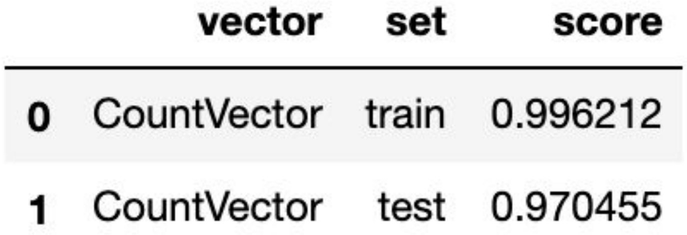
Ball or Bike?

An exploration of NLP

r/motorcyle vs r/soccer

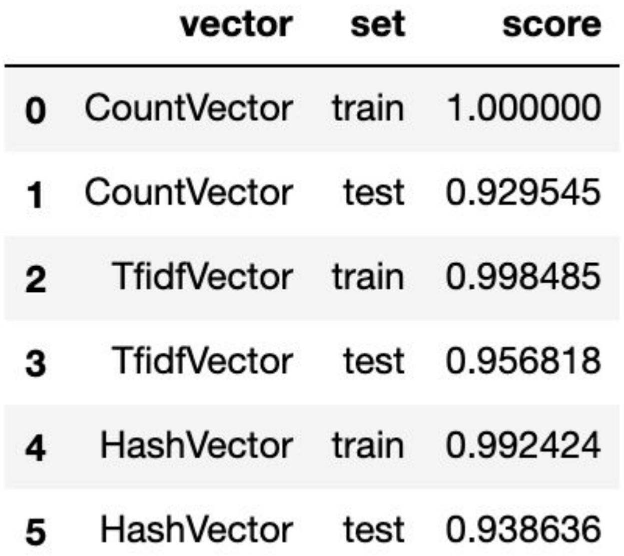
	r/motorcycle	r/soccer
Total Posts	912	848
Vocabulary	2178	3014
Joint	536	

Bayes Naive Multinomial



	vector	set	score
0	CountVector	train	0.996212
1	CountVector	test	0.970455

Logistic Regression



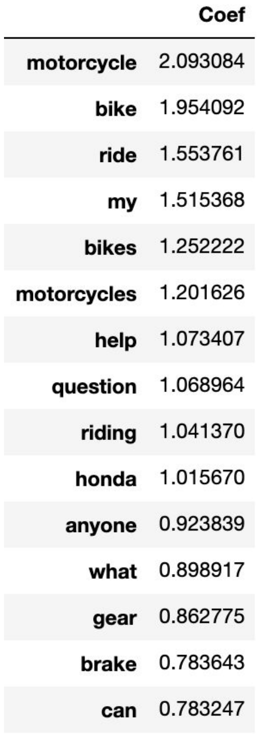
	vector	set	score
0	CountVector	train	1.000000
1	CountVector	test	0.929545
2	TfidfVector	train	0.998485
3	TfidfVector	test	0.956818
4	HashVector	train	0.992424
5	HashVector	test	0.938636

Hello, perfection!

Not a good thing.

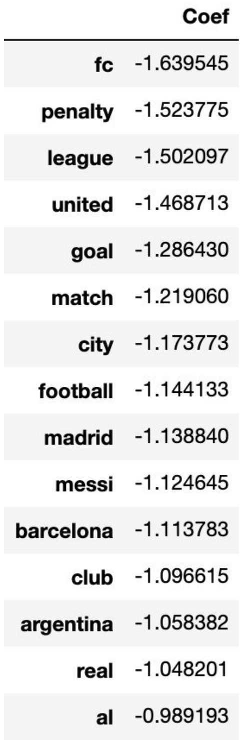
Why so good?

r/motorcycle



	Coef
motorcycle	2.093084
bike	1.954092
ride	1.553761
my	1.515368
bikes	1.252222
motorcycles	1.201626
help	1.073407
question	1.068964
riding	1.041370
honda	1.015670
anyone	0.923839
what	0.898917
gear	0.862775
brake	0.783643
can	0.783247

r/football

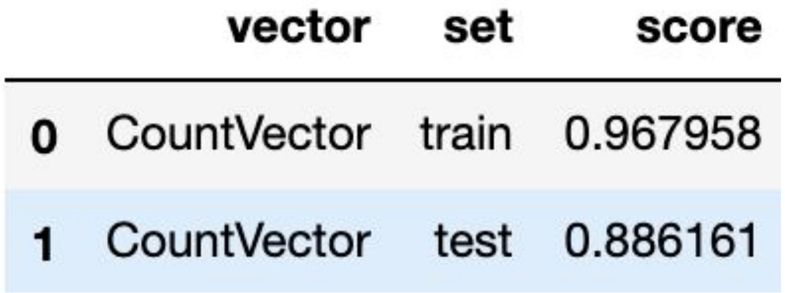


	Coef
fc	-1.639545
penalty	-1.523775
league	-1.502097
united	-1.468713
goal	-1.286430
match	-1.219060
city	-1.173773
football	-1.144133
madrid	-1.138840
messi	-1.124645
barcelona	-1.113783
club	-1.096615
argentina	-1.058382
real	-1.048201
al	-0.989193

r/soccer vs r/MLS

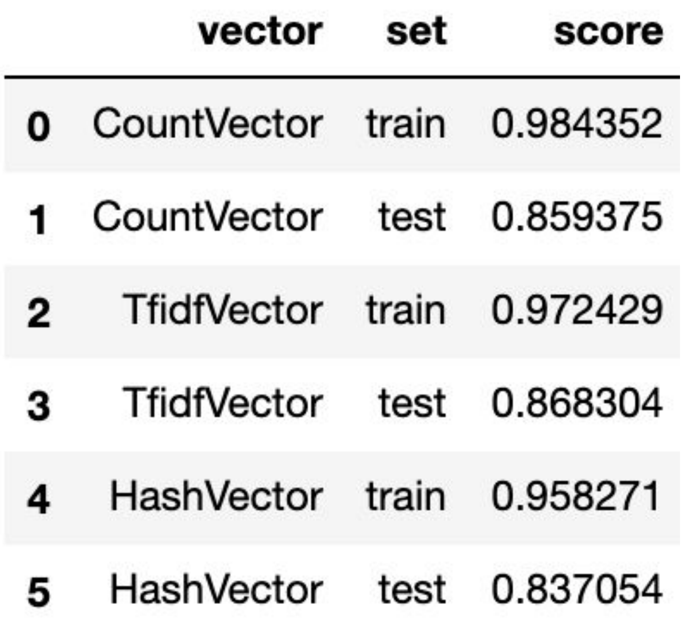
	r/soccer	r/MLS
Total Posts	848	942
Vocabulary	3014	2913
Joint	987	

Bayes Naive Multinomial



	vector	set	score
0	CountVector	train	0.967958
1	CountVector	test	0.886161

Logistic Regression

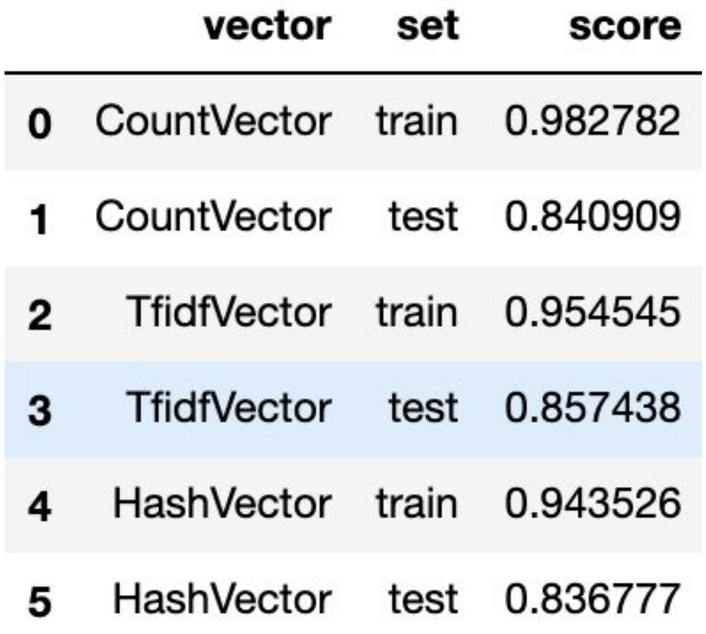


	vector	set	score
0	CountVector	train	0.984352
1	CountVector	test	0.859375
2	TfidfVector	train	0.972429
3	TfidfVector	test	0.868304
4	HashVector	train	0.958271
5	HashVector	test	0.837054

r/MLS vs r/SoundersFC

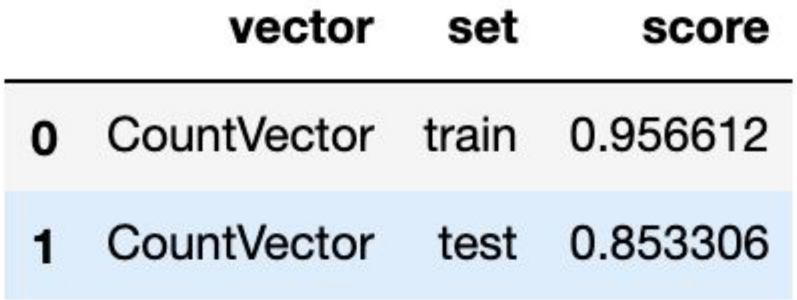
	r/MLS	r/SoundersFC
Total Posts	942	994
Vocabulary	2913	2112
Joint	933	

Logistic Regression



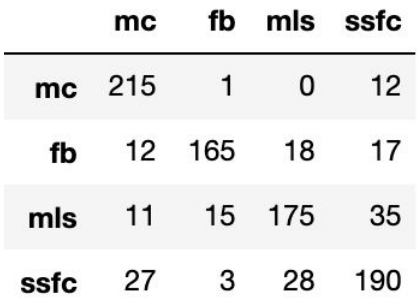
	vector	set	score
0	CountVector	train	0.982782
1	CountVector	test	0.840909
2	TfidfVector	train	0.954545
3	TfidfVector	test	0.857438
4	HashVector	train	0.943526
5	HashVector	test	0.836777

Bayes Naive Multinomial



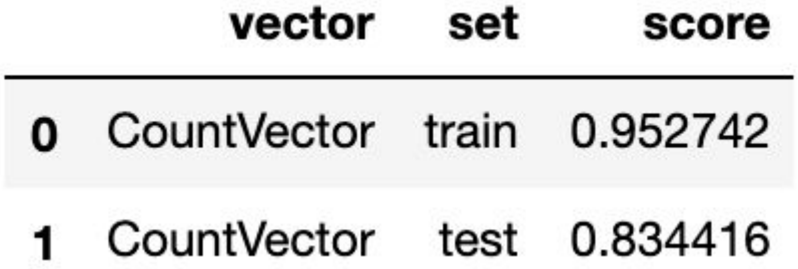
	vector	set	score
0	CountVector	train	0.956612
1	CountVector	test	0.853306

Multiclass Logistic and Bayes Naive



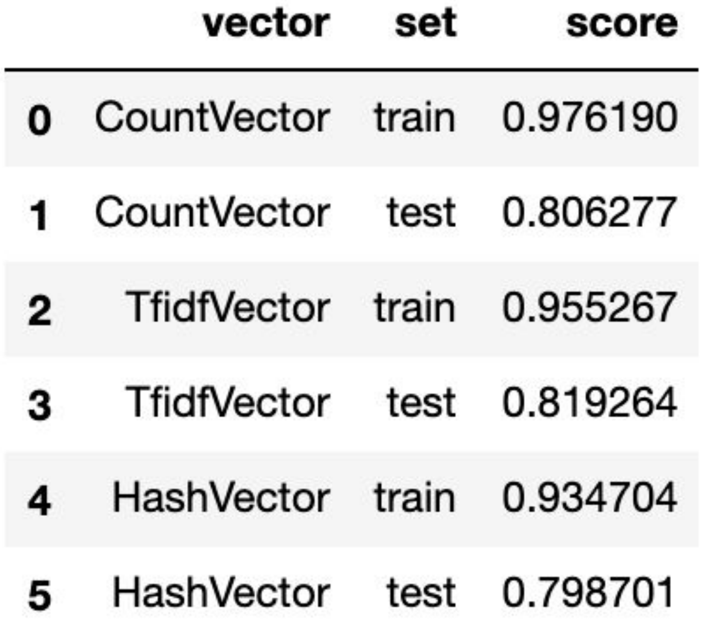
	mc	fb	mls	ssfc
mc	215	1	0	12
fb	12	165	18	17
mls	11	15	175	35
ssfc	27	3	28	190

Bayes Naive Multinomial



	vector	set	score
0	CountVector	train	0.952742
1	CountVector	test	0.834416

Logistic Regression



	vector	set	score
0	CountVector	train	0.976190
1	CountVector	test	0.806277
2	TfidfVector	train	0.955267
3	TfidfVector	test	0.819264
4	HashVector	train	0.934704
5	HashVector	test	0.798701

K-Nearest Neighbors

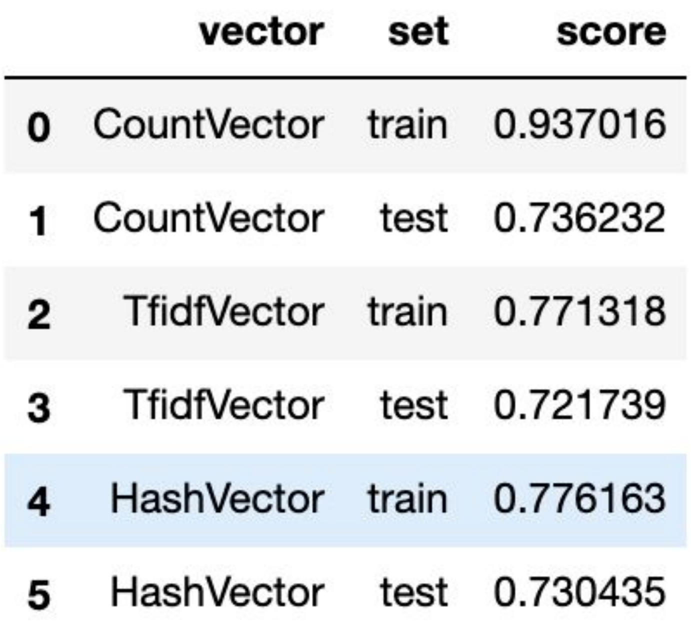
A different tact:

Story time!

r/talesfromretail vs r/talesfromyourserver

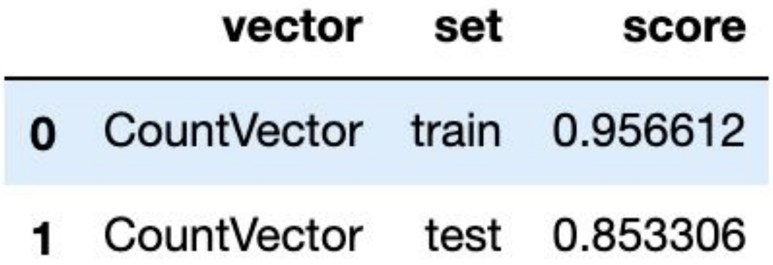
	r/tfr	r/tfys
Total Posts	391	986
Vocabulary	888	1789
Joint	424	

Logistic Regression



	vector	set	score
0	CountVector	train	0.937016
1	CountVector	test	0.736232
2	TfidfVector	train	0.771318
3	TfidfVector	test	0.721739
4	HashVector	train	0.776163
5	HashVector	test	0.730435

Bayes Naive Multinomial

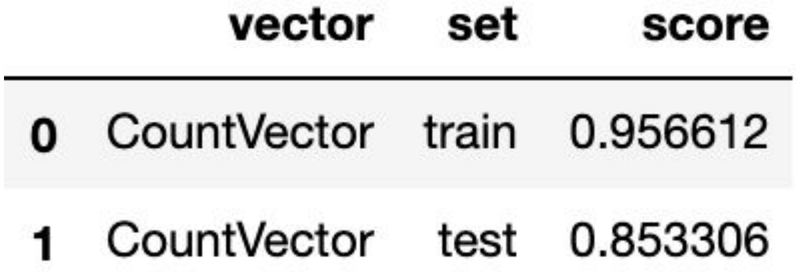


	vector	set	score
0	CountVector	train	0.956612
1	CountVector	test	0.853306

r/talesfromyoursever vs r/bartender

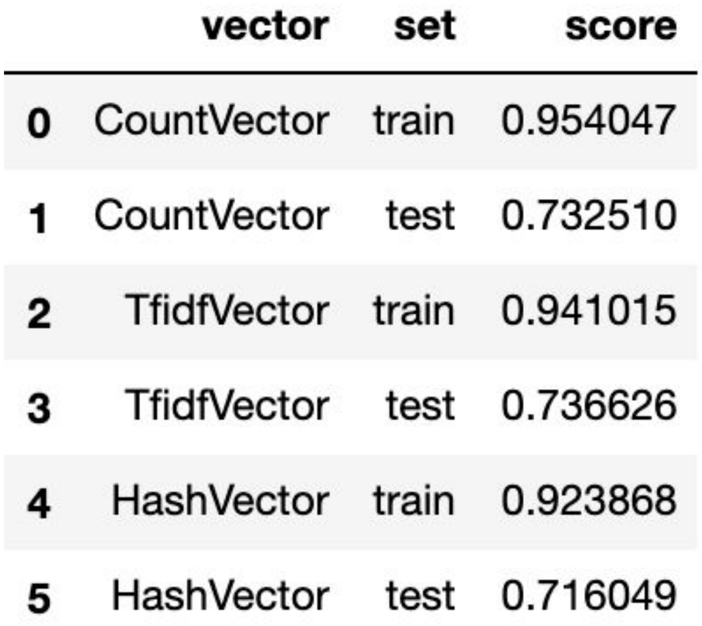
	r/tfys	r/bar
Total Posts	986	958
Vocabulary	1789	2081
Joint	370	

Bayes Naive Multinomial



	vector	set	score
0	CountVector	train	0.956612
1	CountVector	test	0.853306

Logistic Regression

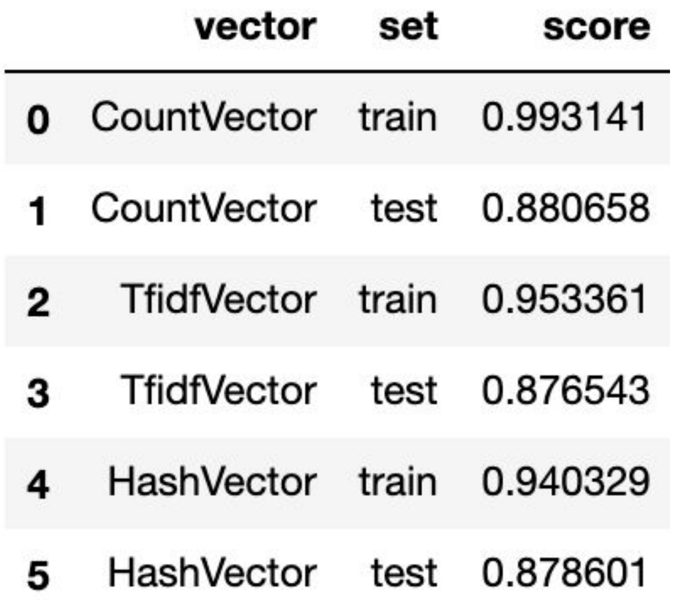


	vector	set	score
0	CountVector	train	0.954047
1	CountVector	test	0.732510
2	TfidfVector	train	0.941015
3	TfidfVector	test	0.736626
4	HashVector	train	0.923868
5	HashVector	test	0.716049

r/tfys vs r/bar title + text

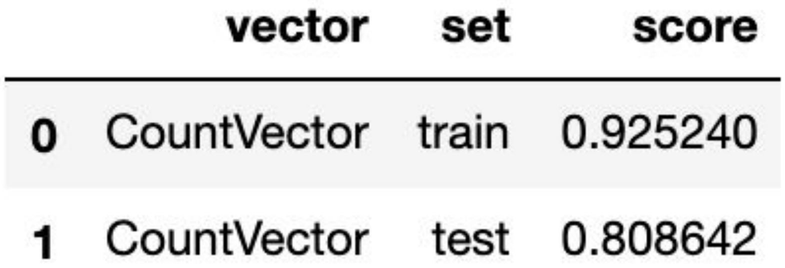
	r/tfys	r/bar
Total Posts	986	958
Vocabulary	1789	6515
Joint	656	

Logistic Regression



	vector	set	score
0	CountVector	train	0.993141
1	CountVector	test	0.880658
2	TfidfVector	train	0.953361
3	TfidfVector	test	0.876543
4	HashVector	train	0.940329
5	HashVector	test	0.878601

Bayes Naive Multinomial



	vector	set	score
0	CountVector	train	0.925240
1	CountVector	test	0.808642

GridSearch

One Model, two results

r/tfys + r/bar, text and title

```
params={
    'cvec__stop_words':    ['english'],
    'cvec__max_features': [2000],
    'cvec__ngram_range': [(1,1)],
    'cvec__max_df':           [.5],
    'Lr__penalty':               ['l2'],
    'Lr__solver':                  ['lbfgs'],
    'Lr__max_iter':              [10]
}
```

Train: .8882

Test: .9012

r/motorcycle, r/soccer

```
params={
    'cvec__stop_words':    [None],
    'cvec__max_features': [3000],
    'cvec__ngram_range': [(1,1)],
    'cvec__max_df':           [0.06]
    'Lr__penalty':               ['l2'],
    'Lr__solver':                  ['sag'],
    'Lr__max_iter':              [100]
}
```

Train: .9318

Test: .9318

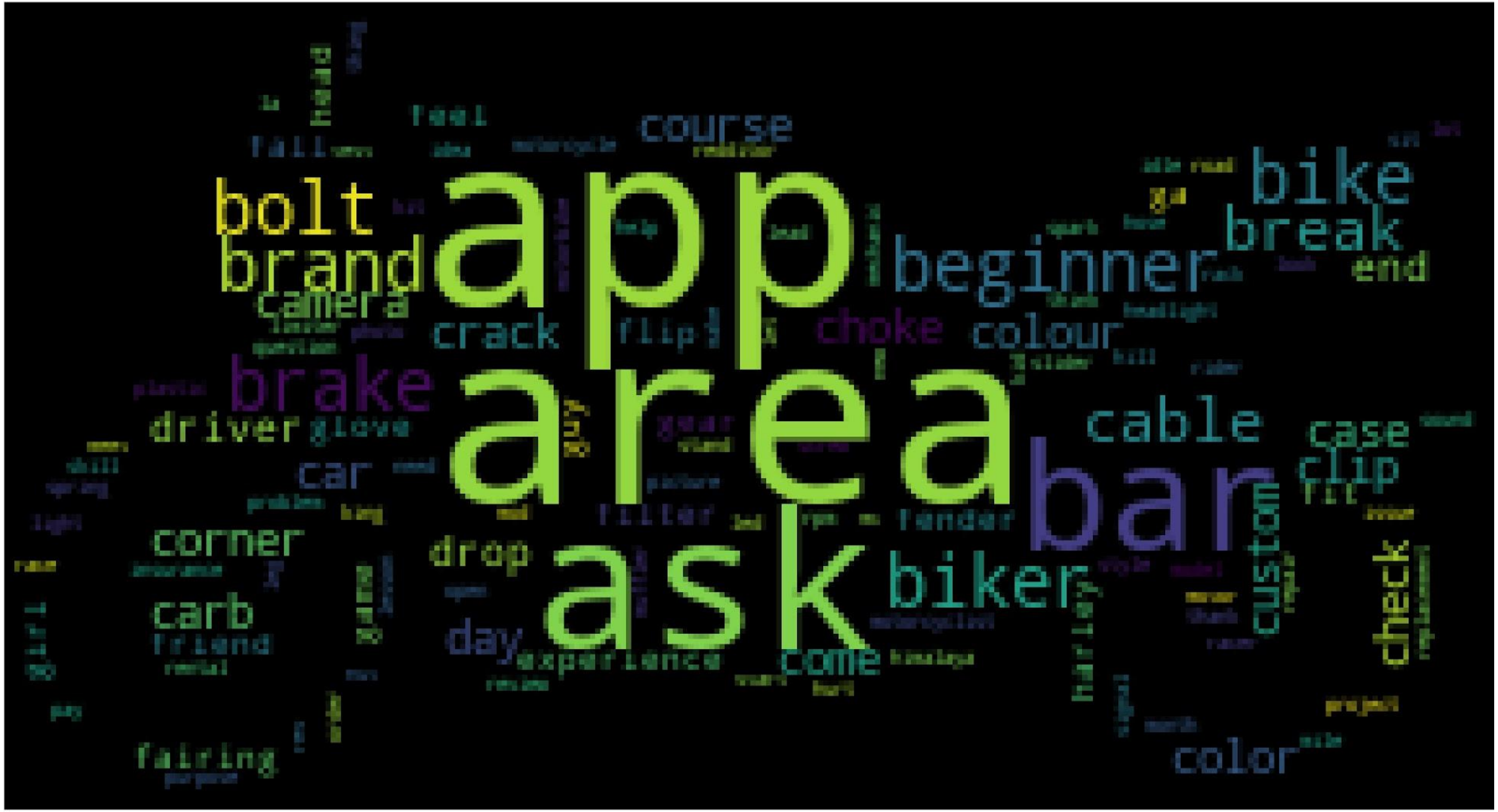
Observations

1. Initial high levels of overfitting
2. Better Test performance on tuned parameters
3. Affect of shared words list size
4. Models are data source specific

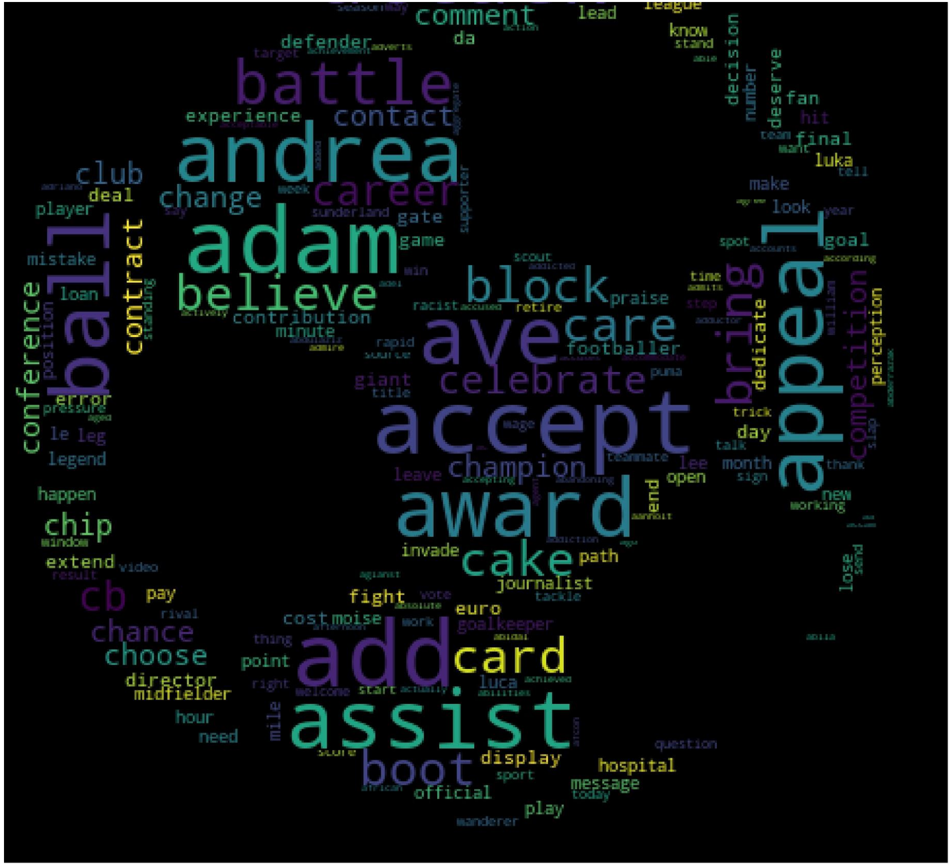
Lessons Learned

1. Better data checking (duplicates)
2. Different models
3. Subject matter didn’t affect as much as anticipated
4. NLP REALLY overfits
5. Multi-class vs. boolean didn’t matter
6. Better understand metrics and what they are saying

I see a motorcycle shaped cloud!



Ball shaped cloud?



The best possible thing….