Social Sentiment Analysis of User Tweets for Arsenal and Tottenham games

Presented for Springboard

By Benjamin Bellman

Outline

- **I. Introduction**
- **II. Data Wrangling and Exploration**
- **III. Exploratory Data Analysis (EDA)**
- **IV. Pre-Processing and Modelling**
- **V. Business Recommendations**

Introduction

Objective:

- Conducting Social Sentiment Analysis of two Premier League Teams.
- Analyze user tweets.
- Build model that predicts team results from tweets.
- Checking which words are most associated with wins or losses.
- Inform Social Media teams of these findings to increase engagement.

Data Wrangling and Exploration

- Collecting Tweets: Access the Twitter API and Twitter Scraper
- Collecting Game Results Data: Access www.football-data.co.uk
- Building Twitter Scraper for multiple Queries
- Query **1000** results for each game.

Introduction

About the Data:

- **370,000** Tweets

- **2** teams: Arsenal & Tottenham.
- **38** games per team per season (except most recent season).
- **1000** tweets per game.
- **5** Seasons: From *2017-2018* to *2021-2022*.

Introduction

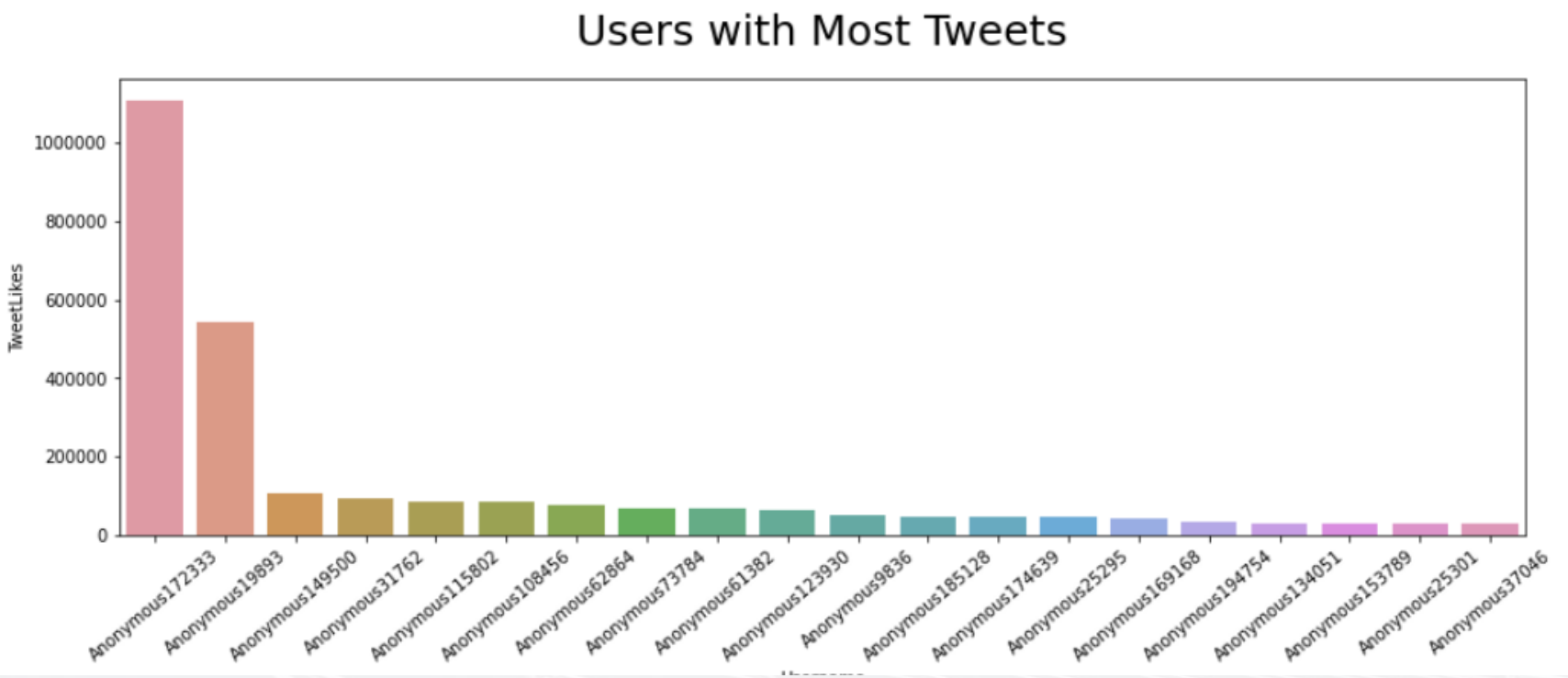
Preview:

	Query2	Date	Username	Tweet	TweetLikes	TweetReplies	RetweetCount	Result	Team
0	Arsenal until:2022-04-24	2022-04-23 23:59:47+00:00	Anonymous19203	@JackAFC01 @LUHG450 @1Thegameis Because you're arsenal and you have no self awareness..in our worst season in years and your best season in years and you only 6 points ahead of us and you think the gap is that big lol 🤣 be real now	1	1	0	1	Arsenal
1	Arsenal until:2022-04-24	2022-04-23 23:59:41+00:00	Anonymous123302	@arsenal_lady bei ihm werde ich einfach immer schwach	1	0	0	1	Arsenal
2	Arsenal until:2022-04-24	2022-04-23 23:59:39+00:00	Anonymous134105	5 games to go\n\n5 cup finals \n\n5 games to UCL or UEL, either way I want my European trips back.\n\nWe can do it @Arsenal	0	0	0	1	Arsenal
3	Arsenal until:2022-04-24	2022-04-23 23:59:37+00:00	Anonymous112922	@Arsenal @HectorBellerin VAMOS @HectorBellerin ! Even if you stay in Spain, you'll always be loved in North London ;-)	18	0	0	1	Arsenal
4	Arsenal until:2022-04-24	2022-04-23 23:59:32+00:00	Anonymous65885	@Cristiano Come to @Arsenal 🦒.. so many assists and crosses with no one to finish/ tap in.	0	0	0	1	Arsenal

	Query2	Date	Username	Tweet	TweetLikes	TweetReplies	RetweetCount	Result	Team	CleanTweet
0	Arsenal until:2022-04-24	2022-04-23 23:59:47+00:00	Anonymous19203	@JackAFC01 @LUHG450 @1Thegameis Because you're arsenal and you have no self awareness..in our worst season in years and your best season in years and you only 6 points ahead of us and you think the gap is that big lol 🤣 be real now	1	1	0	1	Arsenal	Because you re arsenal and you have no self awareness in our worst season in years and your best season in years and you only points ahead of us and you think the gap is that big lol be real now
1	Arsenal until:2022-04-24	2022-04-23 23:59:41+00:00	Anonymous123302	@arsenal_lady bei ihm werde ich einfach immer schwach	1	0	0	1	Arsenal	bei ihm werde ich einfach immer schwach
2	Arsenal until:2022-04-24	2022-04-23 23:59:39+00:00	Anonymous134105	5 games to go\n\n5 cup finals \n\n5 games to UCL or UEL, either way I want my European trips back.\n\nWe can do it @Arsenal	0	0	0	1	Arsenal	games to go cup finals games to UCL or UEL either way I want my European trips back We can do it

Exploratory Data Analysis (EDA)

- 2 users had a much larger fanbase for Tweet Likes



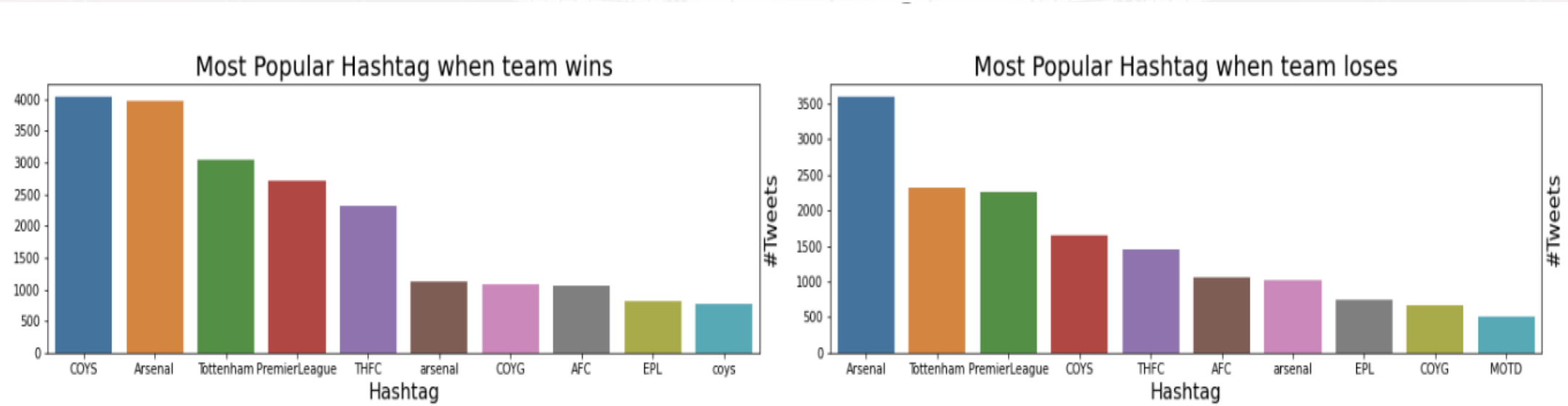
Exploratory Data Analysis (EDA)

- Top 10 tweet by likes all have positive language.

	Tweet	TweetLikes
0	Great start of the year. Let's keep improving together #alltogether #arsenal 2-0 #arsenalvsmanutd https://t.co/PWvqTSyMMm	85832
1	That winnin' feeling 🙌🏻😃 #letskeepthisgoing #YaGunnersYa #M1Ö @arsenal https://t.co/spMq7dx0BR	65144
2	Pochettino was dismissed from Tottenham and won two trophies in his first season with PSG ⏳ https://t.co/O4cj0xHb7Q	40817
3	Incredible. Inconceivable.\n\nWe have no words for this. https://t.co/eQAUaKleRi	36590
4	Great effort from the boys, three points and we keep going! \n#W12 #premierleague #arsenal https://t.co/Ip1jRpsiqG	34758
5	Who's ur ride or die for Super Bowl Sunday Football???\n\n21 Savage: Arsenal \n\nICE: https://t.co/qbvp5ZP8ZW	31388
6	Arsenal using my theme song. The only thing missing, due to COVID, is 30,000 fans chanting "You Suck". #YouSuckCovid #itstrue https://t.co/VQncpBsoin	30610
7	This guy... \n\n😘😘😘 https://t.co/z9e8hfb7oS	30353
8	Unreal, @dele_official! 🔥 \n\n#THFC ⚪ #COYS https://t.co/VWcXKZgpeq	29402
9	Copa del Rey: **Champion** ⏳ \n\nCongratulations, @HectorBellerin ✊ \n\n#CopaDelRey | #BetisAlé https://t.co/1QOHlYSxVu	29200
10	Just gonna leave this here 😍\n\n👋 Have a good night, Gooners! https://t.co/ToFNt09owN	28758

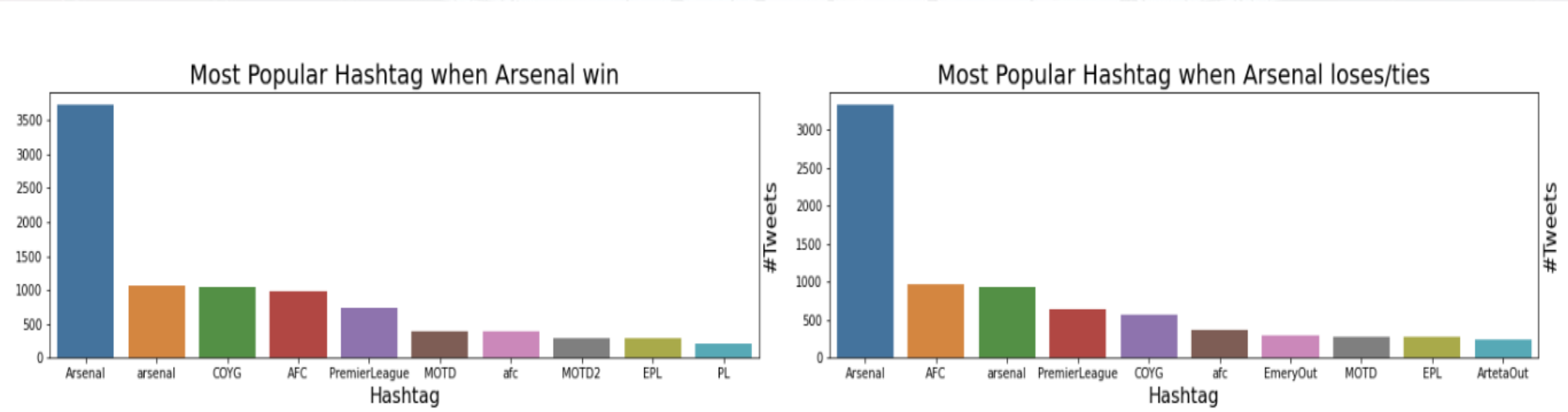
Exploratory Data Analysis (EDA)

- Most Popular Hashtags for both teams



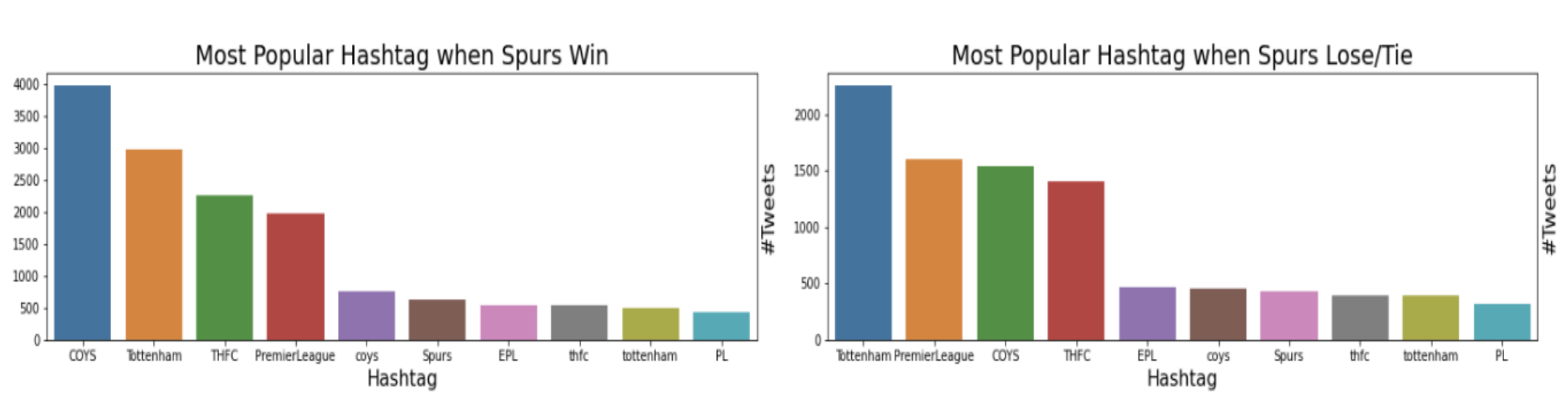
Exploratory Data Analysis (EDA)

- Most Popular Hashtags for Arsenal.



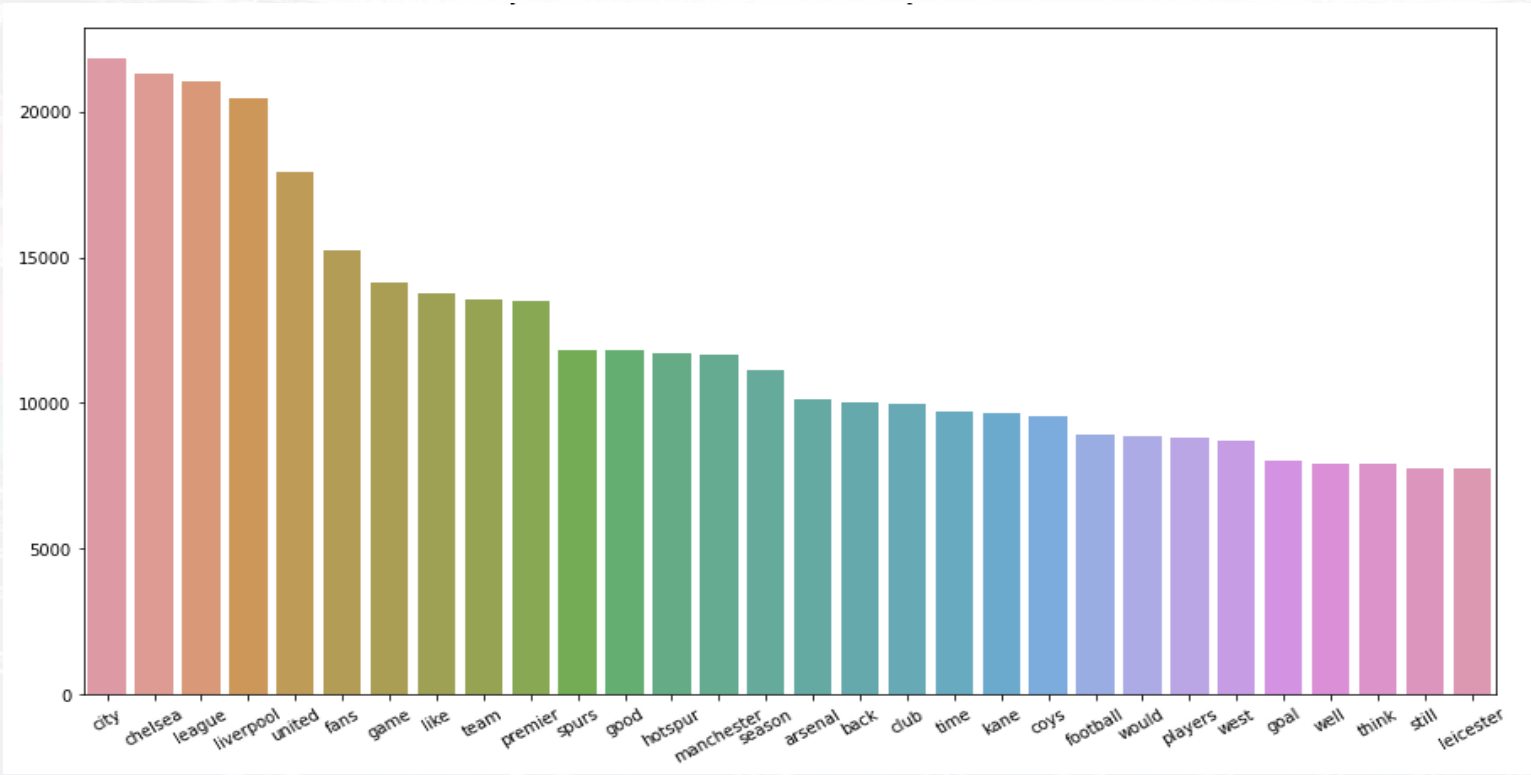
Exploratory Data Analysis (EDA)

- Most Popular Hashtags for Tottenham



Exploratory Data Analysis (EDA)

- Most Popular Non-Stop-Words in Tweets



Pre-Processing and Modelling

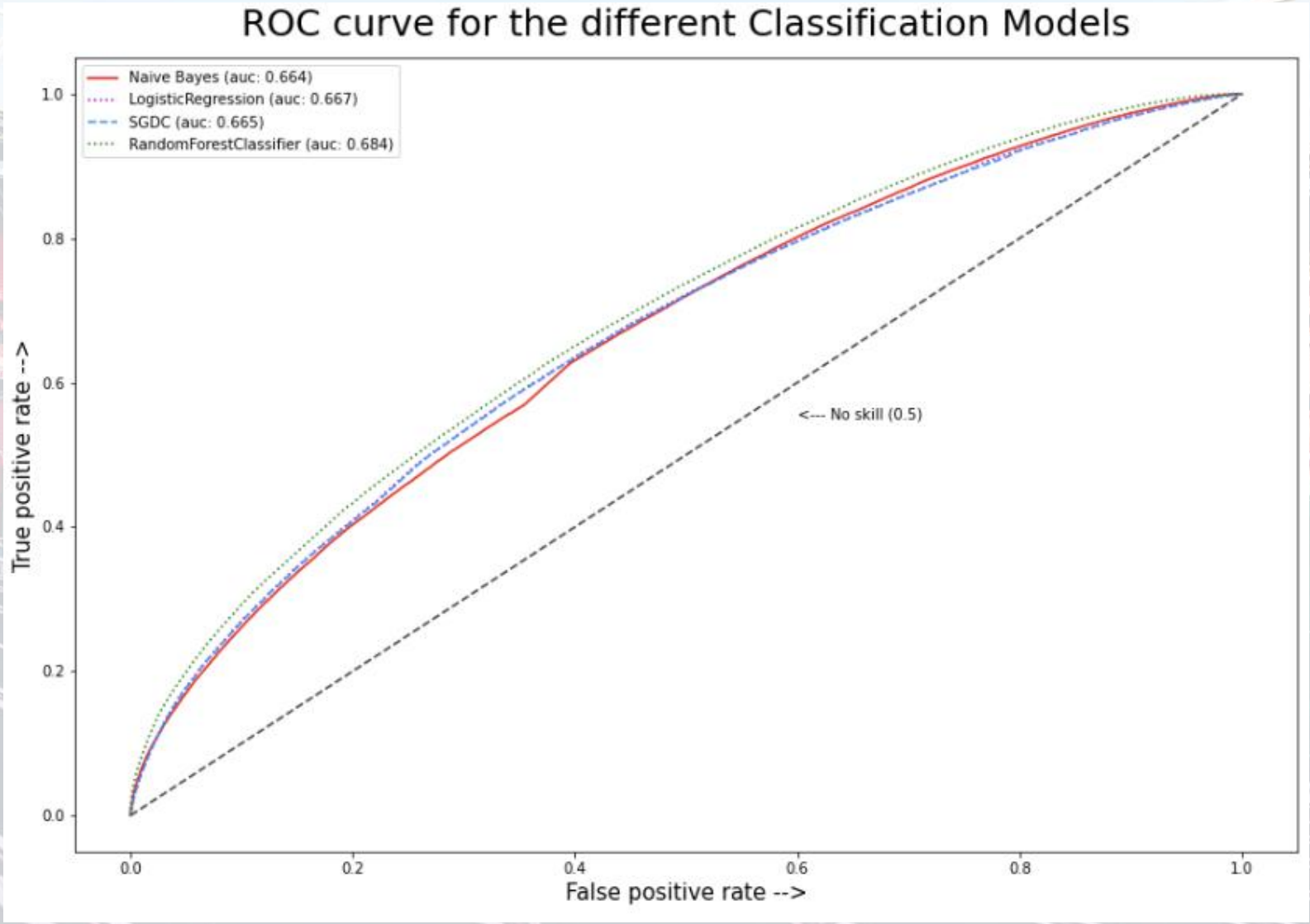
Preprocessing:

- Clean Tweets by removing Usernames and punctuation.
- Remove Stopwords
- Apply TfidVectorizer with *min_df* of 30.
- Use of Grid Search CV

Modelling – Compared 4 models:

1) **Naïve Bayes**
2) **Logistic Regression**
3) **SGDC**
4) **Random Forest Classifier**

Modelling Results – ROC/AUC Curve



Modelling Results – Summary

Model	Best Hyperparameters	ROC_AUC	F1_Score (weighted)	Precision	Recall	Accuracy
Logistic Regression	{'C': 0.1, 'penalty': 'l2'} TF-IDF-Vectorizer {min_df=30}	0.667	0.642	0.614	0.673	0.617
Random Forest Classifier	{'n_estimators': 200, 'min_samples_leaf': 1, 'max_depth': 100, 'bootstrap': True} TF-IDF-Vectorizer {min_df=30}	**0.684**	**0.651**	**0.622**	**0.683**	**0.626**
Naïve Bayes	{'alpha':1} TF-IDF-Vectorizer {min_df=30}	0.664	0.638	0.613	0.664	0.615
SGDC	{'alpha: 0.0001, 'l1_ratio': 0.2 , loss='log} TF-IDF-Vectorizer {min_df=30}	0.665	0.641	0.614	0.67	0.617
No Skill Classifier	NA	0.5	0.51	0.51	0.51	0.51

Model Findings

- Probability that a team won when the word is included:

Good words	P(Good | word)
huddersfield	0.80
contract	0.78
admin	0.77
saka	0.76
cardiff	0.76
coys	0.75
great	0.73
nice	0.73
trophies	0.72
lucas	0.72
harry	0.71
superb	0.71
happy	0.71
love	0.71
brilliant	0.71
stay	0.70
son	0.69
performance	0.69
kane	0.69
clean	0.68
toby	0.68
juan	0.68
sonny	0.68
alli	0.68
goleada	0.68
victory	0.68
newcastle	0.67
gio	0.66
aston	0.66
trophy	0.66

Model Findings

- Probability that a team lost when the word is included:

Bad words	P(Good | word)
emeryout	0.23
var	0.26
draw	0.27
imbang	0.27
lost	0.31
kalah	0.31
sack	0.31
point	0.33
gea	0.33
gomes	0.33
empate	0.35
worse	0.35
blame	0.35
emery	0.37
shit	0.38
coach	0.38
fuck	0.38
poch	0.39
poor	0.39
fault	0.40
lose	0.40
liverpool	0.40
artetaout	0.40
manchester	0.41
klopp	0.41
players	0.41
shambles	0.42
arteta	0.42
loss	0.42
defeat	0.42

Business Recommendations

1) Charge Premium for brands seeking to advertise and get positive coverage.

2) Engage with Indonesian, Pakistani and Spanish speaking audiences.

3) Charge Premiums to advertise posts with Spurs players.

4) For Spurs: Use the COYS tag only when the team wins.

5) For Arsenal: Avoid posting content about manager on games that are losses.

Future Work

Better Data Collection:

- Collect More than 1000 Tweets per game
- Improve Data Cleansing, adding more stop words
- Collect tweets within 3-hour interval of game start and end.
- Collect Tweets for different teams to compare results.

New Models:

- XgBoost, AdaBoost, Gradient Boosting

Further Divide Between Both Teams:

- Test out best models for each team separately.

Use Lang detect to remove non-English tweets.

- Improve model results**.**

Separating Draws from Losses

- See if this further improves the model results.


Thank you!