Multi-class Cyberbullying Tweet Classification

Introduction

As social media usage grows in popularity among people of all ages, the vast majority of citizens rely on it for day-to-day contact. Because of the widespread use of social media, cyberbullying can affect anybody at any time or from any location, and the internet's relative anonymity makes such personal attacks more difficult to stop than conventional bullying. Due to widespread school cancellations, increased screen time, and decreased face-to-face social interaction, UNICEF issued a warning on April 15th, 2020, in reaction to the increased risk of cyberbullying during the COVID-19 pandemic. Cyberbullying statistics are shocking: 36.5 percent of middle and high school kids have been cyberbullied, and 87 percent have witnessed cyberbullying, with consequences ranging from poor academic performance to depression and suicidal ideation.

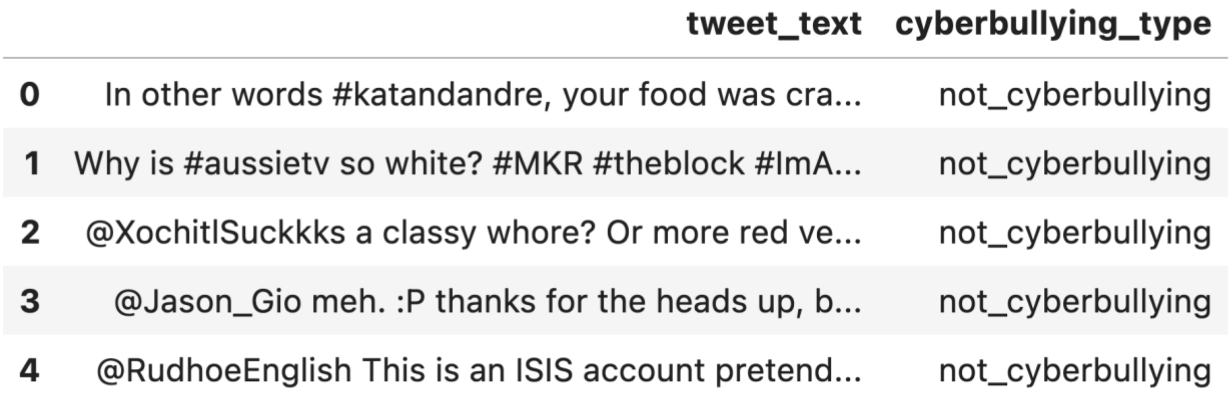
In the project, a multiclassification model has been proposed to classify tweets into different types of cyberbullying. This will help Twitter, law enforcement agencies and general public to identify potential harmful tweets and flag them automatically. A dataset from Kaggle which contains more than 47000 tweets labelled according to the five different classes of cyberbullying and one class of not cyberbullying where the labels are as follows:

1. Age
2. Ethnicity
3. Gender
4. Religion
5. Other type of cyberbullying
6. Not cyberbullying

The raw data has an almost balanced number of tweets approximately 8000 belonging to each class.

Data cleaning

The raw dataset contains 47692 samples of tweets and their respective cyberbullying types. A sample of the dataset showing the first five rows are depicted in Fig. 1. The first column contains the tweet text and the second column contain the cyberbullying type. These seems to be no missing values in the dataset. There are 36 duplicate rows which are dropped from the dataset. However, it seems that there are tweets with the same text but are classified different. Such tweets need to be dropped so as not confuse the machine learning model. After dropping duplicate rows and the rows with same tweet text, the dataset had 44378 samples with unique tweets and six cyberbullying categories.



	tweet_text	cyberbullying_type
0	In other words #katandandre, your food was cra...	not_cyberbullying
1	Why is #aussietv so white? #MKR #theblock #ImA...	not_cyberbullying
2	@XochitlSuckkks a classy whore? Or more red ve...	not_cyberbullying
3	@Jason_Gio meh. :P thanks for the heads up, b...	not_cyberbullying
4	@RudhoeEnglish This is an ISIS account pretend...	not_cyberbullying

Figure 1 Sample data

Exploratory Data Analysis

Importing the data from the previous step, it was found the distribution of classes as shown in Fig. 2. It can be seen that the distribution of tweets across different categories is almost similar with "not_cyberbullying" and "other_cyberbullying" classes with little lesser number of tweets.

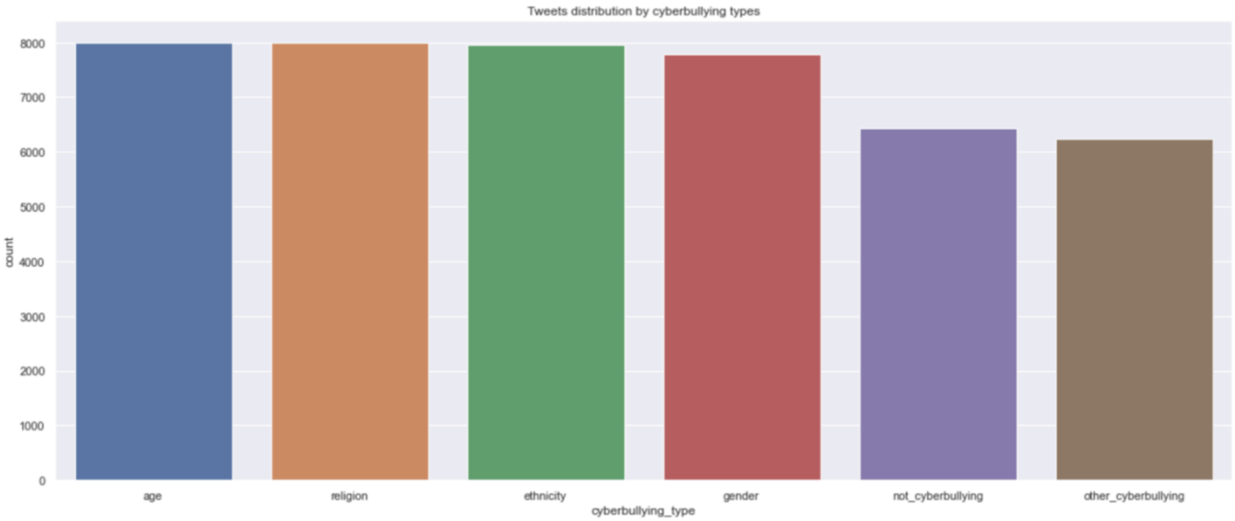


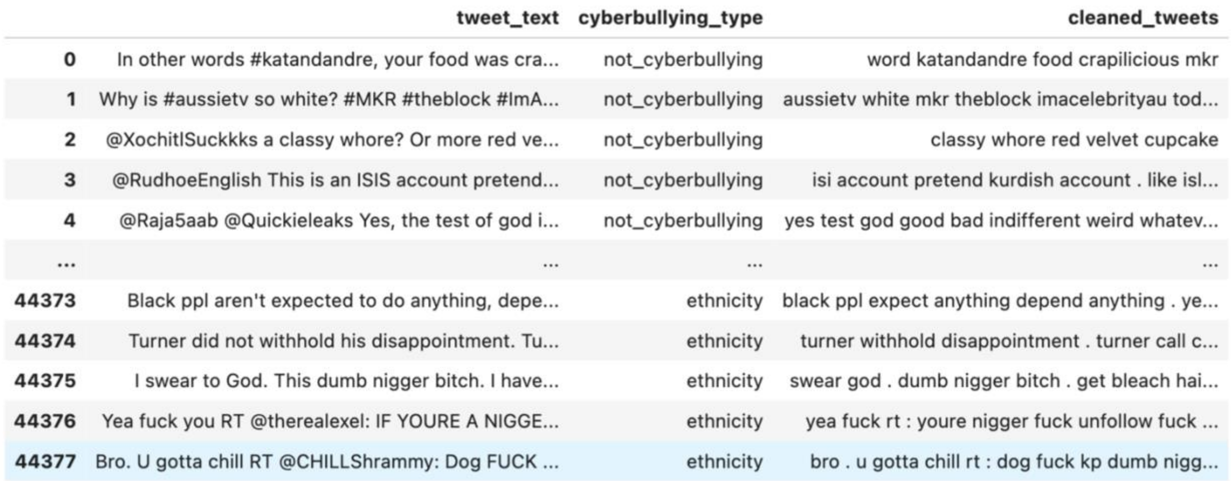
Figure 2. Distribution of tweets across different cyberbullying categories

As expected, the raw tweets contain a lot of text, symbols, numbers and emojoies which are considered not useful in predicting the cyberbullying categories. These includes:

- Removing punctuations, special characters, URLs & hashtags
- Removing leading, trailing & extra white spaces/tabs
- Typos, slangs are corrected, abbreviations are written in their long forms

Further, the stop words which are generic such as 'i','you','a','the','he','which' etc are also removed from the tweets. The words in the tweets are also stemmed which is process of slicing the end or the beginning of words with the intention of removing affixes(prefix/suffix).

Lemmatization on the words of the tweets is also performed which is the process of reducing the words to its base form. The version of tweets after this process has been performed in Fig. 3.



	tweet_text	cyberbullying_type	cleaned_tweets
0	In other words #katandandre, your food was cra...	not_cyberbullying	word katandandre food crapilicious mkr
1	Why is #aussietv so white? #MKR #theblock #ImA...	not_cyberbullying	aussietv white mkr theblock imacelebrityau tod...
2	@XochitlSuckkks a classy whore? Or more red ve...	not_cyberbullying	classy whore red velvet cupcake
3	@RudhoeEnglish This is an ISIS account pretend...	not_cyberbullying	isi account pretend kurdish account . like isl...
4	@Raja5aab @Quickieleaks Yes, the test of god i...	not_cyberbullying	yes test god good bad indifferent weird whatev...
...	...	...	...
44373	Black ppl aren't expected to do anything, depe...	ethnicity	black ppl expect anything depend anything . ye...
44374	Turner did not withhold his disappointment. Tu...	ethnicity	turner withhold disappointment . turner call c...
44375	I swear to God. This dumb nigger bitch. I have...	ethnicity	swear god . dumb nigger bitch . get bleach hai...
44376	Yea fuck you RT @therealexel: IF YOURE A NIGGE...	ethnicity	yea fuck rt : youre nigger fuck unfollow fuck ...
44377	Bro. U gotta chill RT @CHILLShrammy: Dog FUCK ...	ethnicity	bro . u gotta chill rt : dog fuck kp dumb nigg...

Figure 3. Tweets after exploratory data analysis

Fig. 4 depicts the distribution of lengths of tweets across different cyberbullying categories. It can be seen that the tweets belonging to religion and ethnicity categories are longer in general whereas tweets belonging to age are smaller. The tweets from each class where input to a word cloud generator and the results can be seen in Fig. 5. It can be seen that some specific words are prevalent in specific classes such female, joke, rape, gay, bitch are prevalent in Gender class. Age class has girl, school, high, bully, bullied word which make sense considering the age-related cyberbullying might be related to a school environment. Ethnicity class has nigger, nigga, fuck, black and white as prevalent word and religion class is prevailed with Muslim, terrorist, idiot, terrorism, Islam words. It can be seen that the occurrence of certain words clearly defines the class in which the tweet will belong to.

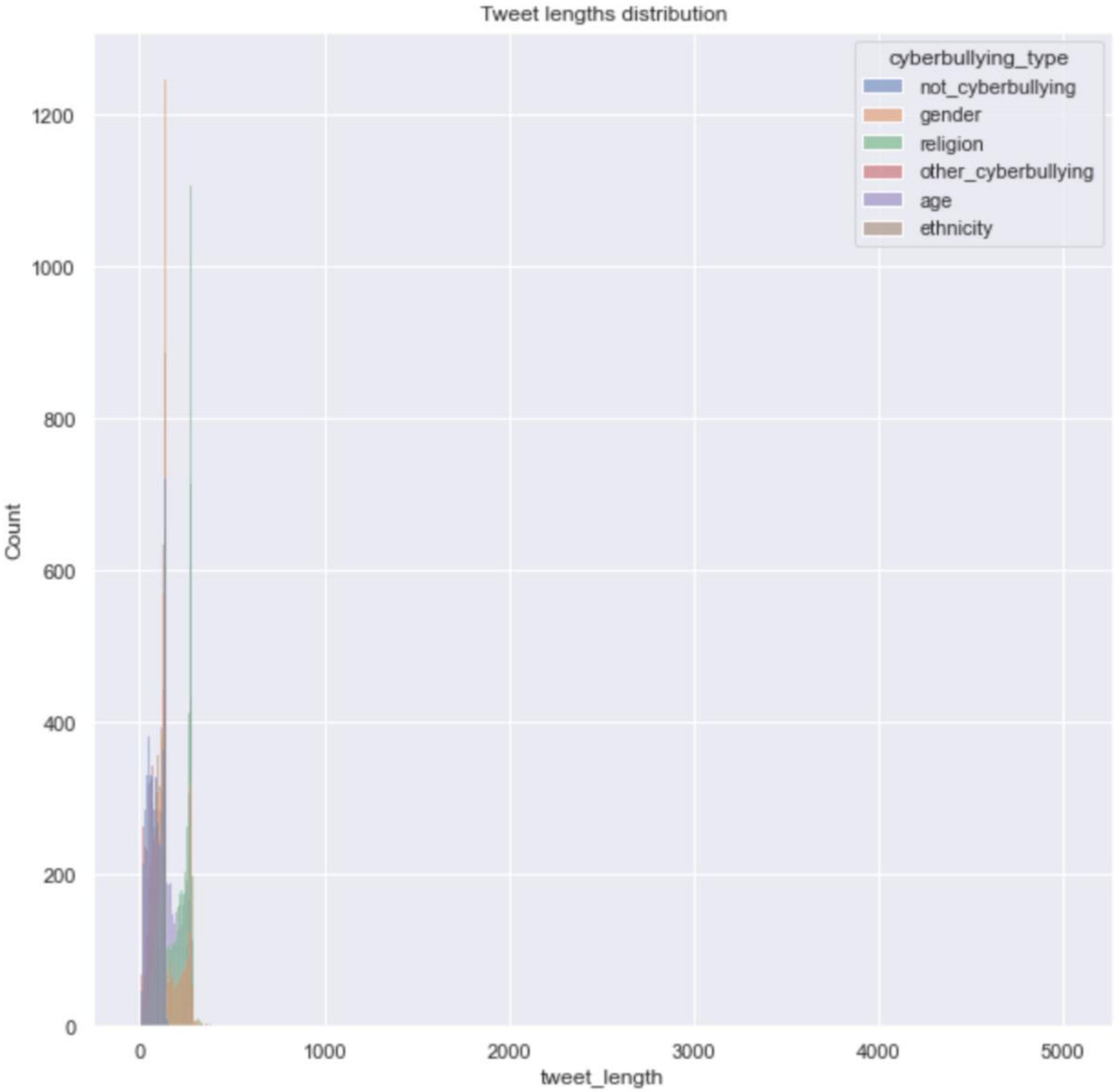
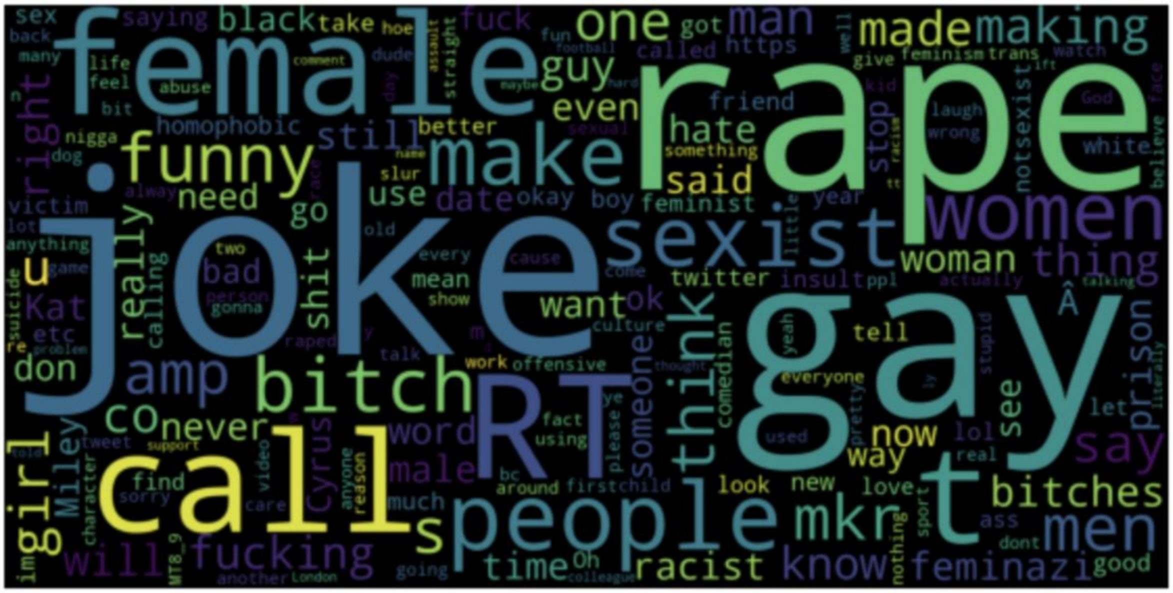
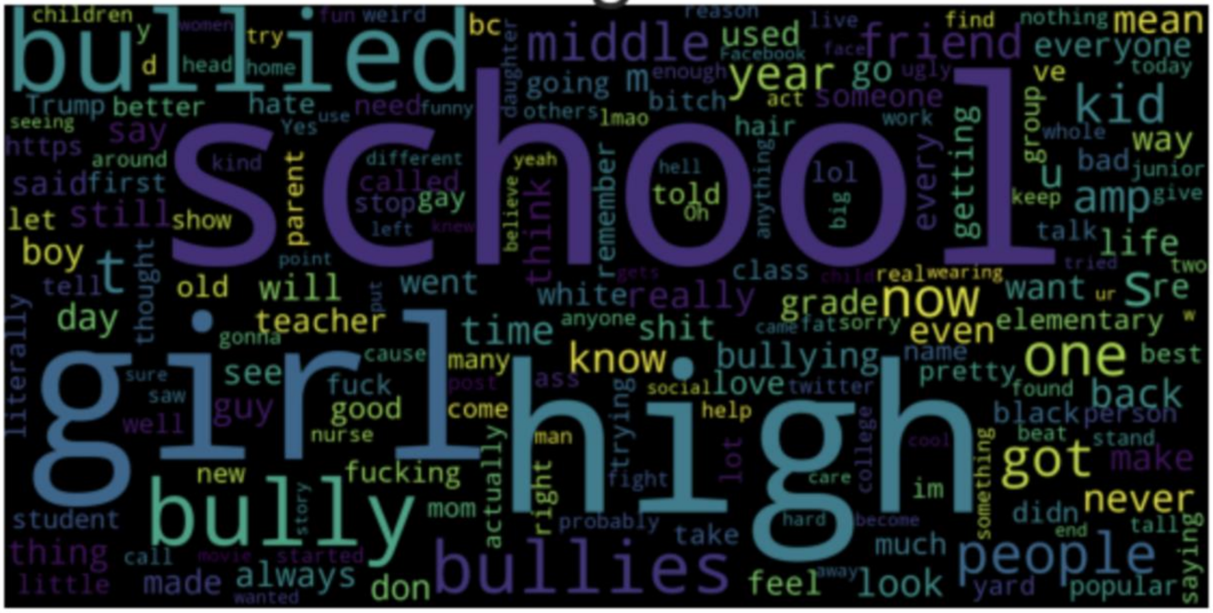


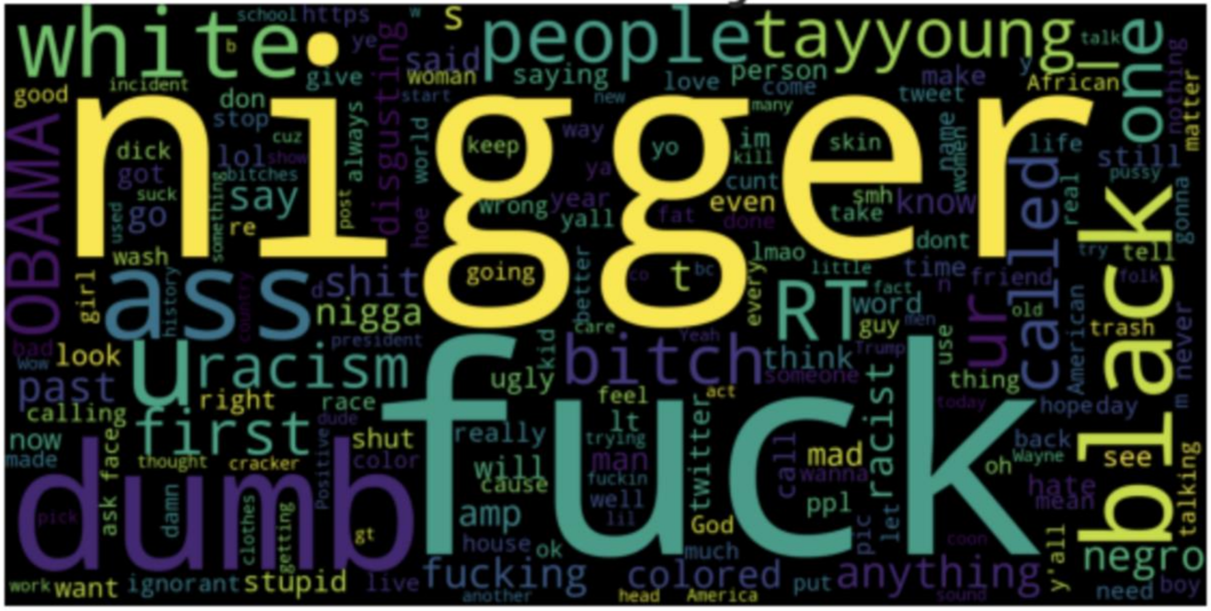
Figure 4 Tweet length distribution across different categories

Gender

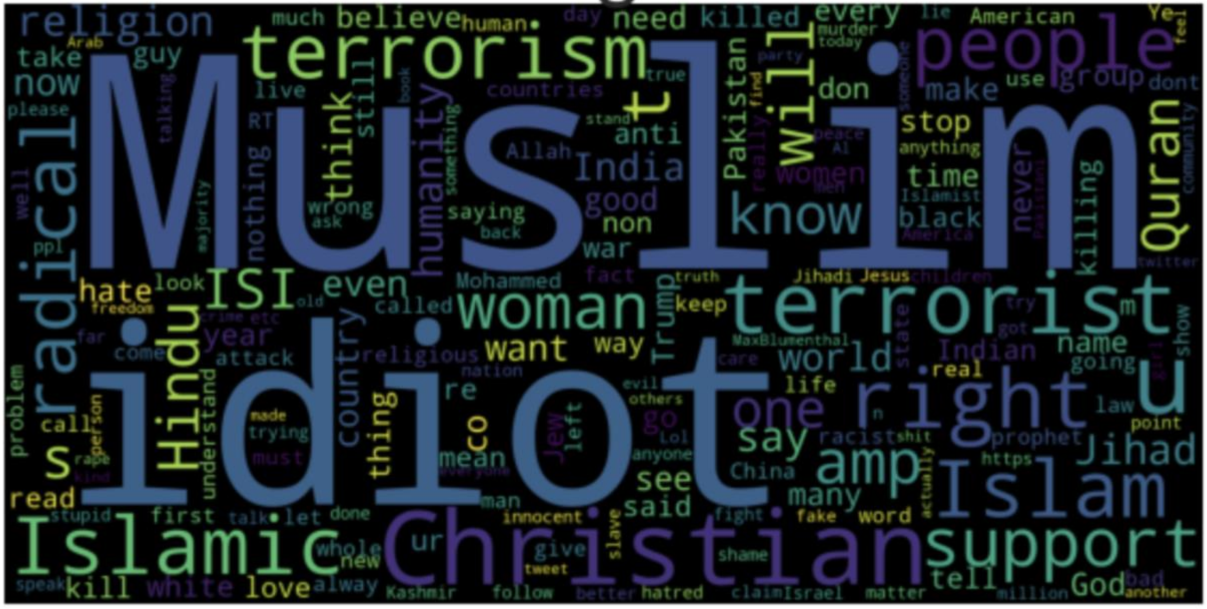

joke
rape
gay
female
call
RT
people
sexist
women
make
funny
bitch
t
s
amp
think
mkr
men
bitches
fucking
racist
know
feminazi
time
will
girl
Miley
Cyrus
prison
say
one
man
made
making
guy
even
hate
said
woman
thing
still
black
take
fuck
sex
saying
got
https
friend
stop
homophobic
right
need
date
okay
boy
feminist
go
use
shit
really
bad
u
Kat
etc
don
co
never
word
male
much
want
ok
someone
comedian
twitter
insult
ppl
tell
see
now
way
lol
love
look
new
find
good
victim
lot
anything
nigga
dog
abuse
feel
life
many
back
bit
suicide
calling
character
tweet
offensive
work
fact
using
please
first
child
around
bc
everyone
yeah
pretty
let
used
real
ass
dont
nothing
sport
going
another
London
colleague
img
MT8_9
Oh
notsexist
white
believe
face
God
kid
laugh
wrong
racism
feminism
trans
watch
give
well
football
hoe
dude
straight
assault
fun
called
something
sexual
better
slur
name
little
year
culture
two
every
cause
mean
show
gonna
person
game
old
raped
talk
stupid
literally
thought
support
video
sorry
care
anyone
reason
maybe
hard
comment
day
race
always
actually
talking
re
problem
y
m
Â
ly
ye
n

Age


school
girl
high
bullied
bully
bullies
middle
friend
kid
year
amp
one
now
people
got
bullying
know
time
went
teacher
will
old
boy
day
t
s
u
re
told
remember
think
getting
every
grade
even
elementary
back
never
didn
much
look
feel
yard
popular
saying
take
good
fuck
see
guy
made
always
don
thing
student
little
call
started
movie
wanted
story
new
fucking
mom
nurse
come
many
shit
white
really
class
want
way
mean
everyone
today
nothing
find
live
used
go
someone
going
m
hate
bitch
better
Trump
home
head
d
y
children
women
fun
weird
try
bc
daughter
others
lmao
enough
Facebook
face
reason
ugly
act
ve
work
group
whole
junior
bad
give
keep
talk
life
two
tried
ur
w
best
found
pretty
name
sorry
fat
came
anyone
real
wearing
child
gets
lol
big
anything
hair
Oh
believe
knew
yeah
hell
different
kind
around
https
seeing
use
Yes
need
say
said
first
called
stop
gay
parent
still
let
show
left
point
put
tell
thought
gonna
literally
sure
saw
well
cause
post
ass
social
love
twitter
help
man
trying
right
actually
fight
lot
care
college
im
something
hard
become
end
tall
probably
cool
black
person
beat
stand
make
away
funny



Ethnicity



Religion

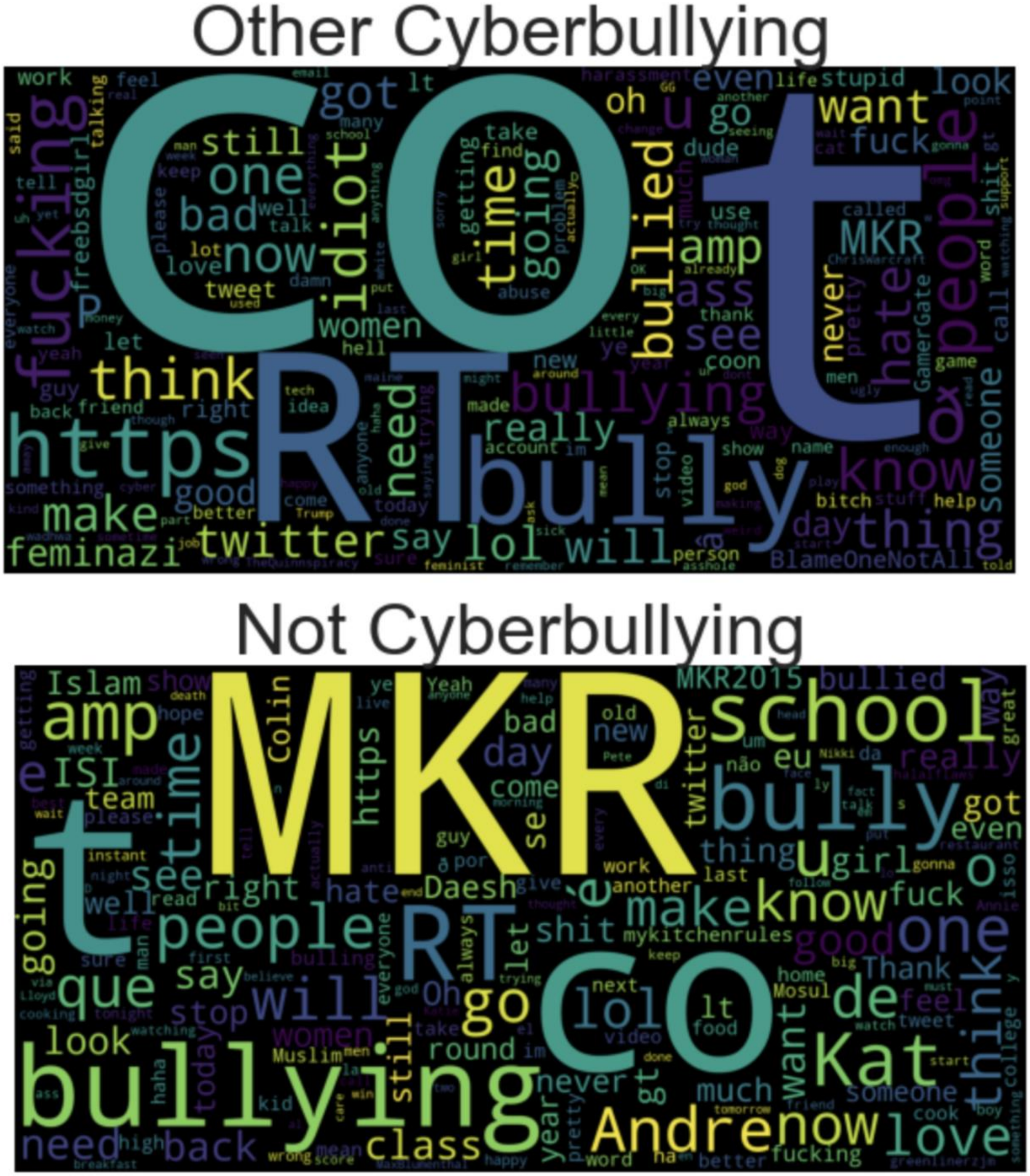


Figure 5 Word cloud for each cyberbullying type

Preprocessing and Modeling

It is important to ensure that there is equitable distribution of tweets in each class after data wrangling and EDA. This is shown in Fig. 6 depicting the count of tweets left in each category after previous steps.

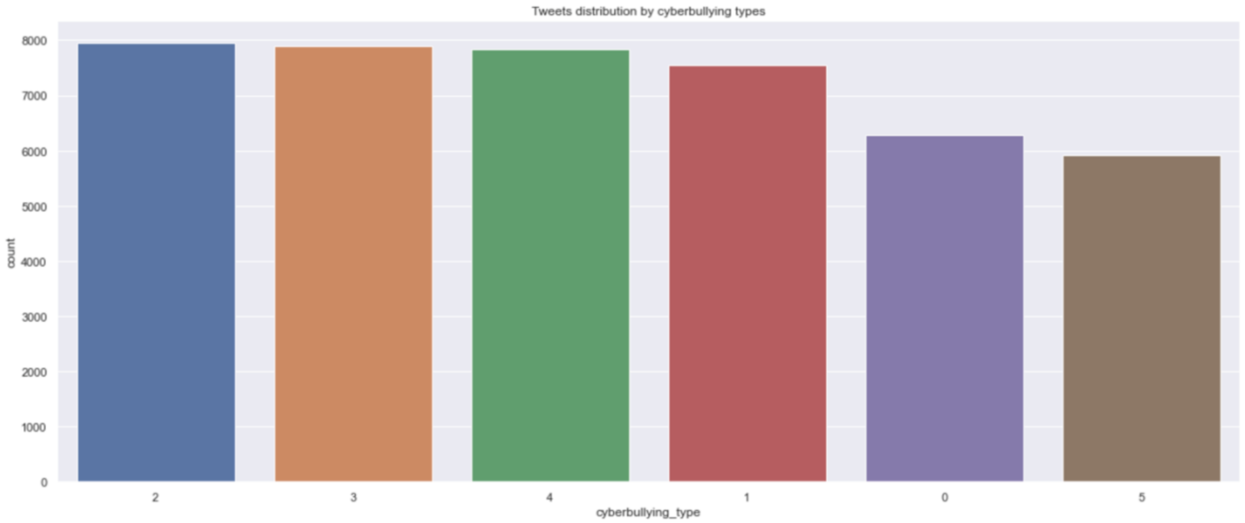
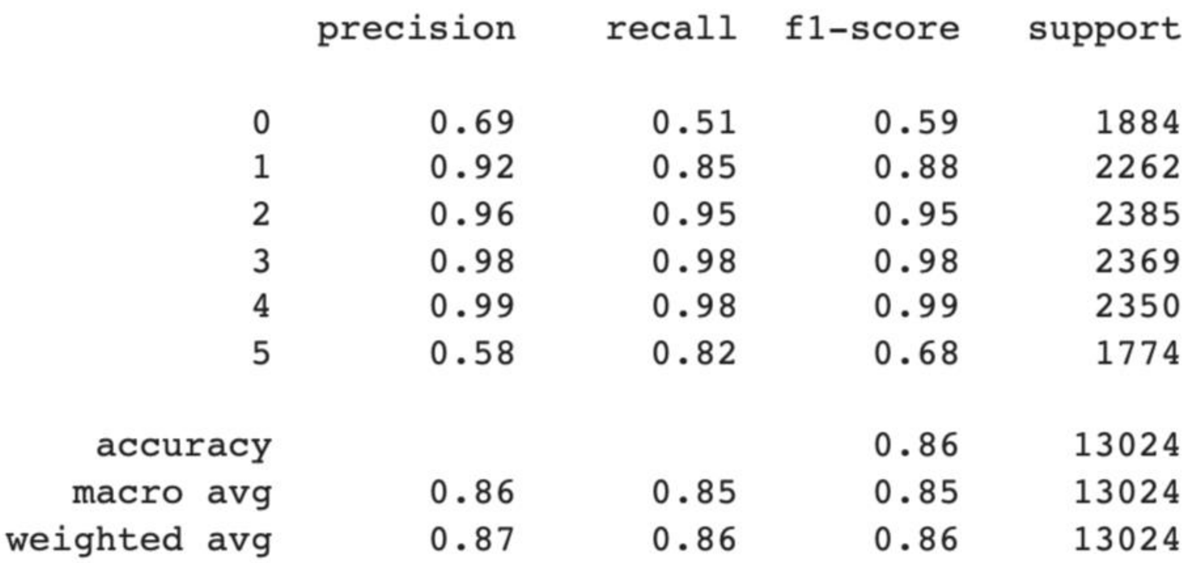


Figure 6 Distribution of tweets after data wrangling and EDA

Next, all the data was divided into training and testing datasets such that training will have 30387 samples and testing will have 13024 samples. The text data has to be converted into well-defined numerical data to be used in machine learning models. The process to convert text data into numerical data/vector, is called vectorization or word embedding. Bag-of-Words(BoW) and Word Embedding (with Word2Vec) are two well-known methods for converting text data to numerical data. However, count vectors might not be the best representation for converting text data to numerical data. So, instead of simple counting, we can also use an advanced variant of the Bag-of-Words that uses the **term frequency–inverse document frequency** (or Tf-Idf)**.** TF-IDF stands for Term Frequency — Inverse Document Frequency and is a statistic that aims to better define how important a word is for a document, while also taking into account the relation to other documents from the same corpus. This is performed by looking at how many times a word appears into a document while also paying attention to how many times the same word appears in other documents in the corpus.

The Tf-Idf vectorizer from the sklearn library has been used to fit and transform the training dataset with maximum number of features equal to 5000 and ngram range of 1-3. This fitted transformer is also used to transform the testing dataset. This is training dataset is then used to trained various classifiers and evaluated over testing dataset. Three classifiers XGBoost, K-NN and logistic regression were experimented with the dataset. The classification report for the performance of these classifiers on the testing dataset is shown in Fig. 7 which includes the precion, recall and F1 score for each class as wells the accuracy values. Fig. 8 ss regrehows the comparison of three metrics the average F1 score, accuracy, and auc-roc for each classifier. It can be seen that XGBoost and logistic regression outperformed KNN classifier significantly. The performance for XGBoost and logistic regression is similar in terms of overall metric however the logistic regression seems to have performed better on the "not_cyberbullying" class where the XGBoost has performed better on the rest of the classes. The auc-roc score for XGBoost is 0.981 which is slightly higher than the score for logistic regression which is 0.978.

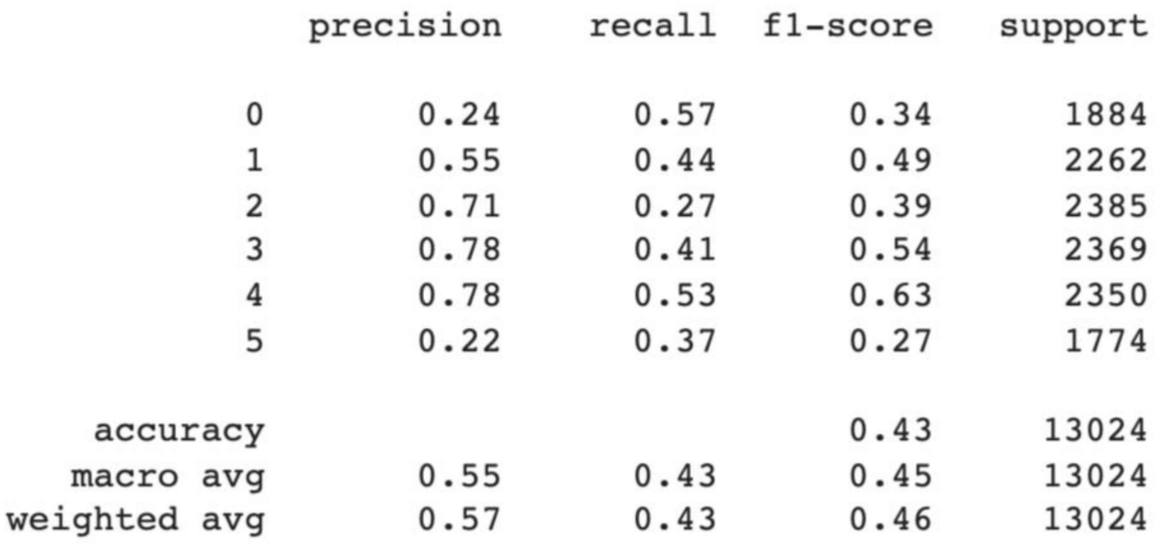


```
              precision    recall  f1-score   support

           0       0.69      0.51      0.59      1884
           1       0.92      0.85      0.88      2262
           2       0.96      0.95      0.95      2385
           3       0.98      0.98      0.98      2369
           4       0.99      0.98      0.99      2350
           5       0.58      0.82      0.68      1774

    accuracy                           0.86     13024
   macro avg       0.86      0.85      0.85     13024
weighted avg       0.87      0.86      0.86     13024
```

(a)

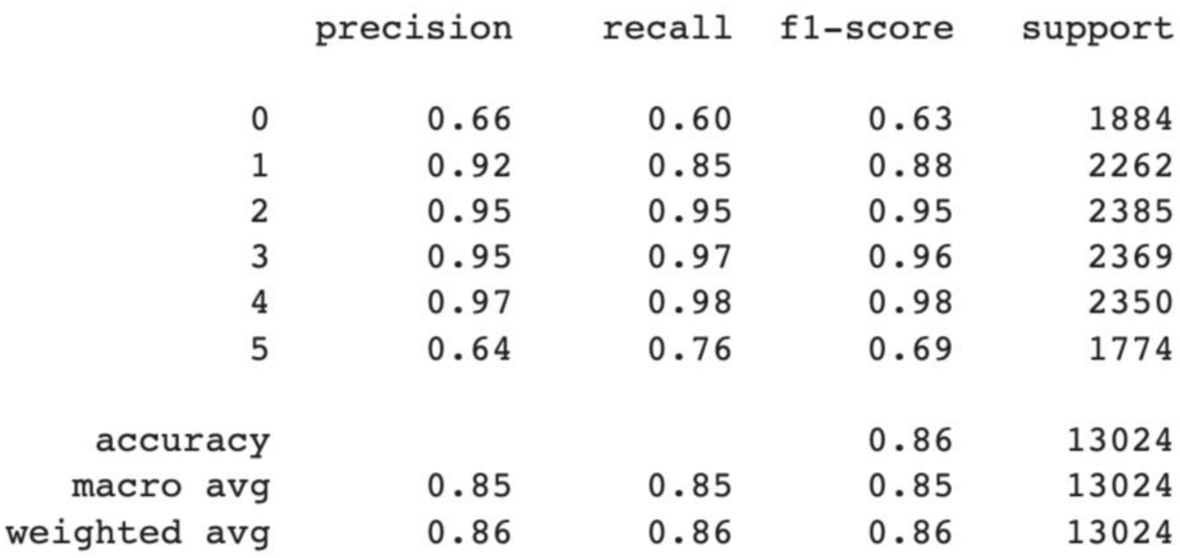


```
              precision    recall  f1-score   support

           0       0.24      0.57      0.34      1884
           1       0.55      0.44      0.49      2262
           2       0.71      0.27      0.39      2385
           3       0.78      0.41      0.54      2369
           4       0.78      0.53      0.63      2350
           5       0.22      0.37      0.27      1774

    accuracy                           0.43     13024
   macro avg       0.55      0.43      0.45     13024
weighted avg       0.57      0.43      0.46     13024
```

(b)



```
              precision    recall  f1-score   support

           0       0.66      0.60      0.63      1884
           1       0.92      0.85      0.88      2262
           2       0.95      0.95      0.95      2385
           3       0.95      0.97      0.96      2369
           4       0.97      0.98      0.98      2350
           5       0.64      0.76      0.69      1774

    accuracy                           0.86     13024
   macro avg       0.85      0.85      0.85     13024
weighted avg       0.86      0.86      0.86     13024
```

(c)

Figure 7 Classification report for (a) XGBoost; (b) KNN; and (c) Logistic regression. 0 = 'not_cyberbullying', 1= 'gender', 2 = 'religion', 3 = 'age',4 = 'ethnicity' 5 = ,'other_cyberbullying'.

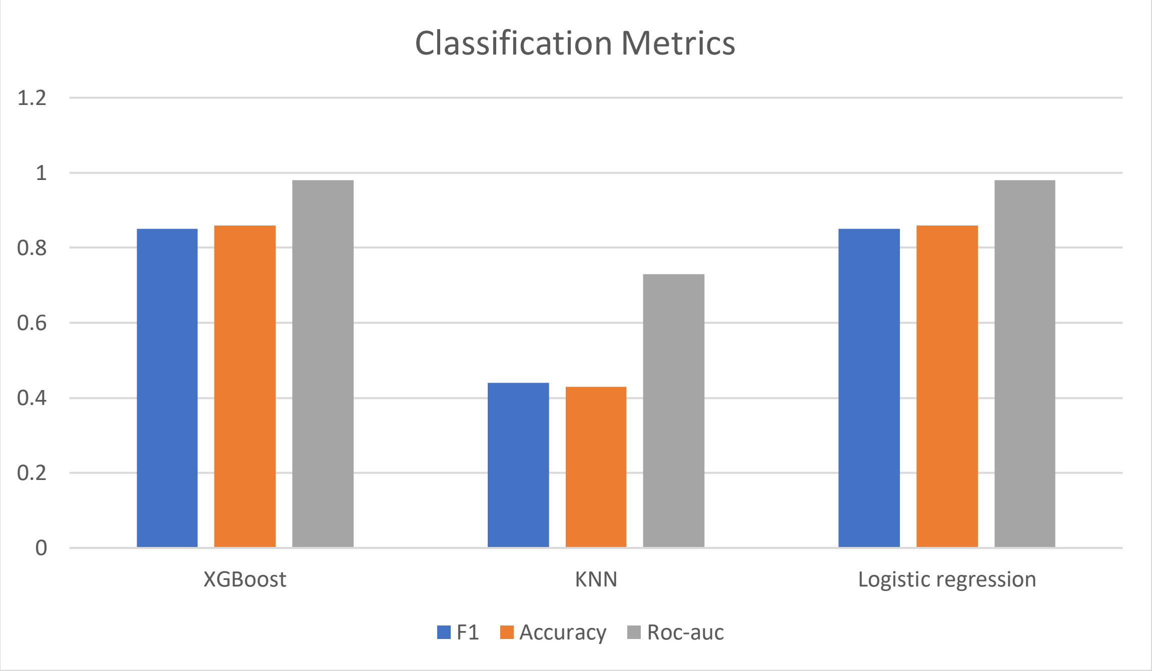


Figure 8. Classification metrics for three classifiers

Conclusion

In this project, a multi-class tweet classification model was built with Tf-idf as word embedding. The performance of classifiers were evaluated with different metrics such as F1 score, auc-roc and accuracy. Finally, the best performing model, XGBoost, was implement in the form of a

Flask API. The performance of this classifier can be further enhanced by using Gridsearch to tune the hyperparameter of the model and using advanced word-embedding methods like GloVe and BERT.