NLP Analysis and Modeling

A case study of two online football communities

Problem Statement

The '12th man' of the team is a common trope used by football pundits, which describes the enormous positive impact a strong supporting crowd can have on the performance of a Football team. Maintaining positive morale and community spirit is a vital component of sporting success. The planning committee of Queen's Park Rangers FC has requested us to look into the knock on effect of poor sporting performance on community activity.

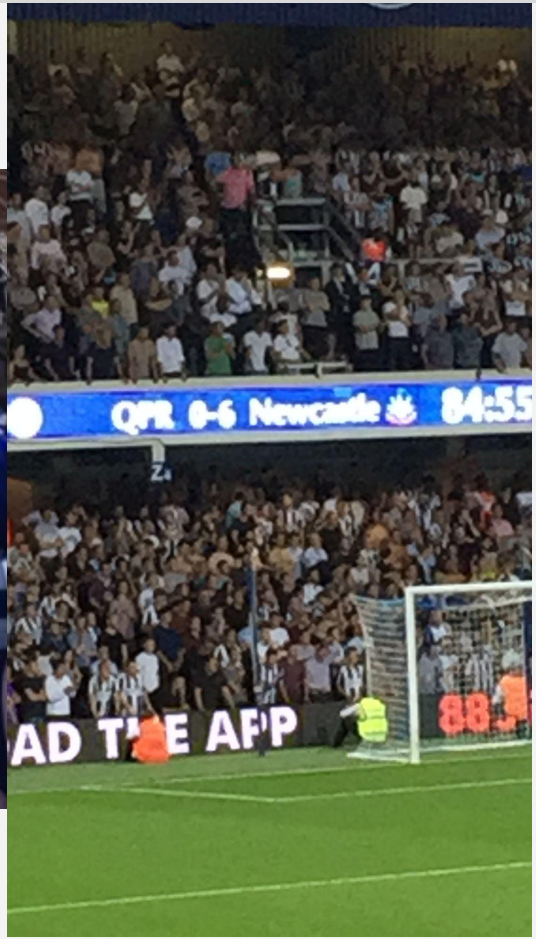
- This project looks at how a teams overall performance can affect the overall activity and language of that teams reddit community, and whether language differences caused by team scale or geography are distinguishable through the balanced accuracy of language processing models.

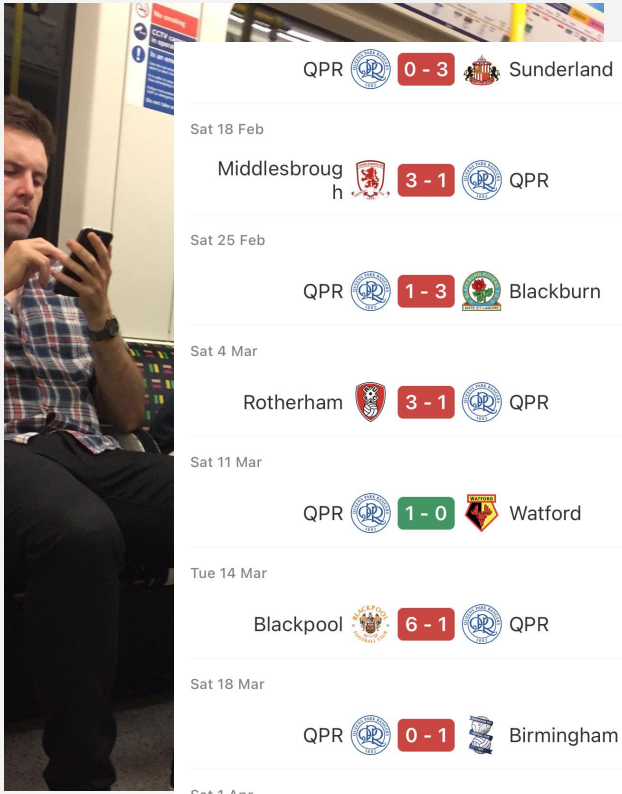
Source:

- thekoptimes
- BT Sport

Source:

- Sky Sports



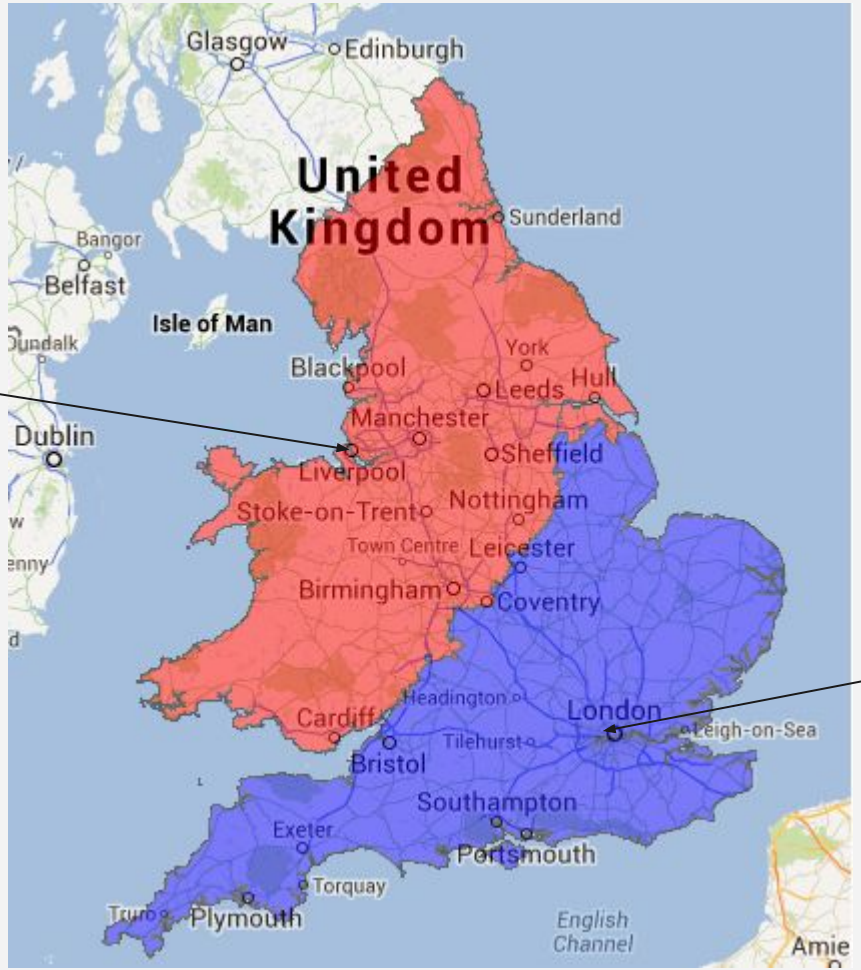


	QPR	0 - 3	Sunderland	
Sat 18 Feb				
	Middlesbrough	3 - 1	QPR	
Sat 25 Feb				
	QPR	1 - 3	Blackburn	
Sat 4 Mar				
	Rotherham	3 - 1	QPR	
Sat 11 Mar				
	QPR	1 - 0	Watford	
Tue 14 Mar				
	Blackpool	6 - 1	QPR	
Sat 18 Mar				
	QPR	0 - 1	Birmingham	
Sat 1 Apr				
	Wigan	1 - 0	QPR	

North South Divide

Liverpool

437k
Kopites

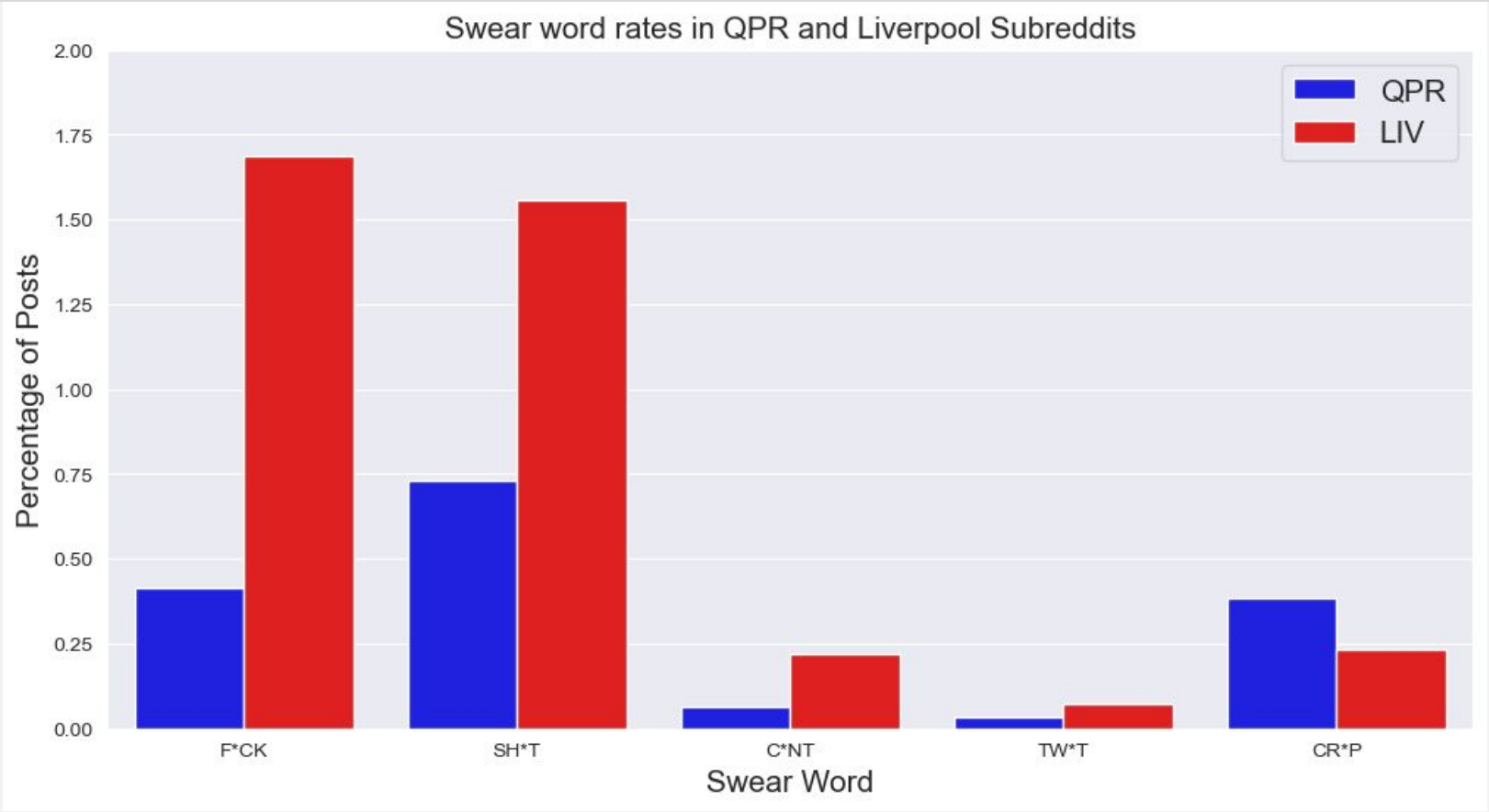


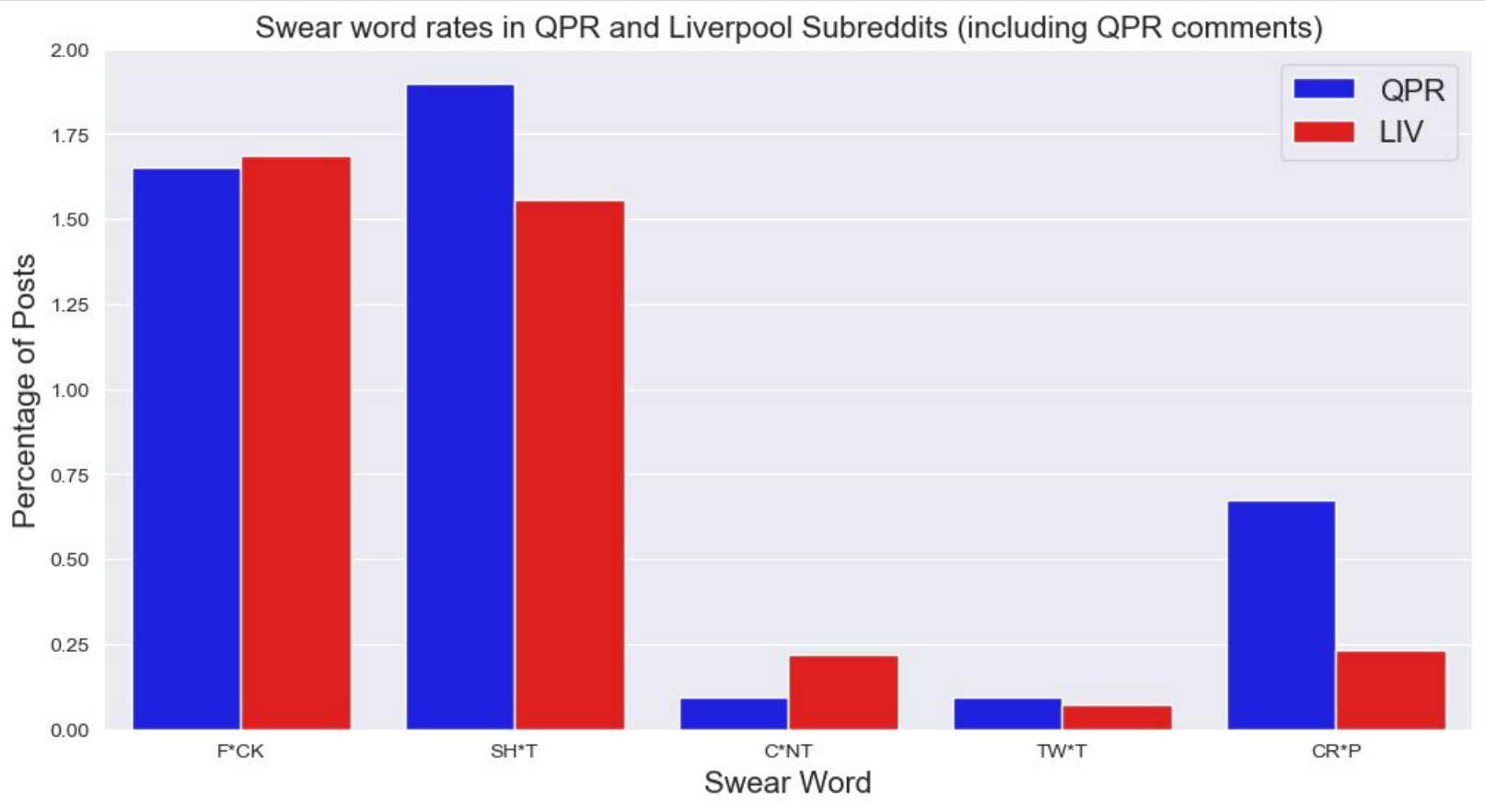
QPR

2.6k
Hoops

Source:

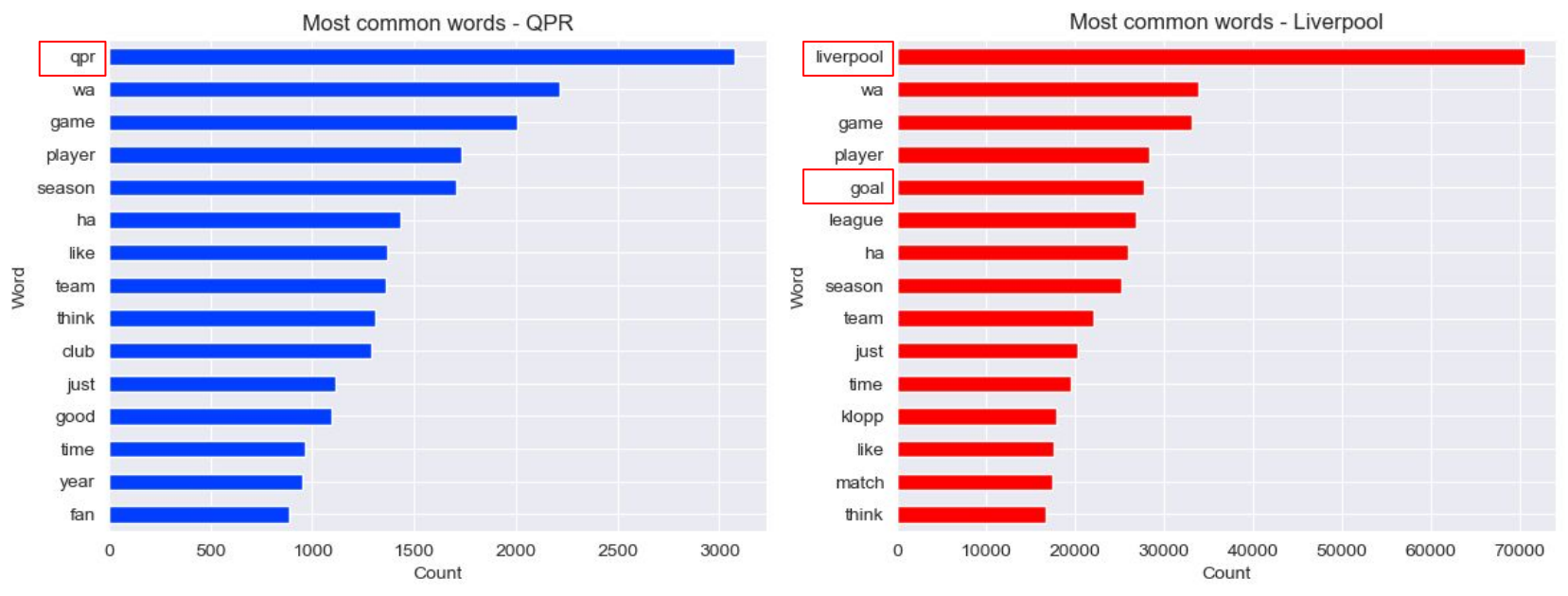
- Brilliant Maps

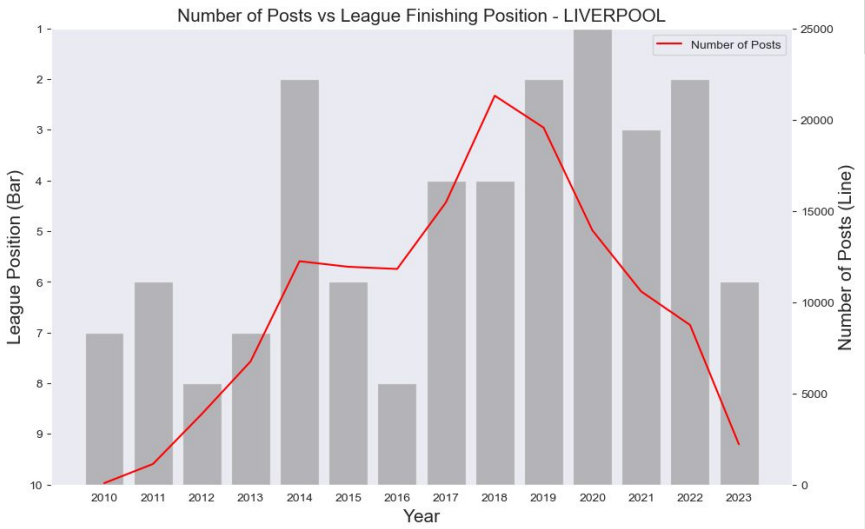

Swear word rates in QPR and Liverpool Subreddits
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Percentage of Posts
QPR
LIV
F*CK
SH*T
C*NT
TW*T
CR*P
Swear Word



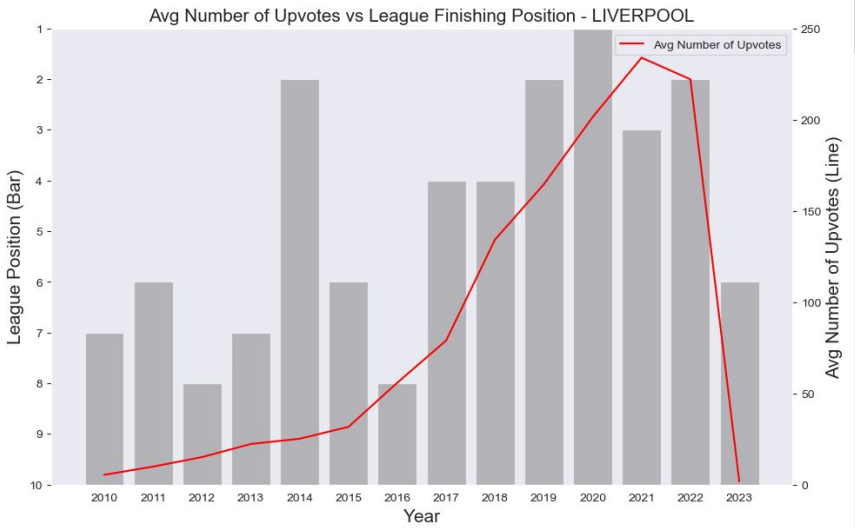
- Comments are far more likely to include strong language than title posts
- Dialect may have an effect on specific frequency of some words

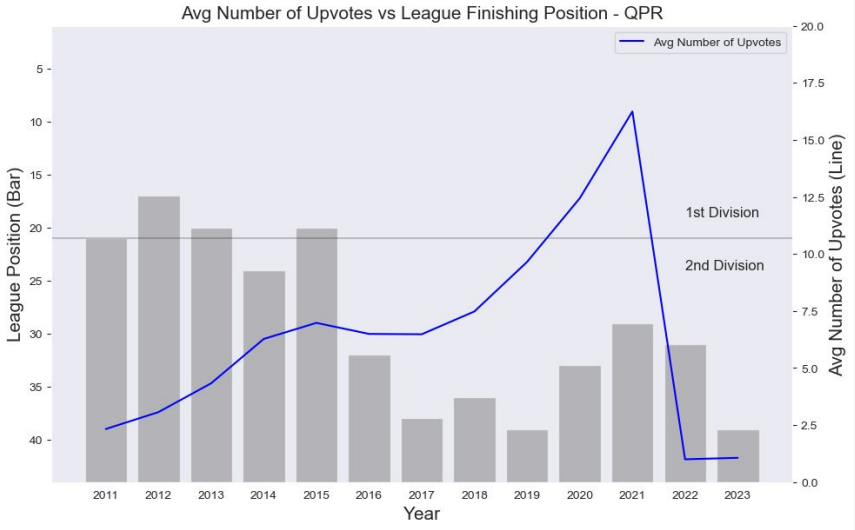
Common Words





- Subreddit activity increases during periods of relative success
- Whilst success tends to peak activity levels, this activity drops over time despite continued success
- Both subreddits show reductions in activity from 2019 - present.

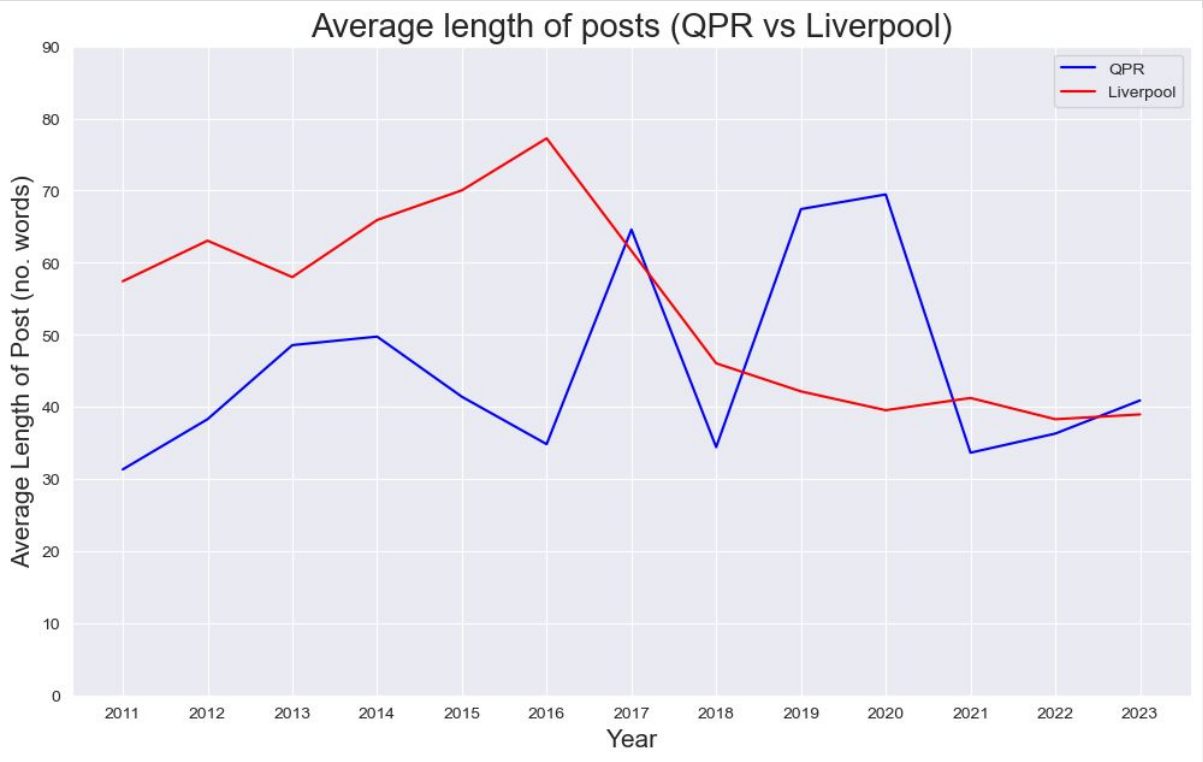




- Post approval has generally increased each year for the last decade

- It is inconclusive that this increase is related to the team performance, since QPRs subreddit also shows a time-wise increase in post approval, whilst team results have worsened.

- Number of subreddit members will usually only increase over time, perhaps explaining this observation



- When LIV performance was the strongest (2018-2022), the average post length decreased in word count.

- Data has been filtered to remove posts shorter than 10 words, so this data is not entirely reliable.

Language Models

- Null model:
    - Balanced Accuracy - 50%
- ExtraTreesClassifier:
    - Balanced Accuracy - 52%
- **Stacked model of Multinomial Naive Bayes, Bernoulli Naive Bayes and Logistic Regression:**
    - **Balanced Accuracy - 92.2%**
- RandomOversampler was the most effective method to combat imbalance in the data

Conclusions

- The success of a football team generally has a positive impact on community activity and post frequency.

- Larger teams are likely to have a more toxic community in terms of swear words used per post than smaller teams. There are exceptions to this however, since some words are more commonly used in certain dialects.

- In general, increased success of a team results in a lower post length on average.

- Language models are shown to be able to predict the original subreddit of a text body with high accuracy, despite initial predictions suggesting that the language used would be too similar to form clear distinctions.

Recommendations

- Collecting data proved to be challenging, since Liverpools subreddit was so large, and QPRs was so small.

    - Collecting comments from Liverpools subreddit would result in too large a data imbalance, but would be required to reliably compare language across all types of text body.

    - Certain language is more likely in comments than in prose paragraphs or titles.

- Comparing language used with a focus on dialect differences would be an interesting follow up route, which was briefly introduced upon inspection of swear word frequency.