WEB APIS & CLASSIFICATION

NURABIDAH JAMIL
27 JUNE 2021

BACKGROUND

My consultancy company has a client who would like to run marketing campaigns for its upcoming football event and wishes to target online audience who is interested in the English Premier League. Hence, he would like to explore classification models that are able to predict whether a particular football-related post is related to the English Premier League.

Problem statement

Data is collected from two subreddits - English Premier League and Champions league.

The goal is to build and compare different classification models that would best predict whether a specific post belongs to the English Premier League subreddit.

Table of Contents

METHODOLOGY

Data collection

- English Premier League – 943 posts
- Champions League – 992 posts

Cleaning & Pre-processing

- Combine post title & content features
- Lemmatize
- Remove URLs, non-relevant words

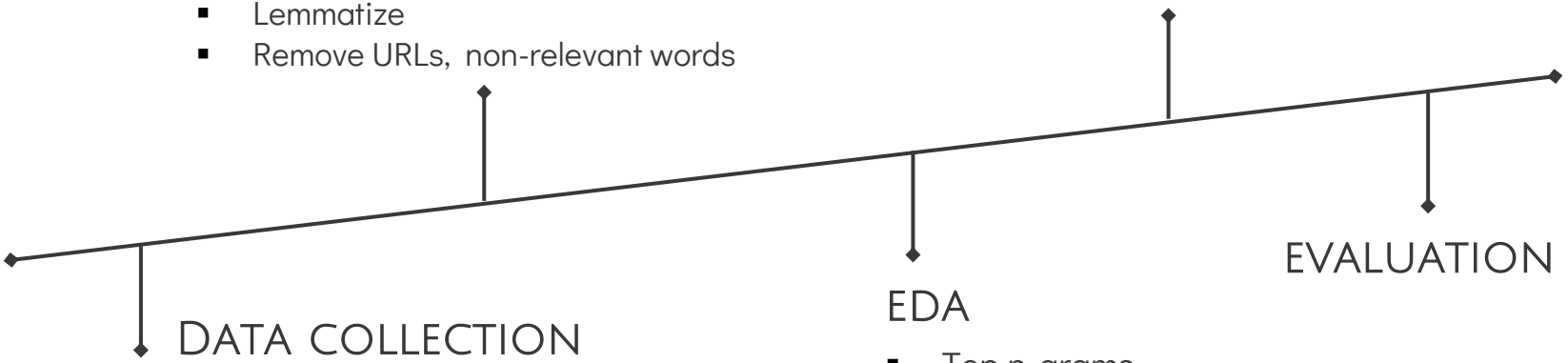
EDA

- Top n-grams

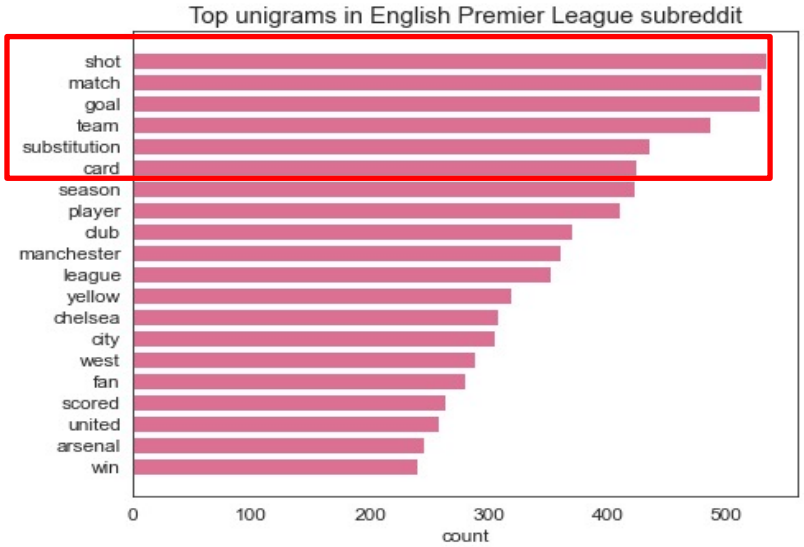
Modeling & tuning

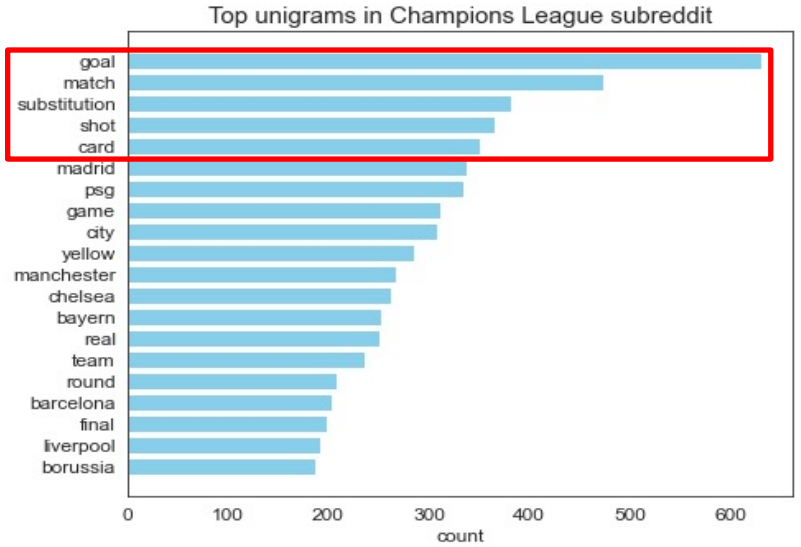
- Multinomial Naïve Bayes
- Logistic Regression
- K-Nearest Neighbours

Evaluation



Top unigrams

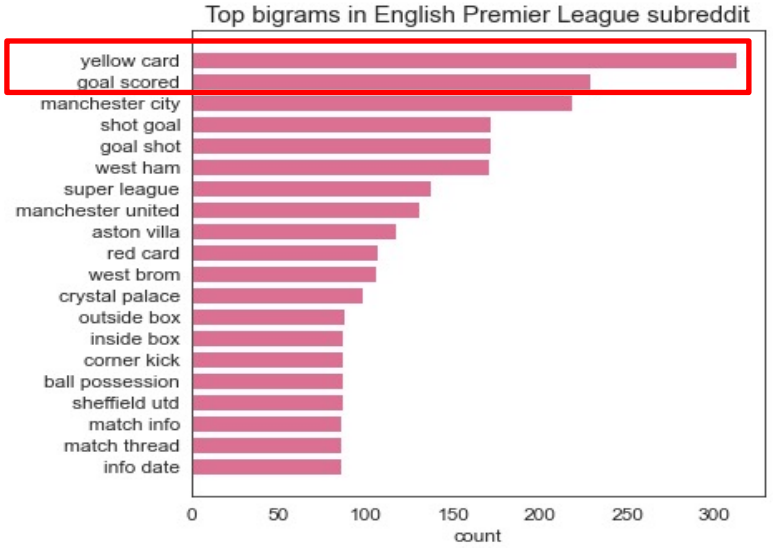


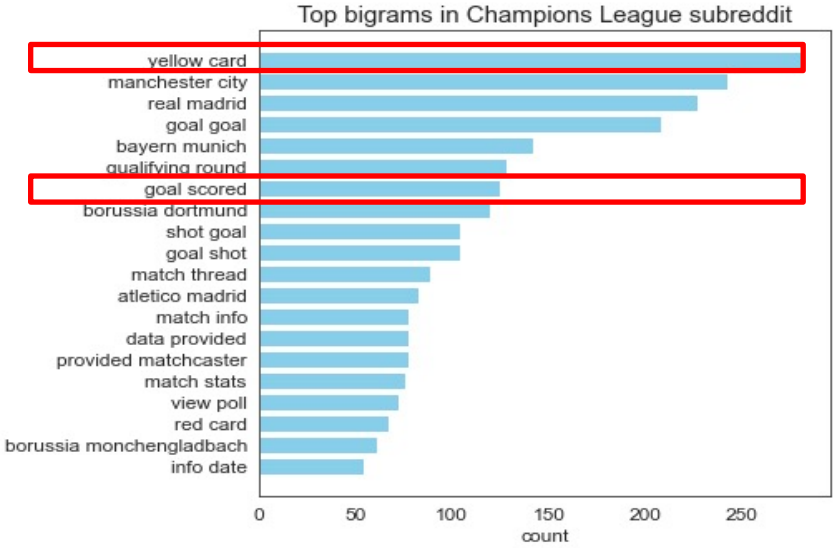


- Top unigrams for EPL subreddit include EPL clubs & common football-related terms
- Common top unigrams found in both subreddits: shot, match, goal

- Top unigrams for CPL subreddit include top-tiered EPL clubs and non-EPL clubs

Top bigrams

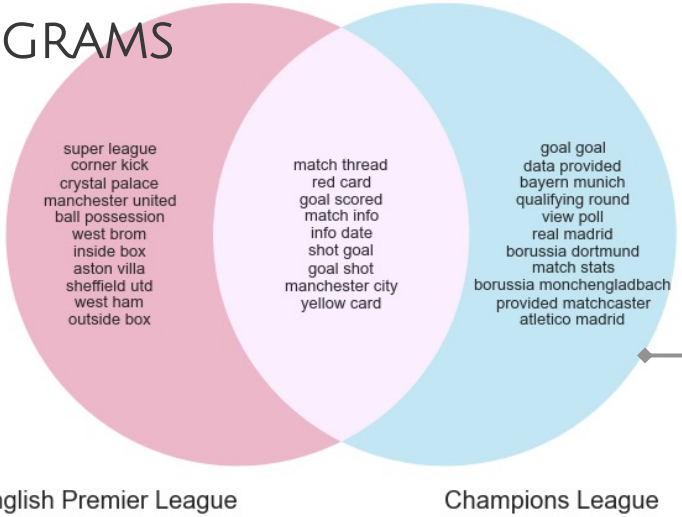




- Top bigrams for EPL subreddit include EPL clubs & common football terms
- Common top bigrams found in both subreddits: yellow card, goal scored, Manchester city
- Top bigrams for CPL subreddit include top-tiered EPL clubs and non-EPL clubs

Top unigrams

Top bigrams



Top unigrams/bigrams

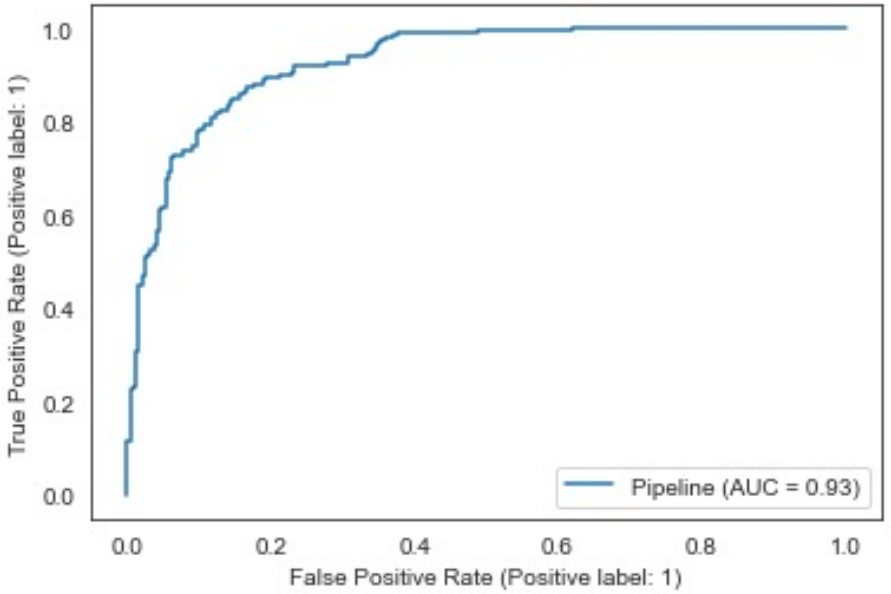
- Words that are specific to the format of the CPL are only found in the CPL subreddit.
- Top-tiered EPL clubs are found in both subreddits while non-EPL clubs are found only in CPL subreddit.

- Words that are specific to the format of the CPL tournament are only found in the CPL subreddit.
- Low/ mid-tiered English clubs are only found in the EPL subreddit.

Results summary

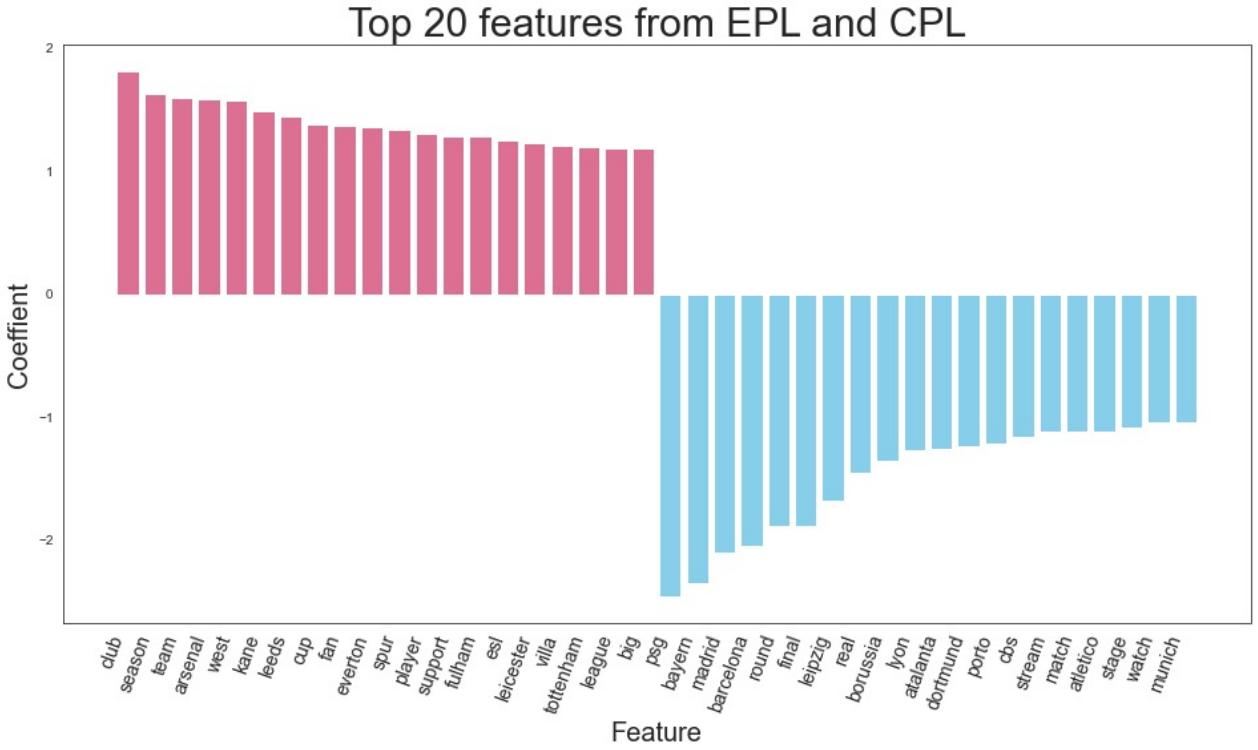
Model	Train	Test	Sensitivity	Specificity	Precision
Logistic Reg + TF-ID	0.938108	0.845833	0.812766	0.877551	0.864253
Naïve Bayes + TF-ID	0.910292	0.841667	0.851064	0.832653	0.829876
Naïve Bayes + Count Vect	0.865090	0.835417	0.914894	0.759184	0.784672
Logistic Reg + Count Vect	0.975661	0.827083	0.757447	0.893878	0.872549
KNN + TF-ID	0.856050	0.735417	0.561702	0.902041	0.846154
KNN + Count Vect	0.769124	0.679167	0.395745	0.951020	0.885714

Model Selected



- ROC AUC is 0.93.
- Positive and negative classes (EPL and CPL posts) are almost perfectly separated.

Model Selected



Team names are important

- Mid/low-tiered clubs for EPL subreddit
- European non-English clubs for CPL subreddit

Terms specific to the league are important

- Final, round and stage for CPL subreddit
- Cup for EPL subreddit

Conclusions & next steps

Many common words

- Top-tiered clubs in both subreddits
- Common football terms in both subreddits

Importance of team names & league specific words

- Mid/low-tiered clubs for EPL subreddit
- Terms associated with the domestic football league like cup.

Incorporate more data & test other models

- Random Forest Classifier