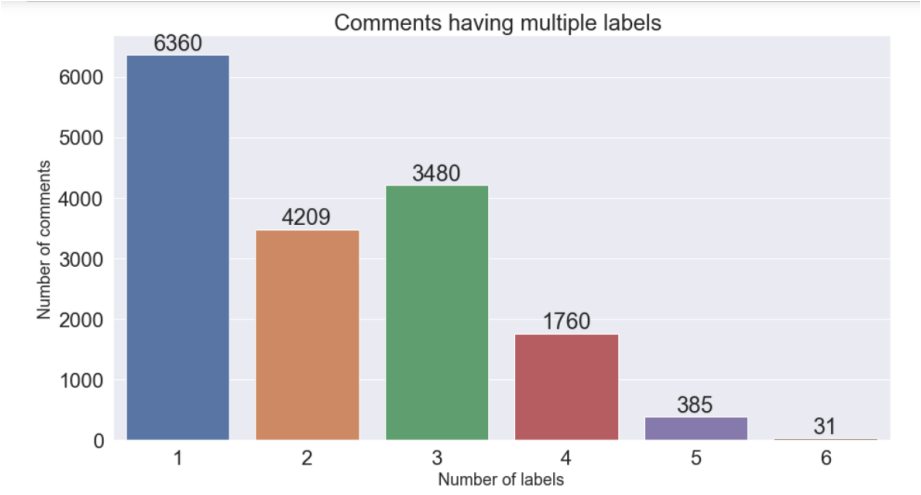
Choosing binary relvance due to highest accuracy score

Metrics obtained: accuracy scores

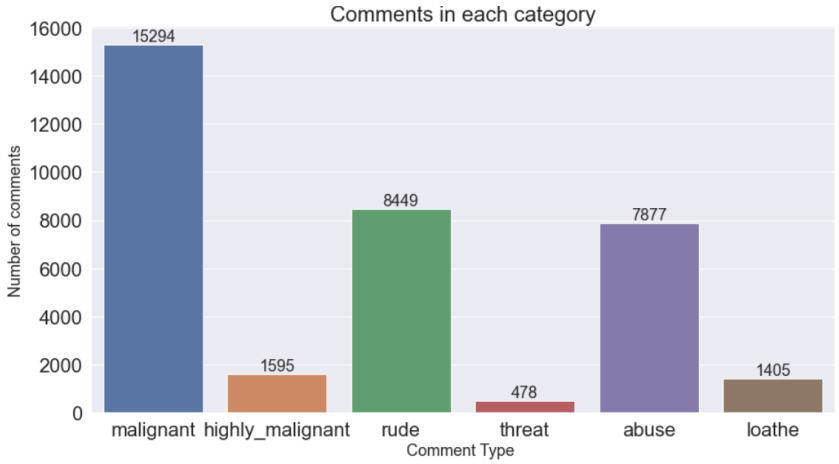
Multiple Binary Classifications One Vs Rest Classifier	**malignant comments** 0.9283333333333333 **highly_malignant comments** 0.9916666666666667

	rude comments 9583333333333334 **threat comments** 0.9933333333333333 **abuse comments** 0.96 **loathe comments** 0.99
Binary Relevance	89.83
Classifier Chains	92.16
Label Powerset	92.16
Adapted Algorithm	91.66

- Visualizations



Maximum comments have one label



Highest comments are for malignant

WordCloud representation of most used words

- **Interpretation of the Results**

  There are two main methods for tackling a multi-label classification problem: problem transformation methods and algorithm adaptation methods.
  Problem transformation methods transform the multi-label problem into a set of binary classification problems, which can then be handled using single-class classifiers.
  Whereas algorithm adaptation methods adapt the algorithms to directly perform multi-label classification. In other words, rather than trying to convert the problem to a simpler problem, they try to address the problem in its full form.

  The model interprets a comment and decides its range of malign

CONCLUSION

- **Key Findings and Conclusions of the Study:**

  The aim and objective of this study is to identify and filter cyber aggressive comments of social media networks in different categories

  Conclusions are:

  1.Most of the comments which are offensive In nature are malignant

  2.Most of the comments can be classified as having only one label

  3.The highest accuracy is for classifier chain and label power set

- **Limitations of this work and Scope for Future Work**

❖ For more speed we could use decision trees and for a reasonable trade-off between speed and accuracy we could also opt for ensemble models.

❖ Other frameworks such as MEKA can be used to deal with multi-label classification problems.