

#### 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.



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

OI MILITIODOLOGI	01	<b>METHODOLOGY</b>
------------------	----	--------------------

- O2 EDA

  Top Unigrams

  Top Bigrams
- O3 MODELS
  Results Summary
  Model Selected
- 04 CONCLUSIONS / NEXT STEPS

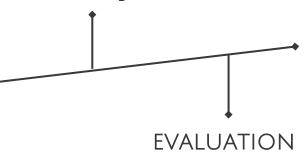
#### **METHODOLOGY**

### CLEANING & PRE-PROCESSING

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

#### MODELING & TUNING

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



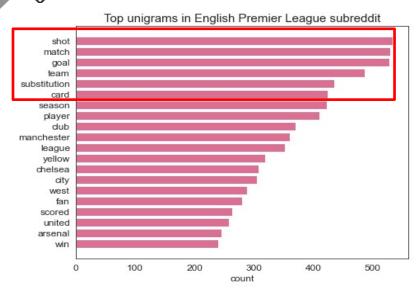
#### DATA COLLECTION

- English Premier League 943 posts
- Champions League 992 posts

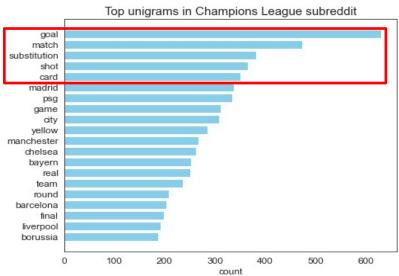
Top n-grams

EDA

#### TOP UNIGRAMS

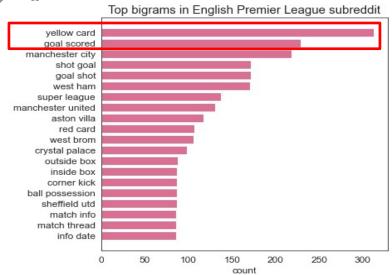


- TOP UNIGRAMS FOR EPL SUBREDDIT INCLUDE EPL CLUBS & COMMON FOOTBALL-RELATED TERMS
- COMMON TOP UNIGRAMS FOUND IN BOTH SUBREDDITS: SHOT, MATCH, GOAL

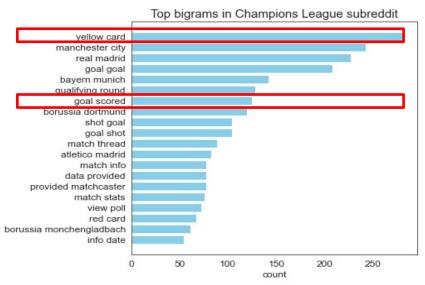


TOP UNIGRAMS FOR CPL SUBREDDIT INCLUDE TOP-TIERED EPL CLUBS AND NON-FPL CLUBS

#### 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

fan arsenal scored west club player win season league united shot card manchester yellow city chelsea team goal match substitution final round game real liverpool psg barcelona madrid bayern borussia

**English Premier League** 

Champions League

#### TOP BIGRAMS

super league corner kick crystal palace manchester united ball possession west brom inside box aston villa sheffield utd west ham outside box

match thread red card goal scored match info info date shot goal goal shot manchester city yellow card goal goal
data provided
bayern munich
qualifying round
view poll
real madrid
borussia dortmund
match stats
borussia monchengladbach
provided match-caster
atletico madrid

English Premier League

Champions League

#### TOP UNIGRAMS/BIGRAMS

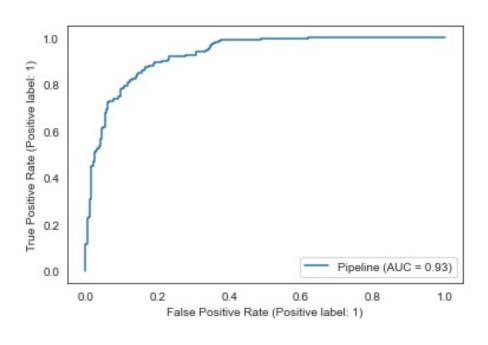
- Words that are specific to the format of the CPL are only found in the CPL subreddit.
- Top-tiered EPL clubs are found in both subreddits while non-EPL clubs are found only in CPL subreddit.

- 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

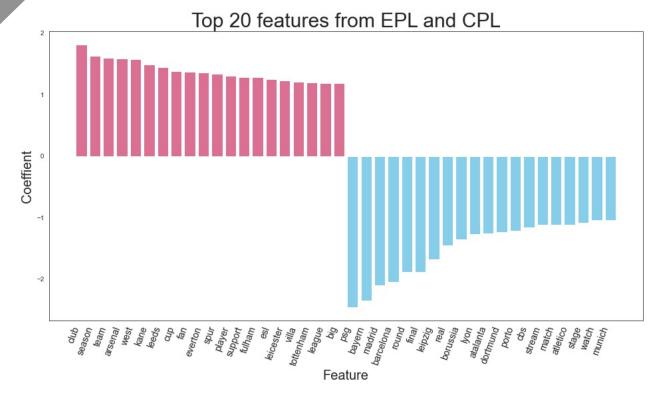
Model	Train	TEST	SENSITIVITY	SPECIFICITY	precision
Logistic Reg + TF-ID	0.938108	0.845833	0.812766	0.877551	0.864253
Naïve Bayes + TF-ID	0.910292	0.841667	0.851064	0.832653	0.829876
Naïve Bayes + Count Vect	0.86509 0	0.835417	0.914894	0.759184	0.784672
Logistic Reg + Count Vect	0.975661	0.827083	0.757447	0.893878	0.872549
KNN + TF-ID	0.85605 0	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 <u>0.93</u>.
- 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