Social Sentiment Analysis of Arsenal and Tottenham Tweets Write Up

By Benjamin N. Bellman

Table of Contents:

I. Introduction

Arsenal Football club and Tottenham Hotspur are both football teams from North London that compete in England's most competitive top-tier division, the Barclays Premier League. Both teams are rivals, and the infamous North London Derby is attended by many fans from all over the world. Many fans watch the games and are vocal on social media about their team's performances. The objective of this project will be conducting a social sentiment analysis of Arsenal and Tottenham. In this analysis, we will use the data science methodology to give recommendations to both the Arsenal and Tottenham social media teams about what to post on social media by analyzing the tweets that have a mention of *"Arsenal"* or *"Tottenham"* in them. We will use Natural Language Processing (NLP) to build a classification model which will be able to predict whether the team won a game or not. In soccer, ties can happen but since we are analyzing two big clubs, we want to focus on the wins as a team is expected to win day in day out. Once we have found the best predictive model. We will extract the words that will yield the highest conditional probability of being associated with a win and loss respectively for both clubs.

II. Data Wrangling & Data Exploration:

To conduct this analysis, we will need to collect data. We will need labelled data for our classification models to be able to predict. The labels in this case will be whether a team won or lost a game. We use the website www.football-data.co.uk to collect data about the results of both teams for the past 5 seasons and store it into an Excel Sheet. For context, there are 20 teams in the Premier League, so in a season, a team plays 38 games (19 home and away games).

After collecting our labelled data, we need to collect data about user sentiment during each of these games for the past 5 seasons . For this, we decided to use the Twitter API to collect tweets for each game.

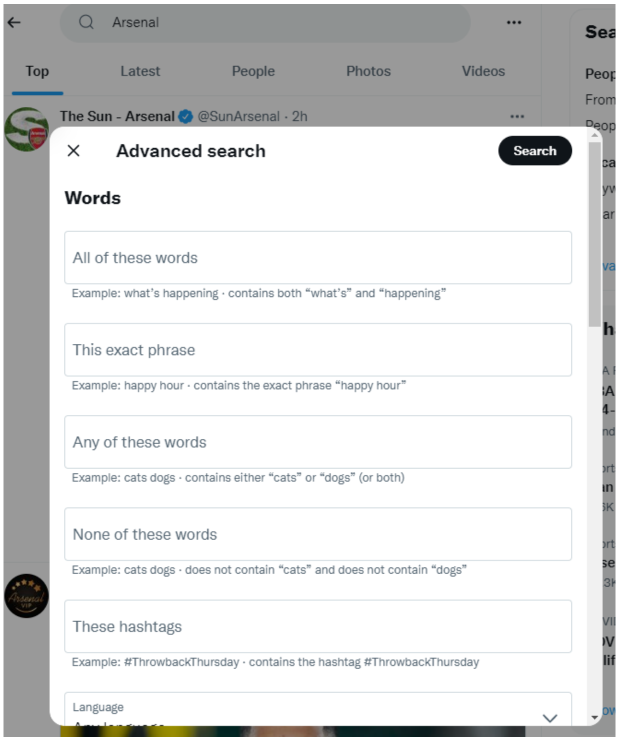
Summary :

- **1000** tweets per game.
- **38** games for each season (except for the last season, where we had data. for only 28 games as we did this analysis in April 2022).
- **5** seasons between *2017/2018* and *2021/2022.*
- **2** teams – Arsenal & Tottenham.

For a total of **370,000** tweets.

When first attempting to collect the data with the Twitter API , we had to create a developer portal and get **elevated** access. When we used the API, we were not very successful in collecting all the data we needed because Twitter limited the number of tweets we could collect. Additionally, we needed to create a program which would be able to run multiple customizable queries from Twitter as we wanted to return 1000 tweets per game per team which was not an easy feature to implement with the API.

As a result, we switched our strategy and used a Python package called *snscrape* which was able to get the results from twitter queries just as if we were querying from Twitter as seen in the figure below.



Advanced search

Words

All of these words
Example: what's happening · contains both "what's" and "happening"

This exact phrase
Example: happy hour · contains the exact phrase "happy hour"

Any of these words
Example: cats dogs · contains either "cats" or "dogs" (or both)

None of these words
Example: cats dogs · does not contain "cats" and does not contain "dogs"

These hashtags
Example: #ThrowbackThursday · contains the hashtag #ThrowbackThursday

Language

To use *snscraper*, we had to create a list of queries to run, use a for loop to run through each of these queries and store each query as a row in a dataframe. The end goal was to convert the list of queries into rows in the dataframe, there would be 370k rows, each representing a tweet a User made. We also wanted to gather information about the tweet, such as the User, the number of likes the tweet received, the number of replies, the number of times it was retweeted and the time the tweet was made.

Upon looking at the snscrape documentation, there was a specific format to collect a certain amount of tweets within a time period. We used an Excel sheet and some formulas which would populate a row in the "*query*" column with the specific format:

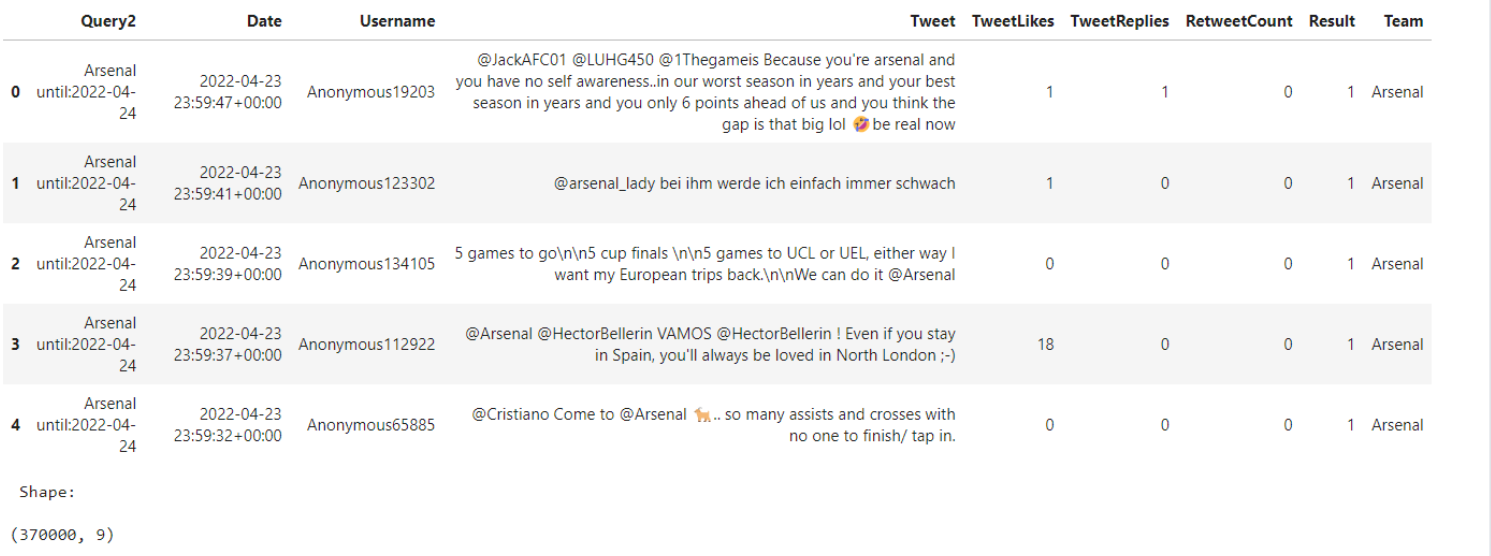
{team name} until: {game date +1}

The reason why we must add an additional day to the game date is that *snsscraper* can run a query that can fetch date up to a limit date. Since game days are not consistent throughout the seasons, we use this method to get tweets about the game day before. For example, if a game happens on the 23rd of May at 3:00 pm, then the query *"Arsenal until: 05/24/21"* will grab the first tweet fetched dated before 05/23/21 11:59:59 PM UTC. We independently verified that all games have at least a minimum of 1000 tweets.

The source code for the Twitter Query scraper code we built can be found here and can be used to run multiple Twitter queries.

Upon using the scraper, we have a complete dataset of 370,000 tweets. The users have been anonymized.

Figure 2: Completed Dataframe



	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

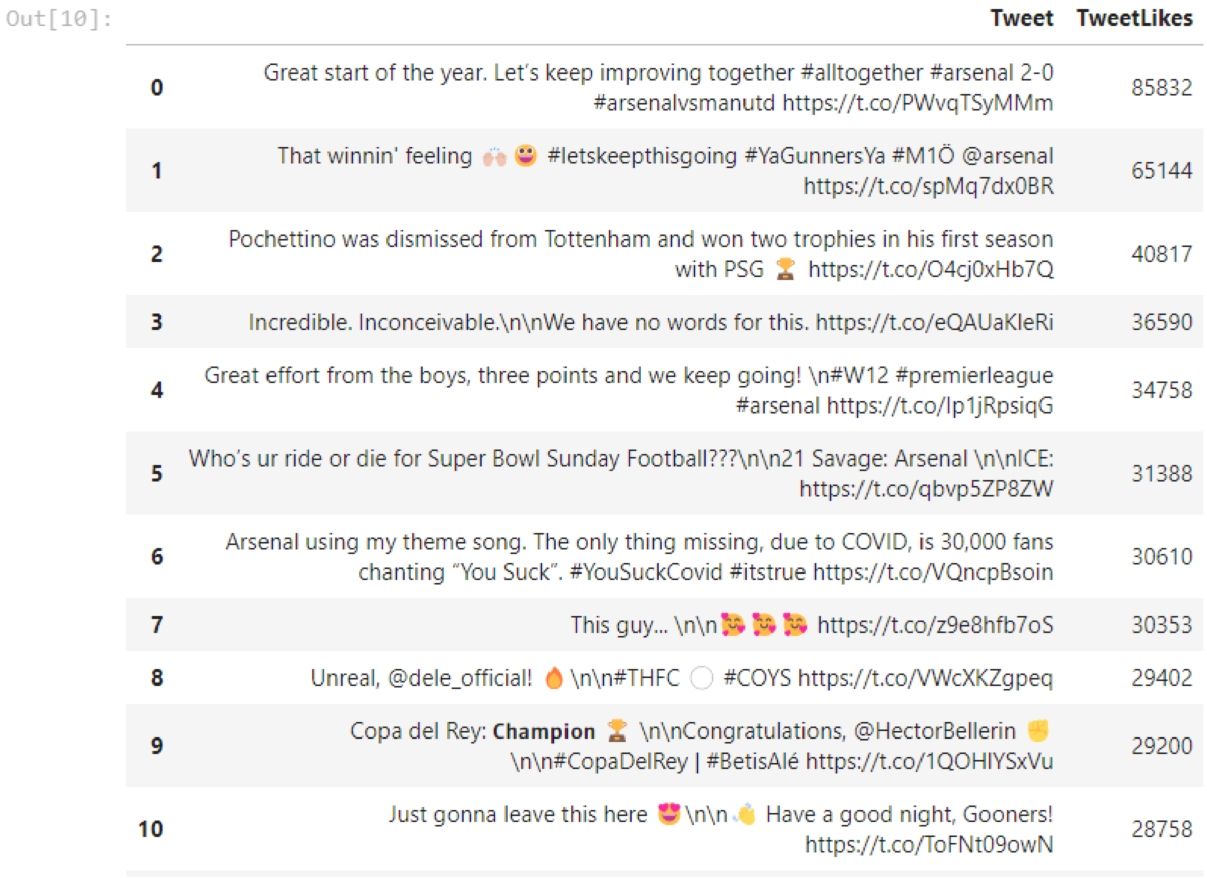
```
Shape:

(370000, 9)
```

Once all the tweets were successfully collected and stored in a pandas dataframe, our next step involved cleaning the tweets before proceeding to exploratory data analysis. The first cleaning step involved removing punctuation and the second involved removing the mentions of users in the tweet, not only to preserve privacy, but because we hypothesize that the username will not be indicative of whether a team won or not. Once this basic cleaning is complete, we proceed to exploratory data analysis where further steps are taken to clean the data and explore patterns / relationships.

III. Exploratory Data Analysis

Now that the tweets went through some cleaning and a new row called "*Clean Tweet*" has been created, we can proceed in conducting exploratory data analysis to get a preview of our data. One of the first things we wanted to check was how the ten tweets with the most likes looked like. Overall, we wanted to see if content that was negative or positive would receive more engagement on the platform as some fans can be very vocal.

Figure 3: Tweets with Most Likes



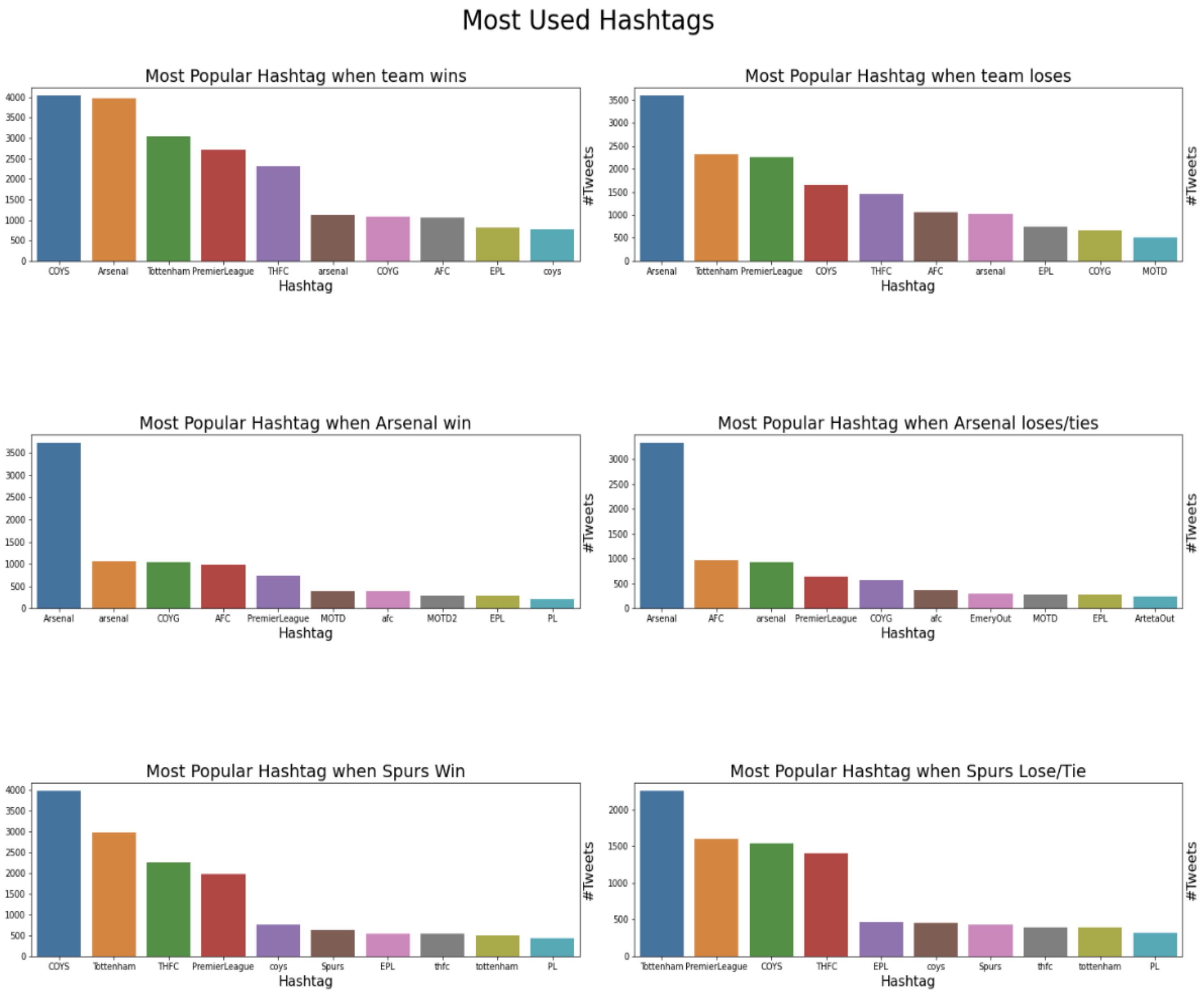
Out[10]:

	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/1QOHIYSxVu	29200
10	Just gonna leave this here 😍 \n\n👋 Have a good night, Gooners! https://t.co/ToFNt09owN	28758

Interestingly, of the top 10 most popular tweets, not one of them was negative. Only tweet #2 as seen above is slightly negative as it is ironically pointing out that a coach that was fired from Tottenham went on to win two trophies at another team. When further digging into the top 20 most popular tweets by likes, they were all positive as well. This is interesting because for all the bad rep fans or Twitter or football fans can get, it seems that the platform tends to be a source of positivity and users are most engaged with positive content about their teams. This does not mean that most tweets are positive, but the ones that are get the most engagement from users.

Once we saw the most popular tweets, we wanted to analyze hashtags that are associated with the games. Specifically, we wanted to look at the most popular hashtags that are used when the teams win or not.

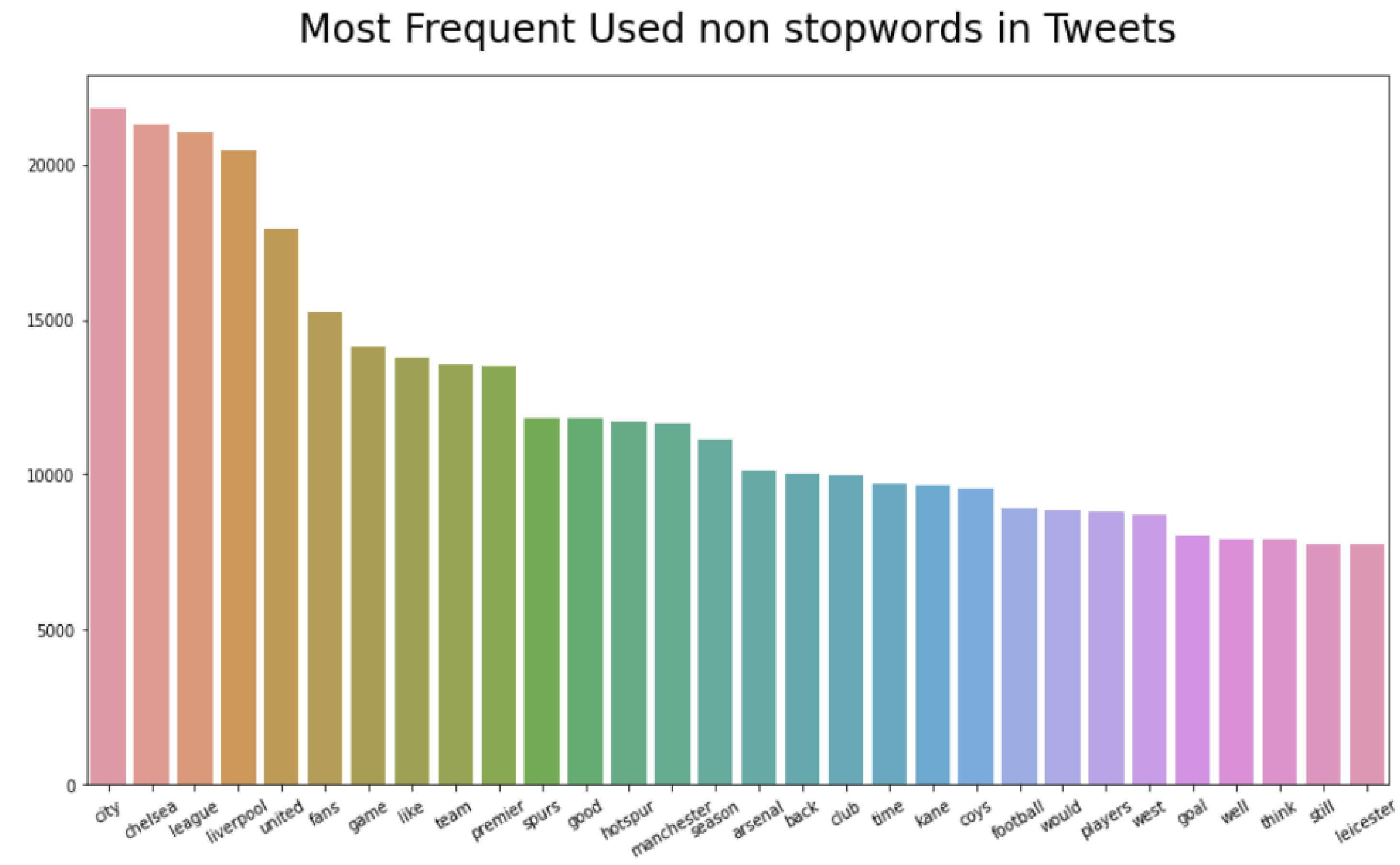
Figure 4: Most Used Hashtags



To get the hashtags, we had to create our own user function called *find_hashtags* that would be able to find words that started with “#”. We then proceeded to filter for whether the teams won or not and further broke it down for both Arsenal and Tottenham to see if there was a difference or not. Something that is more noticeable with Arsenal fans as opposed to Tottenham fans is the level of support for the coach/manager. When Arsenal loses or tie, the #Artetaout #EmeryOut tag is trending, calling for the manager to be relieved from his duties. However, this does not appear to be the case for Tottenham fans as they appear to be more patient with their manager. Furthermore, we can observe that the rallying cry for Spurs and Arsenal, COYS (which means Come on You Spurs!) and COYG (Come on You Gunners!) is most often mentioned when the team wins than loses.

Lastly, to conclude our EDA section, we wanted to observe the most popular frequency of nonstop words. We also removed (most) of our words that contain mentions of *Arsenal* or *Tottenham*.

Figure 5: Most Frequent nonstop words



From what we observe, most non-stop words are associated with other teams in the premier league, with the most mentioned ones being the other top teams, “City” (for Manchester City), “Chelsea”, “Liverpool” and “United”. This means that users are more engaged during big games as opposed to games with smaller teams.

What’s interesting in this is that only one player is mentioned “Kane” indicating that there may not be one dominant superstar on either of the two teams that gets a lot of attention as there is hardly any mention of any of the players. This would imply that users are a lot more interested in the teams and performances of the club as opposed to the players on their team and that engagement is maximized when talking about game days instead of players. With this information context, we are ready to proceed with our modelling section.

IV. Pre-Processing and Modeling

To get our data ready for modeling, we first must remove the stopwords in the tweets. We do this with the *NLTK* library and remove all English stopwords. In addition to the given stopwords, we add to the list some additional punctuation as words relating to *Arsenal* and *Tottenham*. We then apply the Count Vectorizer to vectorize the individual tweets and get substring of words.

We then defined a function called *make_xy* which creates our independent variables (vectorized tweets) and the dependent variable (game result). We then use *train_test_split* which separates our data into training and testing.

For the purposes of modeling, we decided to use three different models :

1) **Naïve Bayes**
2) **Logistic Regression**
3) **SGDC Classifier**
4) **RandomForest Classifier.**

For each of these models, we conducted several trials if we would use *Count Vectorizer* or *TfidVectorizer* and the minimum number of words that had to appear for each of these. The *TF-IDF vectorizer* was of better use since it combined focusing the frequency that the word appears and its relative importance. We used the *min_df* setting to 30 to ensure that that terms that appear too infrequently are not considered. We used *GridSearchCV* on our first three models and *RandomizedSearchCV*. Below is the summary table of our results for each.

Figure 6: Summary Results Table

Model	Best Hyperparameters	ROC_AUC	F1_Score (weighted)	Precision	Recall	Accuracy
Logistic Regression	{'C': 0.1, 'penalty': 'l2'}	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}	**0.684**	**0.651**	**0.622**	**0.683**	**0.626**
Naïve Bayes	{‘alpha’:1}	0.664	0.638	0.613	0.664	0.615
SGDC	{'alpha: 0.0001, ‘l1_ratio’: 0.2 , loss=’log}	0.665	0.641	0.614	0.67	0.617
No Skill Classifier	NA	0.5	0.51	0.51	0.51	0.51

Figure 7: ROC AUC curves for 3 different Models

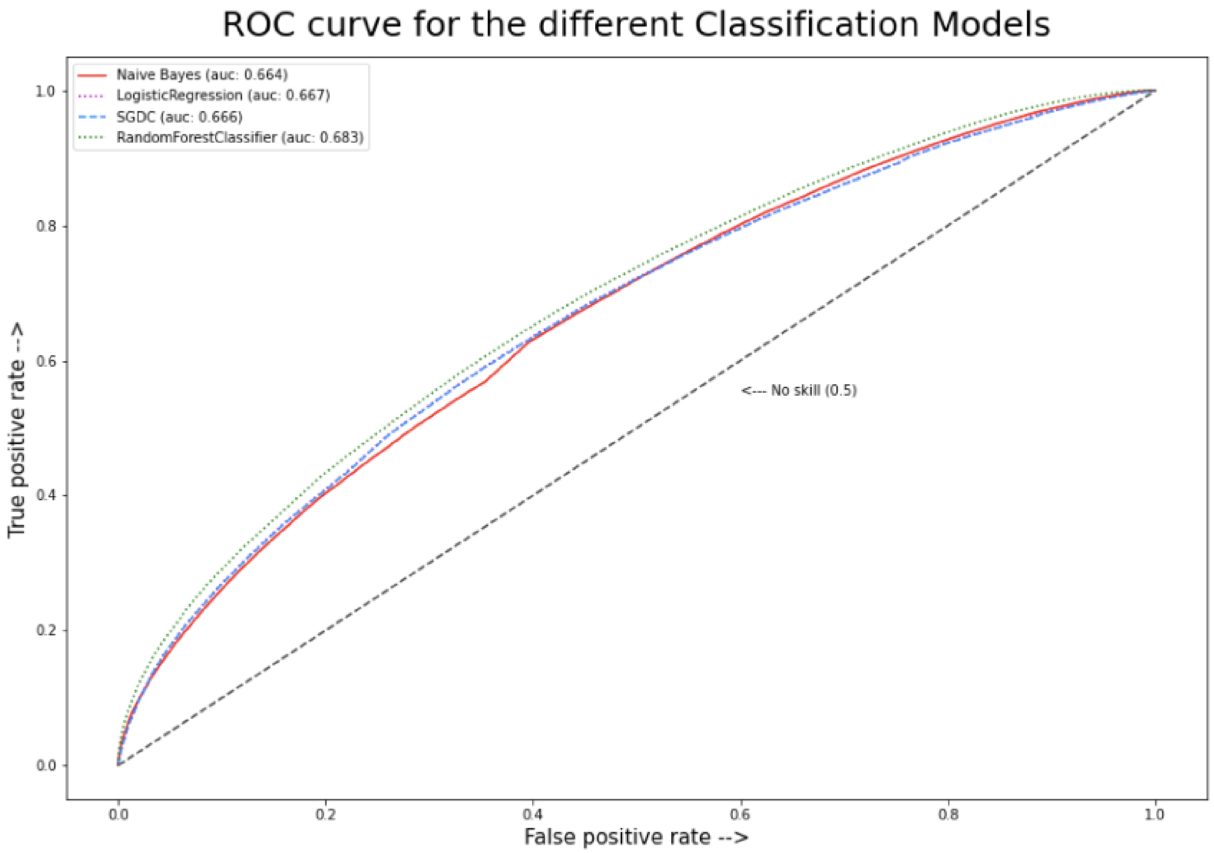
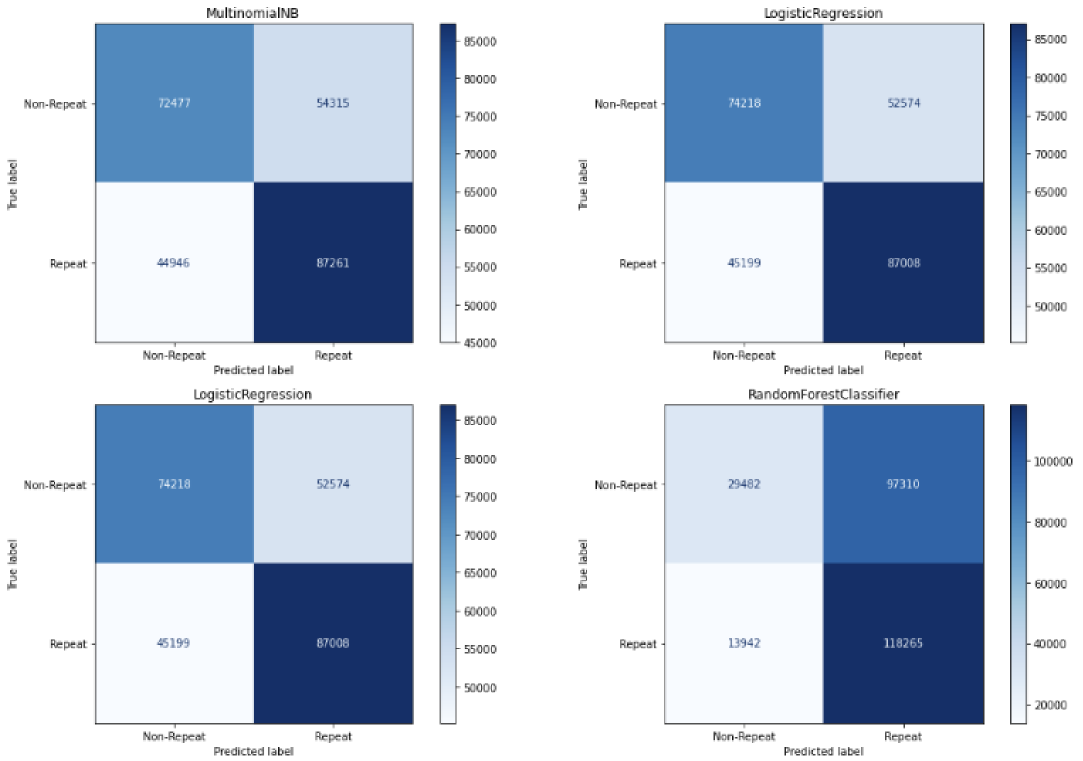


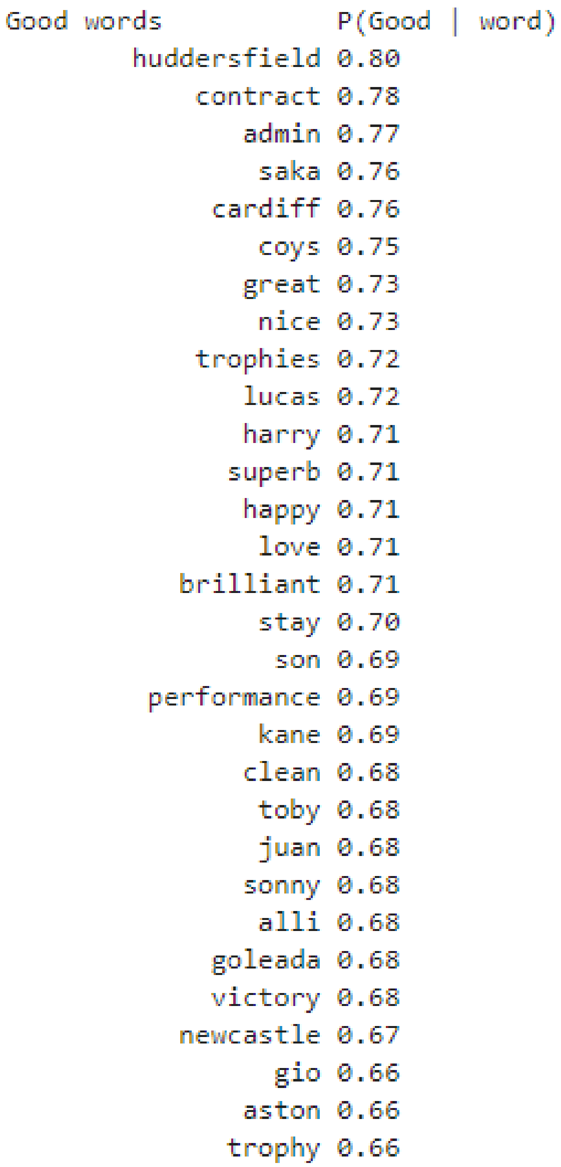
Figure 8: Confusion Matrices



Since the *RandomForest* model yielded the best results for all our metrics, we proceeded to use this model to select the words that had the highest conditional probability of being associated with a win or not. To do this we got the feature names and used *clf.predict_log_proba* and stored our 30 most used positive and negative words.

Here were the results for our list of words most associated with a win for both teams.

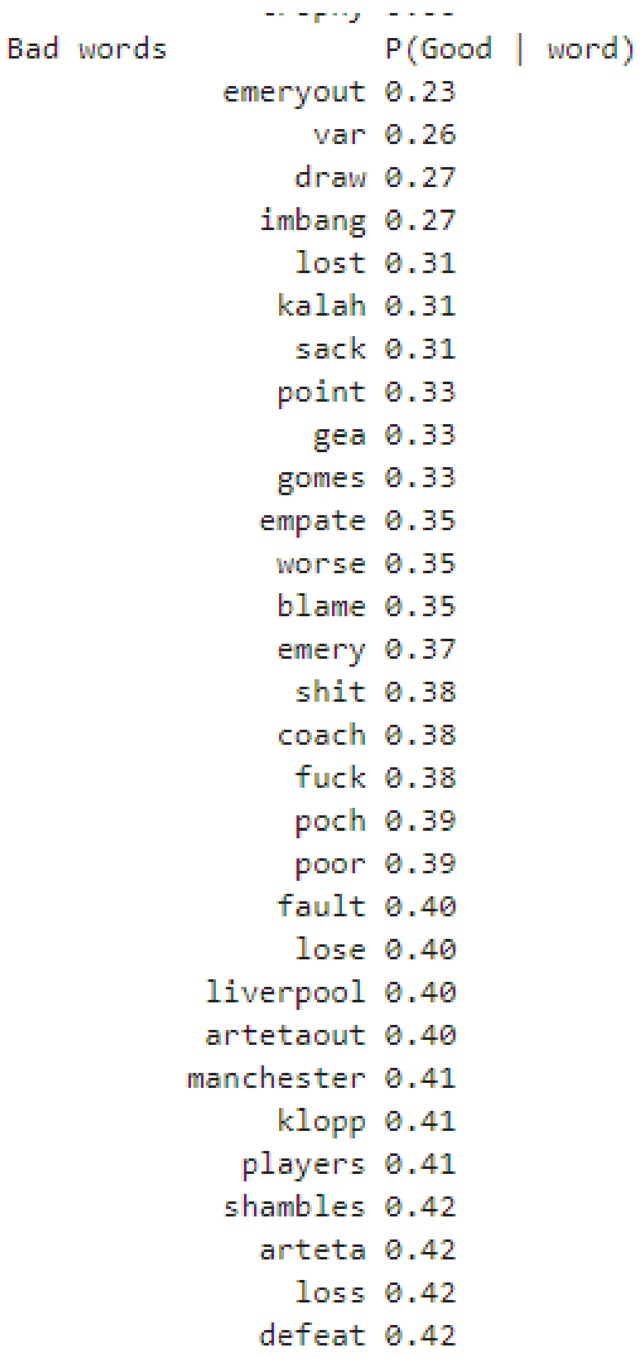
Figure 9: Words most Associated with a Win



```
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
```

Here are the results for words least associated with a win.

Figure 10: Words most associated with a loss:



```
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
```

V. Findings, Business Recommendations and Future Recommendations.

1) Findings

Our model findings ended up being able to discern whether a team won or not based on user social sentiment tweets with moderate success.

When studying the conditional probabilities of words most associated with a win for Arsenal and Tottenham, we see that there are many different Premier league teams listed such as “Huddersfield”, “Cardiff”, “Newcastle” and “Aston” for example. These teams usually rank bottom of the premier league and usually lose when playing against either Arsenal or Spurs so it makes sense that tweets containing these team names would be associated with a win outcome for Arsenal and Tottenham.

There are also positive adjectives that are associated with wins such as “superb”, “brilliant”, “nice” and “great” showing that fans are pleased and satisfied when they’re team wins. Other words indicating victory were “contract”, implying most likely that the fan would like a player that performed well to renew their contract, “trophies” as the fans become optimistic, they can win silverware.

Interestingly, there is no mention of the coaching staff but 8 of the 30 words most associated with a win are players for Arsenal and Tottenham. They are “Saka”, “Lucas”, ”Harry”, ”Sonny”, ”Son”, ”Toby”, ”Juan”, ”Alli”. Interestingly, to the exception of Saka, all the players play for Tottenham perhaps suggesting that Arsenal fans do not have favorite players or do not consider them responsible for helping the team win.

When teams do not win, the story is a bit different. Some of the words associated with non-wins are strong top tier teams such as “Manchester”, possibly referring to both Manchester United and Manchester City as well as “Liverpool”, the latter two having dominating the premier league and being the outright winners for the past 5 years.

Perhaps unsurprisingly, as fans are upset their team did not win, there are a lot more negative words about the team’s performance such as “frustrated“ , “blame “ and “excuses”, “fault”, ”poor”, “shambles” as well as some cuss words.

Another interesting finding is that there are a lot more non-English words associated with non-wins that appear such as “kalah” which means “lost” and “imbang”/ “empate” which means “draw”. Additionally, we find that there are many words which either mean “draw” / “tie” or “lose” in foreign languages.

We also have indications that the fans are upset with the manager when they team does not win. The word with the highest conditional probability of being associated with a loss is “emery out”, referring to Unai Emery, former head coach of Arsenal. The words “sack” and “coach” appear as well as other former coaches of both teams, “Arteta” and “poch”.

One thing that is interesting is that no player from either team is featured in the words with the highest conditional probability of predicting a loss, indicating that there aren’t players who disappoint the fans. This is particularly surprising as a fan myself as I hear many negative mentions of certain players on the squad, but the mention of their names is not necessarily associated with a loss or tie, indicating that some of these fans must be a vocal minority or that overall, the majority of the fans stick behind their players and put responsibility on the coach.

2) Business Recommendations:

From what we were able to observe, we would be able to present a couple of recommendations to both the Arsenal and Tottenham Social Media teams.

1) For both teams, we found that there are certain teams that are usually associated with wins. This means that if consumer /product brands were partnering up with the Arsenal and Tottenham social media teams for sponsorship and the brand is more focused on getting happy

fans, we suggest that the social media team charges a premium for advertising during the games that are likely to be wins for Arsenal and Tottenham, so the fans associate the product to their team winning. This would be for the four teams we found above, for example.

2) Additionally, as certain players are popular amongst Tottenham fans, we can charge a premium to brands who want to advertise with a social media post on gameday with one of their star players we listed above.

3) For the Arsenal Social media team, we recommend that when the team loses to avoid speaking or posting content about the manager / coach but that when they win, to mention the hard work the coach is doing. We mostly suggest this recommendation as Arsenal fans tend to be a lot more vocally critical of their coach as opposed to the Tottenham fan with their coach.

3) Future Work:

Further work can be done to improve this analysis. Given more time and resources, we would try several experiments to collect tweets within a two-hour timeframe of a game's finish time as to collect more accurate tweets of how people are feeling right after the game is. With the methodology we are using, we are collecting tweets that are centered towards the end of the day, and this may not be capturing the full positivity / negativity of the game.

Additional work which could help in generating better results would be to improve our data cleaning process. One technique we can use in future analyses Is *Langdetect* on the tweets in order to remove all the tweets that are not in English. As we saw, many tweets were in Spanish or in Indonesian. Another technique would be to include more words in the stopwords such as "draw" and "loss" as they would be very indicative of the outcome of the game. We could also keep repeating this analysis removing positive and negative adjectives which do not deliver any results.

Lastly, another action which can be repeated would be to separate out wins and ties to see if indeed the fans react differently to these, as opposed to bundling them together in the “did no win” category.

END