MAN UTD VS MAN CITY

(Subreddit edition)

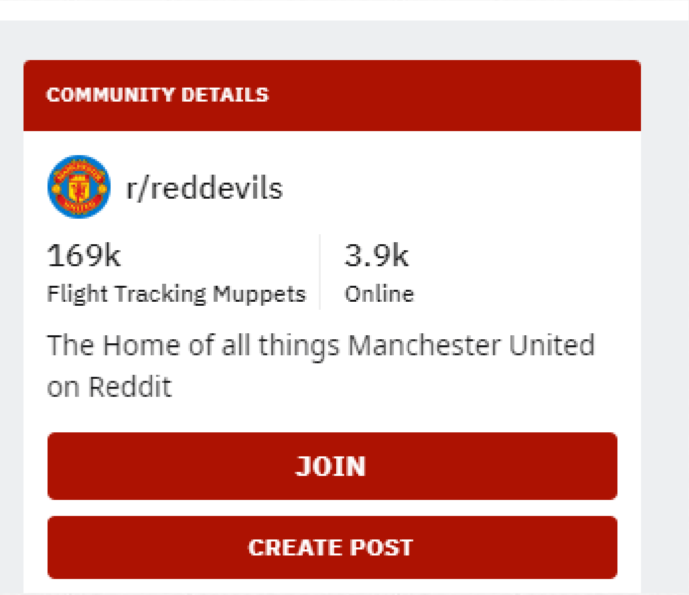
By Adefolafemi Adenugba

Agenda

- Problem Statement
- Data collection Process
- Models used
- Conclusions and final thoughts

Problem Statement

- Are we able to predict the location of a reddit post to either r/MCFC (Manchester City's subreddit) or r/reddevils (Manchester United's subreddit)?

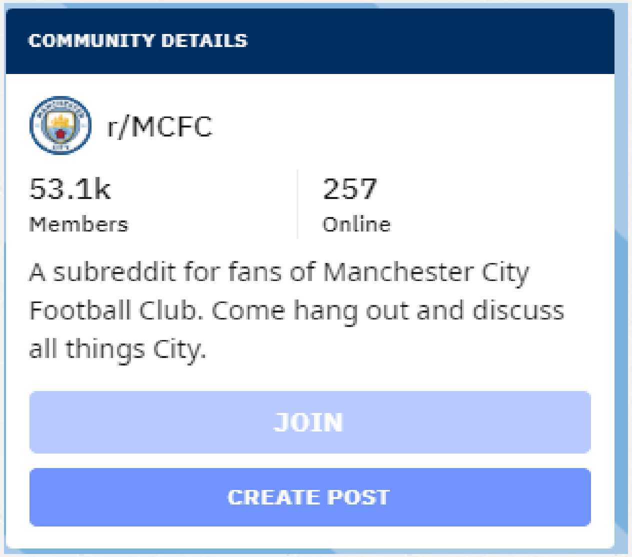


COMMUNITY DETAILS

r/reddevils

169k Flight Tracking Muppets | 3.9k Online

The Home of all things Manchester United on Reddit



COMMUNITY DETAILS

r/MCFC

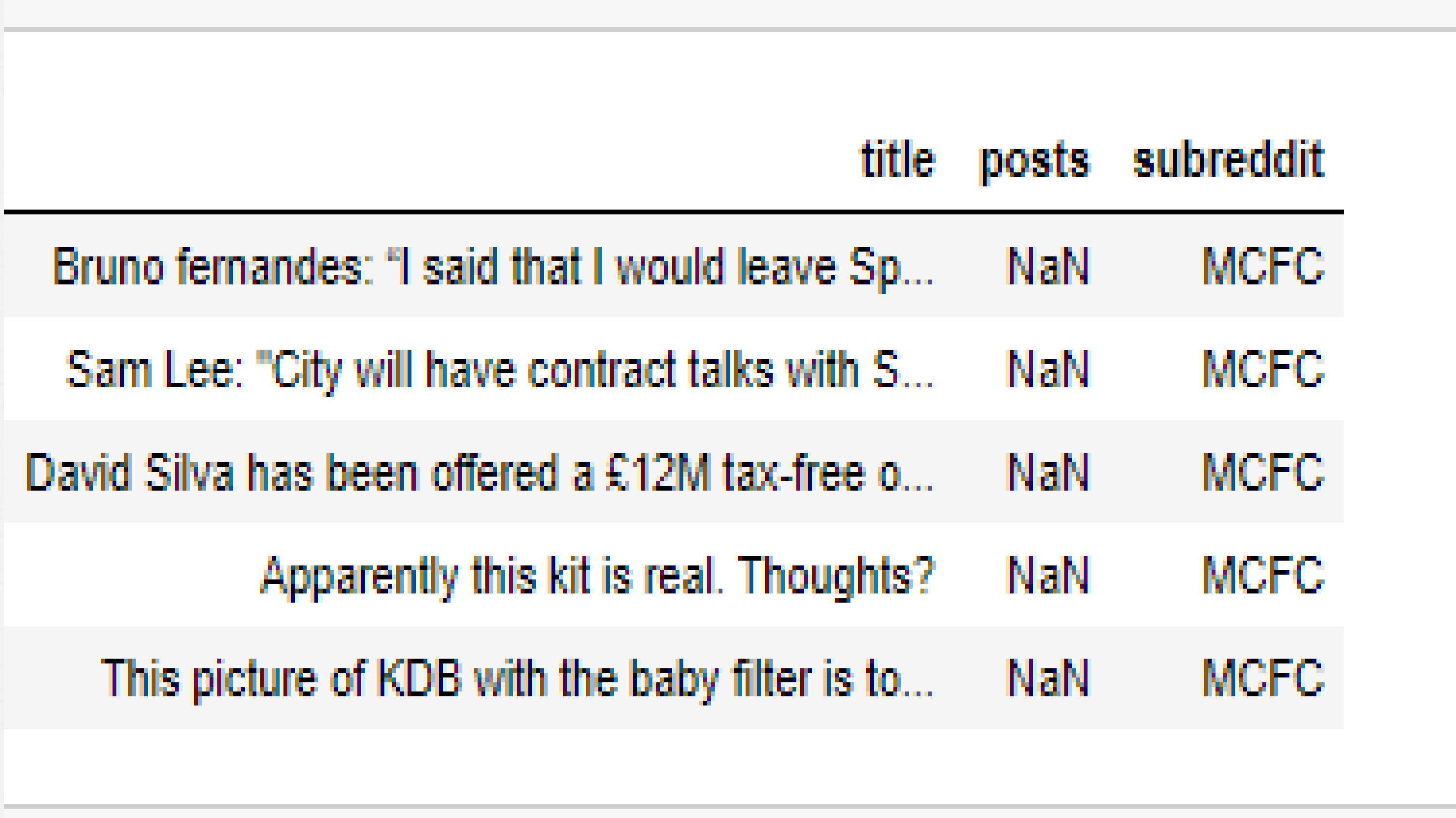
53.1k Members | 257 Online

A subreddit for fans of Manchester City Football Club. Come hang out and discuss all things City.

Two teams, One City

- Man Utd – Known as the biggest team in England and undisputed top 3 in global soccer
- Man City – Owned by Sheikh Mansour

MCFC Dataframe example

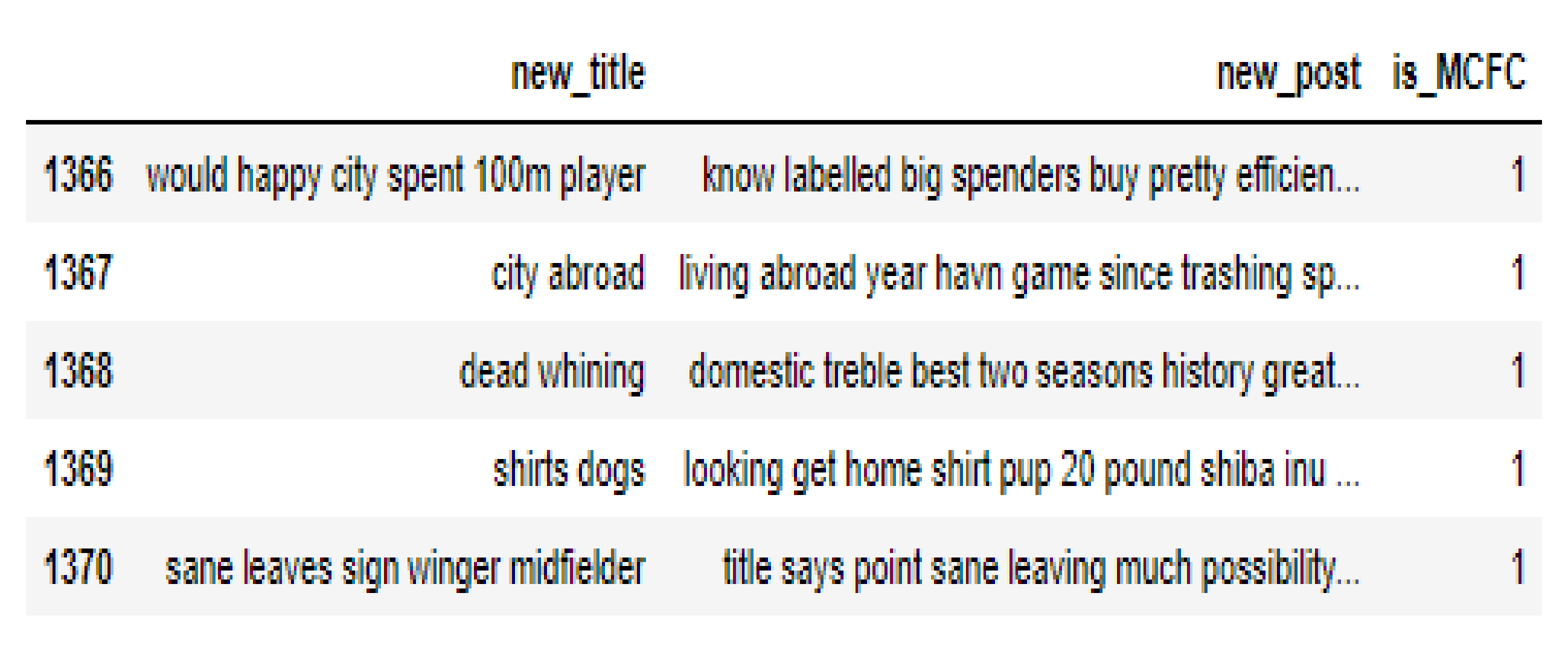


title	posts	subreddit
Bruno fernandes: "I said that I would leave Sp...	NaN	MCFC
Sam Lee: "City will have contract talks with S...	NaN	MCFC
David Silva has been offered a £12M tax-free o...	NaN	MCFC
Apparently this kit is real. Thoughts?	NaN	MCFC
This picture of KDB with the baby filter is to...	NaN	MCFC

Webscraping and Data Cleaning

- An initial 4949 posts → 1371 posts
- Created two separate dataframes for the subreddits and combined them
- Cleaned the dataframes for special characters with regex
- Lower-cased all words

Final cleaned Dataframe



	new_title	new_post	is_MCFC
1366	would happy city spent 100m player	know labelled big spenders buy pretty efficien...	1
1367	city abroad	living abroad year havn game since trashing sp...	1
1368	dead whining	domestic treble best two seasons history great...	1
1369	shirts dogs	looking get home shirt pup 20 pound shiba inu ...	1
1370	sane leaves sign winger midfielder	title says point sane leaving much possibility...	1

Models used

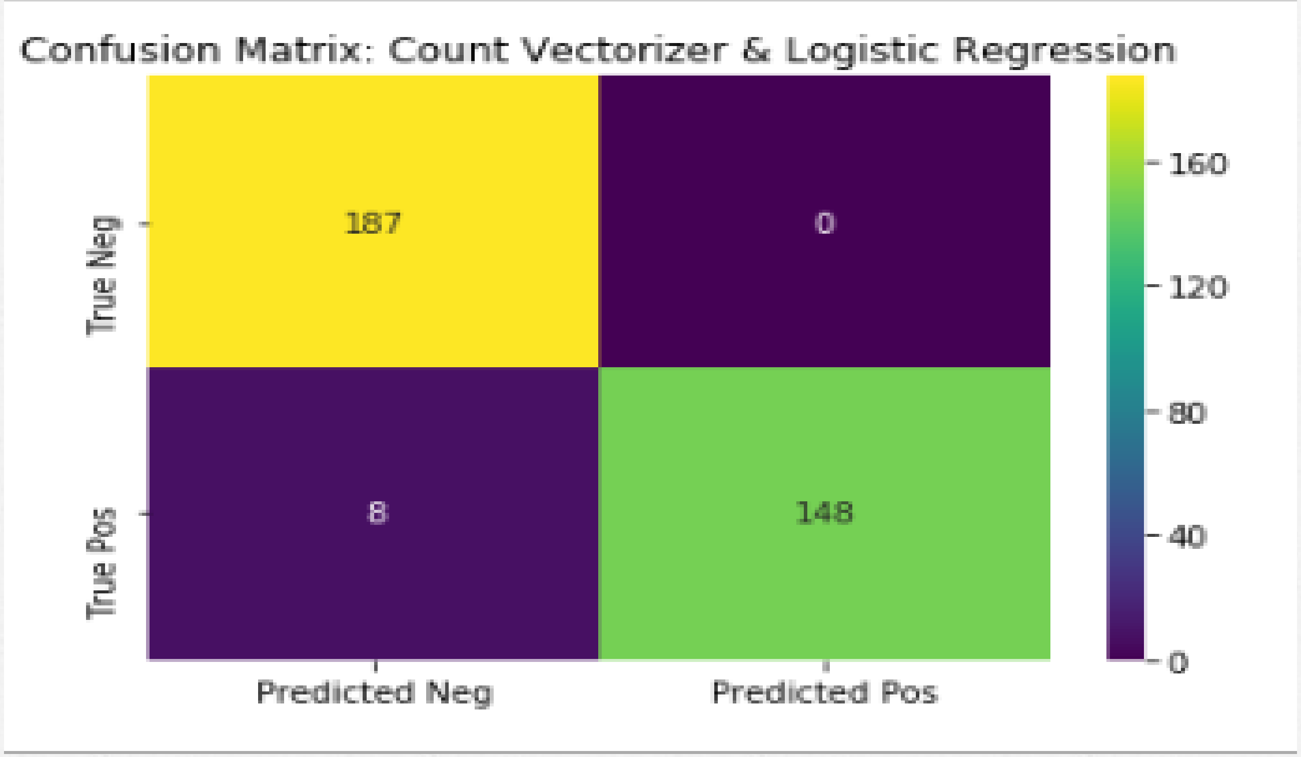
- Logistic Regression
- Naive Bayes

- Baseline: 54.75% accuracy score

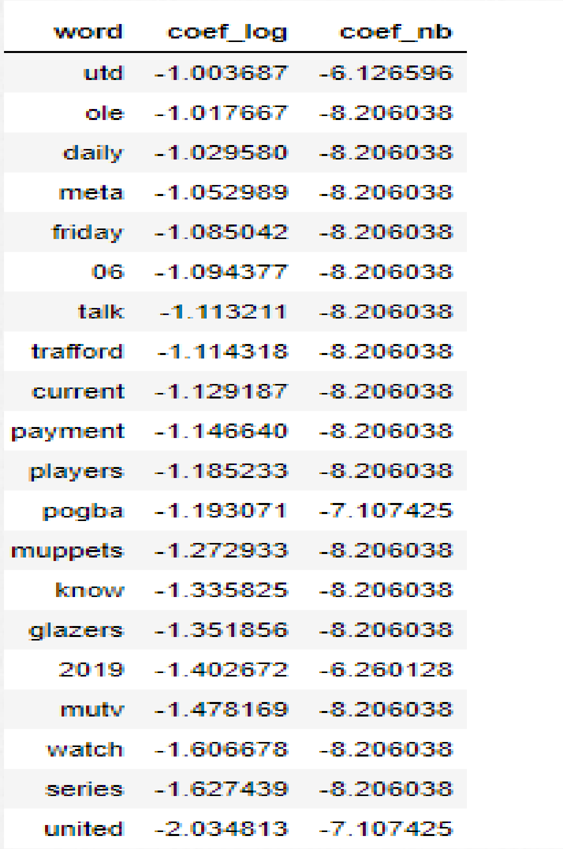
Results

Models	CountVectorizer	TF-IDF
Logistic Regression	Training: 99.71%	Training: 99.22%
	Test: 97.67%	Test: 96.50%
Naïve Bayes	Training: 97.56%	Training: 98.34%
	Test: 94.17%	Test: 95.04%

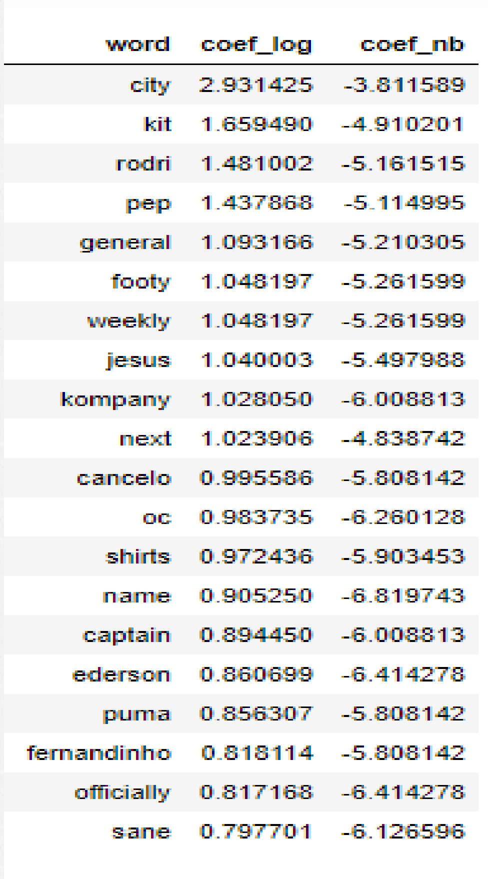
Best model


Confusion Matrix: Count Vectorizer & Logistic Regression
True Neg
True Pos
187
0
8
148
Predicted Neg
Predicted Pos
160
120
80
40
0

Familiar words



word	coef_log	coef_nb
utd	-1.003687	-6.126596
ole	-1.017667	-8.206038
daily	-1.029580	-8.206038
meta	-1.052989	-8.206038
friday	-1.085042	-8.206038
06	-1.094377	-8.206038
talk	-1.113211	-8.206038
trafford	-1.114318	-8.206038
current	-1.129187	-8.206038
payment	-1.146640	-8.206038
players	-1.185233	-8.206038
pogba	-1.193071	-7.107425
muppets	-1.272933	-8.206038
know	-1.335825	-8.206038
glazers	-1.351856	-8.206038
2019	-1.402672	-6.260128
mutv	-1.478169	-8.206038
watch	-1.606678	-8.206038
series	-1.627439	-8.206038
united	-2.034813	-7.107425



word	coef_log	coef_nb
city	2.931425	-3.811589
kit	1.659490	-4.910201
rodri	1.481002	-5.161515
pep	1.437868	-5.114995
general	1.093166	-5.210305
footy	1.048197	-5.261599
weekly	1.048197	-5.261599
jesus	1.040003	-5.497988
kompany	1.028050	-6.008813
next	1.023906	-4.838742
cancelo	0.995586	-5.808142
oc	0.983735	-6.260128
shirts	0.972436	-5.903453
name	0.905250	-6.819743
captain	0.894450	-6.008813
ederson	0.860699	-6.414278
puma	0.856307	-5.808142
fernandinho	0.818114	-5.808142
officially	0.817168	-6.414278
sane	0.797701	-6.126596

Further Analysis

- Future comparison within the subreddits
- Test out other models
- Looking at other columns
- Removing player names on the roster