English Premier League Tweets:
A Sentiment Analysis

By: Kyle Demers

Abstract

Data analysis for teams in the English Premier League (EPL) provides excellent feedback for club teams about the form and general state that they are in. Web scraping techniques were used to pull 100 tweets from each Premier League team and then data wrangling was used to preprocess the tweets. The ambition is to conduct a sentiment analysis on each of the tweets pulled in order to correlate with form, result, and state of the club.

Being able to watch every game for every team is unrealistic. Not only that but keeping up with the drama of each team is nearly impossible. Being able to visualize where each team is throughout the season allows for easy interoperable access without the massive time commitment.

The sentiment analysis modeling uses a natural language processing transformer Hugging Face. The goal is to correlate the predicted sentiment with the performance of the team. Sentiment should also reflect big decisions happening by the clubs. Looking at how sentiment changes with respect to the lead up of a manager getting sacked, the manager getting sacked, and the result within a few weeks of the sacking should also be prevalent.

Tableau dashboards will be used to provide insight into the results of the model and its correlations. There is definitely a correlation between the sentiment of the tweets per team with the results of their respective match in any given match week. Positive sentiment correlates to

wins, neutral sentiment correlates to draws, and negative sentiment correlates to losses. These dashboards also reflect manager sackings with their sentiments changing with the state of the club during that time. Polarizing club decisions like Fulham raising their season tickets are also clear within the dashboard.

Introduction

The English Football League (EFL) is England's professional and not professional hierarchy of league structure. At the end of each season a set number of teams go down a league due to their poor performance and are replaced by teams from the lower league who performed the best. This implies that at the top of the EFL, the EPL the teams there are the best, hence more viewership and therefore leading to more advertisement income.

Unlike most American sports, the culture is that the fans in the English Premier League have a significant amount of power. Controlling most of the clubs' revenue through TV viewership, ticket sales, club merchandise and more, the fans have the ability to make change within the club through their collective voice. In this season alone 14 Premier League manager have been sacked with only 20 total teams. If owners could understand the sentiment of clubs and appease them early on, this could increase revenue through the club through ticket sales or help the team in general perform better. It's also important to look at how the performance of the club is reflected by the fans' sentiment. How do wins draws and losses look for teams in different positions with different objectives, and how does the fan sentiment reflect that?

Literature Review

There is very little natural language processing work being done in this field. There are only a few other sources that really tackles sentiment analysis on the EPL is a Kaggle dataset that

is outdated from 2019 and a few papers on it. This dataset also has an incredible amount of noise by using each club's hashtag as a search term. This allows for tweets from news sources to creep into the data quite often as well as pull tweets that merely use the hashtag but might not be referencing the club. The papers also appear to be a bit outdated.

Dataset

With a lack of information out there on a clean dataset for analyzing the sentiment of current EPL teams, Twitter was the go to place to pull information. Twarc was the python package of choice using a free student developer account to parse through Twitter's very limited API. With the purchase of Twitter by Elon Musk, Tweets being pulled tended to be quite limited in quantity. A safe number was about 100 tweets per team resulting in 2000 tweets every match week.

To avoid noisy data as discussed in the Literature Review section, the search term would be the EPL team's twitter handle. This meant that the tweet being pulled must be someone trying to communicate with the club itself. While the code looped through every team, it would pull the 100 relevant tweets, store all the metadata in a JSONL file and then run a new script to pull the raw tweet from that JSONL file into a txt file. The txt file was simply a list of tweets with each team getting their own JSONL and txt files.

The next step was pulling the tweets from each txt file and storing them in a list. This was then used as the value for each team in a dictionary. Since the team's name was stored as the name of the txt files this step was quite simple by indexing the file name which was the team's name and by reading in the txt file the tweets were available. Next was a tricky bit of feature engineering to limit noise. When Twarc pulls replies it grabs every twitter handle in the thread

being replied to. This means that the Premier League team will be pulled even if someone is in a twitter argument five replies deep. This tweet being pulled is noise as it is likely not about the Premier League team any more, but it is more likely to be noisy data that will negatively impact the results of my model. A regular expression was used to only select tweets with the twitter handle of the Premier League team being used while no other twitter handles are mentioned in the tweet.

The final few steps of preprocessing the data are key for just storing the relevant information into a data frame that can be stored as a csv file until all match weeks for the timeline had been played. The data frame stored the tweet, which team it was about, and the result of the match during that specific week. The team and tweet would get stored first and the only manual update would be to a dictionary storing the result of each team for the match week. Then the results would be plugged in, and the data frame would be stored away.

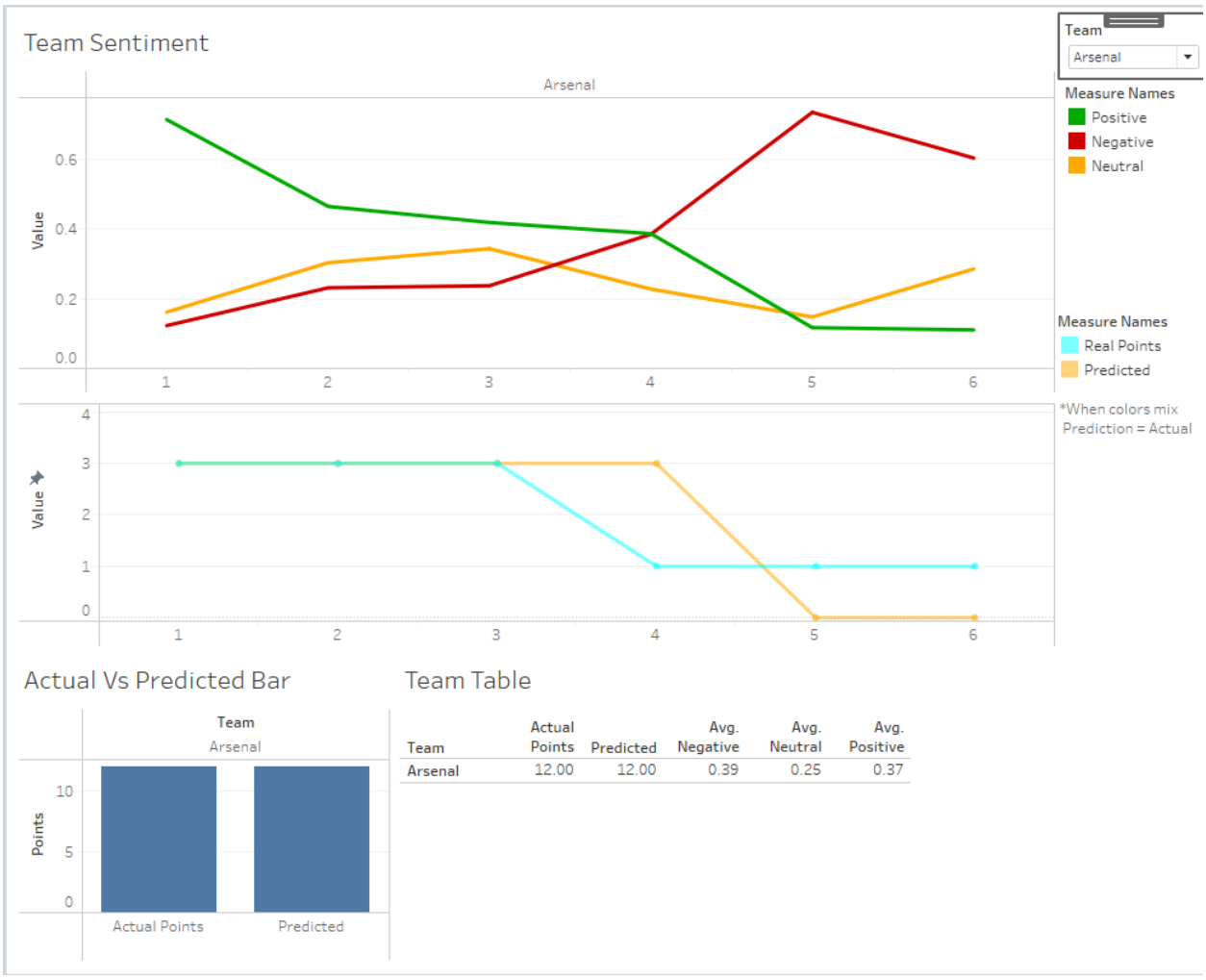
The next step was the actual sentiment analysis. The model used a hugging face transformer called cardiffnlp twitter xlm Roberta base sentiment. This is a pretrained model on an immensely large dataset primed for twitter sentiment analysis. Using this model will provide the scores for the classes positive, neutral, and negative. Reading in the txt files from before, this sentiment analysis can be used on each tweet per match week. In order to eliminate even more noise from the model, the scores will be added for each class and divided by the number of tweets for each team. This should give us the average probabilities for each class for each team every match week. We can use the soft max of the class prediction for the result of the predicted outcome of the game. This can be used to explore how sentiment reflects results of the match week.

Results and Discussion

The full results are stored in a Tableau workbook. The first dashboard reflects the average sentiment as described above for each sentiment graph. Below this is another line chart reflecting the predicted points, based on the class, relative to the actual amount of points earned for the match week. There is a bar chart in the bottom left showing the comparison of predicted vs actual points, as well as a table to the right giving general raw numbers for the team. This dashboard can be filtered by any of the teams in the premier league.

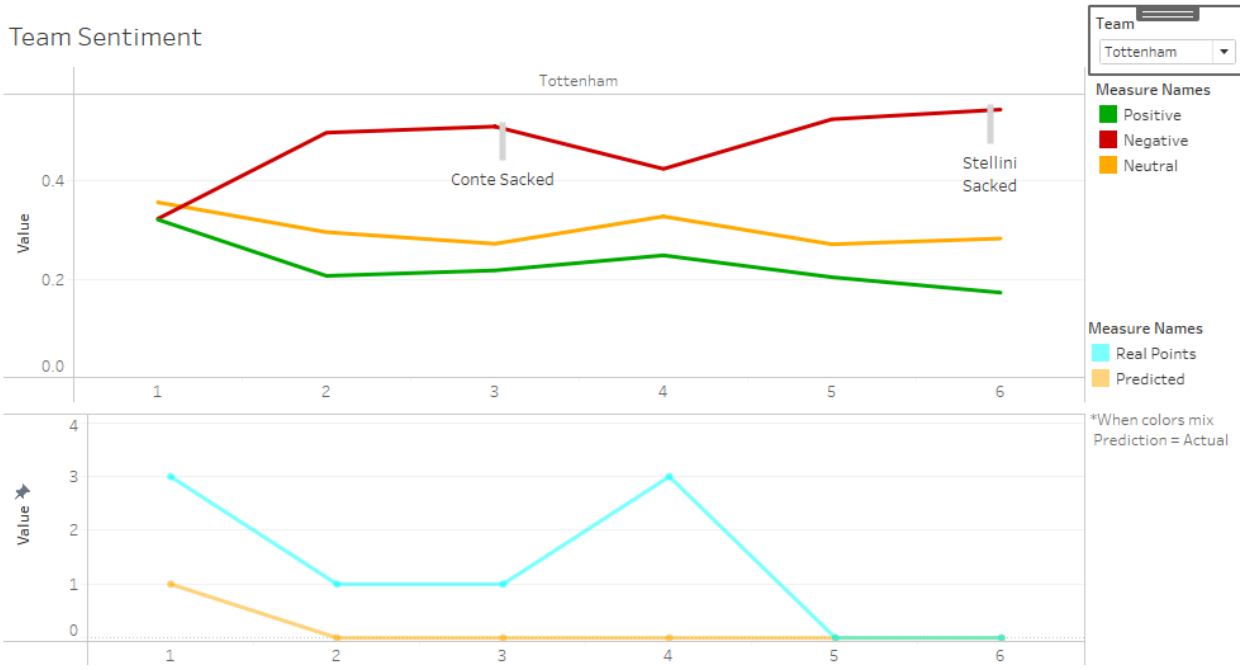
The dashboard is also very interactable with everything having a tool tip. This allows extra details for the specific visual components such as the actual values of the sentiment and hidden things that can't be seen like what match week the game was in. This varies as sometimes teams have to replay matches for events such as the passing of the queen or European events like the Champions League. The time metric was games played since the start of data collection in order to keep time continuity.

There are many important insights that can be taken from this dashboard. Ex-league leaders Arsenal can pull some great insights. When leading the league, it is important to win every game possible. When Arsenal takes their first draw during the studied time frame, there is a sharp rise in negative comments. This shows the two trains of thought. There is the group saying, "Can't win them all," and a group saying, "We can't drop points in a title race." This is followed by them drawing 2 more times, and the severity of that is shown by the sentiment.

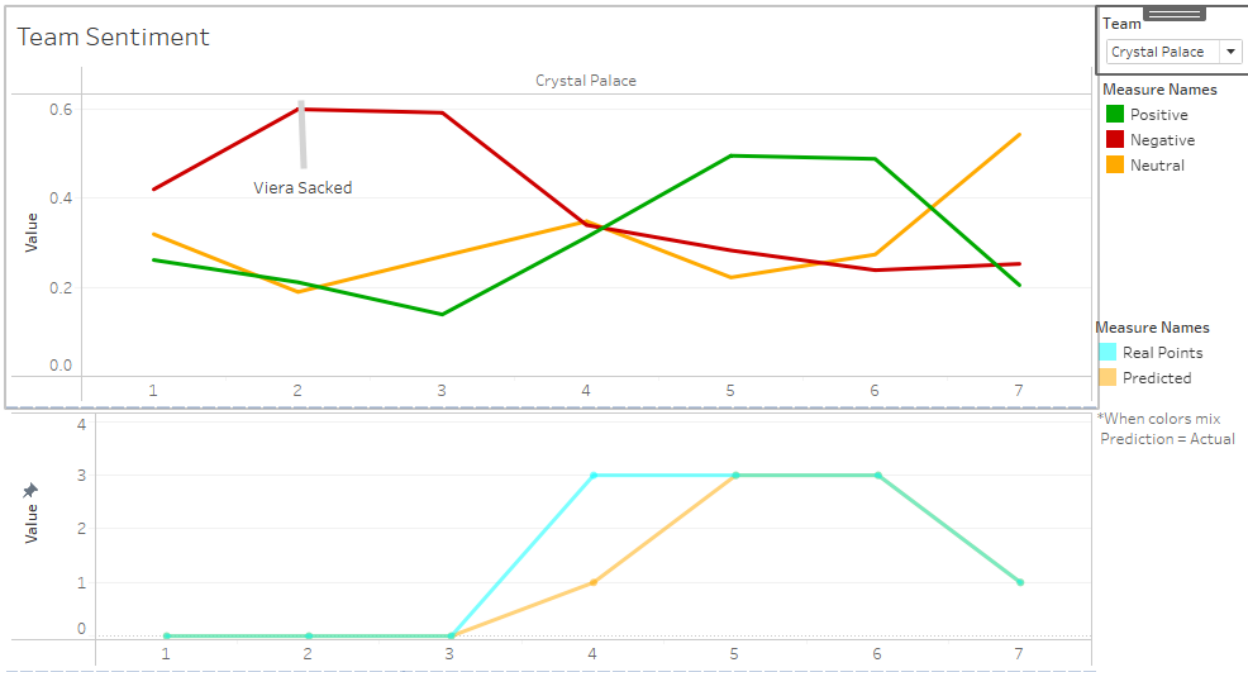


Arsenal's Sentiment

When looking at teams who have made big changes like sacking a manager, the trends become quite clear. Looking at teams like Crystal Palace, Tottenham, and Chelsea who have all sacked their manager, we can see peak negativity is when managers get sacked. This makes sense as it is a sort of last straw. If the manager can make a positive impact like in Crystal Palaces case, sentiment follows normal trends. If the manager can't make that impact we will see negative stay dominant. This occurs in Chelsea and Tottenham whose club situation can only be described as turmoil.

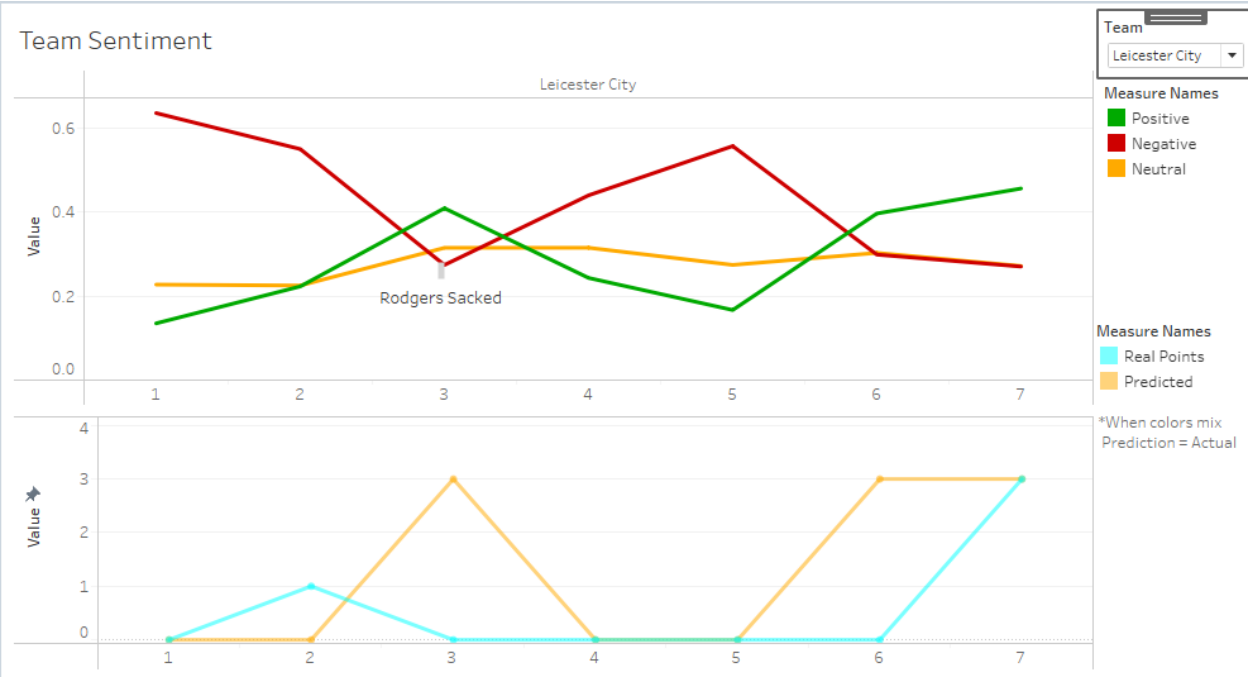


Tottenham's Sentiment



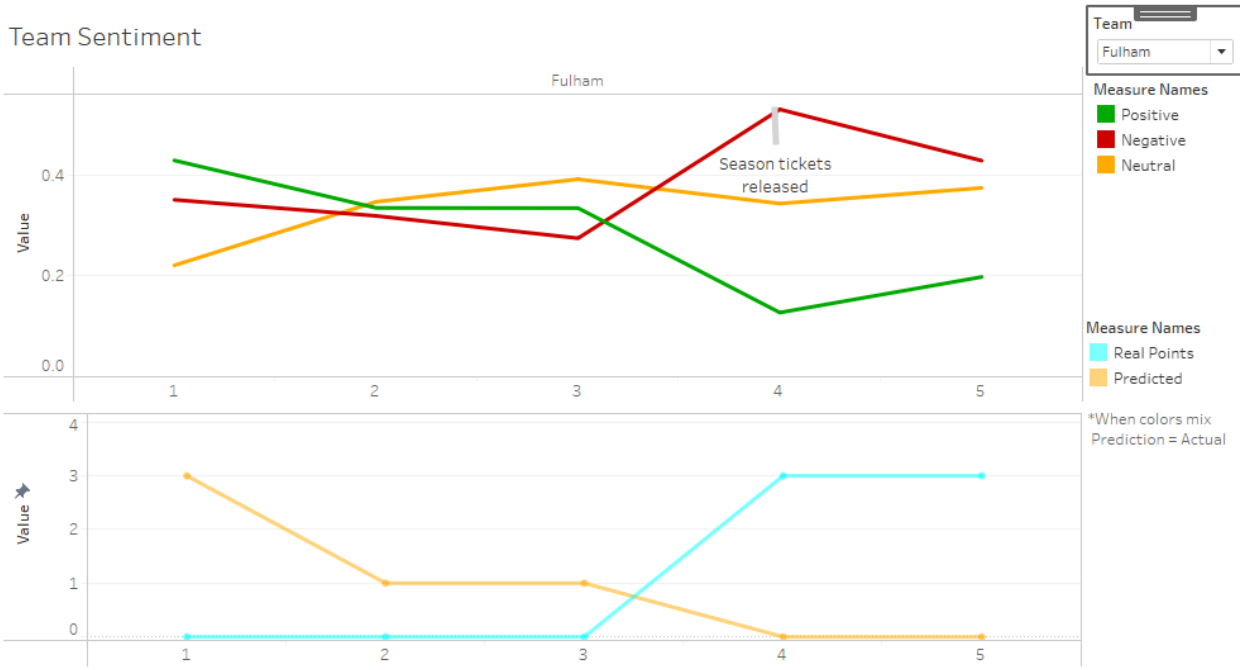
Crystal Palace's Sentiment

It is also interesting to see when Leicester sacked Brendan Rodgers, the positive sentiment actually rose. This is likely due to the fact that he should have been sacked about 20 match weeks prior to when he actually was sacked. This is quite interesting to see how something that usually is correlated with peak negativity is actually showing positive sentiment.



Leicester City's Sentiment

Other club decisions also have an impact on sentiment. This is reflected by Fulham's dashboard. We can see the point at which they released their most expensive season tickets for £3,000 ($3756.55). The common most expensive season ticket price is around £1,000 ($1,252.18). Even though the club won the game for the match week, the negative sentiment was the highest for Fulham over the studied time frame.



Fulham's Sentiment

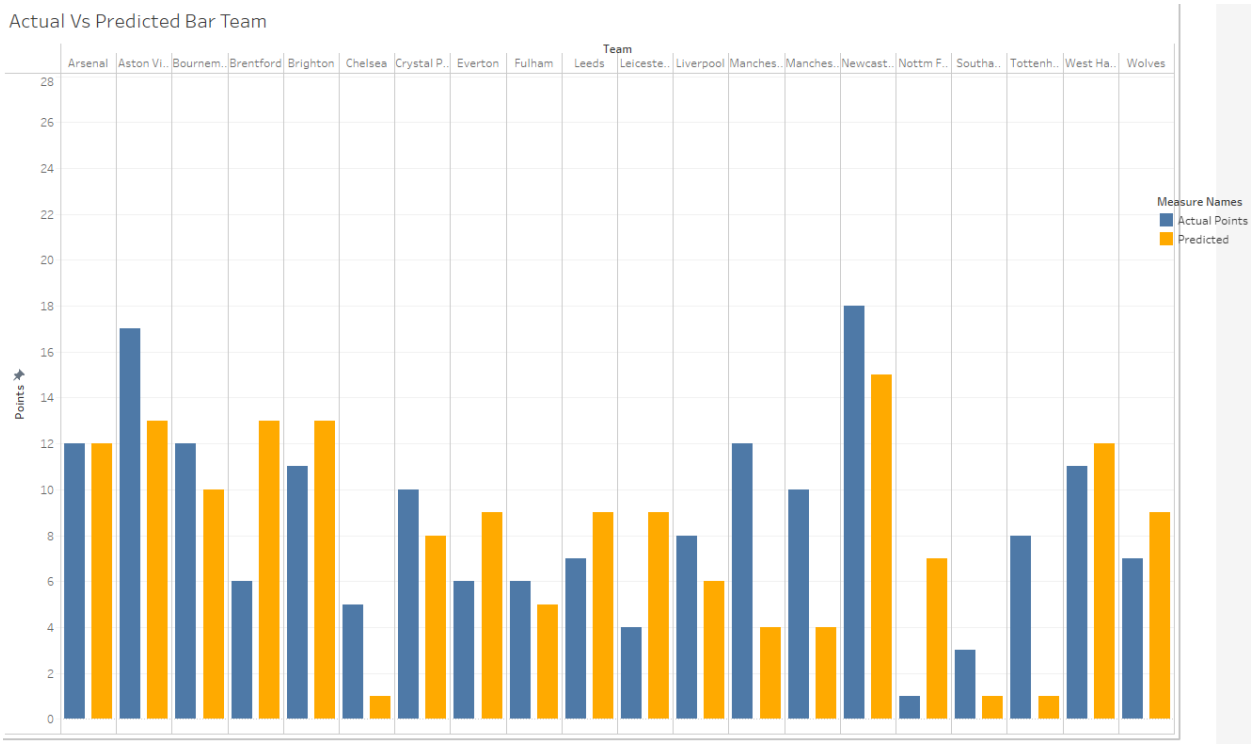
Another interesting dashboard is using the predicted points to see where each club team would be on the table relative to where they actually were. This can provide insight into the performance of the model as well as fun insight into a hypothetical alternative reality were points earned are based on fan sentiment. The model here predicts fairly well with most teams being close to their actual table placement.

Info

Team	Actual	Predicted	Point Differ..	Table Differ..
Arsenal	75	75	0	0
Manchester City	70	62	-8	0
Newcastle	59	56	-3	0
Manchester United	59	53	-6	0
Tottenham	53	46	-7	-4
Aston Villa	51	47	-4	-2
Liverpool	50	48	-2	0
Brighton	49	51	2	2
Fulham	45	44	-1	-1
Brentford	44	51	7	5
Chelsea	39	35	-4	-2
Crystal Palace	37	35	-2	-2
Wolves	34	36	2	2
Westham	34	35	1	-2
Bournemouth	33	31	-2	-3
Leeds	29	31	2	-3
Leicester City	28	33	5	1
Everton	28	36	8	7
Nttm Forest	27	33	6	2
Southampton	24	22	-2	0

Predicted EPL Table Using Sentiment

The last dashboard is also used to evaluate the correlation between points earned and fan sentiment. With this we can see whose fan base is following the anticipated trends. Those fan bases that don't follow the trends are pretty easy to see here such as Tottenham and Chelsea. Using this dashboard insights can be identified by looking into the current status of those teams. These clubs may sack managers or make financial decisions that impact the fans.



Difference Between Actual and Predicted Points

These Tableau dashboards are an advancement in the field of premier league sentiment analysis. There is little information in this field with a few papers and a dataset on Kaggle. These dashboards provide quick filterable graphs for the studying of sentiment over separate match weeks.

Conclusion

From the dashboards, it is easy to see that there is correlation between the average sentiments with the team's performance. When a team wins draws or loses there is typically a reflection of this by the sentiment. This isn't always the case however since the stakes are higher for some teams. When looking at Ex-league leaders Arsenal even drawing a game felt like a loss. With teams whose clubs aren't being well run such as Tottenham and Chelsea, negative sentiment is

high regardless of the results of the match. When clubs make rash financial decisions such as Fulham's season tickets prices being incredibly high, negative sentiment is likely to rise.

In the future another model can be written to identify strictly match outcomes. Sentiment is currently being used as the only predictor for match outcomes, so everything is just the correlation between sentiment and results. Training a model whose sole purpose is identifying outcomes would require more tweets and gold labeling. These two things would allow for a pure prediction model to be built and not rely on sentiment.

Note:

All code can be found and reproduced here:

https://github.com/Kyle-Demers08/NLP_340/tree/main/Final%20Project