ML Project

Session-1(1-19 slides)
Session-2(20-32 slides)
Session-3(33-53 slides)

ASK REDDIT

By,
Fahed Shaikh, IMT2019079
Likhiteswar, IMT2019510

Exploratory Data Analysis.

TRAIN DATA SET

df = pd.read_csv('../input/askreddit-dataset/AskReddit Dataset/train.csv')

Understanding the Data.

df.describe()

count	653061.00000

df.shape

(653061,3)

df.columns

Index(['qid', 'question_text', 'target'], dtype='object')

df.head()

	Qid	question_text	Target
0	a3dee568776c08512c89	What is the role of Lua in Civ4?	0
1	bdb84f519e7b46e7b7bb	What are important chapters in Kannada for 10 ...	0
2	29c88db470e2eb5c97ad	Do musicians get royalties from YouTube?	0
3	3387d99bf2c3227ae8f1	What is the difference between Scaling Social ...	0
4	e79fa5038f765d0f2e7e	Why do elevators go super slow right before th...	0

df.tail()

	Qid	question_text	Target
0	a3dee568776c08512c89	What is the role of Lua in Civ4?	0
1	bdb84f519e7b46e7b7bb	What are important chapters in Kannada for 10 ...	0
2	29c88db470e2eb5c97ad	Do musicians get royalties from YouTube?	0
3	3387d99bf2c3227ae8f1	What is the difference between Scaling Social ...	0
4	e79fa5038f765d0f2e7e	Why do elevators go super slow right before th...	0

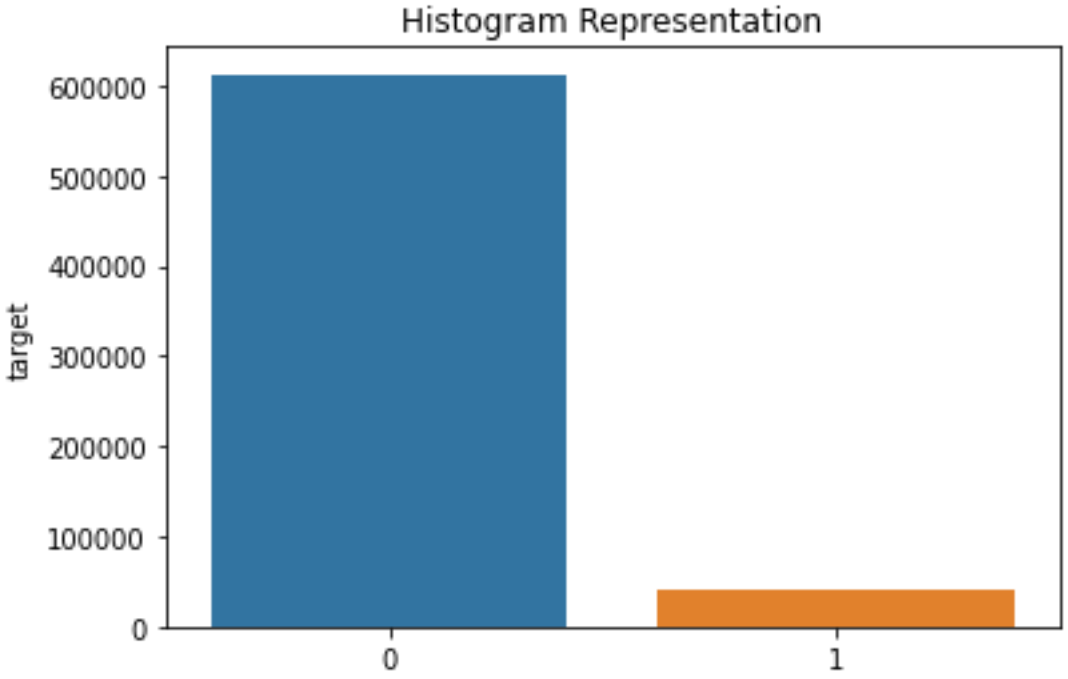
Managing the Data.

df.nunique()

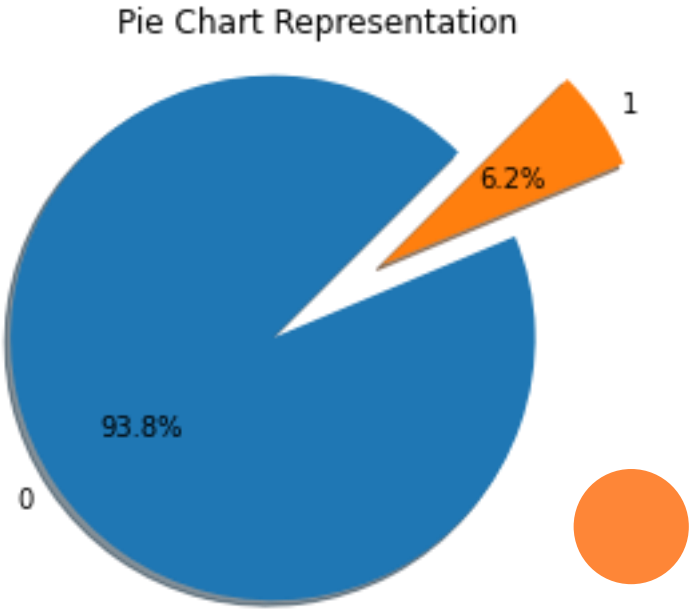
Qid	653061
question_text	653061
Target	2
Dtype	int64

df.groupby('target').describe()

	qid			question_text				
	count	**unique**	**top**	**freq**	**count**	**unique**	**top**	**freq**
target								
0	612656	612656	a3dee568776c08512c89	1	612656	612656	What is the role of Lua in Civ4?	1
1	40405	40405	19af3f158b9e37398746	1	40405	40405	What stupid things do Indians do when in your ...	1



0	Non-Troll Questions	Blue
1	Troll Questions	Orange



```
df["length"] = df["question_text"].apply(len)
df.sort_values(by = "length", ascending = True).head(10)
```

	qid	question_text	target	length
306520	527aac2ce6f12f789fe5	"	1	1
241825	3a9ae962f1094242e36f	If	1	3
644893	0f5a41d6752d5d667895	Is	1	3
333070	2cfd7dec2231e47afd6c	I 12?	0	5
334296	0c2a113858db20e0a4db	Quora:	1	7
580505	4a5c932c3b57957e71c8	Islam:	1	7
606611	6adc80c68b1f75e4540e	India:	1	7
237431	1e52e57a821c597eee0c	Dowry:	1	7
532195	83d01336b3406133723e	Bye Bye?	1	8
73729	955bcd9278b7810cd39a	Incest:	1	8

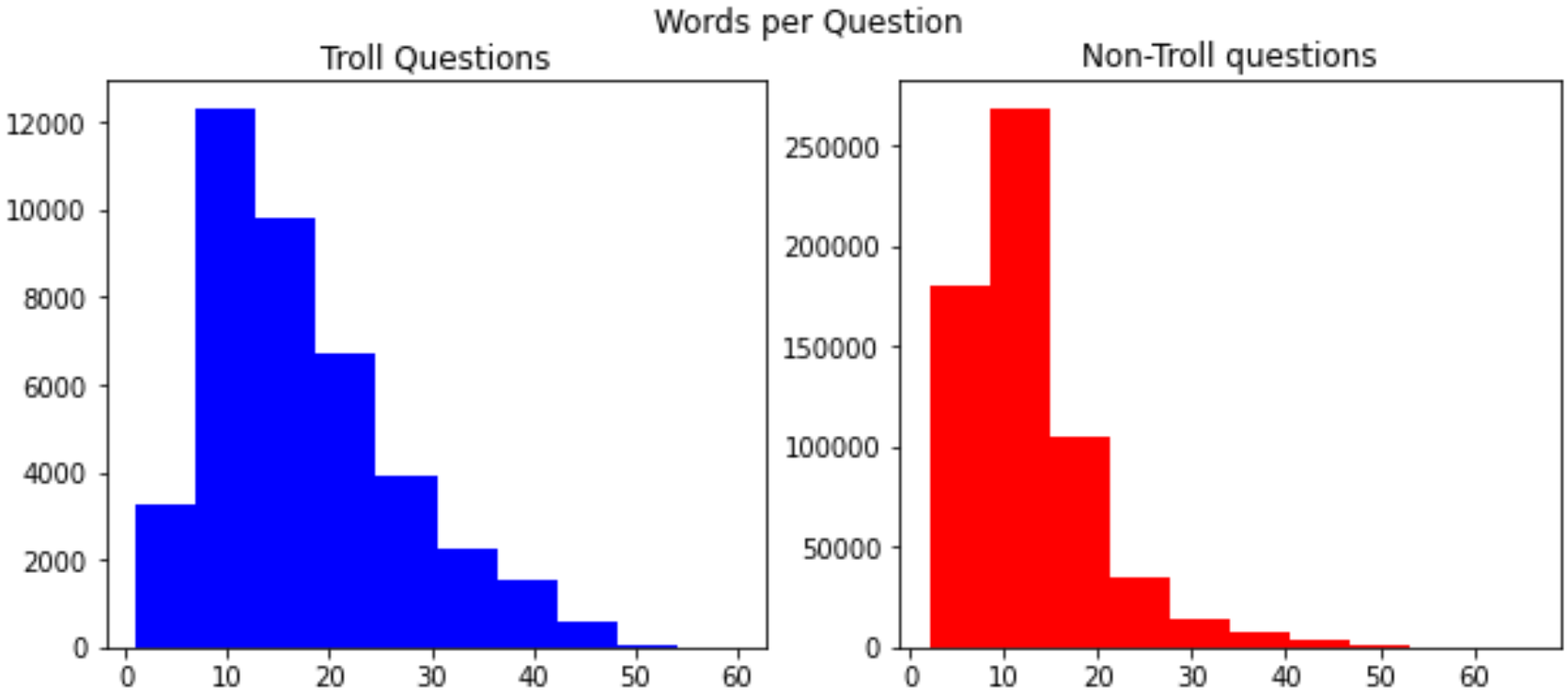
Target	Word Mean
1	17.283059027348102
0	12.505327622678958

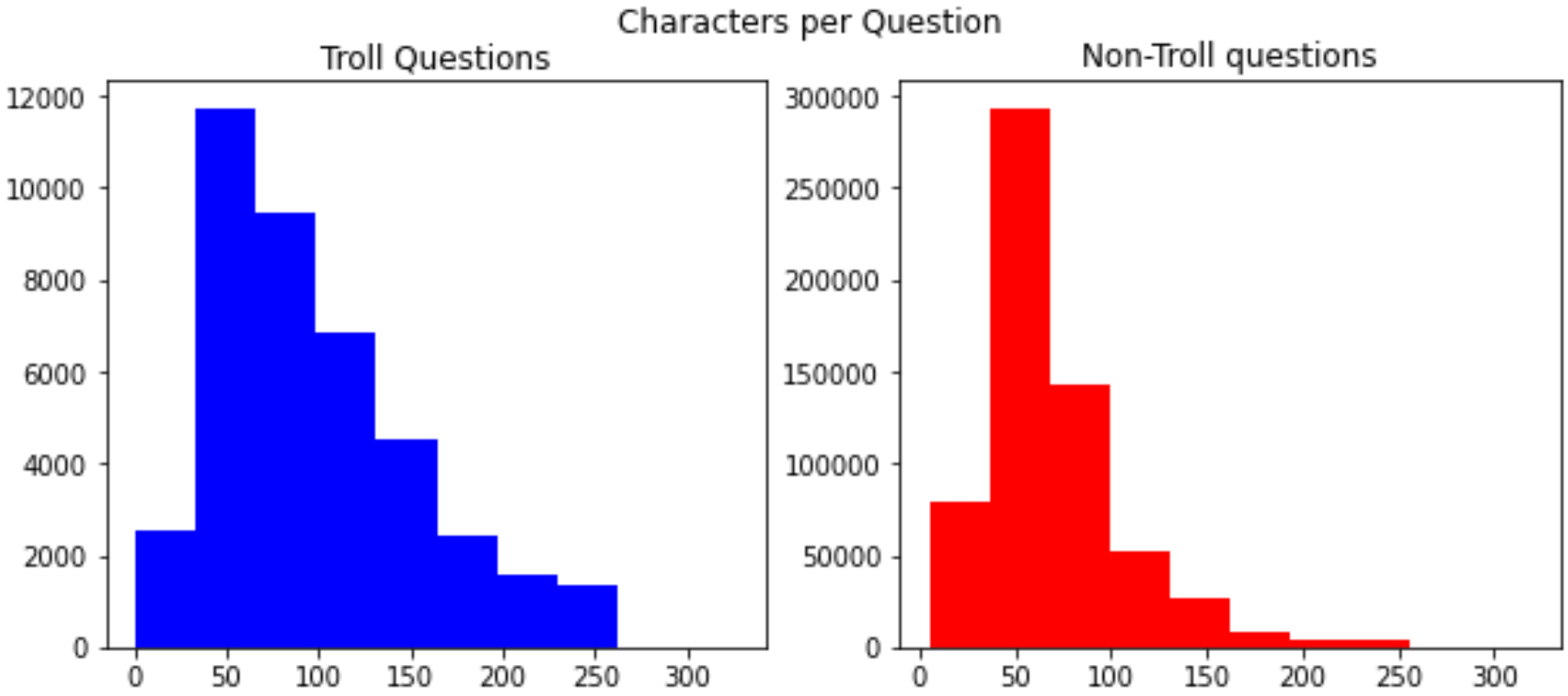
Target	Character Mean
1	98.07437198366539
0	68.86373593011413

df.isna().sum()

qid	0
question_text	0
target	0
length	0
dtype	int64

Not dropping any columns since there are no Null Values.


Words per Question
Troll Questions
Non-Troll questions
12000
10000
8000
6000
4000
2000
0
250000
200000
150000
100000
50000
0
0
10
20
30
40
50
60


Characters per Question
Troll Questions
Non-Troll questions
12000
10000
8000
6000
4000
2000
0
300000
250000
200000
150000
100000
50000
0
0
50
100
150
200
250
300

Text Pre-Processing.

Initial Table.

	qid	question_text	target	length	word_count	char_count
0	a3dee568776c08512c89	What is the role of Lua in Civ4?	0	32	8	32
1	bdb84f519e7b46e7b7bb	What are important chapters in Kannada for 10 ...	0	56	10	56
2	29c88db470e2eb5c97ad	Do musicians get royalties from YouTube?	0	40	6	40
3	3387d99bf2c3227ae8f1	What is the difference between Scaling Social ...	0	81	11	81
4	e79fa5038f765d0f2e7e	Why do elevators go super slow right before th...	0	59	11	59

Tokenization.

```
import re

def Tokenize(question):
    question = re.sub("[^a-zA-Z]", " ", question)
    Tokens = nltk.word_tokenize(question)
    return Tokens

df["Tokenized"] = df["question_text"].apply(lambda j: Tokenize(j))
df.head()
```

	qid	question_text	target	length	word_count	char_count	Tokenized
0	a3dee568776c08512c89	What is the role of Lua in Civ4?	0	32	8	32	[What, is, the, role, of, Lua, in, Civ4]
1	bdb84f519e7b46e7b7bb	What are important chapters in Kannada for 10 ...	0	56	10	56	[What, are, important, chapters, in, Kannada, ...
2	29c88db470e2eb5c97ad	Do musicians get royalties from YouTube?	0	40	6	40	[Do, musicians, get, royalties, from, YouTube]
3	3387d99bf2c3227ae8f1	What is the difference between Scaling Social ...	0	81	11	81	[What, is, the, difference, between, Scaling, ...
4	e79fa5038f765d0f2e7e	Why do elevators go super slow right before th...	0	59	11	59	[Why, do, elevators, go, super, slow, right, b...

Removing Stop Words.

```
import nltk
Unwanted = nltk.corpus.stopwords.words("english")

def RemoveStopWords(Tokens):
    NoStopWords = [word for word in Tokens if word not in Unwanted]
    return NoStopWords

df["Removed StopWords"] = df["Tokenized"].apply(lambda k: RemoveStopWords(k))
df.head()
```

	qid	question_text	target	length	word_count	char_count	Tokenized	Removed StopWords
0	a3dee568776c08512c89	What is the role of Lua in Civ4?	0	32	8	32	[What, is, the, role, of, Lua, in, Civ4]	[What, role, Lua, Civ4]
1	bdb84f519e7b46e7b7bb	What are important chapters in Kannada for 10 ...	0	56	10	56	[What, are, important, chapters, in, Kannada, ...	[What, important, chapters, Kannada, 10, ICSE,...
2	29c88db470e2eb5c97ad	Do musicians get royalties from YouTube?	0	40	6	40	[Do, musicians, get, royalties, from, YouTube]	[Do, musicians, get, royalties, YouTube]
3	3387d99bf2c3227ae8f1	What is the difference between Scaling Social ...	0	81	11	81	[What, is, the, difference, between, Scaling, ...	[What, difference, Scaling, Social, Enterprise...
4	e79fa5038f765d0f2e7e	Why do elevators go super slow right before th...	0	59	11	59	[Why, do, elevators, go, super, slow, right, b...	[Why, elevators, go, super, slow, right, doors...

Lower Casing.

```
df['Lower Case'] = df['Removed StopWords'].apply(lambda s: [word.lower() for word in s])
df.head()
```

	qid	question_text	target	length	word_count	char_count	Tokenized	Removed StopWords	Lower Case
0	a3dee568776c08512c89	What is the role of Lua in Civ4?	0	32	8	32	[What, is, the, role, of, Lua, in, Civ4]	[What, role, Lua, Civ4]	[what, role, lua, civ4]
1	bdb84f519e7b46e7b7bb	What are important chapters in Kannada for 10 ...	0	56	10	56	[What, are, important, chapters, in, Kannada, ...	[What, important, chapters, Kannada, 10, ICSE,...	[what, important, chapters, kannada, 10, icse,...
2	29c88db470e2eb5c97ad	Do musicians get royalties from YouTube?	0	40	6	40	[Do, musicians, get, royalties, from, YouTube]	[Do, musicians, get, royalties, YouTube]	[do, musicians, get, royalties, youtube]
3	3387d99bf2c3227ae8f1	What is the difference between Scaling Social ...	0	81	11	81	[What, is, the, difference, between, Scaling, ...	[What, difference, Scaling, Social, Enterprise...	[what, difference, scaling, social, enterprise...
4	e79fa5038f765d0f2e7e	Why do elevators go super slow right before th...	0	59	11	59	[Why, do, elevators, go, super, slow, right, b...	[Why, elevators, go, super, slow, right, doors...	[why, elevators, go, super, slow, right, doors..

Stemming.

```
ps = nltk.PorterStemmer()

def stemming(SmallCaps):
    stem = [ps.stem(word) for word in SmallCaps]
    return stem

df["Stemmed Text"] = df["Lower Case"].apply(lambda st: stemming(st))
df.head()
```

	qid	question_text	Tokenized	Removed StopWords	Lower Case	Stemmed Text
0	a3dee568776c08512c89	What is the role of Lua in Civ4?	[What, is, the, role, of, Lua, in, Civ4]	[What, role, Lua, Civ4]	[what, role, lua, civ4]	[what, role, lua, civ4]
1	bdb84f519e7b46e7b7bb	What are important chapters in Kannada for 10 ...	[What, are, important, chapters, in, Kannada, ...	[What, important, chapters, Kannada, 10, ICSE,...	[what, important, chapters, kannada, 10, icse,...	[what, import, chapter, kannada, 10, ics, 2018]
2	29c88db470e2eb5c97ad	Do musicians get royalties from YouTube?	[Do, musicians, get, royalties, from, YouTube]	[Do, musicians, get, royalties, YouTube]	[do, musicians, get, royalties, youtube]	[do, musician, get, royalti, youtub]
3	3387d99bf2c3227ae8f1	What is the difference between Scaling Social ...	[What, is, the, difference, between, Scaling, ...	[What, difference, Scaling, Social, Enterprise...	[what, difference, scaling, social, enterprise...	[what, differ, scale, social, enterpris, socia...
4	e79fa5038f765d0f2e7e	Why do elevators go super slow right before th...	[Why, do, elevators, go, super, slow, right, b...	[Why, elevators, go, super, slow, right, doors...	[why, elevators, go, super, slow, right, doors...	[whi, elev, go, super, slow, right, door, open]

Lemmatization.

```
wnl = nltk.WordNetLemmatizer()

def Lemmatizing(stemmed):
    lemm = [wnl.lemmatize(word) for word in stemmed]
    return lemm

df["Lemmatized Text"] = df["Stemmed Text"].apply(lambda lm: Lemmatizing(lm))
df.head()
```

	qid	question_text	No Punctuations	Tokenized	Removed StopWords	Lower Case	Stemmed Text	Lemmatized Text
0	a3dee568776c08512c89	What is the role of Lua in Civ4?	What is the role of Lua in Civ4	[What, is, the, role, of, Lua, in, Civ4]	[What, role, Lua, Civ4]	[what, role, lua, civ4]	[what, role, lua, civ4]	[what, role, lua, civ4]
1	bdb84f519e7b46e7b7bb	What are important chapters in Kannada for 10 ...	What are important chapters in Kannada for 10 ...	[What, are, important, chapters, in, Kannada, ...	[What, important, chapters, Kannada, 10, ICSE,...	[what, important, chapters, kannada, 10, icse,...	[what, import, chapter, kannada, 10, ics, 2018]	[what, import, chapter, kannada, 10, ic, 2018]
2	29c88db470e2eb5c97ad	Do musicians get royalties from YouTube?	Do musicians get royalties from YouTube	[Do, musicians, get, royalties, from, YouTube]	[Do, musicians, get, royalties, YouTube]	[do, musicians, get, royalties, youtube]	[do, musician, get, royalti, youtub]	[do, musician, get, royalti, youtub]
3	3387d99bf2c3227ae8f1	What is the difference between Scaling Social ...	What is the difference between Scaling Social ...	[What, is, the, difference, between, Scaling, ...	[What, difference, Scaling, Social, Enterprise...	[what, difference, scaling, social, enterprise...	[what, differ, scale, social, enterpris, socia...	[what, differ, scale, social, enterpris, socia...
4	e79fa5038f765d0f2e7e	Why do elevators go super slow right before th...	Why do elevators go super slow right before th...	[Why, do, elevators, go, super, slow, right, b...	[Why, elevators, go, super, slow, right, doors...	[why, elevators, go, super, slow, right, doors...	[whi, elev, go, super, slow, right, door, open]	[whi, elev, go, super, slow, right, door, open]

Vectorization.

```
[ ]: vectorizer = CountVectorizer(analyzer= 'word', tokenizer = tokenize, lowercase= True, stop_words= 'english', max_features= 85)
```

```
[ ]: features = vectorizer.fit_transform(df.question_text.tolist())
```

```
[ ]: features_nd = features.toarray()
```

Additional Work:
Performed Simple Logistic Regression.

[99]:
```
X_train, X_test, y_train, y_test = train_test_split(features_nd[0:len(df)], df.target, test_size= 0.3, random_state= 1234)
```

[100]:
```
from sklearn.linear_model import LogisticRegression

log_model = LogisticRegression()
```

[101]:
```
log_model = log_model.fit(X_train, y_train)
```

[102]:
```
y_pred = log_model.predict(X_test)
```

[ ]:
```
print(y_pred)
```

```
from sklearn.metrics import classification_report

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.95      0.99      0.97    183919
           1       0.57      0.12      0.20     12000

    accuracy                           0.94    195919
   macro avg       0.76      0.56      0.59    195919
weighted avg       0.92      0.94      0.92    195919
```

SESSION-2

BERNOULLI MODEL :

- Bernoulli model is a variant of Naïve-Bayes classifier.

CODE:

#BERNOULLI MODEL.

```
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.metrics import f1_score
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, TfidfTransformer
from sklearn.naive_bayes import BernoulliNB
from sklearn.model_selection import cross_val_predict
```

+ Code + Markdown

[20]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size= 0.2, random_state= 0)
```

[21]:
```
pipeline_model = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', BernoulliNB())])
```

[22]:
```
pipeline_model.fit(X_train, y_train)
```

[22]:
```
Pipeline(steps=[('vect', CountVectorizer()), ('tfidf', TfidfTransformer()),
                ('clf', BernoulliNB())])
```

+ Code + Markdown

[23]:
```
y_pred = pipeline_model.predict(X_test)
f1_score(y_test, y_pred, average='macro')
```

[23]:
```
0.7426241778994002
```

PRE-PROCESSING OF TEST DATA:

TOKENIZING THE DATA:

TEST DATA

[24]:
```
db = pd.read_csv('../input/testdataset/test.csv')
```

[25]:
```
db["Tokenized"] = db["question_text"].apply(lambda j: Tokenize(j))
db.head()
```

[25]:

	qid	question_text	Tokenized
0	0a824224322f0a36025f	Why is my fish tank so cloudy?	[Why, is, my, fish, tank, so, cloudy]
1	28af14c4e4777ce1273e	Are AAP supporters/leaders hypocrites?	[Are, AAP, supporters, leaders, hypocrites]
2	6892a52c51103dd95044	Can you still get a ticket if you shut off you...	[Can, you, still, get, a, ticket, if, you, shu...
3	badd9e8886d73fc1fe4e	Why should any liberal or caring person want t...	[Why, should, any, liberal, or, caring, person...
4	4ef178f82a465e4804ae	How can I know who got into my PC using anydesk?	[How, can, I, know, who, got, into, my, PC, us...

LOWER CASING :

[26]:
```
db['Lower Case'] = db['Tokenized'].apply(lambda s: [word.lower() for word in s])
db.head()
```

[26]:

	qid	question_text	Tokenized	Lower Case
0	0a824224322f0a36025f	Why is my fish tank so cloudy?	[Why, is, my, fish, tank, so, cloudy]	[why, is, my, fish, tank, so, cloudy]
1	28af14c4e4777ce1273e	Are AAP supporters/leaders hypocrites?	[Are, AAP, supporters, leaders, hypocrites]	[are, aap, supporters, leaders, hypocrites]
2	6892a52c51103dd95044	Can you still get a ticket if you shut off you...	[Can, you, still, get, a, ticket, if, you, shu...	[can, you, still, get, a, ticket, if, you, shu...
3	badd9e8886d73fc1fe4e	Why should any liberal or caring person want t...	[Why, should, any, liberal, or, caring, person...	[why, should, any, liberal, or, caring, person...
4	4ef178f82a465e4804ae	How can I know who got into my PC using anydesk?	[How, can, I, know, who, got, into, my, PC, us...	[how, can, i, know, who, got, into, my, pc, us...

REMOVING STOP WORDS:

[27]:
```
db["Removed StopWords"] = db["Lower Case"].apply(lambda k: RemoveStopWords(k))
db.head()
```

[27]:

	qid	question_text	Tokenized	Lower Case	Removed StopWords
0	0a824224322f0a36025f	Why is my fish tank so cloudy?	[Why, is, my, fish, tank, so, cloudy]	[why, is, my, fish, tank, so, cloudy]	[fish, tank, cloudy]
1	28af14c4e4777ce1273e	Are AAP supporters/leaders hypocrites?	[Are, AAP, supporters, leaders, hypocrites]	[are, aap, supporters, leaders, hypocrites]	[aap, supporters, leaders, hypocrites]
2	6892a52c51103dd95044	Can you still get a ticket if you shut off you...	[Can, you, still, get, a, ticket, if, you, shu...	[can, you, still, get, a, ticket, if, you, shu...	[still, get, ticket, shut, car, cop, flips, li...
3	badd9e8886d73fc1fe4e	Why should any liberal or caring person want t...	[Why, should, any, liberal, or, caring, person...	[why, should, any, liberal, or, caring, person...	[liberal, caring, person, want, stay, country,...
4	4ef178f82a465e4804ae	How can I know who got into my PC using anydesk?	[How, can, I, know, who, got, into, my, PC, us...	[how, can, i, know, who, got, into, my, pc, us...	[know, got, pc, using, anydesk]

STEMMING:

[28]:
```
db["Stemmed Text"] = db["Removed StopWords"].apply(lambda st: stemming(st))
db.head()
```

[28]:

	qid	question_text	Tokenized	Lower Case	Removed StopWords	Stemmed Text
0	0a824224322f0a36025f	Why is my fish tank so cloudy?	[Why, is, my, fish, tank, so, cloudy]	[why, is, my, fish, tank, so, cloudy]	[fish, tank, cloudy]	fish tank cloudi
1	28af14c4e4777ce1273e	Are AAP supporters/leaders hypocrites?	[Are, AAP, supporters, leaders, hypocrites]	[are, aap, supporters, leaders, hypocrites]	[aap, supporters, leaders, hypocrites]	aap support leader hypocrit
2	6892a52c51103dd95044	Can you still get a ticket if you shut off you...	[Can, you, still, get, a, ticket, if, you, shu...	[can, you, still, get, a, ticket, if, you, shu...	[still, get, ticket, shut, car, cop, flips, li...	still get ticket shut car cop flip light
3	badd9e8886d73fc1fe4e	Why should any liberal or caring person want t...	[Why, should, any, liberal, or, caring, person...	[why, should, any, liberal, or, caring, person...	[liberal, caring, person, want, stay, country,...	liber care person want stay countri nine mont...
4	4ef178f82a465e4804ae	How can I know who got into my PC using anydesk?	[How, can, I, know, who, got, into, my, PC, us...	[how, can, i, know, who, got, into, my, pc, us...	[know, got, pc, using, anydesk]	know got pc use anydesk

LEMMATIZING :

```
db["Lemmatized Text"] = db["Stemmed Text"].apply(lambda lm: Lemmatizing(lm))
db.head()
```

[29]:

	qid	question_text	Tokenized	Lower Case	Removed StopWords	Stemmed Text	Lemmatized Text
0	0a824224322f0a36025f	Why is my fish tank so cloudy?	[Why, is, my, fish, tank, so, cloudy]	[why, is, my, fish, tank, so, cloudy]	[fish, tank, cloudy]	fish tank cloudi	f i s h t a n k c l o u d i
1	28af14c4e4777ce1273e	Are AAP supporters/leaders hypocrites?	[Are, AAP, supporters, leaders, hypocrites]	[are, aap, supporters, leaders, hypocrites]	[aap, supporters, leaders, hypocrites]	aap support leader hypocrit	a a p s u p p o r t l e a d e r h y p...
2	6892a52c51103dd95044	Can you still get a ticket if you shut off you...	[Can, you, still, get, a, ticket, if, you, shu...	[can, you, still, get, a, ticket, if, you, shu...	[still, get, ticket, shut, car, cop, flips, li...	still get ticket shut car cop flip light	s t i l l g e t t i c k e t s h u t ...
3	badd9e8886d73fc1fe4e	Why should any liberal or caring person want t...	[Why, should, any, liberal, or, caring, person...	[why, should, any, liberal, or, caring, person...	[liberal, caring, person, want, stay, country,...	liber care person want stay countri nine mont...	l i b e r c a r e p e r s o n w a n t...
4	4ef178f82a465e4804ae	How can I know who got into my PC using anydesk?	[How, can, I, know, who, got, into, my, PC, us...	[how, can, i, know, who, got, into, my, pc, us...	[know, got, pc, using, anydesk]	know got pc use anydesk	k n o w g o t p c u s e a n y d e s k

PREDICTING THE RESULT:

[30]:
```
db_X = db['Stemmed Text']
```

[31]:
```
pipeline_model.fit(X, y)
```

[31]:
```
Pipeline(steps=[('vect', CountVectorizer()), ('tfidf', TfidfTransformer()),
                ('clf', BernoulliNB())])
```

[32]:
```
result = pipeline_model.predict(db_X)
```

+ Code + Markdown

[33]:
```
db["target"] = result

to_submit = db[['qid',"target"]]

to_submit.to_csv("Pipeline-Bernoulli-Stemmed.csv", index = False)
```

[34]:
```
to_submit["target"].value_counts()
```

[34]:
```
0    613075
1     39986
Name: target, dtype: int64
```

DECISION TREE CLASSIFIER:

USING PIPELINE-LR ON TRAIN DATA:

[14]:
```
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.metrics import f1_score
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, TfidfTransformer
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
```

[15]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size= 0.2, random_state= 0)
```

[16]:
```
pipeline_model = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('dtc', DecisionTreeClassifier(random_state = 0))])
```

+ Code + Markdown

```
pipeline_model.fit(X_train, y_train)
```

[17]:
```
Pipeline(steps=[('vect', CountVectorizer()), ('tfidf', TfidfTransformer()),
                ('dtc', DecisionTreeClassifier(random_state=0))])
```

+ Code + Markdown

[18]:
```
y_pred = pipeline_model.predict(X_test)
f1_score(y_test, y_pred, average='macro')
```

[18]:
```
0.6969481011479202
```

USING MODEL ON TEST DATA:

```
db_X = db['Lemmatized Text']
```

```
pipeline_model.fit(X, y)
```

```
result = pipeline_model.predict(db_X)
```

```
db["target"] = result

to_submit = db[['qid',"target"]]

to_submit.to_csv("Pipeline-LR.csv", index = False)
```

+ Code + Markdown

```
to_submit["target"].value_counts()
```

SESSION-3

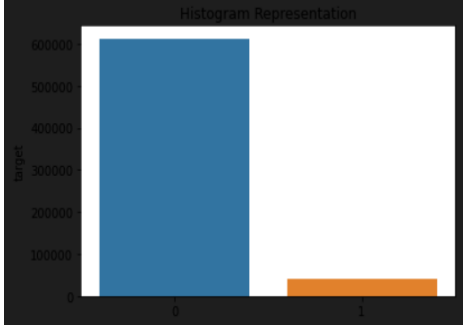
EDA:

- We added few more things in the EDA process like new plots to understand the data better.

BAR PLOT OF VALUE COUNT:

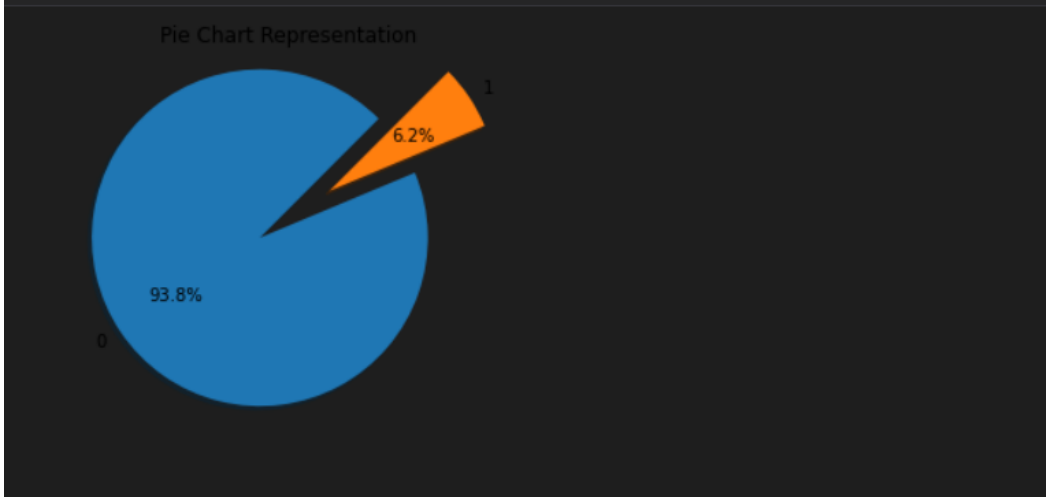
```
x = df['target'].value_counts()

print(x)

sns.barplot(x.index, x)
plt.title("Histogram Representation")
```

```
0    612656
1     40405
Name: target, dtype: int64
C:\Users\fahed\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
Text(0.5, 1.0, 'Histogram Representation')
```



PIE CHART REPRESENTATION OF VALUE COUNT:

```
explode = (0.1, 0.4)

fig, ax = plt.subplots()
ax.pie(x,
       explode=explode,
       labels=x.index,
       autopct='%1.1f%%',
       shadow=True,
       startangle=45)
ax.axis('equal')  # Equal aspect ratio ensures the pie chart is circular.
ax.set_title('Pie Chart Representation')


plt.show()
```



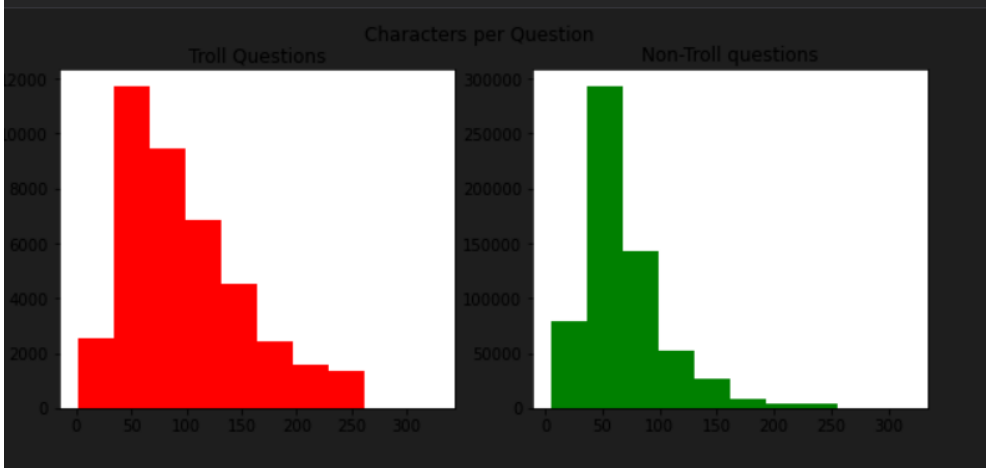
SORTING DATA REGARDING LENGTH:

```
df["length"] = df["question_text"].apply(len)
df.sort_values(by = "length", ascending = True).head(10)
```

	qid	question_text	target	length
306520	527aac2ce6f12f789fe5	"	1	1
241825	3a9ae962f1094242e36f	If	1	3
644893	0f5a41d6752d5d667895	Is	1	3
333070	2cfd7dec2231e47afd6c	I 12?	0	5
334296	0c2a113858db20e0a4db	Quora:	1	7
580505	4a5c932c3b57957e71c8	Islam:	1	7
606611	6adc80c68b1f75e4540e	India:	1	7
237431	1e52e57a821c597eee0c	Dowry:	1	7
532195	83d01336b3406133723e	Bye Bye?	1	8
73729	955bcd9278b7810cd39a	Incest:	1	8

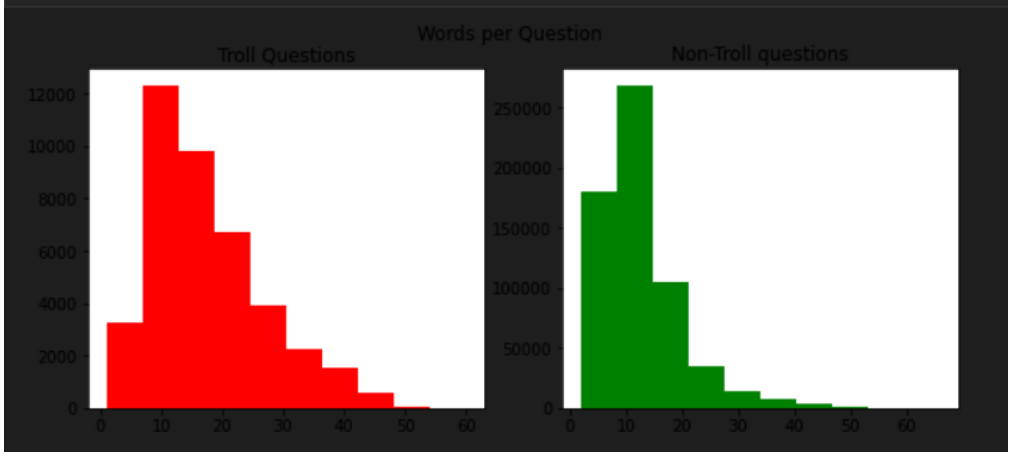
HISTOGRAM OF TROLL AND NON-TROLL QUESTIONS:

```
fig,(ax1,ax2)=plt.subplots(1,2,figsize=(10,4))

train_words=df[df['target']==1]['length']

ax1.hist(train_words,color='red')

ax1.set_title('Troll Questions')

train_words=df[df['target']==0]['length']

ax2.hist(train_words,color='green')

ax2.set_title('Non-Troll questions')

fig.suptitle('Characters per Question')

plt.show()
```



WORD COUNT IN TROLL AND NON-TROLL QUESTIONS:

```
fig,(ax1,ax2)=plt.subplots(1,2,figsize=(10,4))
train_words=df[df['target']==1]['word_count']
ax1.hist(train_words,color='Red')
ax1.set_title('Troll Questions')
train_words=df[df['target']==0]['word_count']
ax2.hist(train_words,color='Green')
ax2.set_title('Non-Troll questions')
fig.suptitle('Words per Question')
plt.show()
```



PRE PROCESSING:

- Did a few changes in the following.

REMOVING STOPWORDS:

```
    Unwanted = set(stopwords.words('english'))
    #Unwanted = nltk.corpus.stopwords.words("english")

    def RemoveStopWords(Tokens):
        NoStopWords = [word for word in Tokens if word not in Unwanted]
        return NoStopWords

    df["Removed StopWords"] = df["Lower Case"].apply(lambda k: RemoveStopWords(k))
    df.head()
```

```
Output exceeds the size limit. Open the full output data in a text editor
                   qid                                      question_text  \
0  a3dee568776c08512c89                    What is the role of Lua in Civ4?
1  bdb84f519e7b46e7b7bb  What are important chapters in Kannada for 10 ...
2  29c88db470e2eb5c97ad           Do musicians get royalties from YouTube?
3  3387d99bf2c3227ae8f1  What is the difference between Scaling Social ...
4  e79fa5038f765d0f2e7e  Why do elevators go super slow right before th...

   target  length  word_count  char_count  \
0       0      32           8          32
1       0      56          10          56
2       0      40           6          40
3       0      81          11          81
4       0      59          11          59

                                      Tokenized  \
0           [What, is, the, role, of, Lua, in, Civ]
1  [What, are, important, chapters, in, Kannada, ...
2     [Do, musicians, get, royalties, from, YouTube]
3  [What, is, the, difference, between, Scaling, ...
4  [Why, do, elevators, go, super, slow, right, b...

                                      Lower Case  \
0           [what, is, the, role, of, lua, in, civ]
1  [what, are, important, chapters, in, kannada, ...
2     [do, musicians, get, royalties, from, youtube]
...

                                Removed StopWords
0                                [role, lua, civ]
1              [important, chapters, kannada, icse]
2              [musicians, get, royalties, youtube]
3  [difference, scaling, social, enterprises, soc...
4   [elevators, go, super, slow, right, doors, open]
```

PORT-STEMMER:

```
    ps = nltk.PorterStemmer()

    def stemming(SmallCaps):
        stem = [ps.stem(word) for word in SmallCaps]
        return stem

    df["Stemmed Text"] = df["Removed StopWords"].apply(lambda st: stemming(st))
    df.head()
```

```
Output exceeds the size limit. Open the full output data in a text editor
                    qid                                      question_text  \
0  a3dee568776c08512c89                  What is the role of Lua in Civ4?
1  bdb84f519e7b46e7b7bb  What are important chapters in Kannada for 10 ...
2  29c88db470e2eb5c97ad           Do musicians get royalties from YouTube?
3  3387d99bf2c3227ae8f1  What is the difference between Scaling Social ...
4  e79fa5038f765d0f2e7e  Why do elevators go super slow right before th...

   target  length  word_count  char_count  \
0       0      32           8          32
1       0      56          10          56
2       0      40           6          40
3       0      81          11          81
4       0      59          11          59

                                      Tokenized  \
0           [What, is, the, role, of, Lua, in, Civ]
1  [What, are, important, chapters, in, Kannada, ...
2     [Do, musicians, get, royalties, from, YouTube]
3  [What, is, the, difference, between, Scaling, ...
4  [Why, do, elevators, go, super, slow, right, b...

                                      Lower Case  \
0           [what, is, the, role, of, lua, in, civ]
1  [what, are, important, chapters, in, kannada, ...
2     [do, musicians, get, royalties, from, youtube]
...

                                    Stemmed Text
0                                  [role, lua, civ]
1                      [import, chapter, kannada, ics]
2                     [musician, get, royalti, youtub]
3  [differ, scale, social, enterpris, social, fra...
4          [elev, go, super, slow, right, door, open]
```

LEMMATIZER:

```
wnl = nltk.WordNetLemmatizer()

def Lemmatizing(stemmed):
    lemm = [wnl.lemmatize(word) for word in stemmed]
    string = ""
    for s in lemm:
        string = string + ' ' + s
    return string


df["Lemmatized Text"] = df["Stemmed Text"].apply(lambda lm: Lemmatizing(lm))
df.head()
```

```
Output exceeds the size limit. Open the full output data in a text editor
                   qid                                      question_text  \
0  a3dee568776c08512c89                  What is the role of Lua in Civ4?
1  bdb84f519e7b46e7b7bb  What are important chapters in Kannada for 10 ...
2  29c88db470e2eb5c97ad           Do musicians get royalties from YouTube?
3  3387d99bf2c3227ae8f1  What is the difference between Scaling Social ...
4  e79fa5038f765d0f2e7e  Why do elevators go super slow right before th...

   target  length  word_count  char_count  \
0       0      32           8          32
1       0      56          10          56
2       0      40           6          40
3       0      81          11          81
4       0      59          11          59

                                      Tokenized  \
0           [What, is, the, role, of, Lua, in, Civ]
1  [What, are, important, chapters, in, Kannada, ...
2     [Do, musicians, get, royalties, from, YouTube]
3  [What, is, the, difference, between, Scaling, ...
4  [Why, do, elevators, go, super, slow, right, b...

                                      Lower Case  \
0           [what, is, the, role, of, lua, in, civ]
1  [what, are, important, chapters, in, kannada, ...
2     [do, musicians, get, royalties, from, youtube]
...

                                Lemmatized Text
0                                  role lua civ
1                        import chapter kannada ic
2                       musician get royalti youtub
3   differ scale social enterpris social franchis
4               elev go super slow right door open
```

SPLITTING DATA:

```
qt = df['question_text']
target = df['target']
```

```
#get all words from spam and non-spam datasets
train_spam_words = ' '.join(df[df.target == True].question_text).split(' ')
train_non_spam_words = ' '.join(df[df.target == False].question_text).split(' ')

common_words = set(train_spam_words).intersection(set(train_non_spam_words))
```

```
df
```

```
..                                          question_text  target  length  \
0                                           role lua civ       0      32
1                                 import chapter kannada ic       0      56
2                                musician get royalti youtub       0      40
3             differ scale social enterpris social franchis       0      81
4                       elev go super slow right door open       0      59
...                                                    ...     ...     ...
653056                    coach centr best delhi ia prepar       0      59
653057   api check cibil score download credit report ...       0      74
653058   paranorm show spirit talk medium use imag ins...       0      82
653059                           josh gordon well develop ab       0      50
653060                               mani way methan extract       0      42

        word_count  char_count
0                8          32
1               10          56
2                6          40
3               11          81
4               11          59
...            ...         ...
653056          10          59
653057          14          74
653058          14          82
653059           9          50
653060           8          42

[653061 rows x 5 columns]
```

DIVIDING 1'S AND 0'S:

```
ones = df[df.target == 1]
zeroes = df[df.target == 0]
```

```
len(zeroes)
```

```
...  612656
```

```
len(ones)
```

```
...  40405
```

RESAMPLING:

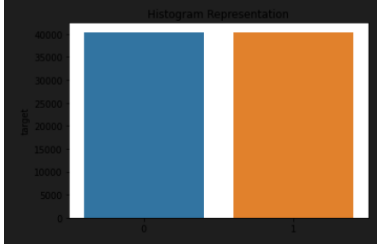
```python
from sklearn.utils import resample

balanced_df = pd.concat([resample(zeroes, replace=True, n_samples=len(ones)), ones])
```

Python

```python
x = balanced_df['target'].value_counts()

print(x)

sns.barplot(x.index, x)
plt.title("Histogram Representation")
```

Python

```
0    40405
1    40405
Name: target, dtype: int64
C:\Users\fahed\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:

Pass the following variables as keyword args: x, y. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.

Text(0.5, 1.0, 'Histogram Representation')
```



```
balanced_df
```

```
                                        question_text  target  length  \
274442                       aftermath sink rm lusitania       0      59
457865                        safe carri lpg cylind elev       0      48
647862                                       use protocol       0      29
472028     industri design motorcycl protect shield cover       0      75
550800   salari aircraft maintan engin ame dgca certif...       0      98
...                                                    ...     ...     ...
652967        liber understand differ pollut climat chang       1      89
653021   unattract averag look men ever get girlfriend...       1     144
653029                                 grab aunti boob p       1      36
653034                       girl like treat like sex toy       1      43
653049   liber also concern lot quora question phrase ...       1     138

        word_count  char_count
274442          11          59
457865          10          48
647862           6          29
472028          10          75
550800          16          98
...            ...         ...
652967          13          89
653021          25         144
653029           8          36
653034           9          43
653049          23         138

[80810 rows x 5 columns]
```

```
X =balanced_df["question_text"]
y = balanced_df["target"]
```

BAG OF WORDS:

```
# Bag Of Words
```

```
FRAC_troll_TEXTS = balanced_df.target.mean()
```

```
#get all words from troll and non-troll datasets
train_troll_words = ' '.join(balanced_df[balanced_df.target == True].question_text).split(' ')
train_non_troll_words = ' '.join(balanced_df[balanced_df.target == False].question_text).split(' ')

common_words = set(train_troll_words).intersection(set(train_non_troll_words))
```

```
train_troll_bow = dict()
for w in common_words:
    train_troll_bow[w] = train_troll_words.count(w) / len(train_troll_words)

train_non_troll_bow = dict()
for w in common_words:
    train_non_troll_bow[w] = train_non_troll_words.count(w) / len(train_non_troll_words)
```

VECTORIZER:

```
# Vectorizer

vectorizer = CountVectorizer(max_features=1000)

X = vectorizer.fit_transform(balanced_df.question_text.tolist())
qt = vectorizer.fit_transform(df.question_text.tolist())
```

LOGISTIC MODEL USING PIPELINE:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size= 0.25, random_state= 0)
```

```
# pipeline_model = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', SVC())])
# pipeline_model = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()), ('clf', MultinomialNB())])
```

```
pipeline_model = LogisticRegression(class_weight = 'balanced', max_iter=1000, solver="lbfgs", penalty="l2")
```

```
pipeline_model.fit(X_train, y_train)
```

```
... LogisticRegression(class_weight='balanced', max_iter=1000)
```

```
y_pred = pipeline_model.predict(X_test)
f1_score(y_test, y_pred, average='macro')
```

```
... 0.8523551035595869
```

- Similarly all the test data was pre processed as the above mentioned methods.
- Then we predicted the data using logistic regression model with pipeline.

FINAL RESULT:

```
result = pipeline_model.predict(Xdb_X)
```

```
db["target"] = result

to_submit = db[['qid',"target"]]

to_submit.to_csv("submission.csv", index = False)
```

```
to_submit["target"].value_counts()
```

```
0    551058
1    102003
Name: target, dtype: int64
```

ML Project

Session-1

ASK REDDIT

Thank You

By,
Fahed Shaikh, IMT2019079
Likhiteswar , IMT2019510