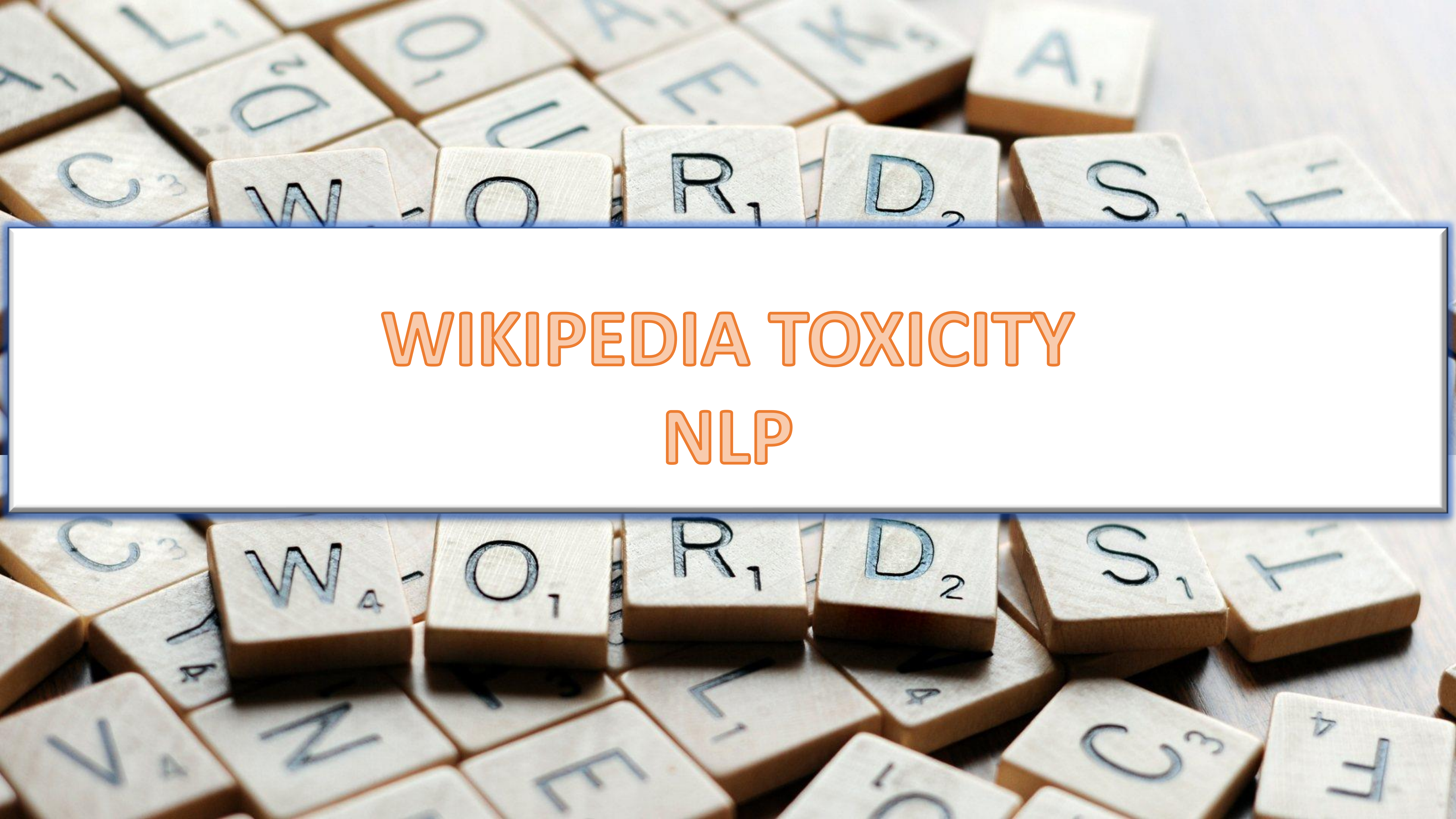


# WIKIPEDIA TOXICITY NLP

# Introduction

Multilingual online encyclopedia based on open collaboration

Web-based applications like web browsers

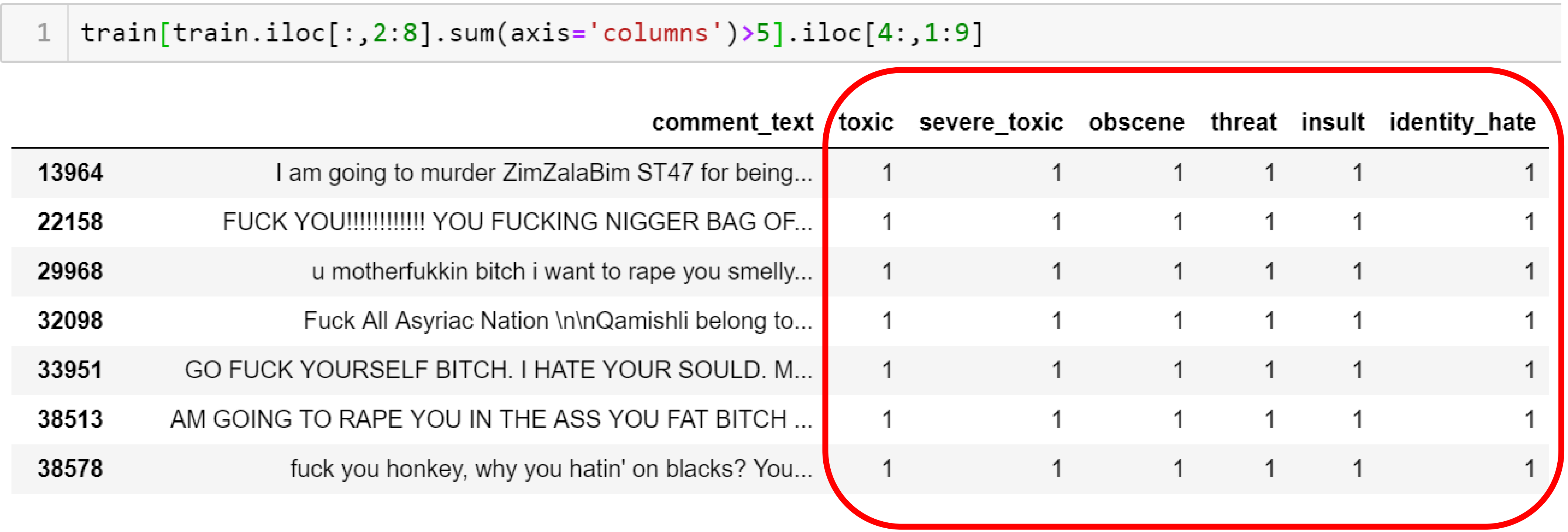
Largest and most popular general reference work on the WWW

# Data set

*Profane, vulgar, or offensive text*

The number of observations in the dataset: 150 K+

A comment may be any combination of the following:

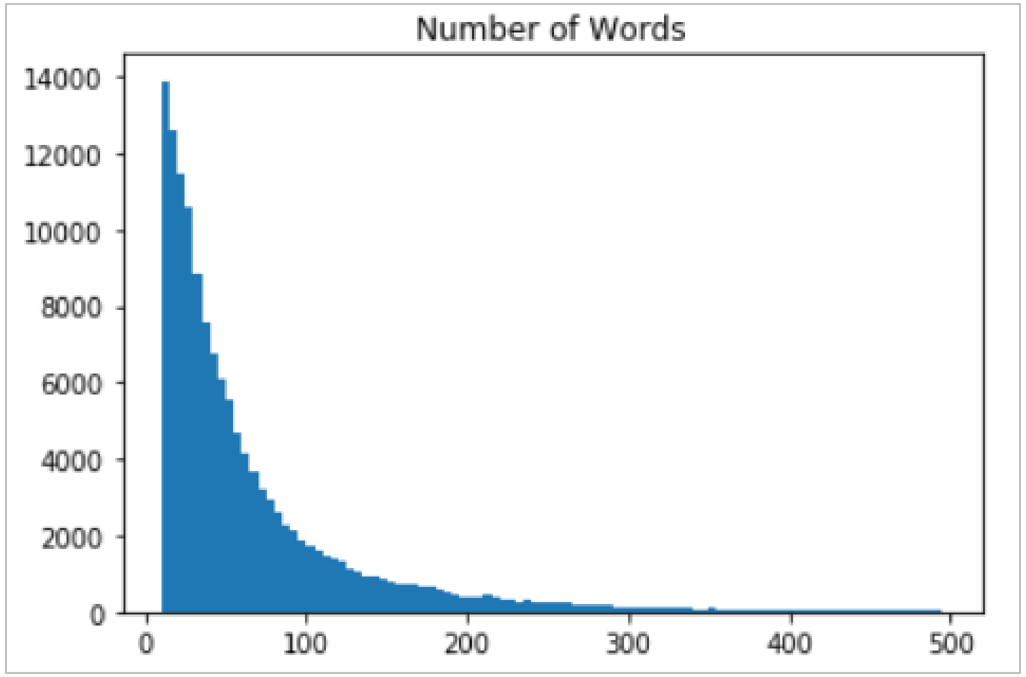


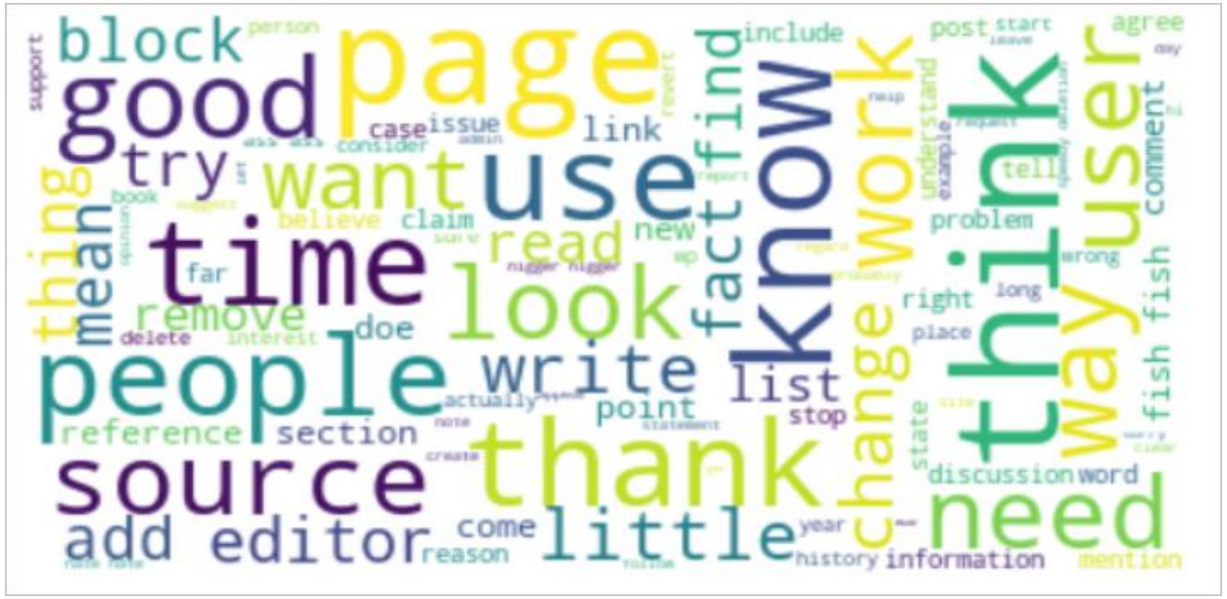
```python
train[train.iloc[:,2:8].sum(axis='columns')>5].iloc[4:,1:9]
```

| | comment_text | toxic | severe_toxic | obscene | threat | insult | identity_hate |
|---|---|---|---|---|---|---|---|
| **13964** | I am going to murder ZimZalaBim ST47 for being... | 1 | 1 | 1 | 1 | 1 | 1 |
| **22158** | FUCK YOU!!!!!!!!!!!! YOU FUCKING NIGGER BAG OF... | 1 | 1 | 1 | 1 | 1 | 1 |
| **29968** | u motherfukkin bitch i want to rape you smelly... | 1 | 1 | 1 | 1 | 1 | 1 |
| **32098** | Fuck All Asyriac Nation \n\nQamishli belong to... | 1 | 1 | 1 | 1 | 1 | 1 |
| **33951** | GO FUCK YOURSELF BITCH. I HATE YOUR SOULD. M... | 1 | 1 | 1 | 1 | 1 | 1 |
| **38513** | AM GOING TO RAPE YOU IN THE ASS YOU FAT BITCH ... | 1 | 1 | 1 | 1 | 1 | 1 |
| **38578** | fuck you honkey, why you hatin' on blacks? You... | 1 | 1 | 1 | 1 | 1 | 1 |

# Data Exploration

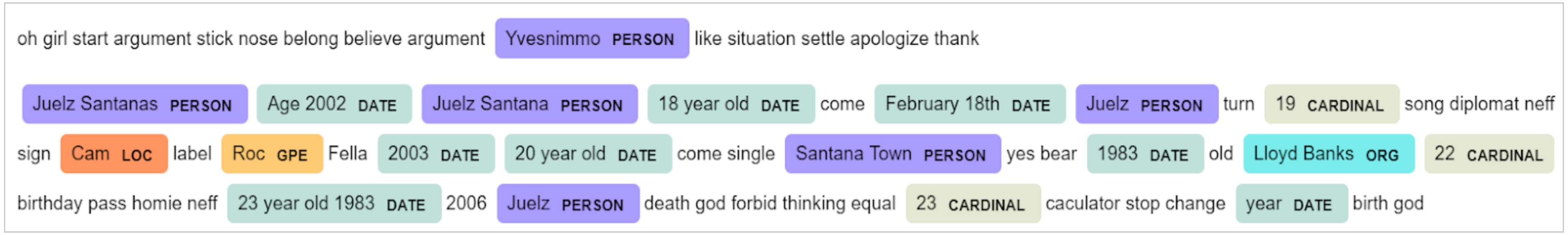
- The number of words in 159’571 documents: 7’153’449
- Number of unique words: 210’337.





# Text Pre-processing

- Regex
  - Remove Accents
  - Remove URLs
  - Remove Special Chars etc
- Spacy
  - English Model
  - Tokenize
  - Lemmatize
  - POS/NER
- Gensim
  - Word2Vec

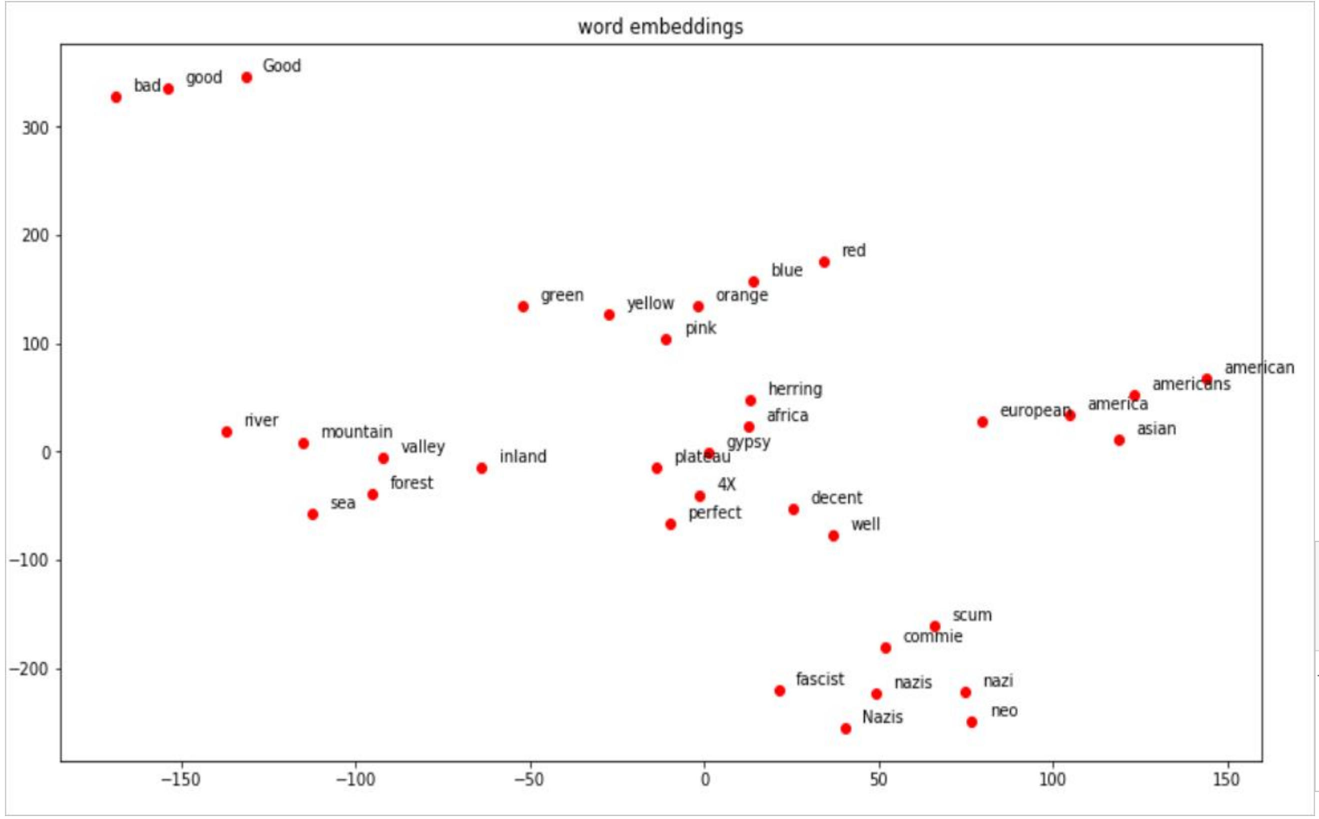


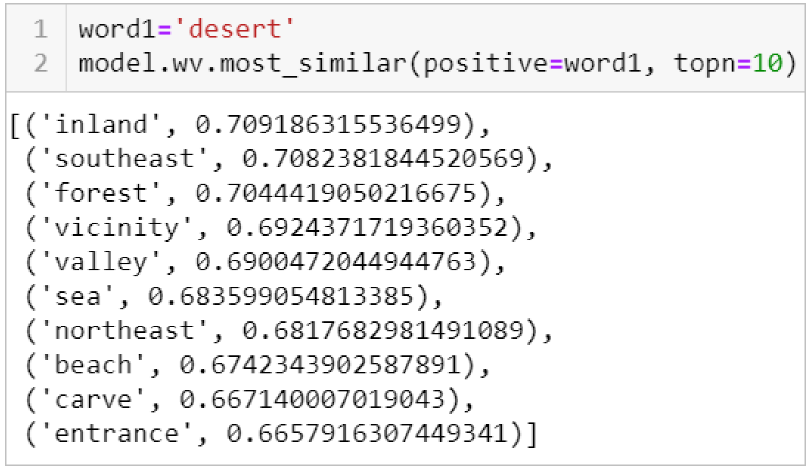
oh girl start argument stick nose belong believe argument Yvesnimmo PERSON like situation settle apologize thank

Juelz Santanas PERSON Age 2002 DATE Juelz Santana PERSON 18 year old DATE come February 18th DATE Juelz PERSON turn 19 CARDINAL song diplomat neff sign Cam LOC label Roc GPE Fella 2003 DATE 20 year old DATE come single Santana Town PERSON yes bear 1983 DATE old Lloyd Banks ORG 22 CARDINAL birthday pass homie neff 23 year old 1983 DATE 2006 Juelz PERSON death god forbid thinking equal 23 CARDINAL caculator stop change year DATE birth god

# Similarities

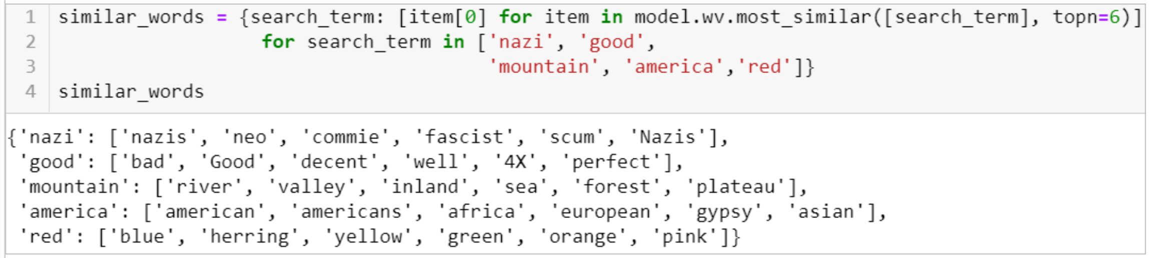
- Trained own vectors via Gensim Word2Vec





```
word1='desert'
model.wv.most_similar(positive=word1, topn=10)
```

```
[('inland', 0.709186315536499),
 ('southeast', 0.7082381844520569),
 ('forest', 0.7044419050216675),
 ('vicinity', 0.6924371719360352),
 ('valley', 0.6900472044944763),
 ('sea', 0.683599054813385),
 ('northeast', 0.6817682981491089),
 ('beach', 0.6742343902587891),
 ('carve', 0.667140007019043),
 ('entrance', 0.6657916307449341)]
```

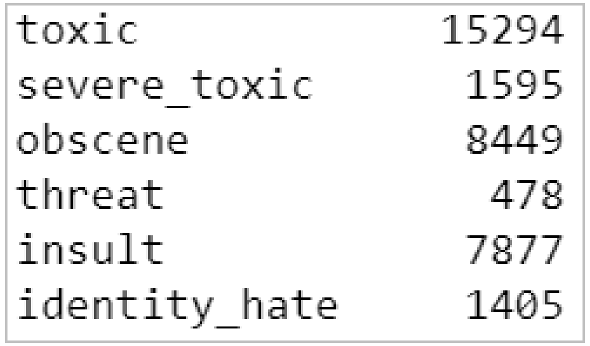


```
similar_words = {search_term: [item[0] for item in model.wv.most_similar([search_term], topn=6)]
                     for search_term in ['nazi', 'good',
                                         'mountain', 'america','red']}
similar_words
```

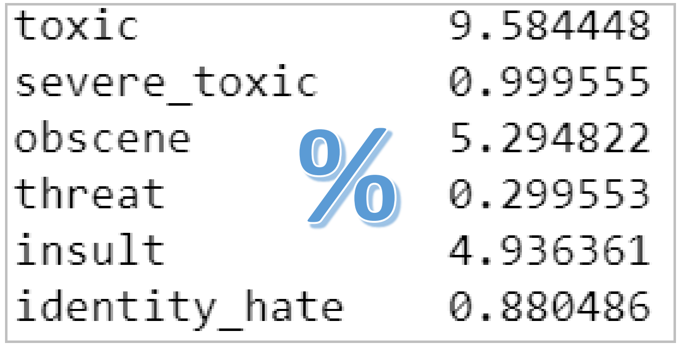
```
{'nazi': ['nazis', 'neo', 'commie', 'fascist', 'scum', 'Nazis'],
 'good': ['bad', 'Good', 'decent', 'well', '4X', 'perfect'],
 'mountain': ['river', 'valley', 'inland', 'sea', 'forest', 'plateau'],
 'america': ['american', 'americans', 'africa', 'european', 'gypsy', 'asian'],
 'red': ['blue', 'herring', 'yellow', 'green', 'orange', 'pink']}
```

# Baseline Models

- TFIDF Vectorizer, Count Vectorizer
- Multinomial Naive Bayes
- Logistic Regression
- Support Vector Classiffier



```
toxic            15294
severe_toxic      1595
obscene           8449
threat             478
insult            7877
identity_hate     1405
```



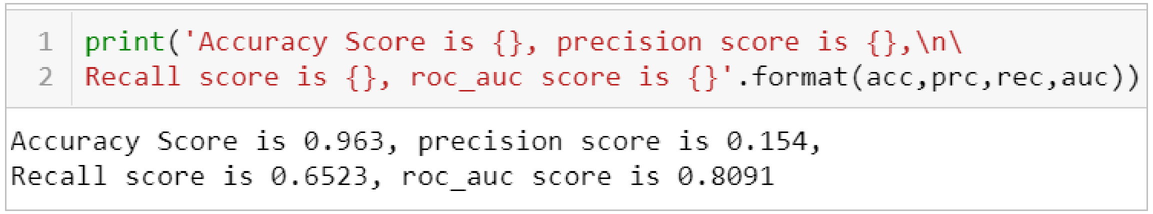
```
toxic               9.584448
severe_toxic        0.999555
obscene             5.294822
threat              0.299553
insult              4.936361
identity_hate       0.880486
```

%

Heavy Class Imbalance

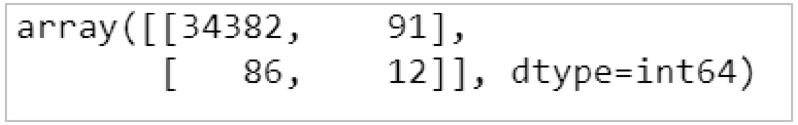
# MNB

## Confusion matrix and false negatives for threat… Low precision!

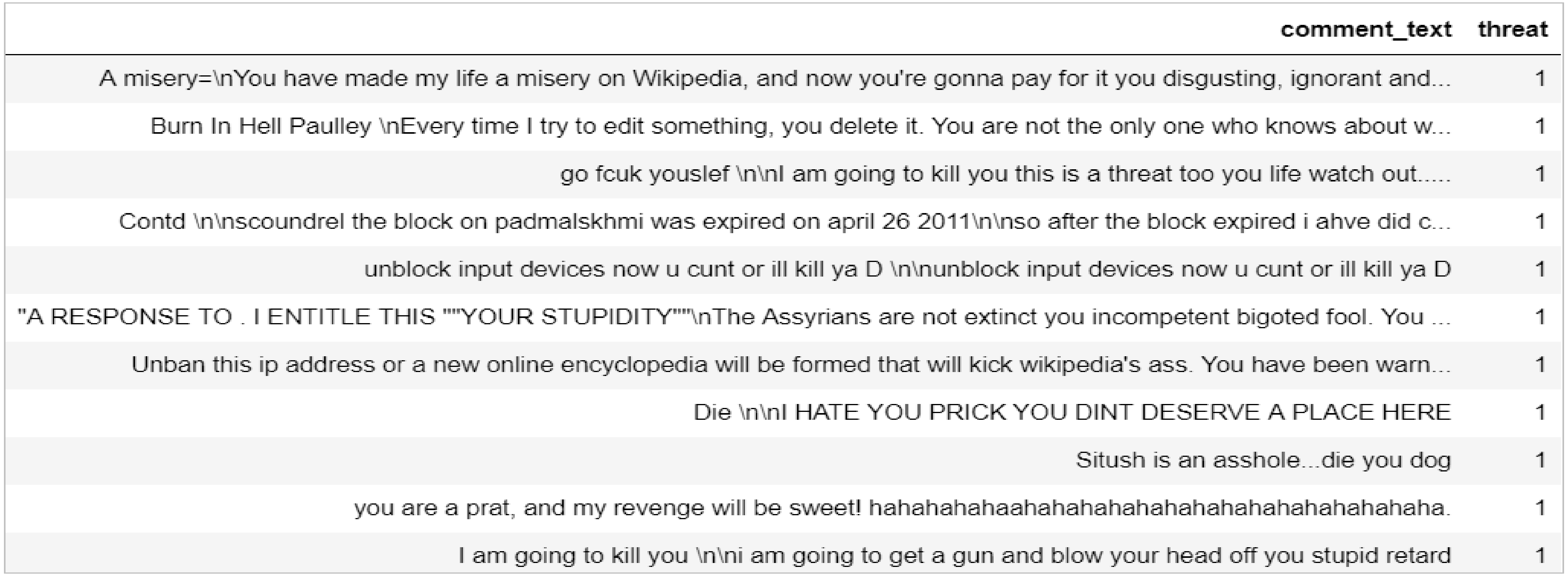


```
print('Accuracy Score is {}, precision score is {},\n\
Recall score is {}, roc_auc score is {}'.format(acc,prc,rec,auc))
```

```
Accuracy Score is 0.963, precision score is 0.154,
Recall score is 0.6523, roc_auc score is 0.8091
```



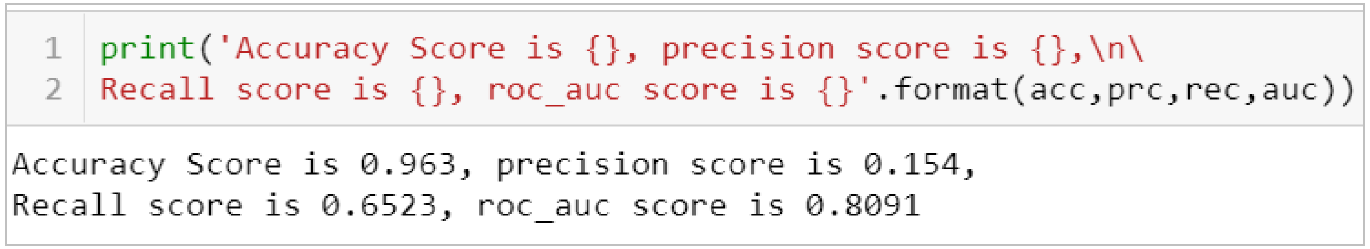
```
array([[34382,    91],
       [   86,    12]], dtype=int64)
```



| comment_text | threat |
|---|---|
| A misery=\nYou have made my life a misery on Wikipedia, and now you're gonna pay for it you disgusting, ignorant and... | 1 |
| Burn In Hell Paulley \nEvery time I try to edit something, you delete it. You are not the only one who knows about w... | 1 |
| go fcuk youslef \n\nI am going to kill you this is a threat too you life watch out..... | 1 |
| Contd \n\nscoundrel the block on padmalskhmi was expired on april 26 2011\n\nso after the block expired i ahve did c... | 1 |
| unblock input devices now u cunt or ill kill ya D \n\nunblock input devices now u cunt or ill kill ya D | 1 |
| "A RESPONSE TO . I ENTITLE THIS ""YOUR STUPIDITY""\nThe Assyrians are not extinct you incompetent bigoted fool. You ... | 1 |
| Unban this ip address or a new online encyclopedia will be formed that will kick wikipedia's ass. You have been warn... | 1 |
| Die \n\nI HATE YOU PRICK YOU DINT DESERVE A PLACE HERE | 1 |
| Sitush is an asshole...die you dog | 1 |
| you are a prat, and my revenge will be sweet! hahahahahaahahahahahahahahahahahahahaha. | 1 |
| I am going to kill you \n\ni am going to get a gun and blow your head off you stupid retard | 1 |

# Logistic Regression

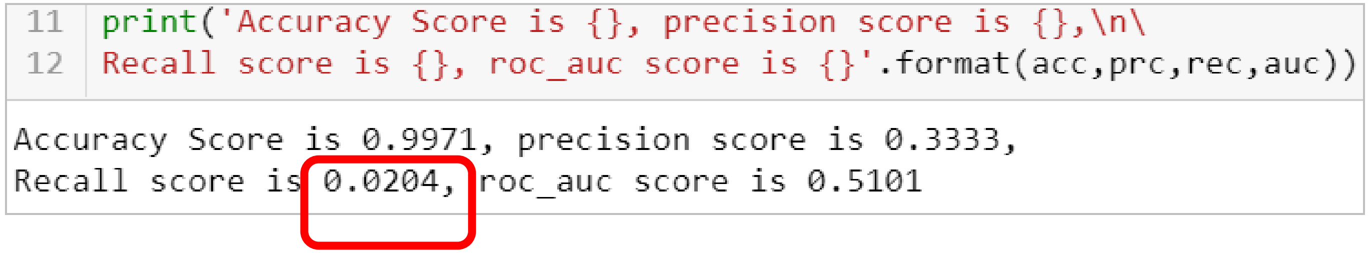
- Toxic



```
print('Accuracy Score is {}, precision score is {},\n\
Recall score is {}, roc_auc score is {}'.format(acc,prc,rec,auc))
```

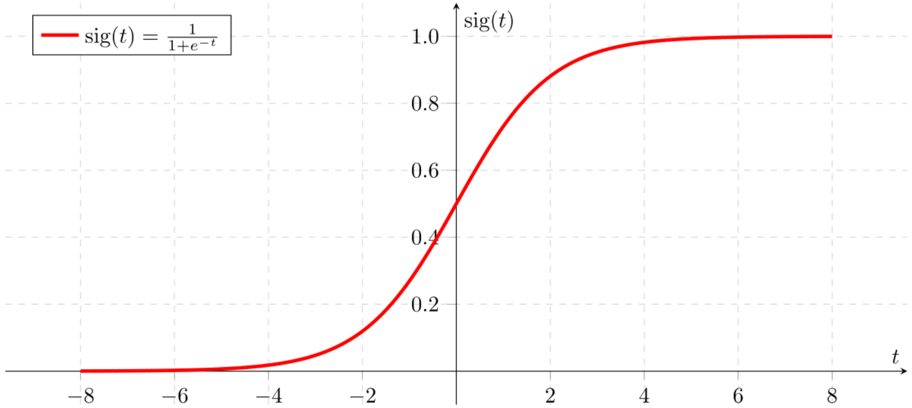
```
Accuracy Score is 0.963, precision score is 0.154,
Recall score is 0.6523, roc_auc score is 0.8091
```

- Threat



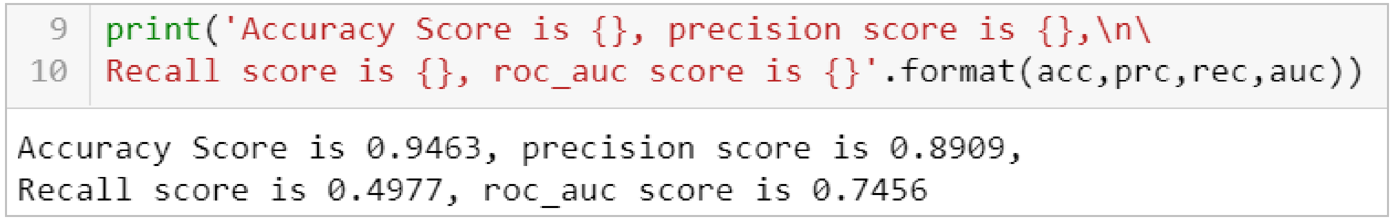
```
print('Accuracy Score is {}, precision score is {},\n\
Recall score is {}, roc_auc score is {}'.format(acc,prc,rec,auc))
```

```
Accuracy Score is 0.9971, precision score is 0.3333,
Recall score is 0.0204, roc_auc score is 0.5101
```



# Support Vector Classification

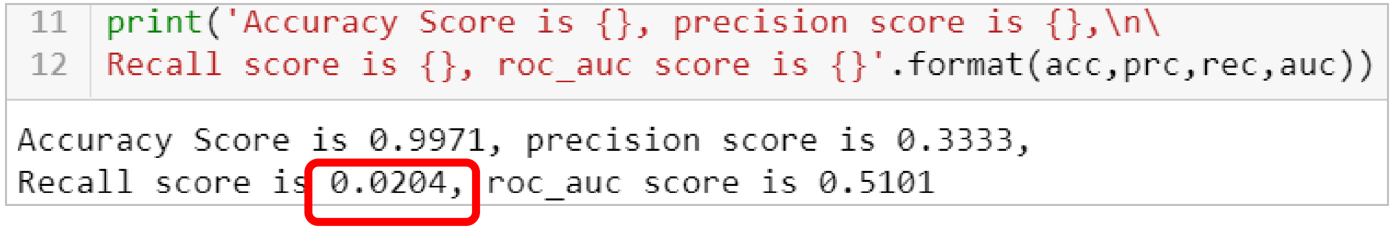
- Toxic



```
print('Accuracy Score is {}, precision score is {},\n\
Recall score is {}, roc_auc score is {}'.format(acc,prc,rec,auc))
```

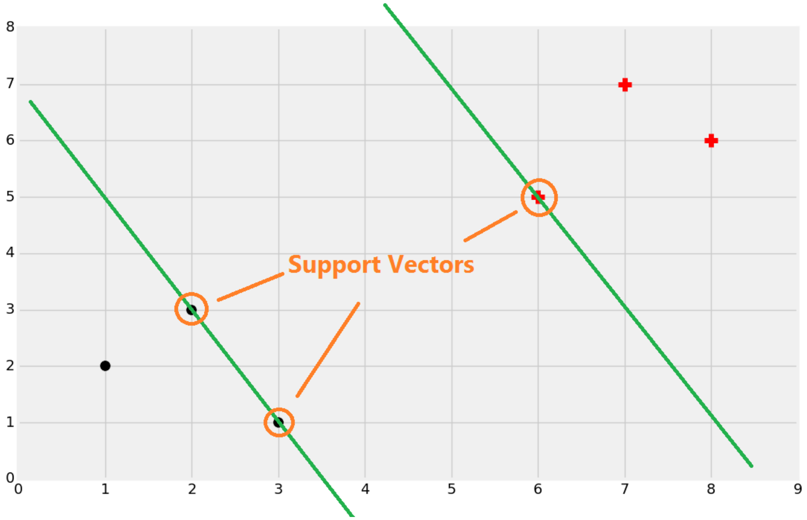
```
Accuracy Score is 0.9463, precision score is 0.8909,
Recall score is 0.4977, roc_auc score is 0.7456
```

- Threat



```
print('Accuracy Score is {}, precision score is {},\n\
Recall score is {}, roc_auc score is {}'.format(acc,prc,rec,auc))
```

```
Accuracy Score is 0.9971, precision score is 0.3333,
Recall score is 0.0204, roc_auc score is 0.5101
```



# Conclusion

- Baseline models successfully explained “toxic, severe toxic, obscene”, but failed for“threat”.
- Recurrent neural-network models will be used to fill this gap and provide a better solution to the problem.