Sorting Readers' Comments for EuroFootball Magazine

Loh Chuan Hui
for the EuroFootball Data Science Team

Contents

Introduction

EuroFootball

Monthly European Football Magazine with Reader's Comment Section

Editorial Team

- Review Readers' Comments
- Unsorted comments increase editorial team workload

Data Science Team

Tasked with training a model capable of classifying comments as Champions League or Premier League related using Reddit posts from r/PremierLeague and r/championsleague

Background


United Kingdom
WALES

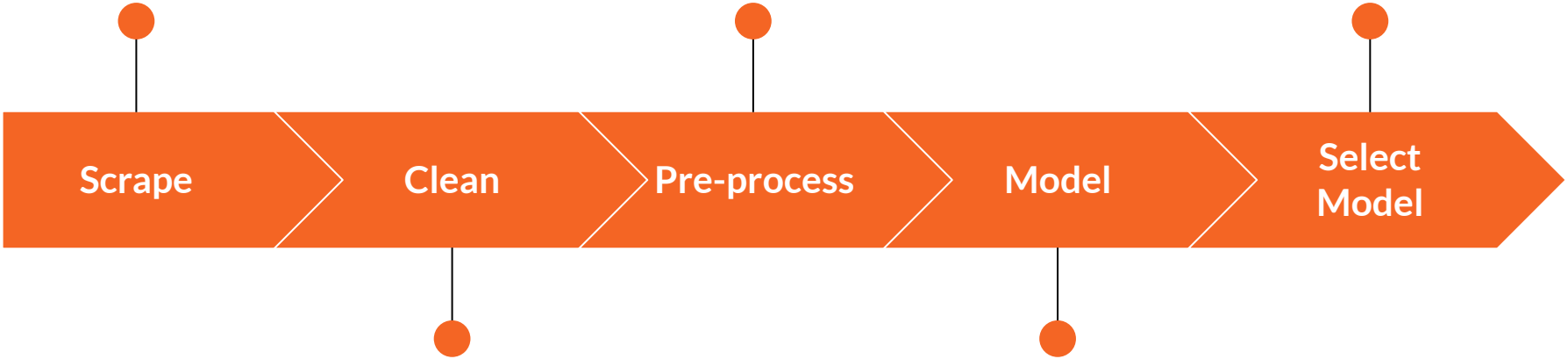

Leeds
Newcastle
Aston Villa
West Ham
Liverpool
Man City
Man United
Chelsea
Real Madrid
Barcelona
Juventus
Bayern Munich


UEFA
CHAMPIONS
LEAGUE

Scrape posts from r/championsleague and r/PremierLeague

Tokenize and Lemmatize texts

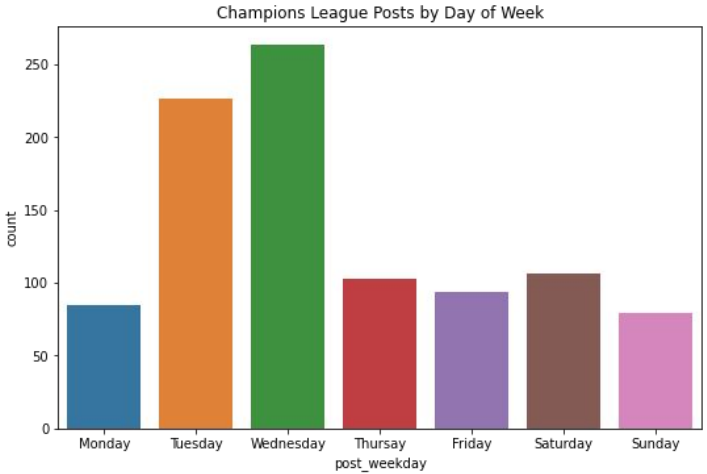
Pick the best model

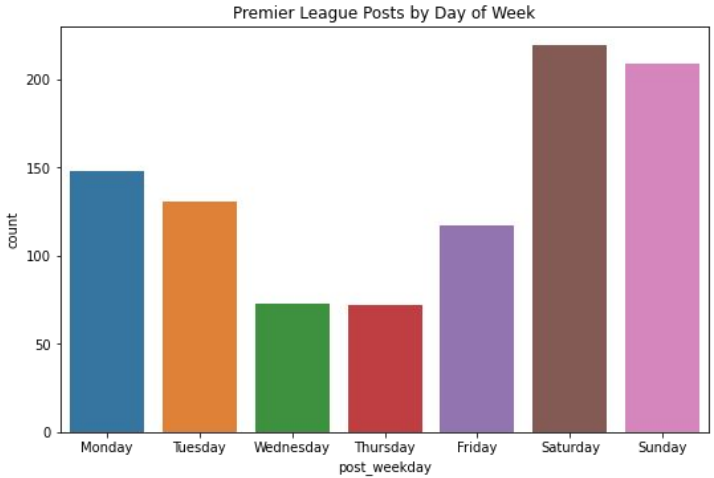


Remove urls and duplicates

Fit and evaluate several models

Posts by Day of Week

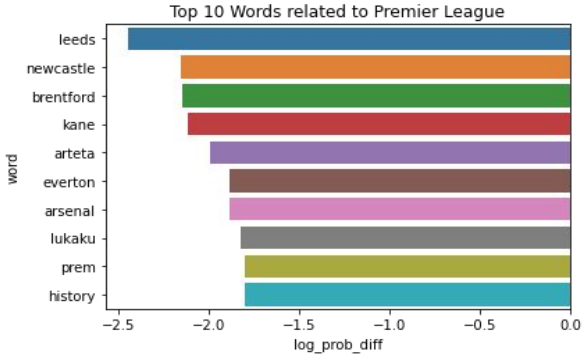


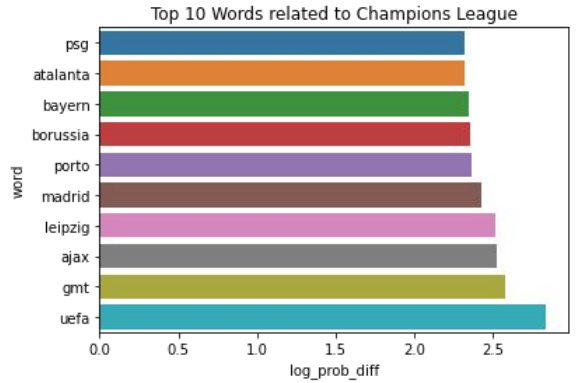


Accuracy:
0.878

Model 1 - Multinomial Naive Bayes

- English club names not in CL are important features for predicting a Premier League subreddit
- Likewise, European clubs for Champions League





Accuracy:
0.875

Model 2 - Log Reg with L1 Pen

- Coefficient of Log Reg a measure of words feature importance
- L1 Penalization shuts off coefficient of word features aggressively

Champions League

champion	2.110390e+01
view	1.691131e+01
uefa	1.607233e+01
psg	1.549558e+01
ucl	1.532187e+01
gmt	1.497678e+01
final	1.423930e+01
leipzig	1.280796e+01
tournament	1.246166e+01
canada	1.191024e+01

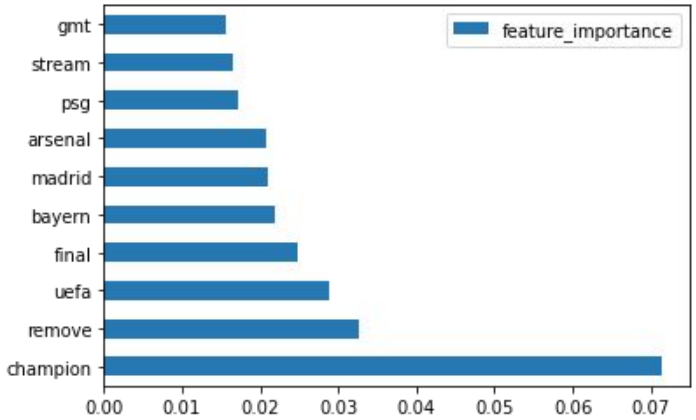
Premier League

kane	-9.726392e+00
prem	-9.728505e+00
history	-9.903072e+00
newcastle	-1.062111e+01
leeds	-1.066110e+01
quality	-1.105094e+01
reason	-1.114178e+01
arsenal	-1.231157e+01
final win	-1.554898e+01
champion champion	-2.169564e+01

Accuracy:
0.885

Model 3 - Random Forest

- Feature importance of Random Forest identify words which contribute most to gini impurity
- Curtailing Random Forest max depth reduced accuracy of model



Observations

- Model confirms that word features of importance tend to be names of football clubs that is not an English Club playing in the Champions League
- Misclassified posts tends to be:
  - Too short and lacking important words
  - Talking about the Champions League in a Premier League subreddit and vice-versa
  - Has nothing to do with the subreddit. Mostly request for links of pirated live streams of football matches
- Random Forest can handle high dimensional and complex data structures vs algorithms which work on the assumption that classes can be separated by a straight line (e.g.Logistic Regression)

Conclusion & Recommendation

- Recommend Model 3 (TfidfVectorizer with Random Forest) which is 88.5% accurate
- Follow-up Actions to improve model:
  - Scrape more posts for training model from high quality football journal website where focus is not on sharing video clips
  - Remove short post and post that are irrelevant

Thank you