Binary Classification with NLP

- Maryam Amjad

Problem Statement

To build multiple binary classification models to classify the scraped data from Pushshift's API and compare the models performance. Aslo to do exploratory data analysis.

Web-Scraping

URLs Used

- https://api.pushshift.io/reddit/search/comment?subreddit=iphone
- https://api.pushshift.io/reddit/search/comment?subreddit=android

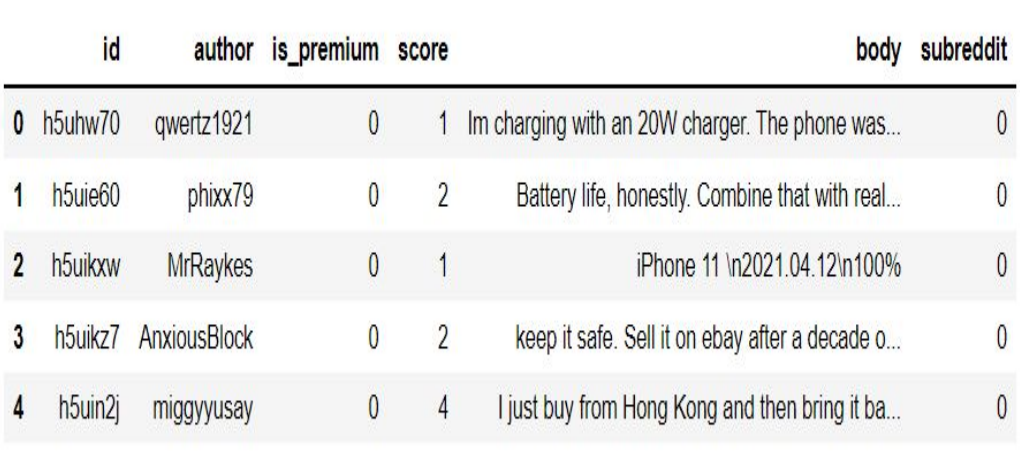
The data collected is from the past 75 days, 100 comments for each subreddit category from each day. It is done with help of adding parameters to the URL

Important features such as Author name, comment score and comment body are collected.

Data Cleaning

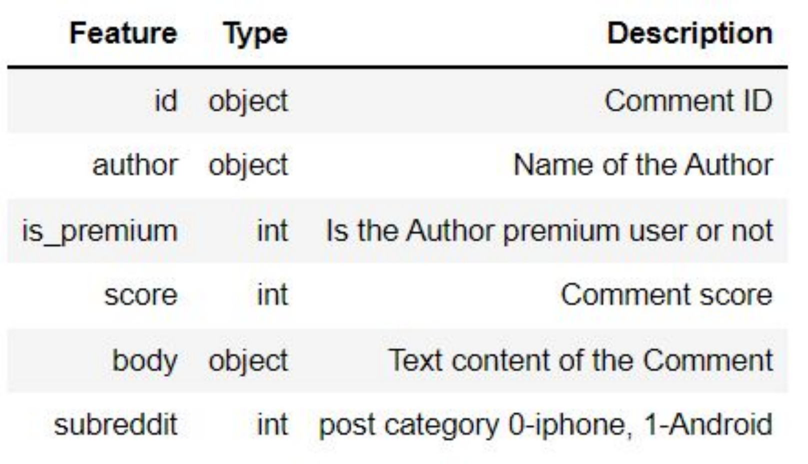
- Null check is done and removed if any
- Data types are converted appropriately
  - The ‘subreddit’ column is mapped as
    - 0 - iphone
    - 1 - Android
  - The ‘author_premium’ column is converted from bool to int
- Column renaming is done
- The Cleaned data is then exported as .csv file

Data Frame



	id	author	is_premium	score	body	subreddit
0	h5uhw70	qwertz1921	0	1	Im charging with an 20W charger. The phone was...	0
1	h5uie60	phixx79	0	2	Battery life, honestly. Combine that with real...	0
2	h5uikxw	MrRaykes	0	1	iPhone 11 \n2021.04.12\n100%	0
3	h5uikz7	AnxiousBlock	0	2	keep it safe. Sell it on ebay after a decade o...	0
4	h5uin2j	miggyyusay	0	4	I just buy from Hong Kong and then bring it ba...	0

Data Dictionary



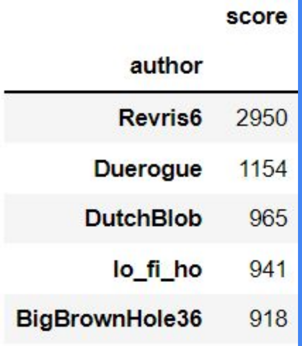
Feature	Type	Description
id	object	Comment ID
author	object	Name of the Author
is_premium	int	Is the Author premium user or not
score	int	Comment score
body	object	Text content of the Comment
subreddit	int	post category 0-iphone, 1-Android

Data Preparation

- NLP is done with the help of countVectorizer
  - Binary countVectorizer is used
  - Punctuations are removed by default pattern
  - Stop words are removed
  - The countVectorizer is fitted with ‘author’ and ‘body’ columns and transformed
  - It is then converted into a DataFrame
- Exploratory Data Analysis is done
- Input and output columns are determined
- Test and train data are splitted in 1:3 ratio

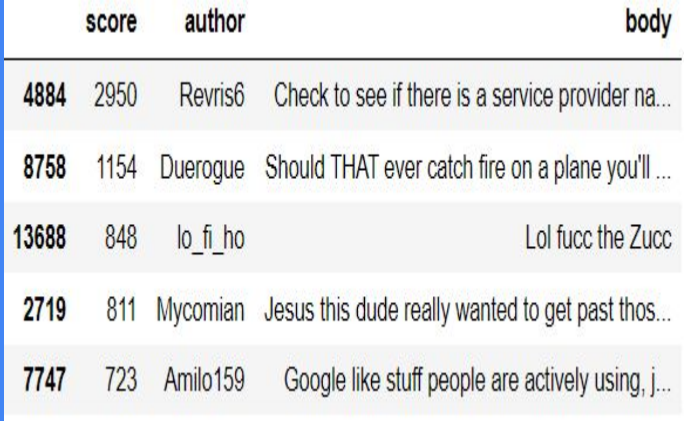
Exploratory Data Analysis

Top 5 Authors by Total Comment Scores



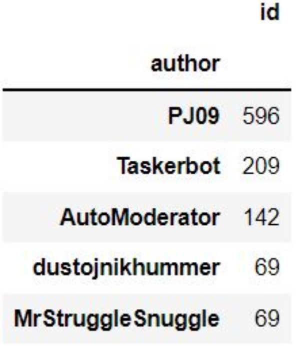
	score
author	
Revris6	2950
Duerogue	1154
DutchBlob	965
lo_fi_ho	941
BigBrownHole36	918

Top 5 Comments by Scores



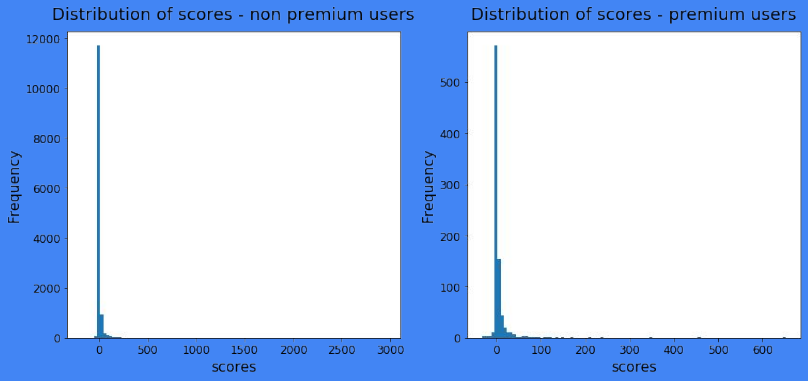
	score	author	body
4884	2950	Revris6	Check to see if there is a service provider na...
8758	1154	Duerogue	Should THAT ever catch fire on a plane you'll ...
13688	848	lo_fi_ho	Lol fucc the Zucc
2719	811	Mycomian	Jesus this dude really wanted to get past thos...
7747	723	Amilo159	Google like stuff people are actively using, j...

Top 5 Authors by Total No.of Comments

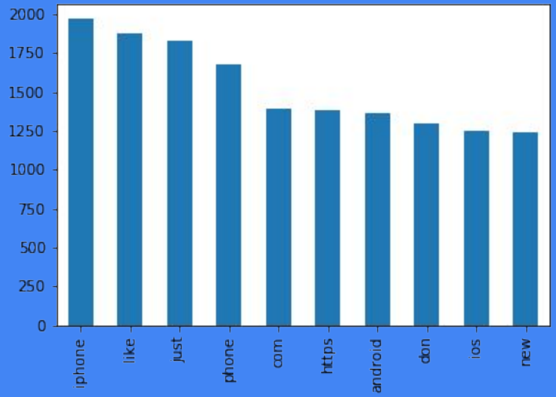


	id
author	
PJ09	596
Taskerbot	209
AutoModerator	142
dustojnikhummer	69
MrStruggleSnuggle	69

EDA - Visualizations



From the Hist plots we can observe that the distribution of scores is right skewed for both premium and non-premium users. We can also note that non-premium users have higher scores.

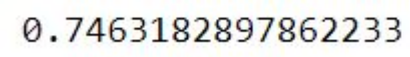


From the Bar graph we can observe the 10 most used words in the dataset and their no.of occurence. ‘iphone ’ is the most frequent word in the dataset.

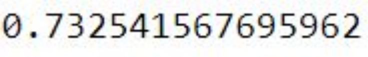
Logistic Regression Model

Random Forest Classifier

Accuracy

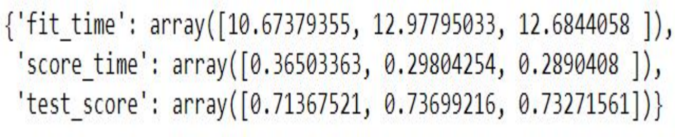


```
0.7463182897862233
```

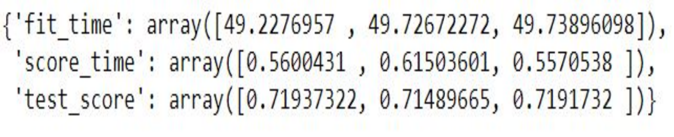


```
0.732541567695962
```

Cross validation

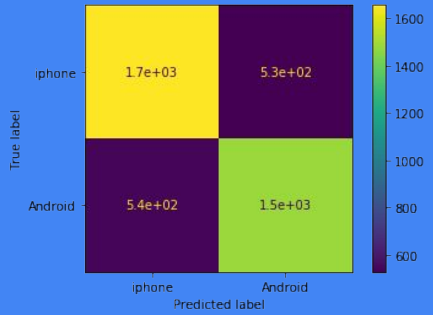


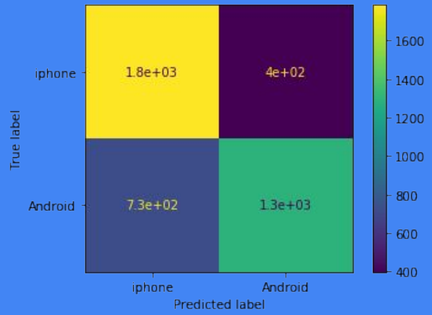
```
{'fit_time': array([10.67379355, 12.97795033, 12.6844058 ]),
 'score_time': array([0.36503363, 0.29804254, 0.2890408 ]),
 'test_score': array([0.71367521, 0.73699216, 0.73271561])}
```



```
{'fit_time': array([49.2276957 , 49.72672272, 49.73896098]),
 'score_time': array([0.5600431 , 0.61503601, 0.5570538 ]),
 'test_score': array([0.71937322, 0.71489665, 0.7191732 ])}
```

Confusion Matrix





From both the confusion Matrix it is observed that false positives and false negatives are comparatively very small to the true values.

Conclusion

From the Observation of various evaluation parameters such as accuracy, cross validation and confusion matrix, it is seen that there is a significant difference in performance of Logistic Regression model and Random Forest Classifier. Both the models performs well with the test data with >70% accuracy. At times we can observe similar performance from both the model. The Logistic regression model performance is consistent and have better accuracy and fitting time.

This solution can be enhanced and used in various social media platforms and forums to classify the genuine post from the violent ones. Further the genuine post can be tagged to the appropriate classes.

Thank you