

El Clásico de Reddit



Bárça vs. Real Madrid

By Christopher Williams

20 December 2018

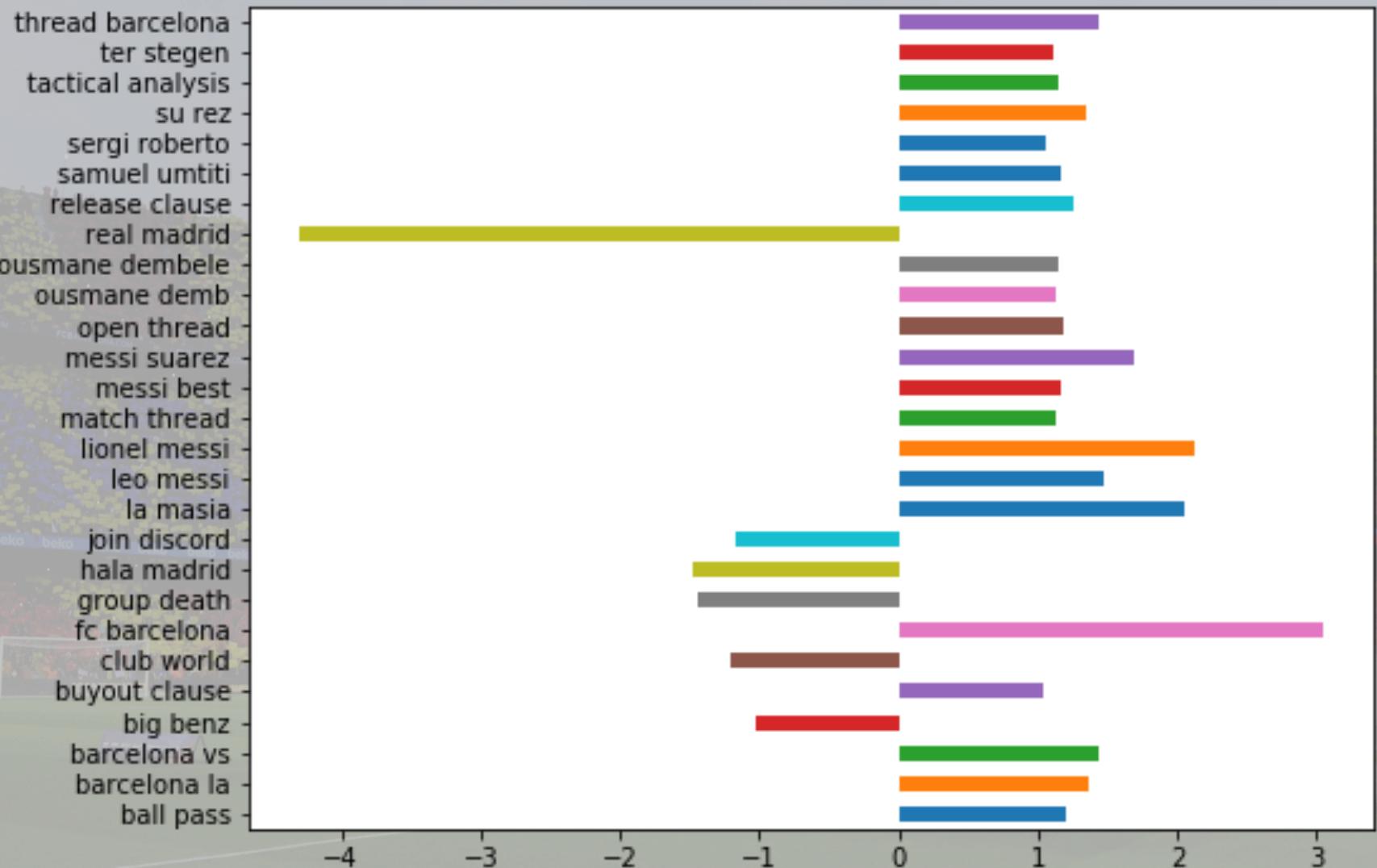
Data Science Problem

- Use NLP models to predict if certain comments / submissions come from one Subreddit or the other
- Subreddit's
 - /r/Barca
 - /r/realmadrid
- The Data Collection Process
 - 2834 from Barca
 - 2747 from realmadrid

Features with Players / Teams



2 grams with Players / Teams



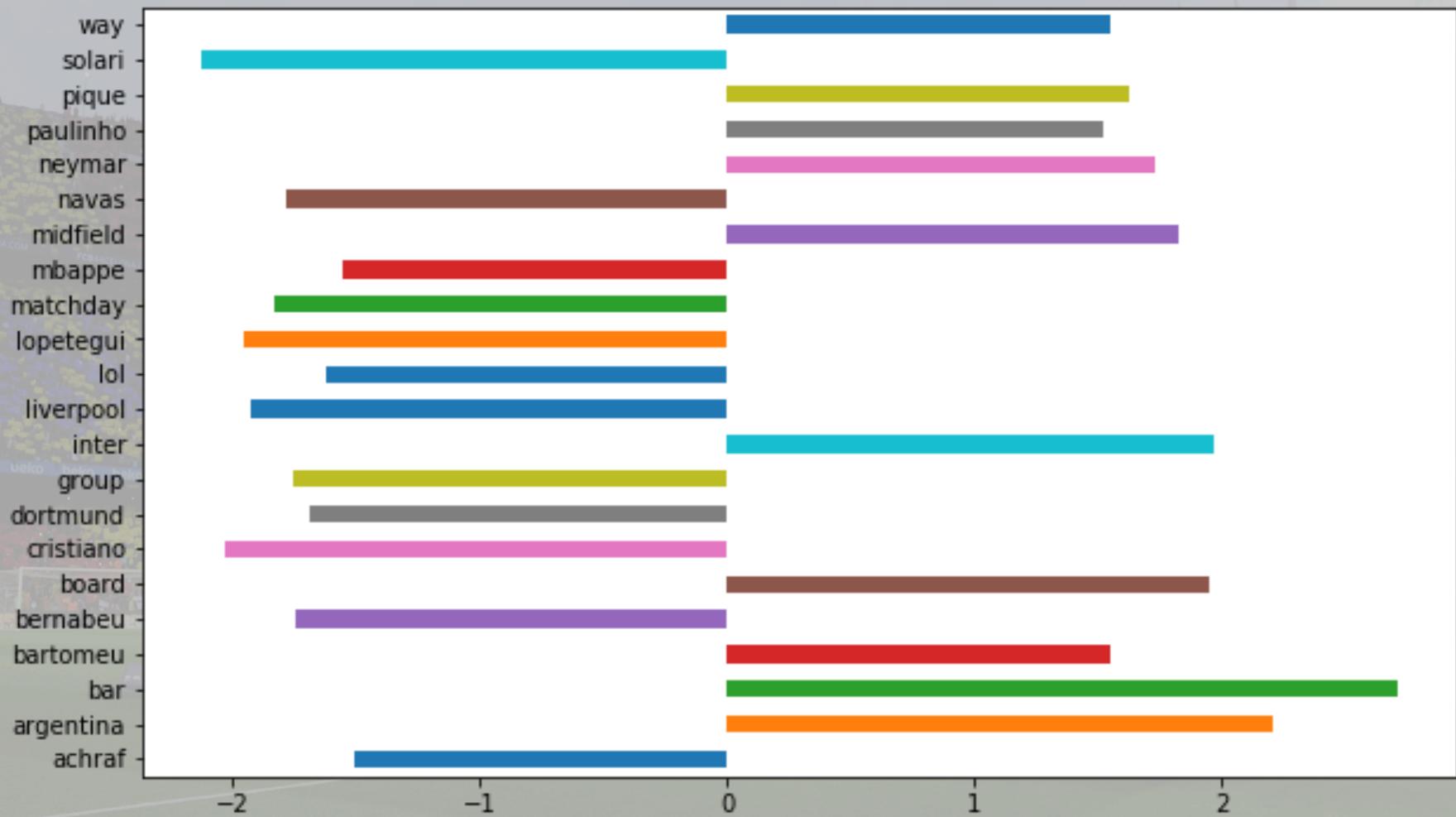
arturo 'semedo'
casemiro 'real'
umtiti 'cuenca'
lenglet 'lunin'
ter '
munir '
adriozala '
daniel 'lioneL'
n  lson 'modric'
cilleissen 'marcos'
debollos 'zidane'
gerrero 'valverde'
haddadi 'el
suarez 'gerard'
samuel 'courtouis'
vidal 'odegaard'
lucas 'asensio'
marc 'alvaro'
thibaut 'casemiro'
su  rez 'casemiro'
dembele 'barca'
arthur 'ivan'
ramos 'varane'
gareth 'raspberry'
marco 'barca'
ezkietar 'casilla'
karim 'malcom'
kiko 'messi'
vinicius 'barca'
mayoral 'barca'
toni 'rakitic'
denis 'carvajal'
luis 'fC'
marcelo 'philippe'
toni 'martin'
ousmane 'jorge'
luca 'andré'
jokin 'coutinho'
tomas 'kroos'
busquets 'isco'
jorge 'vallejo'
ronaldo 'ronaldo'
iniesta 'jordi'
tomas 'ronaldo'

Stop Words:

Took Out Players Names to see performance of models

109 additional stop words along with the sklearn ones

Without Players / Teams



The Models

TfidfVectorizer

LogisticRegression

RandomForestClassifier

ExtraTreesClassifier

TfidfVectorizer and LogisticReg

- TfidfVectorizer - With players names
 - Tfidf training score 0.926
 - Tfidf testing score started at 0.696 accuracy
 - With LogReg in GridSearch with ngrams as (1,3) improved to 0.774 accuracy score

With Players Names

Accuracy: 0.7747014115092291				
	precision	recall	f1-score	support
0	0.75	0.79	0.77	877
1	0.80	0.76	0.78	965
avg / total	0.78	0.77	0.77	1842

Predicted Barca Predicted realmadrid

Actual Barca	697	180
Actual realmadrid	235	730

Without Names

Accuracy: 0.6959826275787188				
	precision	recall	f1-score	support
0	0	0.69	0.69	0.69
1	1	0.70	0.70	0.70
avg / total	0.70	0.70	0.70	1842

Predicted Barca Predicted realmadrid

Actual Barca	628	279
Actual realmadrid	281	654

Forest and Trees

- RandomForestClassifier
 - Best cv scores:
 - Train: 0.732 without | 0.655 with
 - Test: 0.699 without | 0.608 with
- ExtraTreesClassifier
 - Best cv scores:
 - Train: 0.743 without | 0.664 with
 - Test: 0.699 without | 0.604 with
- Without or With Player & Team stop words

Random Forest

- With Player's & Clubs Names
 - Grid Search CV best score: 0.766
 - Best Params: {'max_depth': None, 'max_features': 'auto', 'n_estimators': 43}
- Without Player's & Clubs Names
 - Grid Search CV best score: 0.685
 - Best Params: {'max_depth': None, 'max_features': 'log2', 'n_estimators': 45}

Conclusions

- Top feature is players names
- Take the names out, accuracy score drops substantially
- Regardless due to the nature of comments being so randomly similar within the two subreddits
- To do:
 - To do more model parameters
 - More ngram features (3,6)
 - Put in coaches as stop words