Web APIs & NLP

OR

LIVERPOOL FOOTBALL CLUB

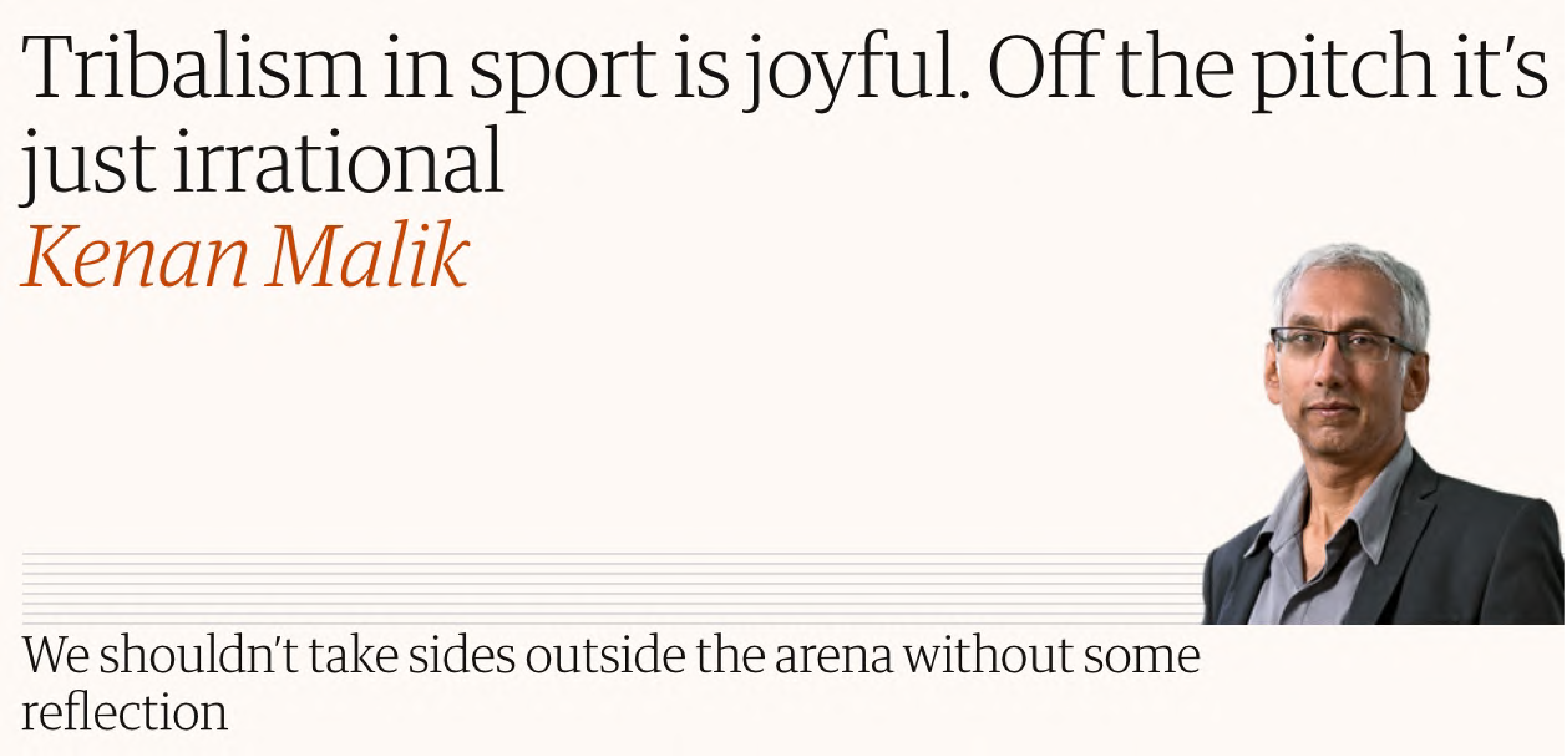
vs

EVERTON

A REDDIT CLASSIFICATION CHALLENGE

"Some people think football is a matter of life and death. I assure you, it's much more important than that. "

Bill Shankly, Manager, Liverpool FC, 1959-1974

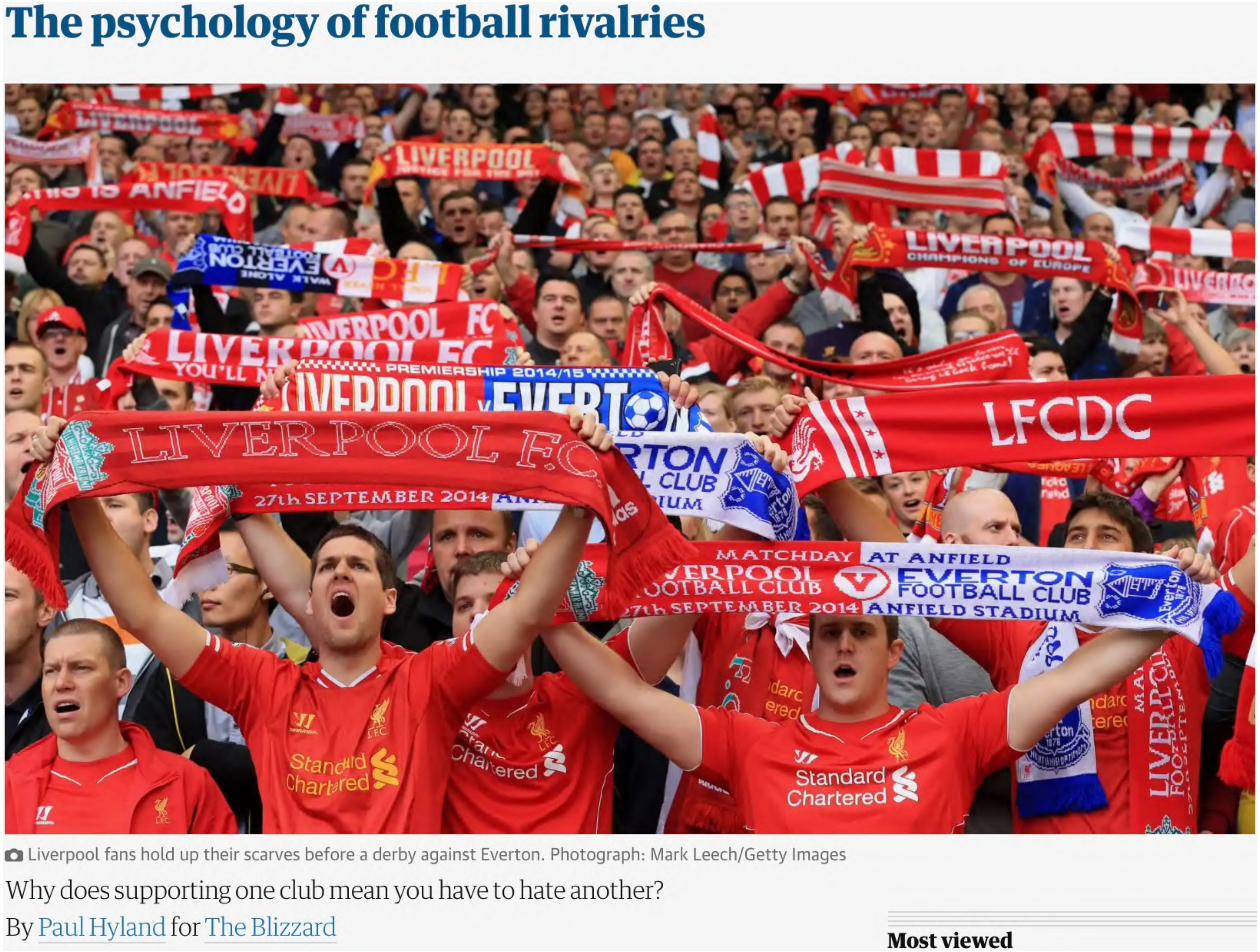


Tribalism in sport is joyful. Off the pitch it's just irrational

Kenan Malik

We shouldn't take sides outside the arena without some reflection

THE GUARDIAN



The psychology of football rivalries

Liverpool fans hold up their scarves before a derby against Everton. Photograph: Mark Leech/Getty Images

Why does supporting one club mean you have to hate another?

By Paul Hyland for The Blizzard

Most viewed

THE GUARDIAN

It's a well-known principle in social psychology that people define themselves in terms of social groupings and are quick to denigrate others who don't fit into those groups. Others who share our particular qualities are our "ingroup," and those who do not are the "outgroup."

PSYCHOLOGY TODAY

Question:

Is it possible to create a binary classification model that can distinguish between posts in rival subreddits with greater than 90% accuracy?

INITIAL CONSIDERATIONS

CHALLENGES

- □ **The obvious:** r/LiverpoolFC and r/Everton are football (soccer) subreddits, so naturally there will be overlap in how they speak about certain things
- □ **The annoying (for my model):** Liverpool and Everton are both based in Liverpool, meaning there are not any easily detectable regional dialect differences
- □ **The hilarious (to me):** There is a night and day difference between total Liverpool and Everton supporters (Liverpool's subreddit has 337k members while Everton's has 30k), which can lead to issues balancing the data.

The Dataset

1. Reddit submissions data acquired using Pushift's API
2. 57, 908 posts
3. 29, 010 from r/LiverpoolFC
4. 28, 889 from r/Everton
5. Initial features were subreddit, selftext, and title

THE PROCESS

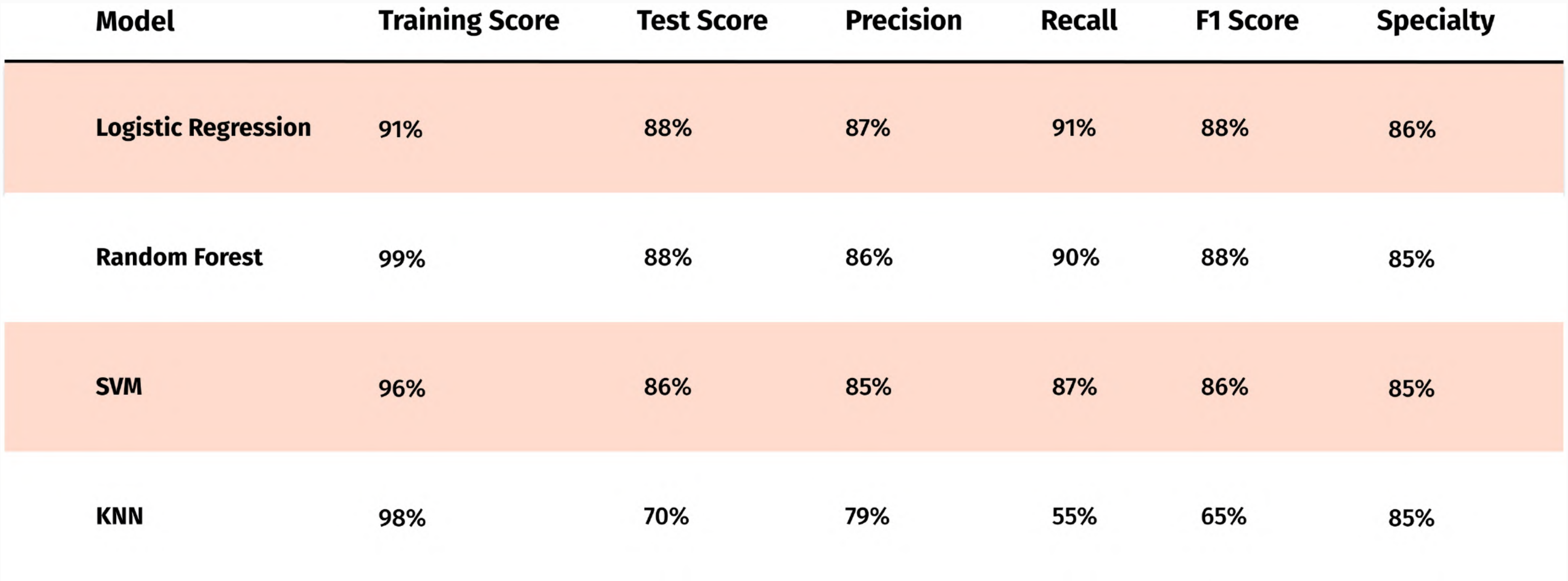
1. Collect data

2. Clean data: removed duplicates, mapped integer values for Liverpool (1) and Everton(0), and imputed text into into empty selftext fields

3. Exploratory data analysis

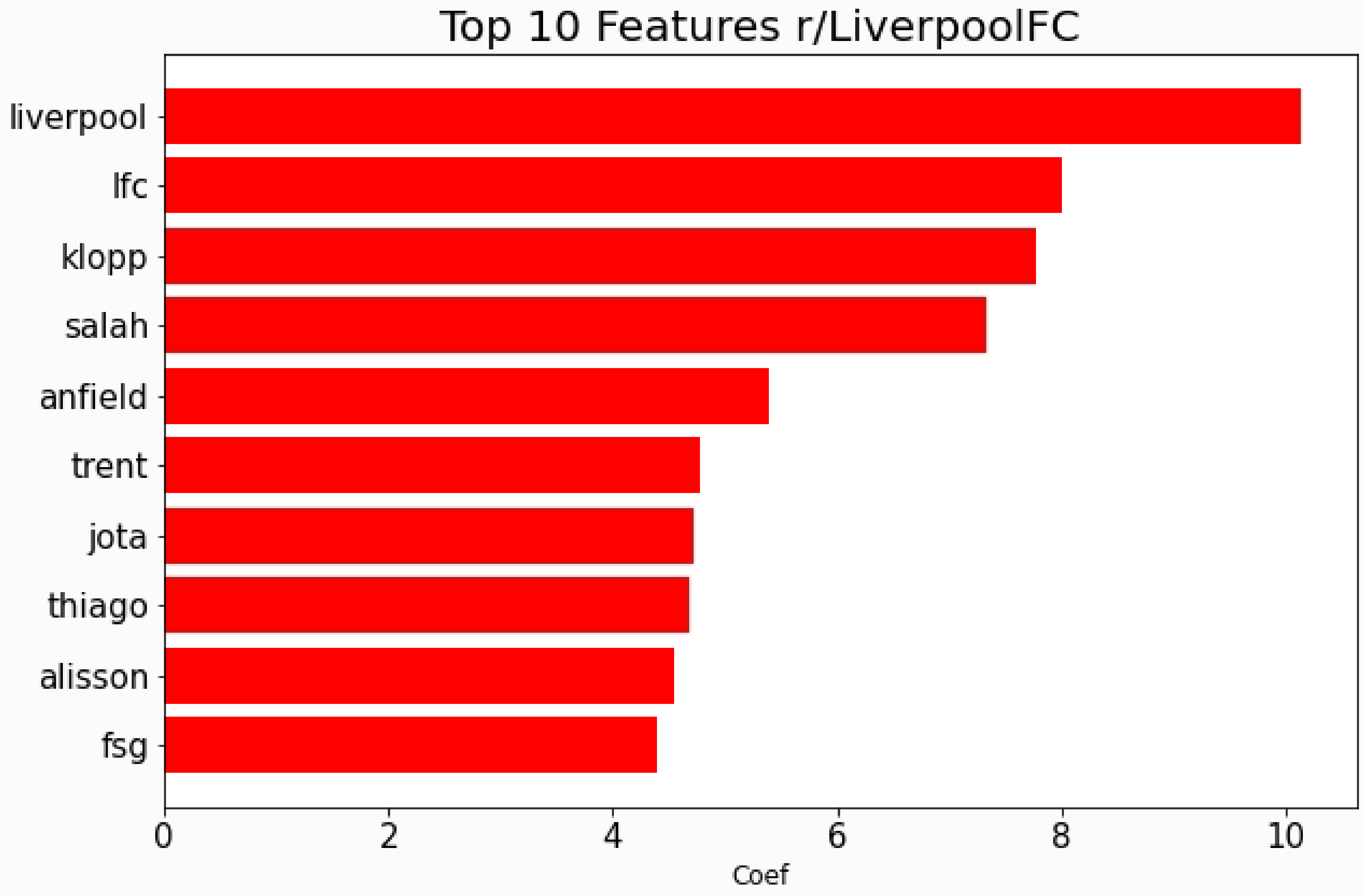
4. Modelling

Model Evaluation

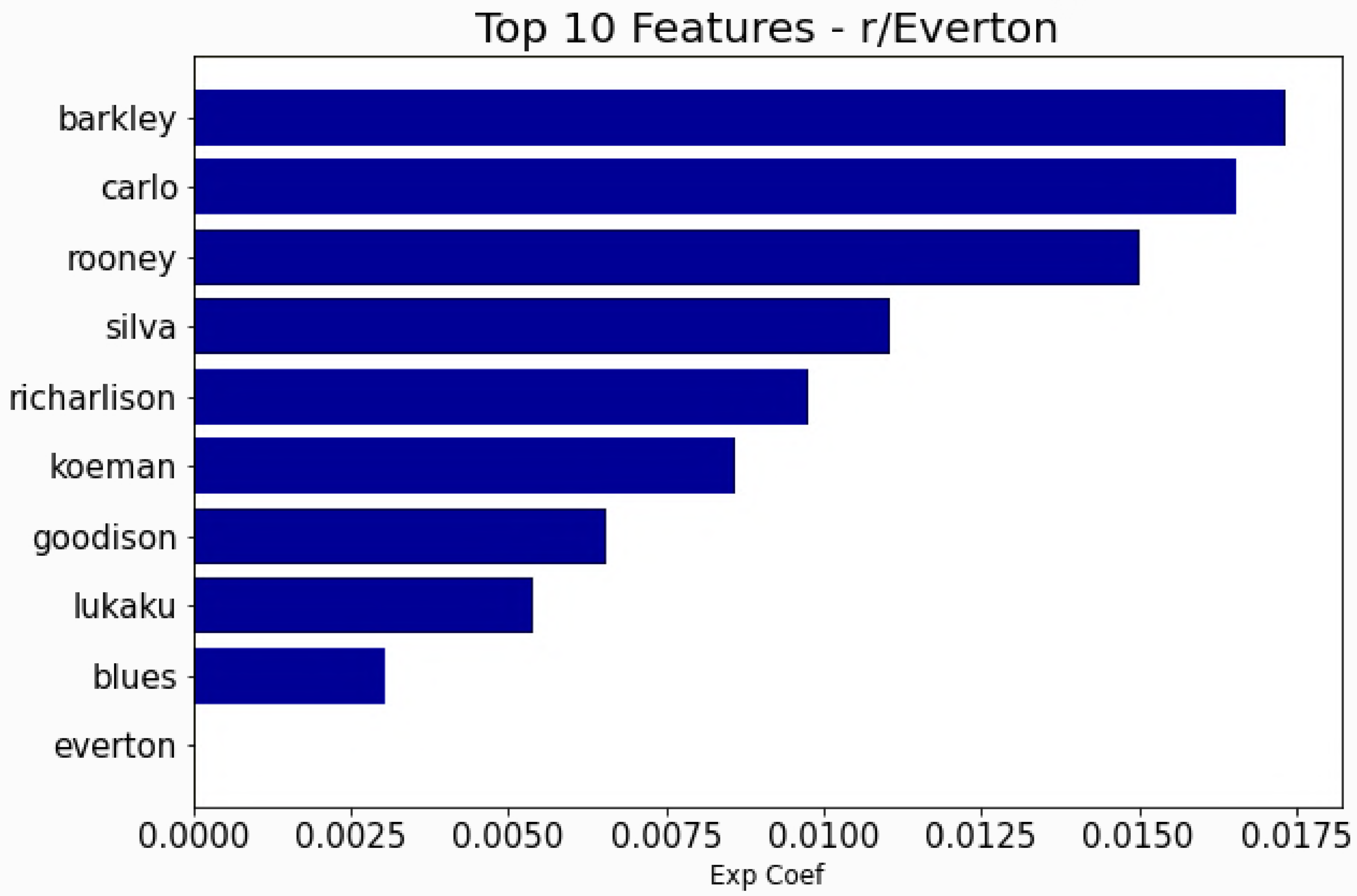


Model	Training Score	Test Score	Precision	Recall	F1 Score	Specialty
Logistic Regression	91%	88%	87%	91%	88%	86%
Random Forest	99%	88%	86%	90%	88%	85%
SVM	96%	86%	85%	87%	86%	85%
KNN	98%	70%	79%	55%	65%	85%

Top Features for r/LiverpoolFC



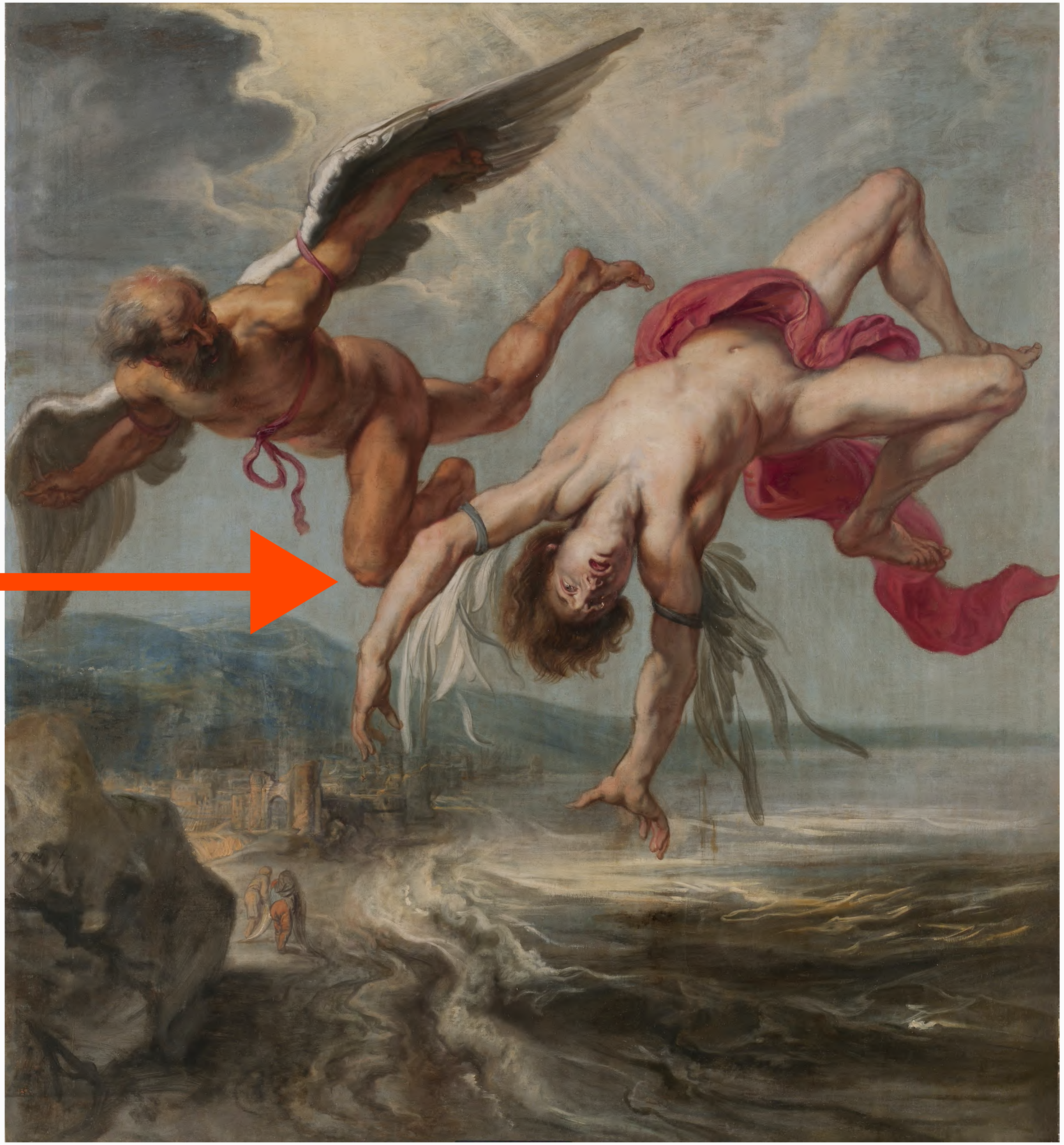
Top Features for r/Everton



Pitch

CONCLUSIONS

This was me



This was also me

1. If there is something inherently and fundamentally different about rival supporters, it isn't easily found with a classification model

2. Listen to Sumit

3. If you don't listen to Sumit, make sure you have a lot of processing power

THANK YOU!