El Clásico de Reddit

Bárça vs. Real Madrid

By Christopher Williams

20 December 2018

Data Science Problem

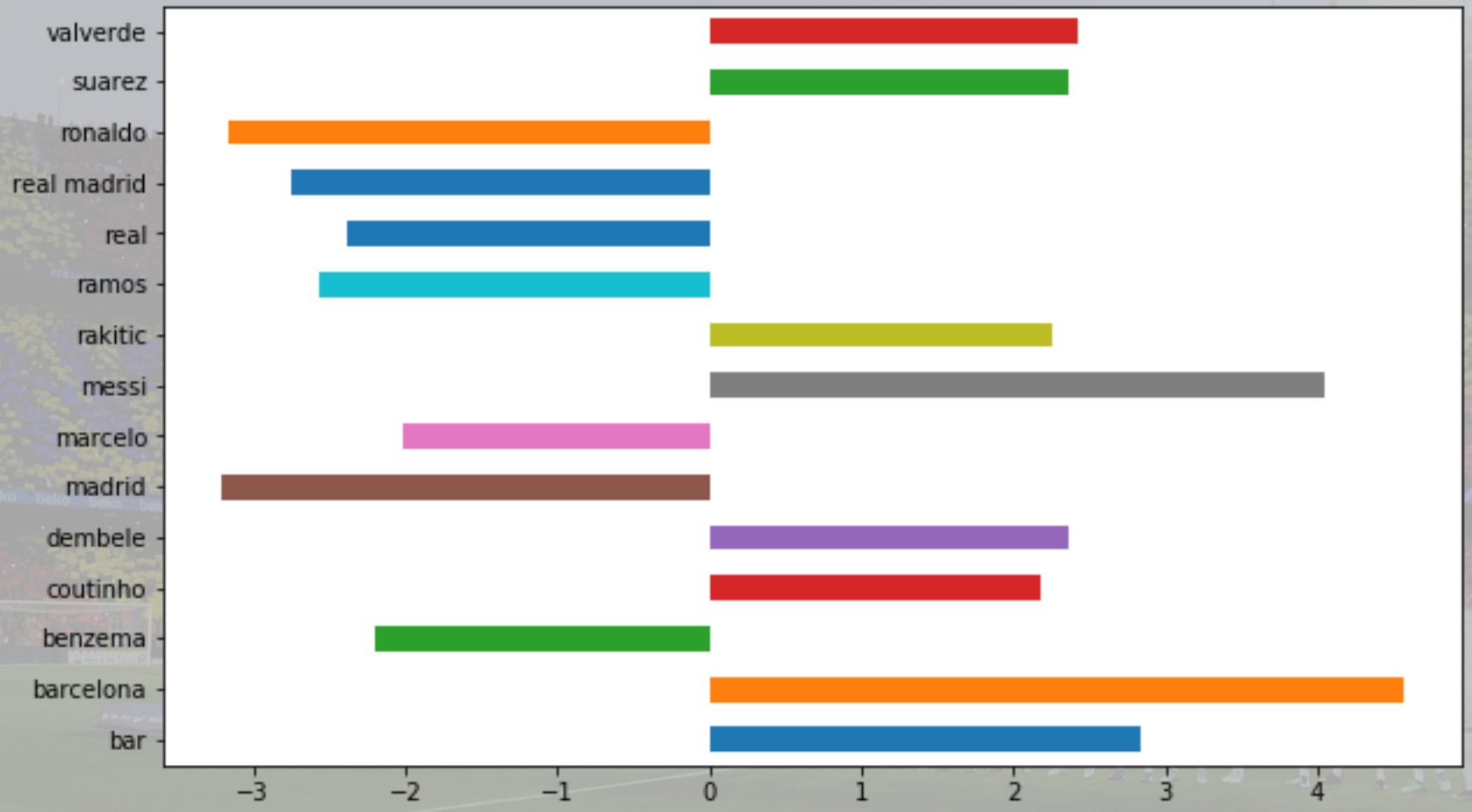
- Use NLP models to predict if certain comments / submissions come from one Subreddit or the other
- Subreddit's
  - /r/Barca
  - /r/realmadrid
- The Data Collection Process
  - 2834 from Barca
  - 2747 from realmadrid

Examples of Subreddits' Data

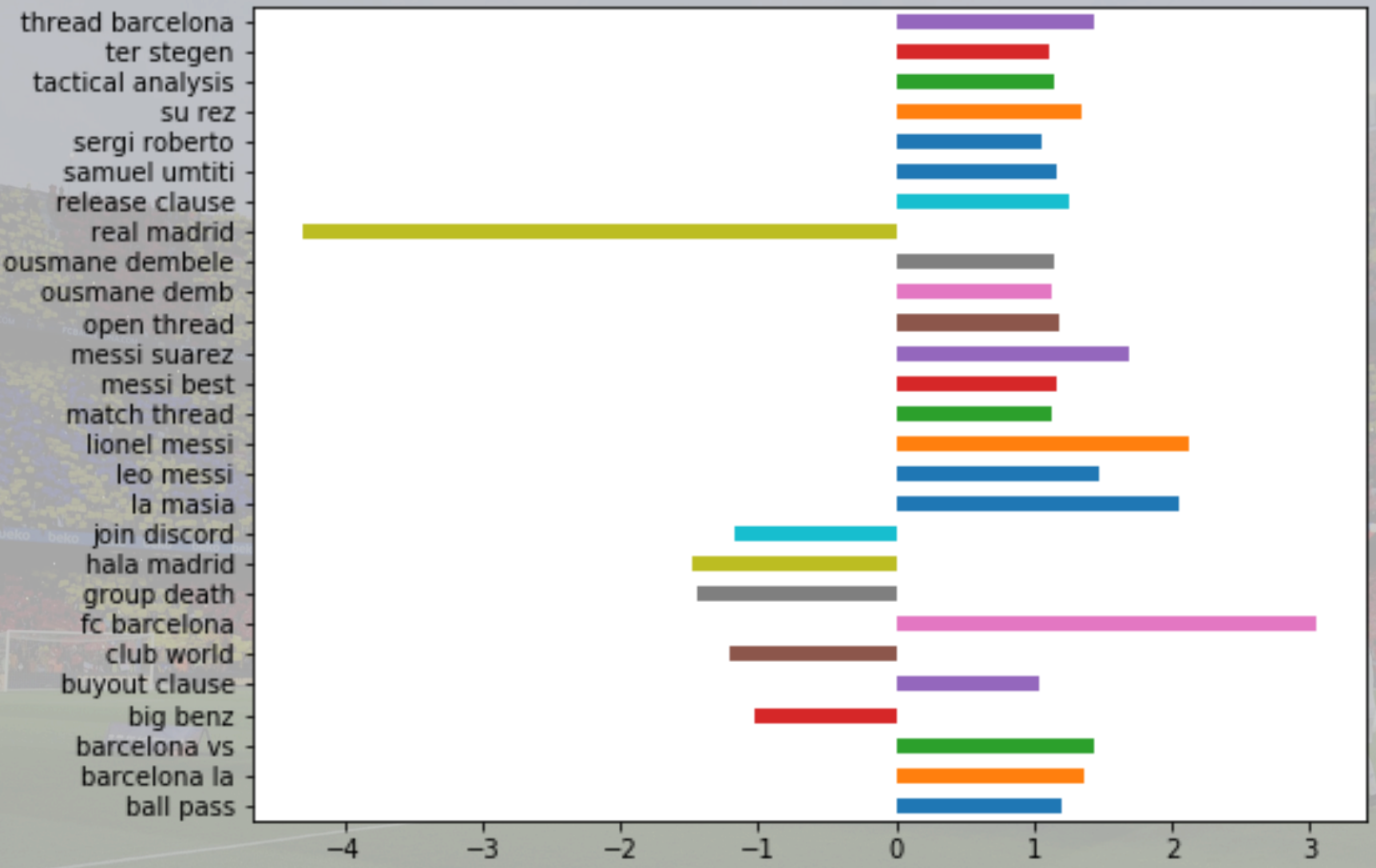
	subreddit	body	created_utc	id	link_id	parent_id
0	Barca	The Tavern on 12th! The F.C. Barcelona Penya, ...	1525573121	dyioab8	t3_8hazib	t3_8hazib
1	Barca	How about Puyol and Raúl?	1525572556	dyints8	t3_8h7que	t1_dyhqxbw
2	Barca	Tap 24 by campus is fun too, just get there ea...	1525572535	dyint5g	t3_8hazib	t3_8hazib
3	Barca	They were a rip off about 1200$ each category 2	1525571525	dyimzkp	t3_8h6e1p	t1_dyimwte
4	Barca	Ah gotcha. How much were the tickets for you i...	1525571433	dyimwte	t3_8h6e1p	t1_dyime71

	subreddit	body	created_utc	id	link_id	parent_id
0	realmadrid	Because I absolutely hate Dutch football. It's...	1525573170	dyiobor	t3_8gue84	t1_dyi9vep
1	realmadrid	It's youtube what did you expect	1525572953	dyio5ir	t3_8h8rxy	t1_dyi1jim
2	realmadrid	I want 5-0 to us, and Keylor Navas scoring a f...	1525572589	dyinuqj	t3_8hafod	t3_8hafod
3	realmadrid	Well,klopp is also not the only reason liverpo...	1525572336	dyinn8x	t3_8h74qb	t1_dyi754j
4	realmadrid	Real Madrid already took the high road when th...	1525572167	dyinibf	t3_8h7hs5	t1_dyhwqi6

Features with Players / Teams



2 grams with Players / Teams

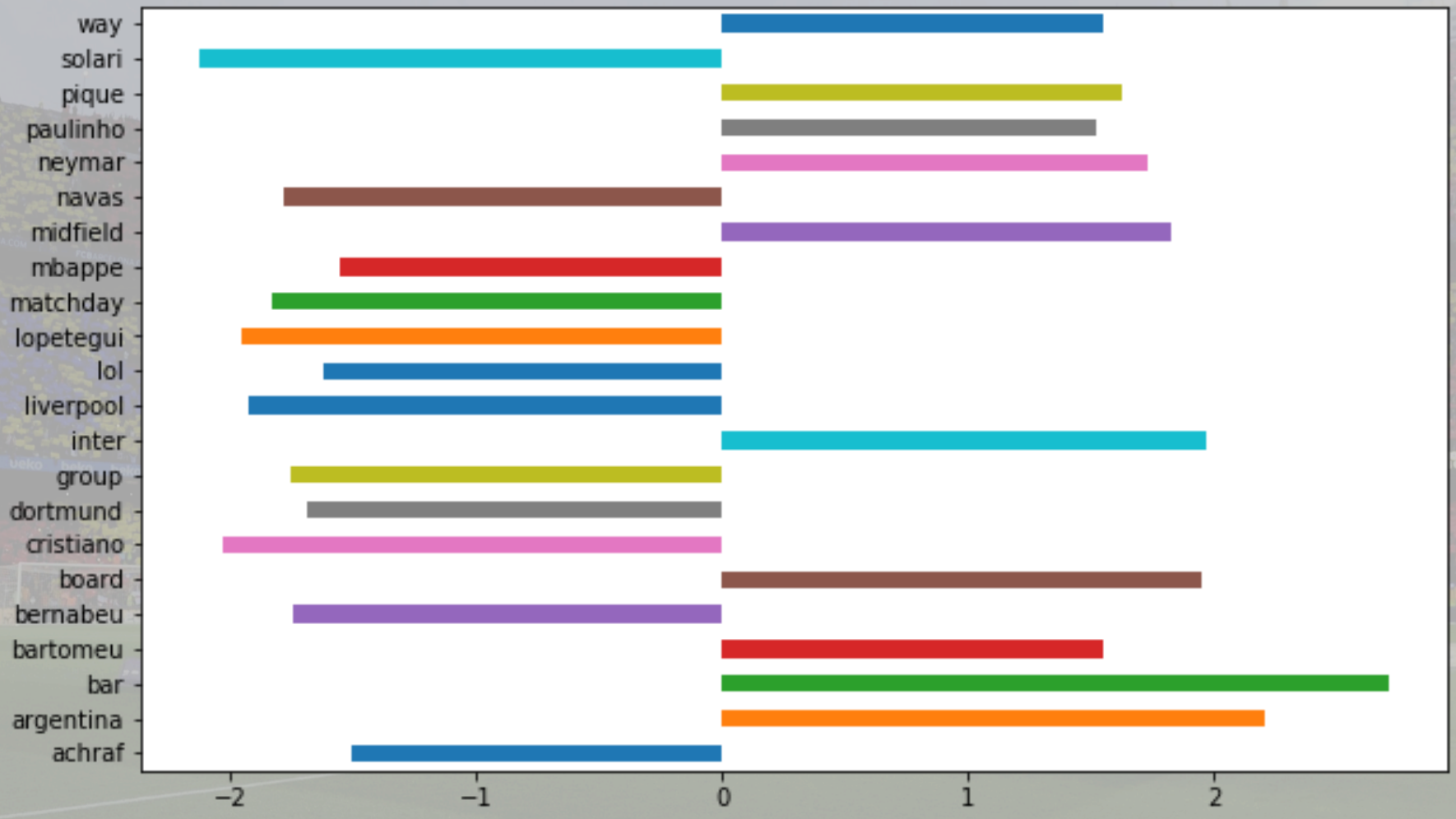


Stop Words:

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	0.69	0.69	0.69	907
1	0.70	0.70	0.70	935
avg / total	0.70	0.70	0.70	1842

	Predicted Barca	Predicted realmadrid
Actual Barca	628	279
Actual realmadrid	281	654

Forest and Trees

- Without or With Player & Team stop words

- RandomForestClassifier
  - Best cv scores:
    - Keeping stop words in model, Train: 0.732 | Test: 0.699
    - Taking stop words out of model, Train: 0.655 |Test: 0.608
- ExtraTreesClassifier
  - Best cv scores:
    - Keeping stop words in model, Train: 0.743 | Test: 0.699
    - Taking stop words out of model, Train: 0.664 |Test: 0.604

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 & Takeaways

- Top feature is clearly players names
- Take the names out, accuracy score drops substantially
- To improve:
  - Gridsearch over more model parameters
  - More ngram features (3,6)
  - Put in coaches as stop words