Predicting Reddit Sports Posts

Elizabeth Sebastian | Dennis Chan | Steven Lee

Agenda

- Background / Objectives
- Web Scraping and Null Values
- Feature Engineering
- EDA: Word Count and Sentiment Analysis
- Pre-processing and Modelling
- Limitations
- Conclusion

Background / Objectives

- Operations executive for a company that runs a sports news feed website that collates information of major sports event from around the world.
- Develop a binary classification model to predict the sports topic of a sub-Reddit forum.
- Primary stakeholder: Operations Management of company website
- Secondary stakeholder: Website subscribers

Web Scraping and Null Values

- Use of web API
- JSON result dataset
- Specify 'User-agent' in the header
- ‘After’ attribute for continuation
- 1000 records limit
- Impute NULL selftext
- “Sleeping” between requests

JSON Data Labels

- **subreddit**: label or name of the subreddit forum
- **subreddit_id**: unique Id of the subreddit forum
- **created_utc**: timestamp of the post
- **title**: title text of the user post
- **selftext**: raw text of the user post
- **name**: unique Id of the parent post
- **score**: number of upvotes of the post has

Feature Engineering

Reasons and explanations:

- Often there is only text in the title column but not on the body columns.
- Thus, we combine all text of the thread into one column called all_text
- Binarised subjects: nba is 0, and epl is 1
- Added word_count column to count total words in all_text

	selftext	title	subreddit	all_text	word_count
0	# Game Threads Index (January 07, 2021):\n\n|T...	Trash Talk Thursday + Game Thread Index	0	Trash Talk Thursday Game Thread Index Game T...	44
1	Here is a place to have in depth, x's and o's,...	[SERIOUS NEXT DAY THREAD] Post-Game Discussion...	0	SERIOUS NEXT DAY THREAD PostGame Discussion Ja...	119
2	Luka Doncic tonight against the Denver Nuggets...	Luka Doncic tonight against the Denver Nuggets...	0	Luka Doncic tonight against the Denver Nuggets...	47
3	Tim Duncan made a rare podcast appearance on t...	Tim Duncan on who he likes to watch in the NBA...	0	Tim Duncan on who he likes to watch in the NBA...	121
4	\n|\t\t\n|:-:|\t\t\n|[](/PHI) **109 - 122** ...	[Post Game Thread] The Brooklyn Nets (5-4) def...	0	Post Game Thread The Brooklyn Nets defeat the...	115

EDA

Basic EDA on Word Count

Nba = 0

EPL = 1

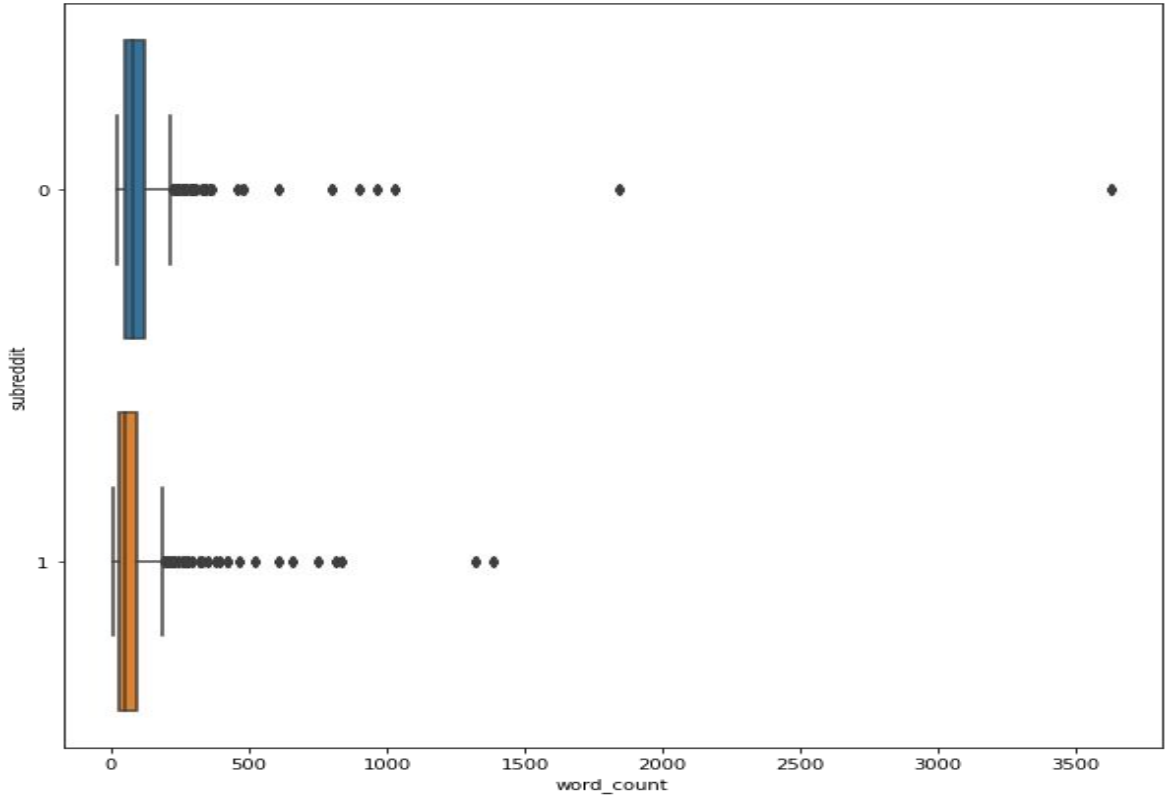
Same total word count:

Balanced Dataset

	subreddit	0	1
word_count	count	2000.00000	2000.000000
	mean	115.16300	87.127500
	std	203.76846	130.315683
	min	19.00000	7.000000
	25%	52.00000	32.000000
	50%	76.00000	52.500000
	75%	119.00000	96.000000
	max	3629.00000	1385.000000

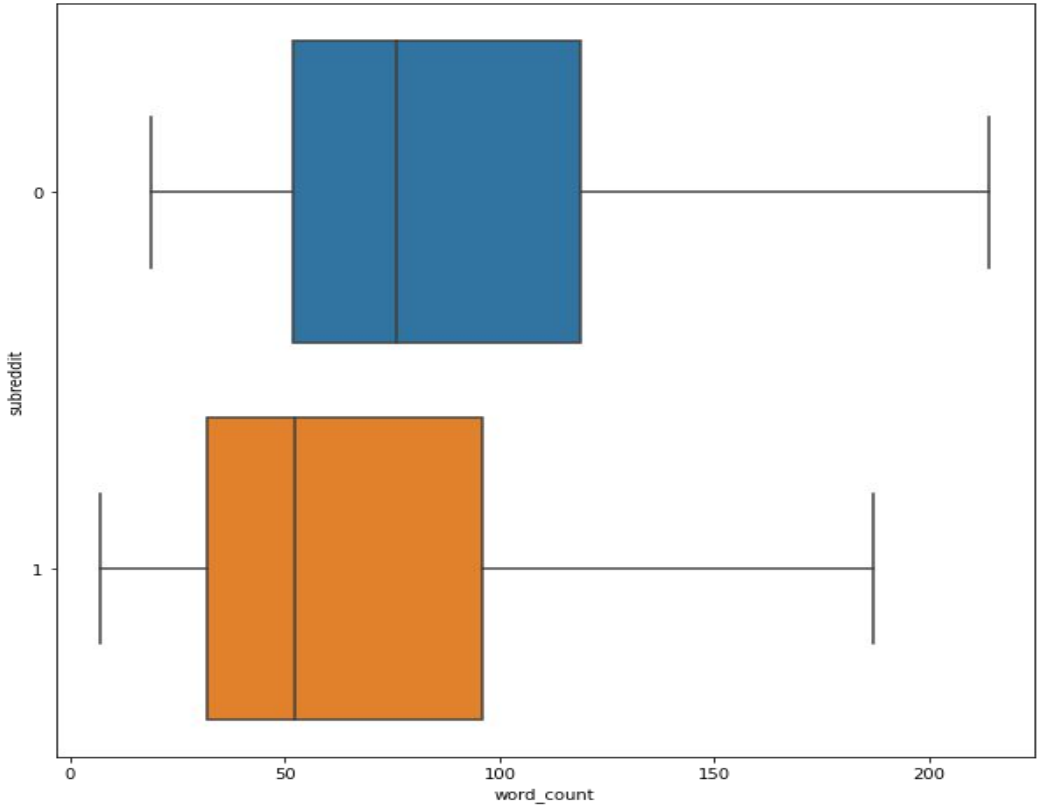
Visualising Word Count using Boxplot

- Many outliers
- Similar interquartile range
- Significantly different actual range due to outliers

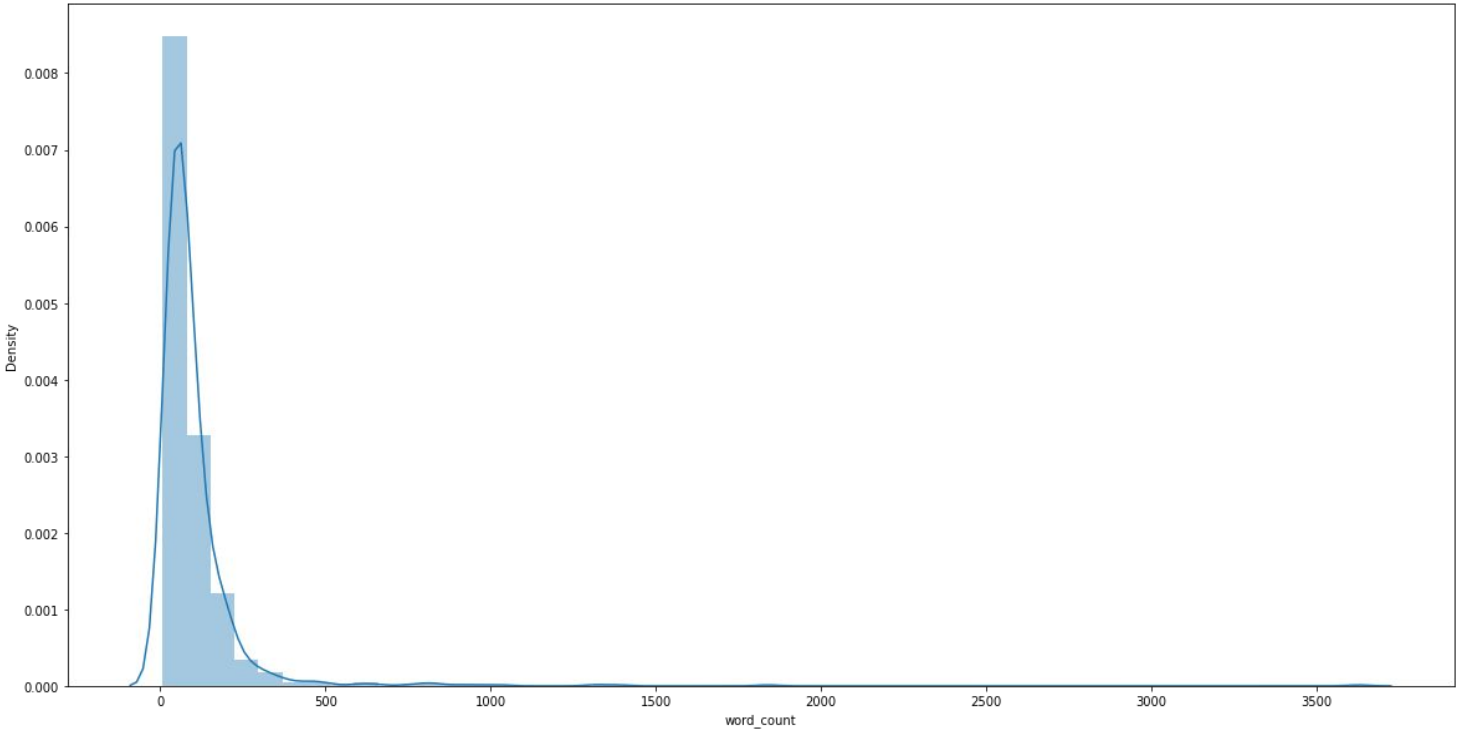


Visualising Word Count using Boxplot (w/o outliers)

- NBA has higher interquartile range than EPL
- NBA has higher range than EPL (without factoring outliers)



Distribution of Word Count



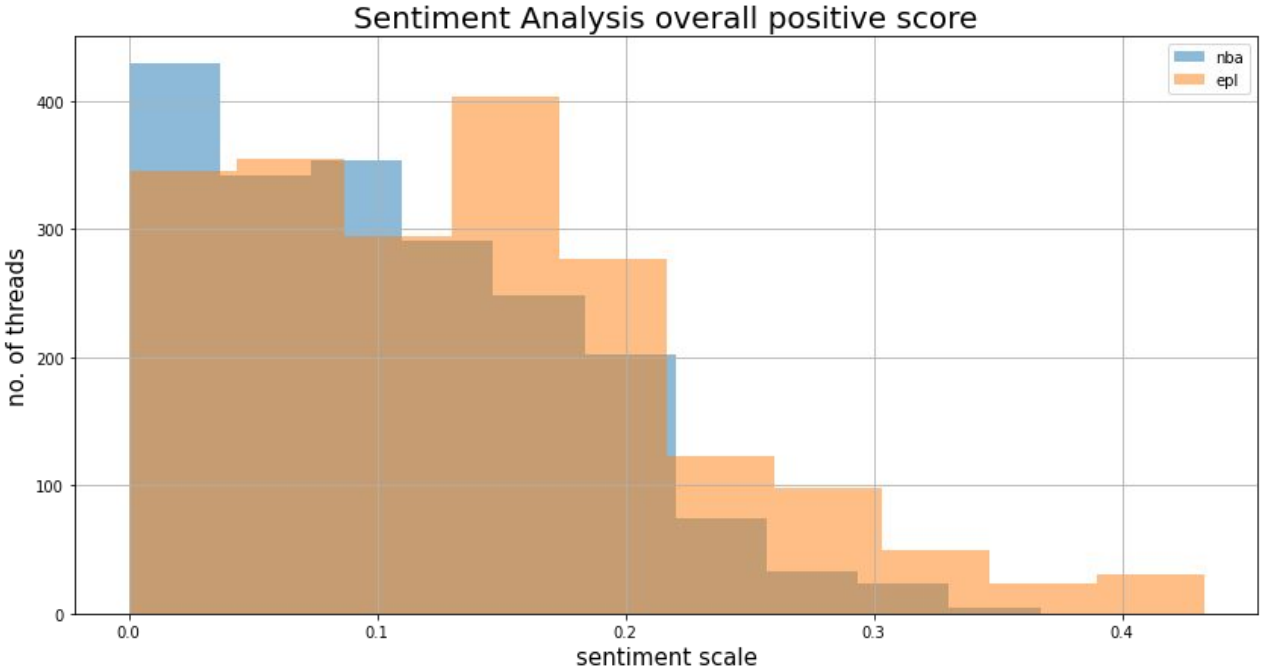
- Right skew
- Unimodal
- Majority of words only appears a few times

EDA: Sentiment Analysis

Evaluate the all_text column based on positive, negative, neutral score

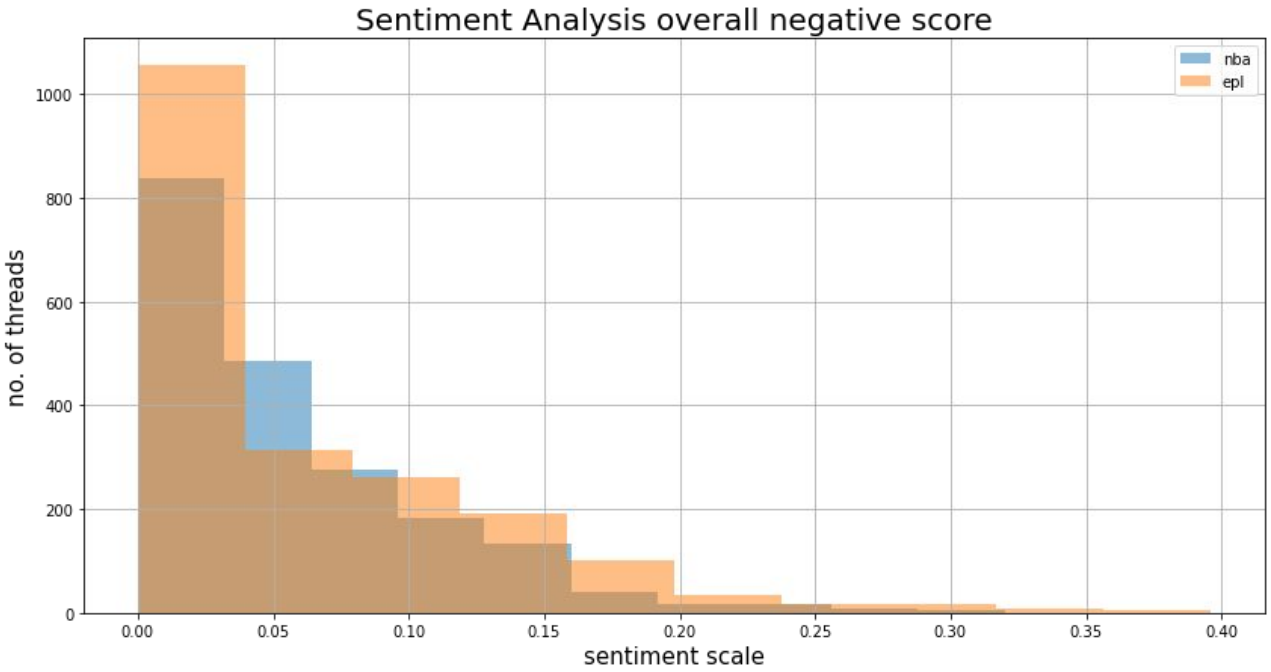
Positive Sentiment Score

- Negatively sloped graph for both
- EPL has a higher positive sentiment score on average
- Right skewed



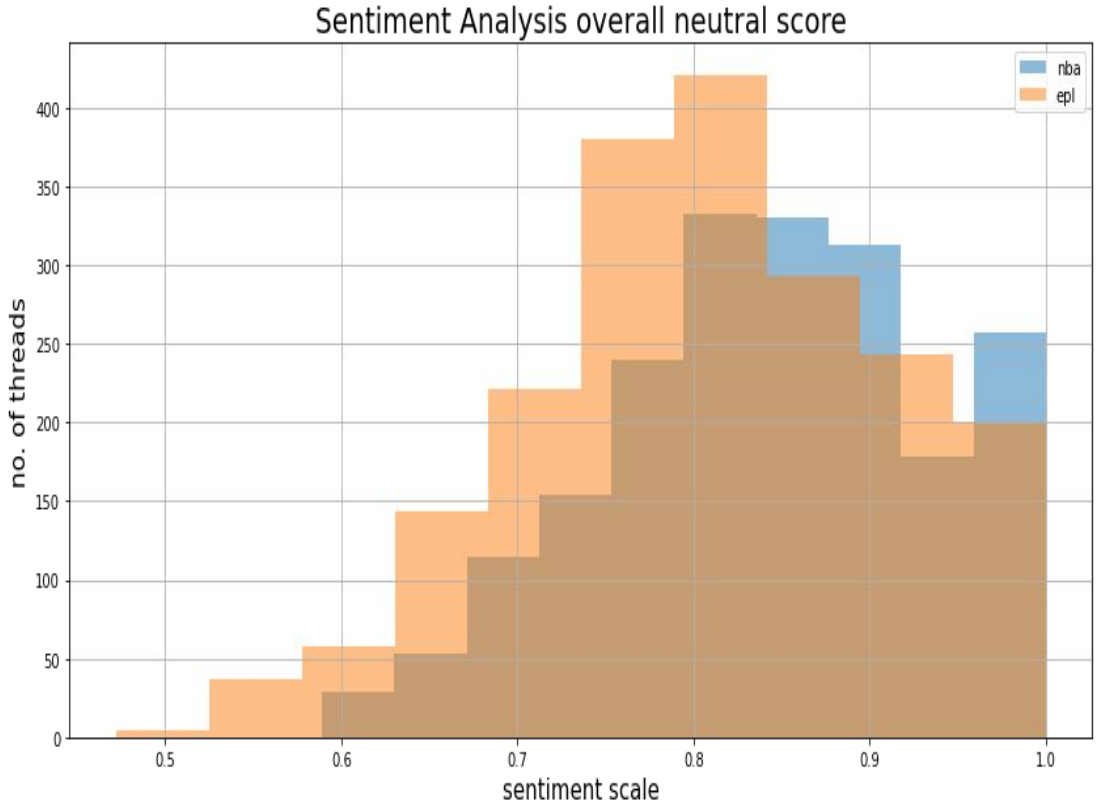
Negative Sentiment Score

- Negatively sloped graph for both
- EPL has a higher negative sentiment score on average



Neutral Sentiment Score

- Left skewed
- Unimodal
- EPL words tend to be more neutral than NBA words



Preprocessing and Modelling

- Split the data into X (input variable) and y (output variable)

```
X = combine.text
y = combine.target
```

- Split our data into training and testing sets

- Turn our text into features
  - ➢ CountVectorizer
  - ➢ TF-IDFVectorizer

Modelling

Classifier Models	
Naive Bayes Model (Multinomial NB)	**Logistic Regression**
Columns are all integer counts	Response default falls into one of two categories.

Modelling-Evaluation

BASELINE SCORE

	EPL	NBA
target	0.527586	0.472414

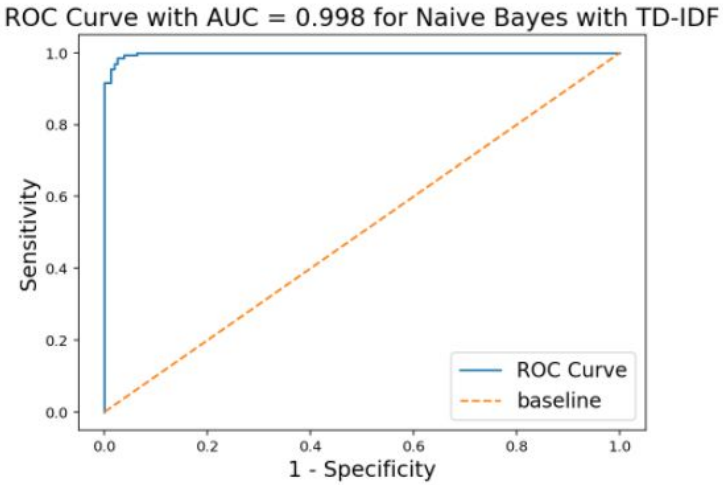
ACCURACY

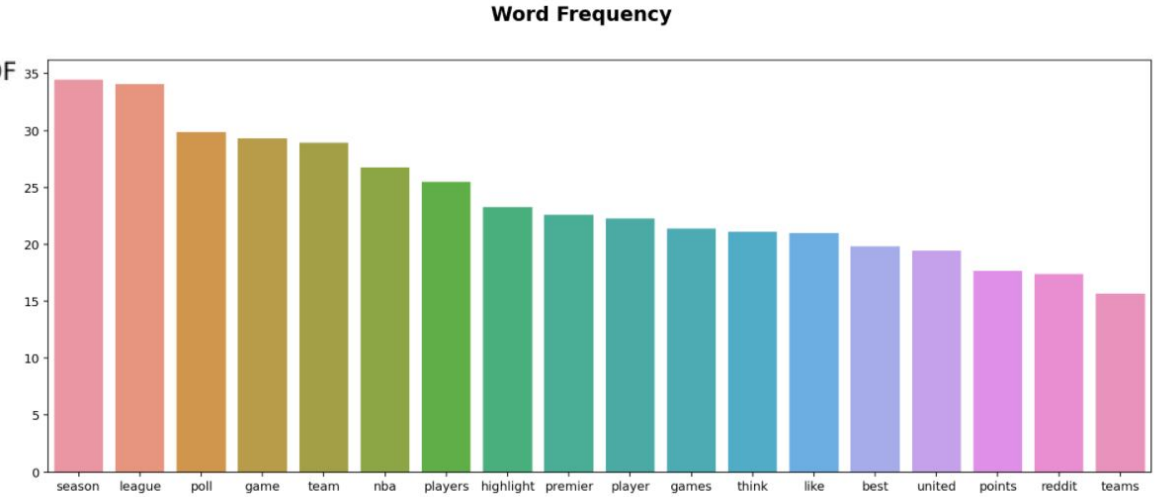
Model	Train Score	Test Score	Model
Lr+CVEC	0.989	0.958	-0.031
Lr+TDIDF	0.984	0.966	-0.018
CVEC+MultiNB	0.993	0.970	-0.023
TDIDF+MultiNB	0.996	0.972	-0.024

CONFUSION MATRIX

	Predicted EPL	Predicted NBA
Actual EPL	152	6
Actual NBA	2	130

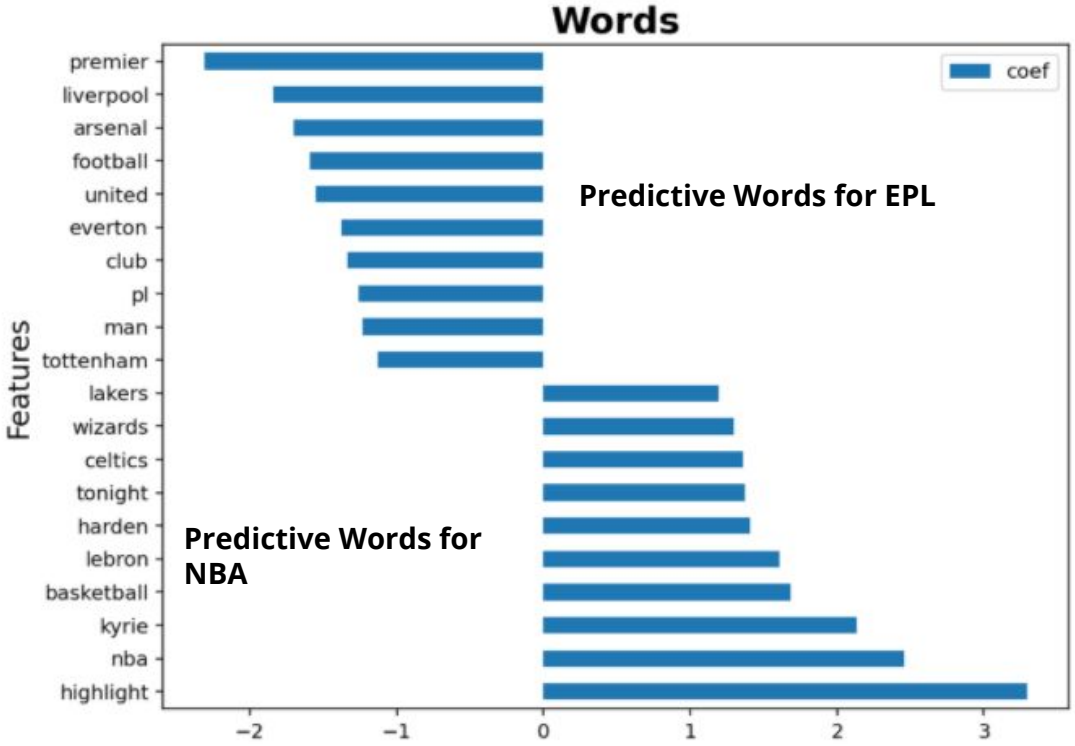
Modelling-Evaluation





- ROC AUC represents the degree or measure of separability of the model into the 2 classes
- Words that appears in the corpus most number of times

Modelling



- Predictive words that belong to each subreddit.

LIMITATIONS OF THE MODEL

- The analysis and modeling is based on ~1000 posts per subreddit which may not be a good representation of the subreddit itself.
- The dataset of that the model is trained is likely not representative of the content of the subreddit in the long run. This is because the content of subreddits changes with current news and events.

- The model can be further enhanced :

    - To eliminate synonyms, post agnostic words, common words
    - To include phrases instead of just words for prediction.

CONCLUSION

- API scraping and machine learning allows for quick classification of sports news and automation of workflow.
- Model can be easily adapted to source relevant and trending content from various websites.