Sentiment and Video Assistant Referees in Premier League Soccer

Dylan Blechner

Introduction

November 18, 2009. It was the all-important second leg of the FIFA Men's World Cup Qualifying playoff between France and the Republic of Ireland where the winner would move on to South Africa to compete for the World Cup and the loser would return home. After France won the first leg 1-0, Ireland leveled the score in the Stade de France sending the match to extra time. What happened in extra time would change the course of soccer history forever. French legend Thierry Henry illegally hit the ball with his hand in the buildup to a William Gallas goal that ended up being the match winner. It was a blatant missed call by the referee and sent the Irish players into a frenzy. Henry later admitted to his "crime" and the debate to introduce video refereeing was fueled.

Ten years later, video assistant referees (VAR) are on hand for almost every professional soccer match and the technology has been used in the FIFA Men's and Women's World Cups, UEFA Champion's League, and each of the top five professional soccer leagues in Europe (English Premier League, German Bundesliga, Spanish La Liga, Italian Serie A, French Ligue 1). While blatant missed calls like the "Henry Incident" have been removed from the game, controversy remains due to the specific rules of VAR.

There are four categories of decisions that can be reviewed by VAR: a goal, a penalty kick, a red card, and mistaken identity (red or yellow card awarded to the wrong player). A specified video assistant referee looks at every event on the field and is in direct contact with the head on-field official during the match. If the video assistant referee feels the head official may want to view a particular call, a signal is made and the game is stopped while the referee looks at the replay. After review, the head official then makes his final call. The majority of controversy surrounding VAR has to do with the idea of a "clear and obvious error." Subjective refereeing errors may be reviewed with VAR, but they can only be overturned if the error made on the field was "clear and obvious." This has led to much controversy among soccer fans and pundits who feel erroneous decisions have been made.

I became very interested in this area of research while watching the 2019 FIFA Women's World Cup. It seemed there were many more VAR decisions than in other professional competitions and there was more controversy surrounding said calls. The Washington Post's Steven Goff[1] noticed this as well and wrote an article during the tournament on the subject of VAR. He noted that the number of VAR calls in the Women's World Cup was higher than those at the Men's World Cup (29 in 44 matches vs. 20 in 64 matches), which was worrying to some. FIFA kept track of their refereeing decisions and made the claim that through 44 matches of the Women's World Cup, referees had been correct on 98.2% their calls which is an improvement from the 92.5% accuracy without VAR. Even with this assertion that VAR is improving the game, some people are upset over the use of the technology to negate goals in tight decisions. I wanted to dive deeper into a relatively new area of the game of soccer to analyze VAR in a novel way.

In this exploratory investigation I will test several hypotheses through the use of various data analyses. The first hypothesis is that VAR decisions during professional soccer matches increase social media activity. Between the sources I have researched and my own experiences as a fan, I believe that the use of VAR has sparked controversy among supporters on social media which has led to more activity on sites such as Twitter. The second hypothesis is that the sentiment of tweets about VAR decisions is more negative rather than positive or neutral. I believe that while the use of VAR has improved the game of soccer, fans are not used to the types of calls that are made with the assistance of the technology. As a result, fans are usually upset with the calls made on the field and take to Twitter to vent their frustrations.

Through my research I hope to analyze VAR decisions in professional soccer to help improve its use in leagues and competitions. Because the use of VAR is quite subjective, I will show that analysis of fan sentiment can be used as a proxy to determine which calls may have been incorrect. There is a lot of work that can be done to improve VAR to make the refereeing of the game as fair and accurate as possible. A better use of this technology can improve television ratings in professional leagues and improve overall fan interest.

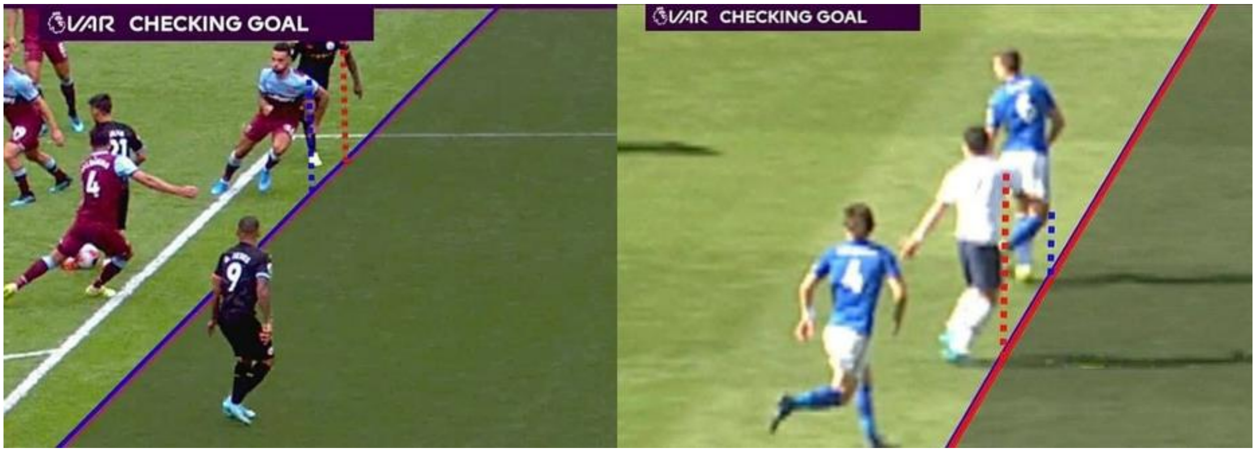
Literature Review

Non-Academic VAR Analysis:

Although little academic research has been done directly in the area of VAR and its various impacts on professional soccer matches and fan experience, many non-academic articles have been written discussing the accuracy of VAR decisions in key matches. Dale Johnson[2] and others wrote numerous articles about the use of VAR in the English Premier League during its first week of existence in 2019. The BBC[3] noted that there were 70 VAR checks that occurred behind the scenes during the first ten matches of the Premier League season. Prominent figures in the Premier League such as former England captain Alan Shearer, former referee Dermot Gallagher, and current Manchester City winger Raheem Sterling were asked about VAR and they gave their opinions about the technology. Shearer was very pleased with the Premier League during the opening weekend, saying, "they are using a high bar for VAR involvement and if it stays like that then it should be a success." Gallagher was very impressed with VAR's use and said "I wish I had it when I was a referee. I can look back at a few mistakes I made. As a referee, why would you not want to have it?" Even Raheem Sterling, whose Manchester City squad were disallowed a wonderfully worked team goal as a result of a fractional offside decision, said "if we're getting the decisions right then it can only be a good thing." While most people within the Premier League agree that VAR has been successfully used, some fans do not concur.

The opening weekend of the 2019-2020 Premier League season had some interesting VAR scenarios that put the technology to the test right off the bat. The most controversial decision from the weekend was in the Manchester City vs. West Ham United game where Gabriel Jesus' goal was disallowed after it was declared Raheem Sterling was marginally offside in the buildup. This tight decision showed how the Premier League would judge offside decisions: as a binary decision that will not be based on "clear and obvious" errors. This led to backlash from fans and much discussion around the league. Similar offsides decisions later in the season against Tottenham's Heung-Min Son and Liverpool's Roberto Firmino were also heavily criticized by supporters. Another call from the Manchester City match raised some eyebrows when Sergio Agüero was allowed to retake his weakly hit penalty because West Ham's Declan Rice stepped into the penalty box before Agüero made contact with the ball. Such a call had rarely been seen before the implementation of VAR, leading to more discussion around the league about how infrequently called infractions would be refereed under the new rules.

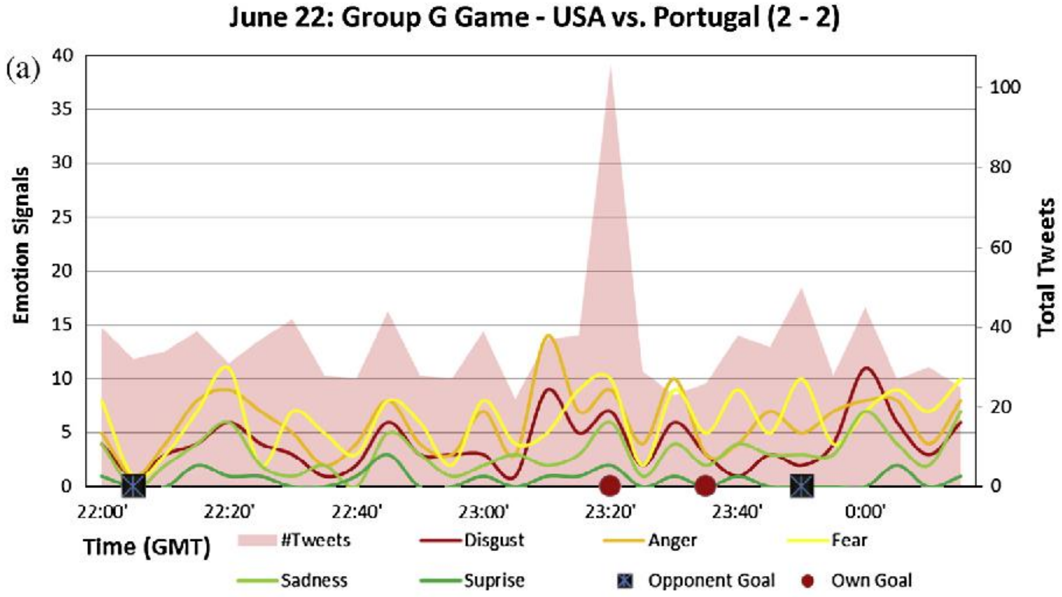
Figure 1: *Two marginal offsides decisions during the 2019-2020 English Premier League season including Manchester City's Raheem Sterling (left) and Tottenham's Heung-Min Son (right)*



Sentiment Analysis in Soccer

The new public discourse regarding referee decisions led me to research sentiment analysis during soccer matches. Is it possible to determine whether or not a VAR call is "correct" by using fan sentiment? In a paper by Yang Yu and Xiao Wang[4], the sentiment of United States soccer fans was analyzed during the 2014 FIFA Men's World Cup. The findings were to be expected as more fear and anger were shown after an opponent goal while joy was the primary sentiment indicator after a US goal. The authors concluded that using sentiment analysis was a generally accurate predictor of fan disposition during a soccer match.

Figure 2: *Sentiment analysis during the 2014 FIFA Men's World Cup Group Stage match between USA and Portugal*[3]



A paper by Corney, Martin, and Goker[5] analyzed matches during the 2012 and 2013 FA Cup Finals. Similar to the previous paper, they were able to identify key match events through spikes in Twitter usage and sentiment. They were also able to classify fans of specific teams through their tweets. The strategy used by both this paper and the previous were to analyze words, bigrams, and trigrams to see which phrases led to spikes in specific sentiment categories. Gratch *et al.*[6] used Twitter data from the 2014 World Cup to analyze fan sentiment during matches in the tournament. This paper specifically wanted to redefine what makes a sporting event exciting. The authors decided to classify all of the tweets into three categories: positive, neutral, and negative. They also studied the pre-match odds as well as in-game events to try and quantify excitement and used a classifier from the SemEval 2014 Twitter sentiment analysis challenge to classify their tweets. The classifier performed very well, and led me to believe that it would provide a good starting place for my own sentiment analysis.

Figure 3: *The number of Tweets from supporters from each team during the 2012 (left) and 2013 (right) FA Cup Finals at various points of the match*[5]

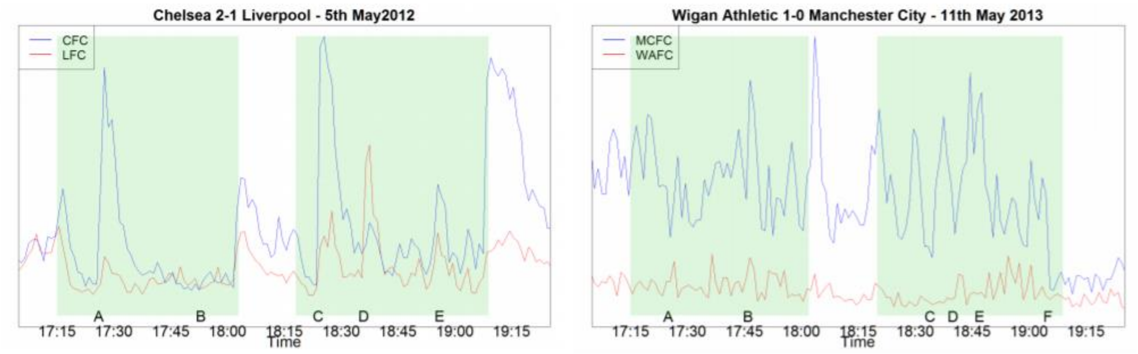
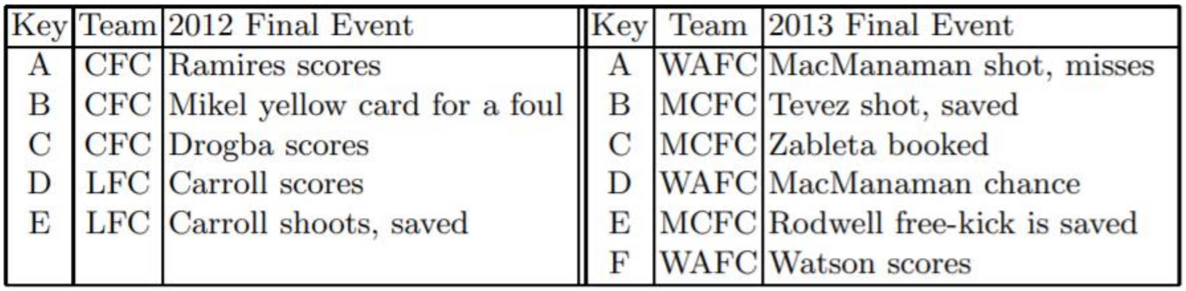


Table 1: *The corresponding labeled match events*[5]



Key	Team	2012 Final Event	Key	Team	2013 Final Event
A	CFC	Ramires scores	A	WAFC	MacManaman shot, misses
B	CFC	Mikel yellow card for a foul	B	MCFC	Tevez shot, saved
C	CFC	Drogba scores	C	MCFC	Zableta booked
D	LFC	Carroll scores	D	WAFC	MacManaman chance
E	LFC	Carroll shoots, saved	E	MCFC	Rodwell free-kick is saved
			F	WAFC	Watson scores

Data and Methodology

Data Summary

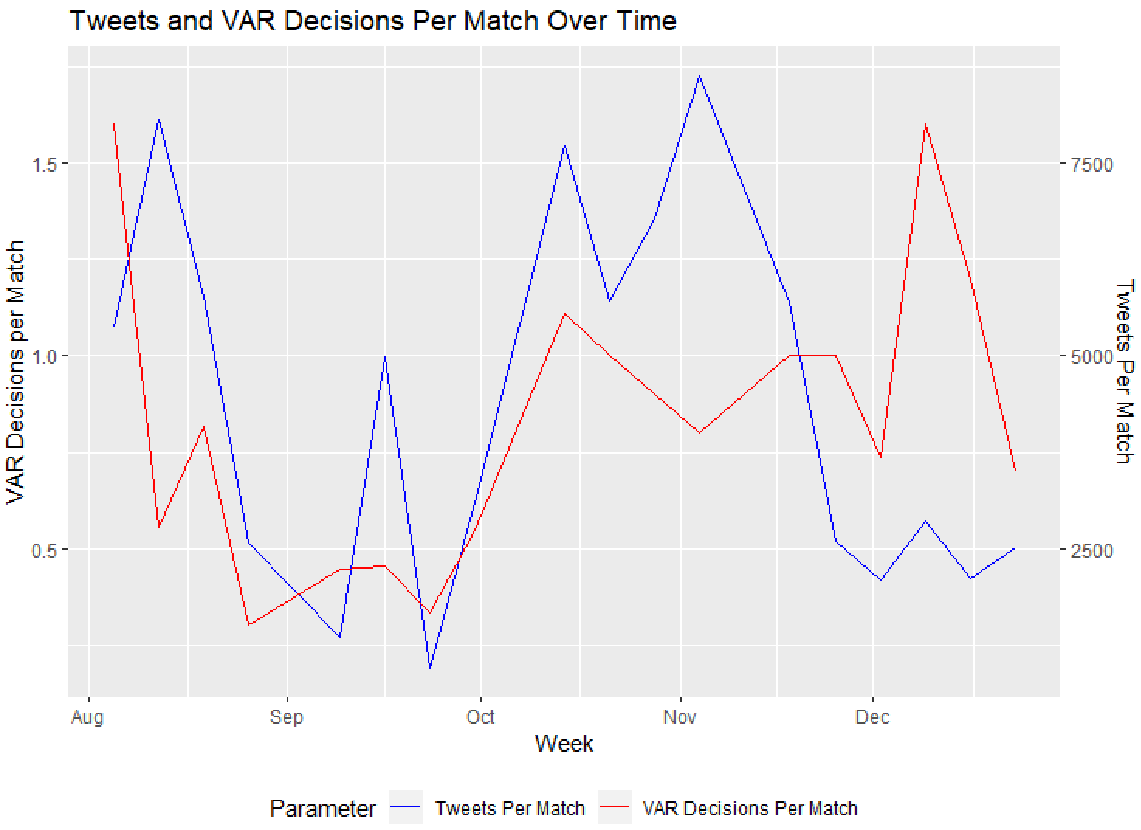
Data was gathered from several different sources for the separate analyses done in this paper. For the sentiment analysis, Twitter data was gathered using the "twitterscraper" module in Python[7]. Tweets were collected in a two-step process. First, tweets written in English containing the term "var" or the hashtag "#var" were scraped for each day that a Premier League game occurred during the first half of the 2019-2020 season (between August 9, 2019 and December 27, 2019). Data was gathered in this manner to evaluate fan sentiment for English Premier League games specifically and eliminate any tweets that were not about VAR. Second, tweets containing the term "var" as well as the corresponding game-specific hashtag were scraped for each game day (e.g. "#LIVMCI" is the game-specific hashtag for a game between Liverpool and Manchester City). These tweets were gathered to analyze sentiment on a game level and determine if any individual VAR decisions were deemed by fans as incorrect. Because of the nature of the scraping process, many of the tweets in this dataset were also included in the first dataset. The game-specific hashtags as well as the EPL schedule were taken from the official Premier League website. Once these tweets were scraped, information about the authors was gathered including the number of followers and location by using the "rtweet" package in R. The account names of these authors will not be shared at any point during this analysis to preserve anonymity.

Commentary data was also gathered for the Premier League from ESPN by using the "fcscrapR" package in R[8]. This commentary data includes information about key events that took place during a given match. VAR decisions are specified in the commentary data and were then separated into the four categories of decisions that can be reviewed by VAR: a goal, a penalty kick, a red card, and mistaken identity. Although there are many VAR decisions during professional matches, ESPN only specifies those that caused a significant stoppage in play due to a referee check.

Sentiment Analysis: VAR

There were 805,548 tweets containing the term "var" or the hashtag "#var" gathered during the first half of the 2019-2020 English Premier League season. Using these tweets, I was able to perform sentiment analysis to evaluate how fans felt about the new use of VAR in the Premier League. First, I looked at how many tweets were posted each week to understand how fans felt about VAR over time. *Figure 4* shows the average number of tweets per match by week from August to December as well as the number of VAR decisions made during that same time. There were 158 VAR decisions made during the first half of the Premier League season. It is clear that fans were quick to tweet about the use of VAR through the early portion of the season, especially since a significant number of calls were being made during that time. Comments about the use of VAR died down as the season went on because fans became more used to the technology and had much less to say. There is a weak positive correlation between the number of VAR calls and the number of tweets about VAR during a given week (R = .206). This means that we would generally expect fans to tweet more about VAR when there are more decisions that involve the technology, which makes sense.

Figure 4: *Number of tweets per match and number of VAR decisions per match over the course of the 2019-2020 English Premier League season*



Twitter Sentiment Background

Next, I calculated the word frequency of every term from the tweets containing "var" in order to take a closer look into the content of such tweets[9]. After tokenizing the text and removing common stop words, I created a list of the top 15 most common words and a word cloud of the top 250 terms (excluding "var" which was the most frequent term). Among the most frequent terms were popular words in soccer such as "goal", "game", "penalty", and "ref" as well as the team names of the top two clubs in the English Premier League: "liverpool" and "city" (short for Manchester City). The casual and informal nature of Twitter shows as profanity was used by tweeters when discussing VAR.

Figure 5: *Word frequency plot of the top 15 most common terms in tweets containing "var"*

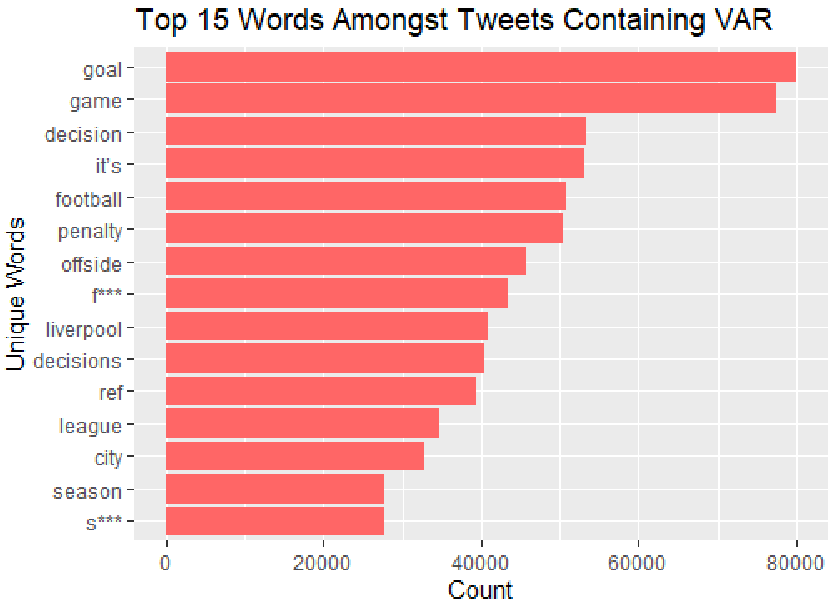
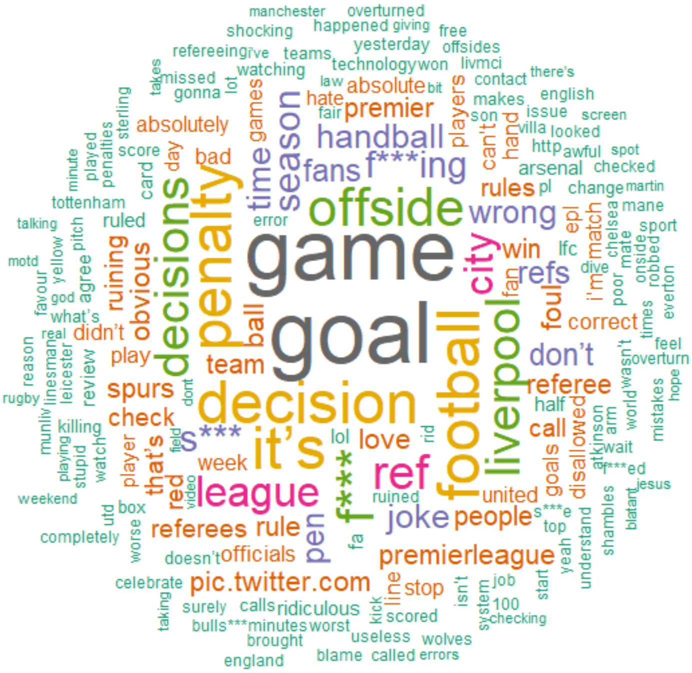
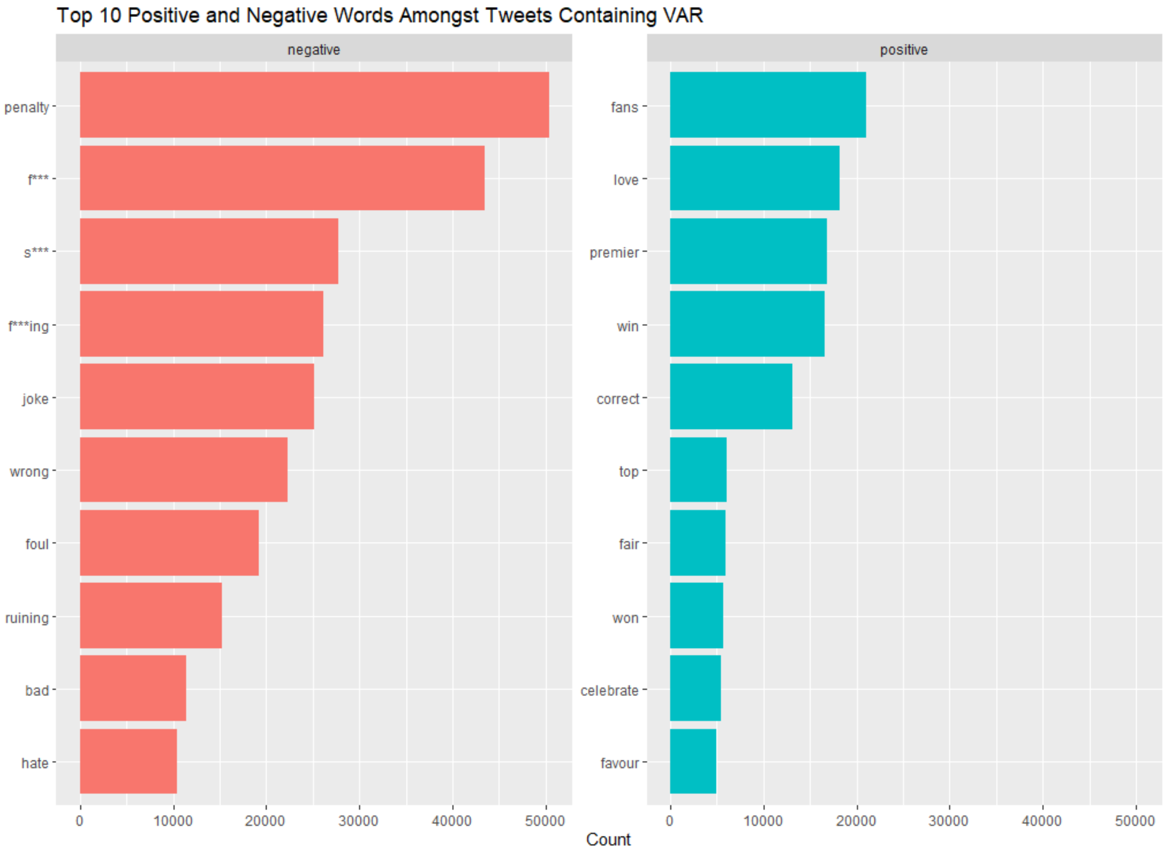


Figure 6: *Word cloud of 250 most frequent terms*



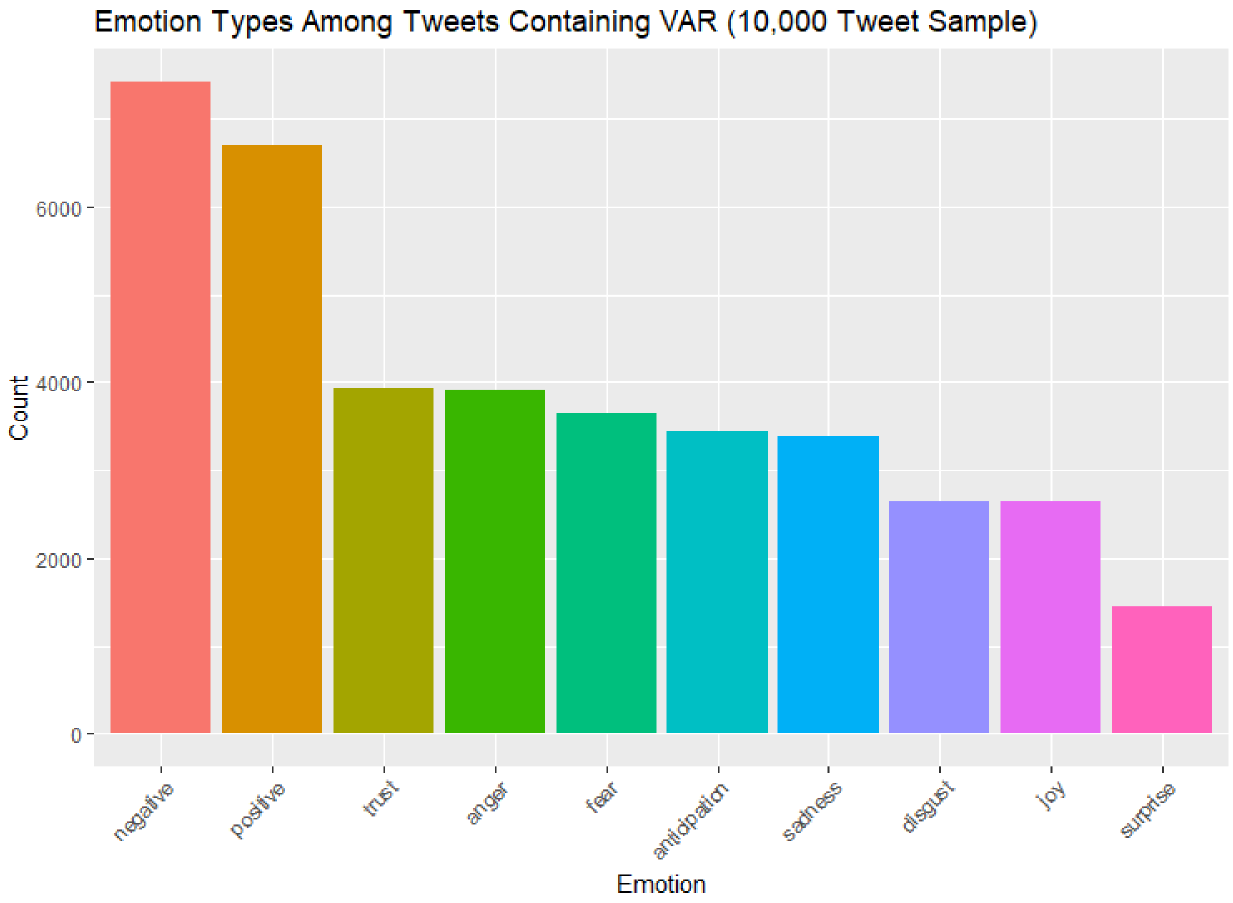
By using the Bing Sentiment Lexicon which categorizes 6,786 words as either positive or negative, we can see how frequently positive and negative words occur in tweets. After merging the Bing Lexicon with the tokenized terms, the word frequency graphs in *Figure 7* can be created with negative terms on the left side and positive terms on the right. Several new terms appear when using the Bing Lexicon. Frequent negative terms such as "joke", "wrong", "bad", and "hate" suggest that some fans are discouraged with the use of VAR and may feel that decisions are not going the right way. Some of the positive terms such as "love", "correct", and "fair" also show the opposite viewpoint of VAR where fans are appreciative of the new technology. It seems that there is a larger portion of negative words in the plot, but this is the case only because the Bing Lexicon is made up of 70% negative words and only 30% positive words. Similarly, 70% of the terms in the VAR tweets picked up by the Bing Lexicon were negative.

Figure 7: *Word frequency plots of top 10 most common positive and negative terms*



Sentiment can be broken down even further than just positive and negative through the use of the NRC Lexicon which categorizes words into eight additional categories: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. A word may fall into zero or multiple categories as well. Because of the large number of tweets in the dataset which all have many terms that must be broken down in the NRC Lexicon, I analyzed a random sample of 10,000 tweets to see which emotions were most prevalent in the data. *Figure 8* shows that the most basic sentiments of negative and positive appeared most frequently in the sample. There were 7,426 occurrences of negative terms followed by 6,697 positive terms which once again shows the negativity shown by tweeters towards VAR. One of the additional sentiment categories that appeared frequently was anger with 3,911 terms showing that a good portion of tweeters do not just have negative feelings towards the use of VAR, but are also angry with the decisions being made. Tweeters showed trust towards VAR as well, which demonstrates the divide amongst fans about the use of the technology. It must be noted that there may be bias in these findings of sentiment because fans are more likely to tweet about the use of VAR if it negatively affects their favorite team. Fans are very unlikely to tweet when they are unaffected by the use of VAR, which may be the cause of high frequencies of positive and particularly negative sentiment. This similar bias has been noted in political tweets as well where there is considerable negative rhetoric amongst opposing politicians[10].

Figure 8: *Frequencies of terms by NRC Lexicon Category*



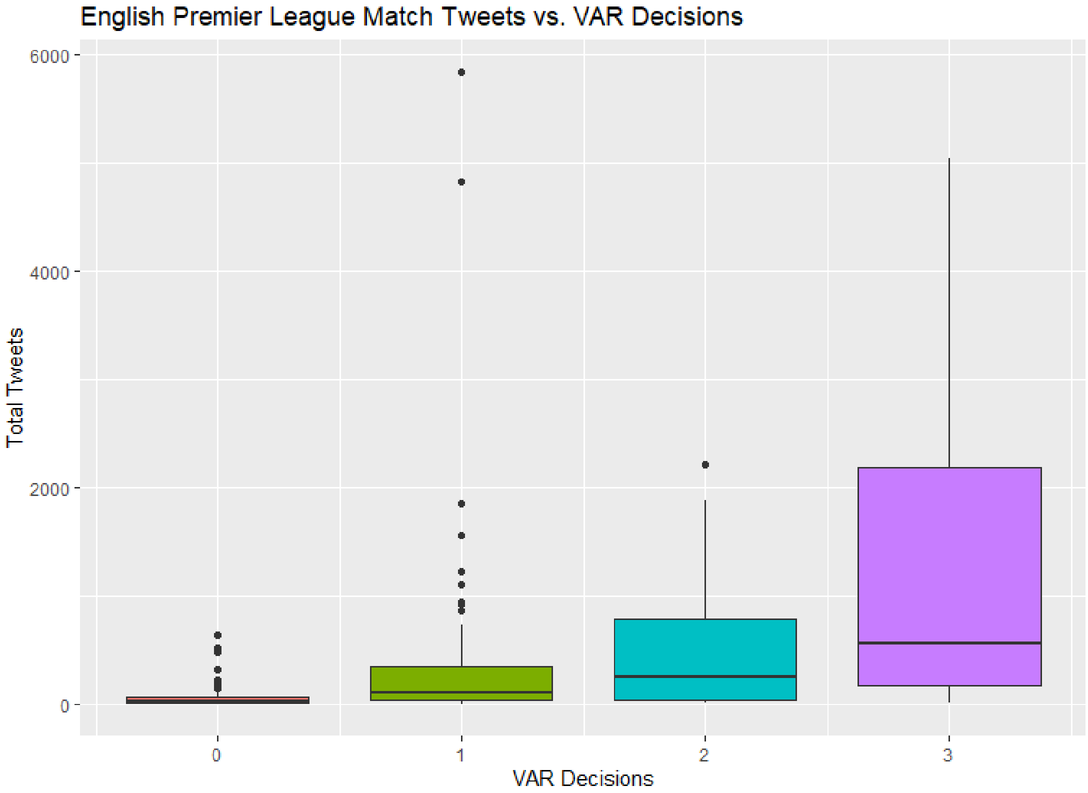
Sentiment Analysis: Game Hashtags

There were 61,127 tweets containing the term "var" and a game-specific hashtag gathered during the first half of the 2019-2020 English Premier League season. Using these tweets, I was able to perform sentiment analysis to evaluate how fans felt about the new use of VAR in specific games during the Premier League season. By looking at tweets on a game level, we can analyze how the number of VAR decisions in a match affects fan sentiment. *Figure 9* plots the total number of tweets against the number of VAR decisions made in the match. A clear relationship can be seen as the number of tweets grows as the number of VAR decisions made in a match increases. The correlation between total tweets and VAR decisions is 0.357 suggesting a weak positive correlation between the two variables. Intuitively, this relationship makes sense as fans are more likely to voice their opinions about VAR when a call occurs during the match they are watching. I also analyzed the relationship between sentiment and number of VAR decisions in a match. The number of positive and negative terms was once again calculated by using the Bing Sentiment Lexicon. There is not a strong relationship between positive and negative sentiment and VAR calls. One may think that a fan may show more negative sentiment towards VAR if more calls are made during the match, but it seems that fans tend to be generally negative no matter how many calls are made.

Table 2: *Average number of tweets and total matches by number of VAR decisions*

VAR Decisions	Average Tweets	Total Matches
0	65.90	82
1	415.18	68
2	465.30	27
3	1240.50	12

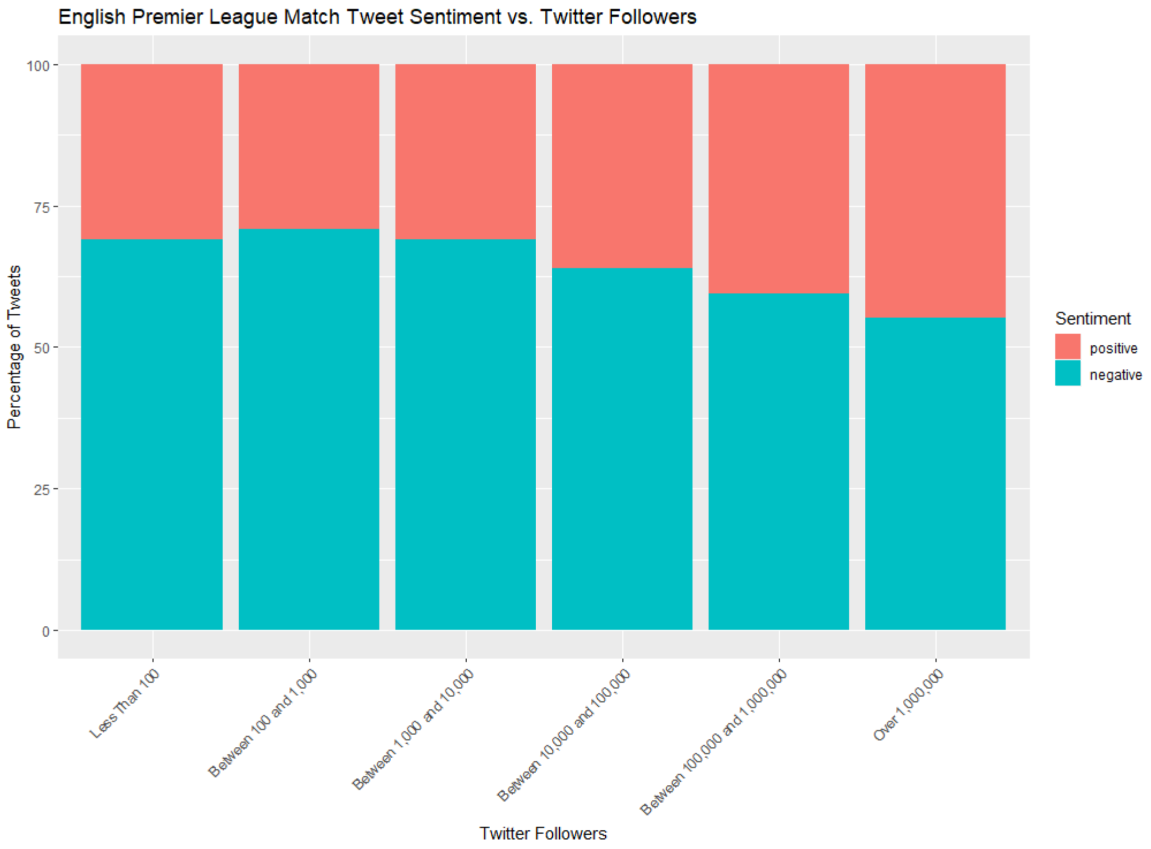
Figure 9: *Total tweets by number of VAR decisions*



Twitter Followers:

I also wanted to take a look at the impact of the number of followers on VAR tweet sentiment. For each tweet that contained VAR during the English Premier League season, I also scraped the user's corresponding information which includes number of followers at the time of the scrape. I once again utilized the Bing Lexicon to analyze the positive and negative tweets made by each user. *Figure 10* shows the percentage of positive and negative tweets for users in several bins of follower counts. There is a clear relationship between number of followers and tweet sentiment: the more followers a user has, the more positive they are. This makes sense because many users with a large number of followers that discuss the use of VAR on Twitter are employed by the Premier League and do not want to make any harsh claims about the technology used in their own league. Common fans with fewer followers do not share this constraint and are free to tweet in a more negative fashion. This disconnect between the common fans and the higher-ups in the Premier League is important because it may make ordinary people feel that they do not have a say in the decisions the league makes. While several current and former players such as Raheem Sterling, Gary Linekar, İlkay Gündoğan, and Peter Schmeichel have spoken about the negatives of VAR, many large corporate accounts have not done the same.

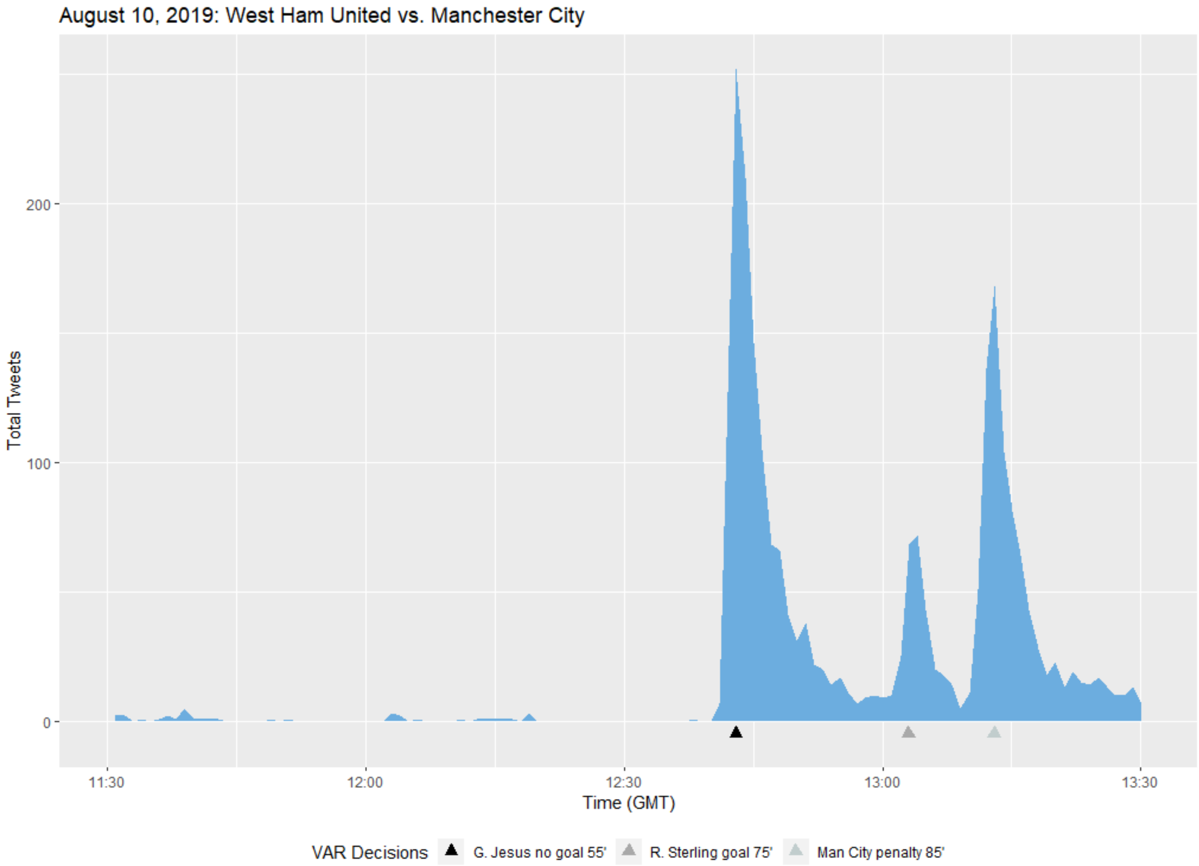
Figure 10: *Twitter sentiment by number of followers*



In-game Analysis:

By using Twitter data from specific games, sentiment can be measured over the course of a match. We can see how fans react when a VAR call is made and evaluate whether a call may have been the correct one. *Figure 11* shows the breakdown of tweets during a selected match from the 2019-2020 English Premier League season. The match was between West Ham United and Manchester City and will go down in Premier League history as the first match to utilize VAR technology. As can be seen in the plot, there was very little talk on Twitter about the use of VAR during the match during the first half as the technology was not needed. A huge spike in tweets can be seen after the 55th minute when Manchester City striker Gabriel Jesus scored the opening goal after a clever assist from Raheem Sterling. VAR was needed to check for offsides in the build-up to the goal and many on Twitter were excited to talk about the use of the technology. Controversially, Sterling was called offsides (see *Figure 1*) and the goal was overturned. Two more instances of VAR were needed during the match including a less-controversial Sterling goal and a penalty retake by Manchester City. Both these instances saw spikes in Twitter activity which was expected.

Figure 11: *Breakdown of tweets during the August 10, 2019 match between West Ham United and Manchester City, the first match in the Premier League where a VAR decision was made*



We can go one step further by analyzing sentiment using the NRC Lexicon to evaluate positive and negative emotions which can be seen in *Figure 12* and *Figure 13*. The same spikes in sentiment occurred during the VAR decisions during the match. It is clear that surprise does not play a significant role in terms of sentiment during these VAR decisions as it does not reach the same high sentiment value as the other categories. Fans did reach high peaks of anger, sadness, and disgust during the first VAR decision, suggesting that this was a bad call. It was a very tight decision and Manchester City supporters as well as neutrals may have disliked the nature of the call. They were less concerned with the second call and reached higher levels of positivity rather than negativity. Interestingly enough, fans reached a peak in terms of fear after the third VAR decision, possibly because of the way the call was made. Manchester City were allowed to retake their penalty because of Declan Rice's early entry into the penalty box. Because such a call had rarely been made before the implementation of VAR, fans may have been fearful that the use of VAR would completely change how matches would be officiated.

Figure 12: *Breakdown of tweets by negative sentiment types during the match*

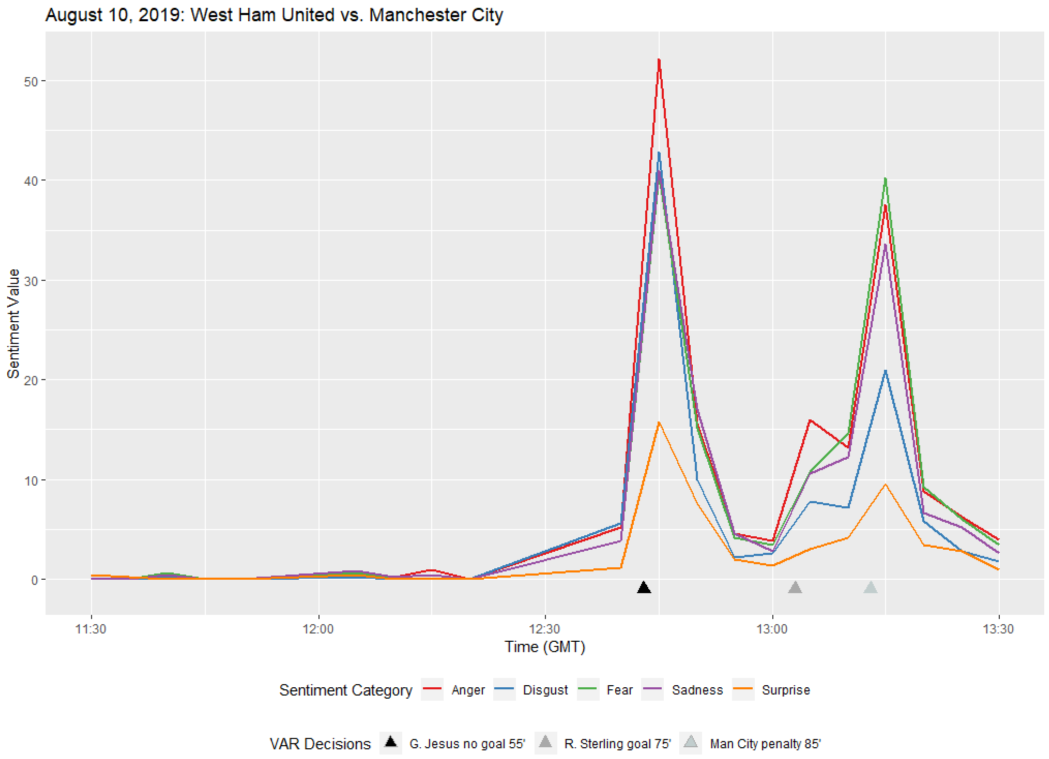
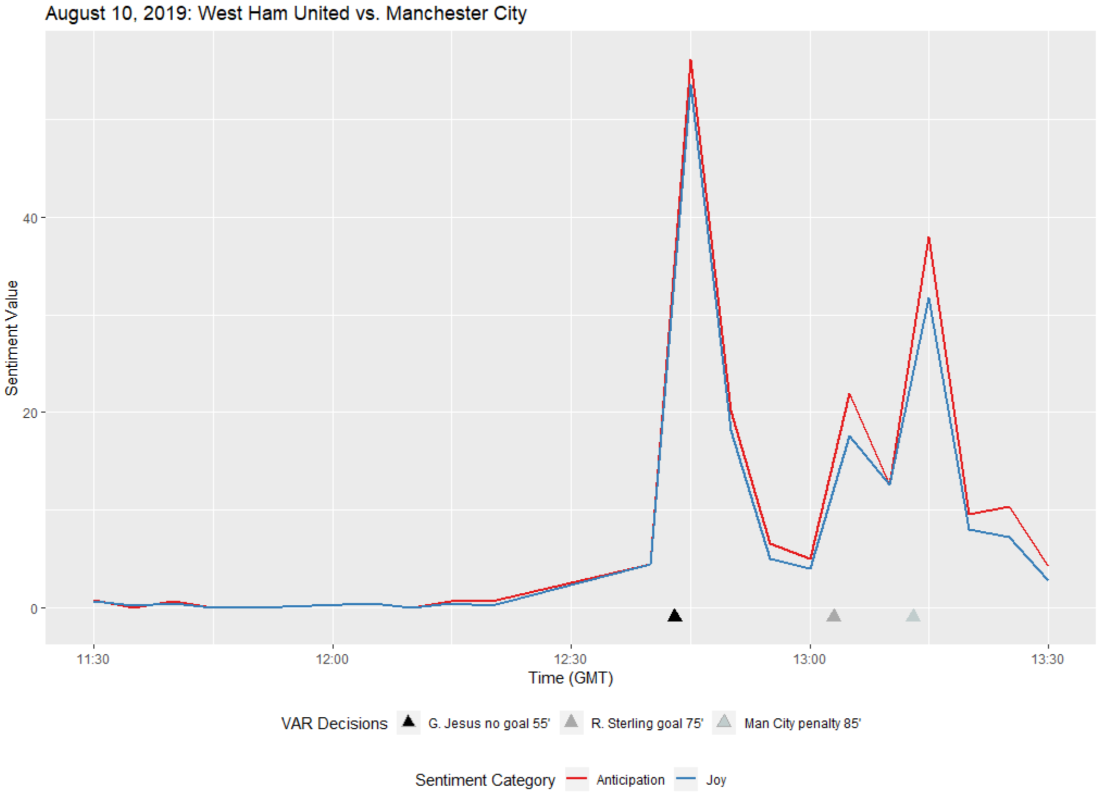


Figure 13: *Breakdown of tweets by positive sentiment types during the match*



Natural Language Processing: Predicting Sentiment

Now that we have analyzed fan sentiment on Twitter through the use of lexicons such as Bing and NRC, I wanted to utilize Natural Language Processing techniques to create a model to predict the sentiment of future tweets concerning VAR. To train my model, I used Twitter data from the SemEval 2014 competition which consisted of approximately 10,000 tweets that were categorized as either positive, negative, or neutral. I ran a Naïve Bayes Classification model in Python with the NLTK package using feature sets to predict the sentiment of a given tweet.

To create my model, I first tokenized the tweets by splitting each phrase into its corresponding words. For each of these strings of tokens, I then removed common stop words (ex. "the", "a", "and") and negation words (ex. "no", "not", "never") that would not have too big an impact on sentiment. Next, I found the 1500 most common words in the SemEval tweets and used those as word features. I originally used the 2000 most common words, but changed to 1500 because I noticed in previous steps that some of the most common words were not used very frequently. The vocabulary used on Twitter is more expansive than that used in normal writing because Twitter uses emoticons, handles, and slang terms. I then wanted to filter out a more specific list of stop words that were more relevant to Twitter. I utilized a list of texting and online chat abbreviations on Webopedia[11] which I filtered out of the 1500 most common features. Then, I used some of the options in the TweetTokenizer[12] function in NLTK to try and improve my model. The default of the "strip_handles" parameter in the TweetTokenizer function is False, but when changed to True, all Twitter handles are removed from the tweet. This removes the Twitter handles in replies to other tweets which are not necessary in this analysis. The default of the "red_length" parameter is False, but when changed to True, all words with more than three consecutive letters are trimmed down to a maximum of three (ex. "waaaaaayyyyyyy" changes to "waaayyy"). This was useful in this analysis as it increased the frequency of some terms and caused them to become common features. Finally, I converted all of the words to lowercase and removed all hyperlinks that would not be helpful in the analysis. I also tried removing hashtags, emoticons, punctuation, and at-symbols, but this led to decreased accuracy in my model. At the end of this preprocessing stage, my model was 58% accurate in predicting the sentiment of a tweet, which is very good, but could be improved even more.

In the experimental portion of my feature engineering, I utilized several popular sentiment lexicons used in Natural Language Processing. I first used the Subjectivity lexicon which analyzes the polarity of subjective expressions. Expressions were labeled as positive, negative or neutral by annotators in the Subjectivity lexicon. I added a "positivecount", "negativecount", and "neutralcount" feature for each word in a given tweet that was also in the Subjectivity lexicon at the noted polarity. Next, I used the LIWC (Linguistic Inquiry and Word Count) Lexicon which lists positive and negative words in a similar fashion to the Bing Lexicon. I did this to catch some of the words that were not in the Subjectivity lexicon. If the word was contained in the positive lexicon, the "positivecount" feature would increase by one, while one would be added to the "negativecount" for the negative words. Finally, I used the VADER[13] (Valence Aware Dictionary and sEntiment Reasoner) Lexicon which specializes in sentiment analysis in social media. In Python, VADER can be used to return a positive, negative, neutral, and compound (aggregate score of sentiment in other popular lexicons) score for a given word or phrase. The sentiment scores for positive, negative, and neutral all add up to one. VADER is useful for Twitter data because it deals with emoticons, punctuation, and slang terms. For the emoticon :) a positive sentiment score of 1 is returned, and for the emoticon :(a negative sentiment score of 1 is returned. Also, the sentiment score for punctuation can differ depending on the number of punctuation symbols in a row (ex. ! is different than !!!). Finally, slang terms that would not be recognized in a sentiment lexicon like Subjectivity or LIWC would be picked up in the VADER lexicon. I added the VADER lexicon to the feature set function by taking the highest sentiment score for each word in a tweet and adding one to the respective feature set ("positivecount", "negativecount", "neutralcount"). For example, the sentiment scores for the word "happy" are {'neg' : 0.0, 'neu' : 0.0, 'pos' : 1.0}, so one would be added to the positive count feature set. I also tried to use part of speech feature sets, but part of speech tagging is difficult with Twitter data. I also attempted to use bigram feature sets, but there were not enough common bigrams to have a significant effect.

In my preliminary analysis of the model, I split the tagged tweets so 90% were in the training set and 10% were in the test set. I then ran the Naïve Bayes Classifier on the training data and evaluated the 30 most informative features.

Table 3: *The 30 most informative features in the Naïve Bayes Classification model*

Feature	Relationship	Feature	Relationship
f***	neg : neu = 66.4 : 1	negativecount = 9	neg : neu = 19.5 : 1
:)	pos : neg = 51.4 : 1	thanks	pos : neu = 19.4 : 1
fun	pos : neu = 49.5 : 1	missing	neg : pos = 18.3 : 1
excited	pos : neu = 47.2 : 1	negativecount = 10	neg : neu = 18.1 : 1
sad	neg : neu = 43.9 : 1	cool	pos : neu = 16.8 : 1
sorry	neg : neu = 43.9 : 1	cant	pos : neu = 15.1 : 1
positivecount = 12	pos : neu = 42.5 : 1	brilliant	pos : neu = 15.1 : 1
great	pos : neu = 38.9 : 1	funny	pos : neu = 14.2 : 1
happy	pos : neu = 35.7 : 1	exciting	pos : neu = 14.2 : 1
amazing	pos : neu = 35.7 : 1	wrong	neg : neu = 14.2 : 1
luck	pos : neu = 28.8 : 1	hate	neg : neu = 13.9 : 1
thank	pos : neu = 25.4 : 1	negativecount = 8	neg : pos = 13.8 : 1
:(	neg : pos = 25.3 : 1	can't	neg : neu = 13.4 : 1
injury	neg : pos = 23.5 : 1	interesting	pos : neu = 13.3 : 1
awesome	pos : neu = 21.9 : 1	pavol	neg : pos = 13.1 : 1

Some of the top features are what we would expect in such a classifier. The :) emoticon and words like "fun", "excited", and "great" led to tweets with more positive sentiment than negative. Words like the f-word, "sad", and "sorry" led to more negative sentiment. Tweets with 12 positive words were usually tagged as positive and tweets with 8-10 negative words were usually tagged as negative. This shows that although there are certain words with strong sentiment that are good predictors of overall tweet sentiment, the number of positive or negative words in a tweet can also be a tool to evaluate sentiment.

I then used cross-validation to find precision (positive predicted value), recall (true positive rate), and F1 (balance between precision and recall) scores. Using 10 folds, I found a mean accuracy of 0.687 or 68.7%.

Table 4: *Sentiment model accuracy scores*

Sentiment Category	Precision	Recall	F1
Neutral	0.693	0.660	0.676
Negative	0.490	0.500	0.495
Positive	0.658	0.695	0.676

As can be seen, the classifier performed worse on tweets with negative sentiment than those with positive or neutral. This is most likely the case because it is harder to pick up on sarcasm in tweets. Overall, the model performed very well and can be used to predict the sentiment of the tweets about VAR in the English Premier League.

After testing the model using cross validation, I wanted to test it on the game-specific tweets from the English Premier League to evaluate the predicted sentiment. *Figure 14* shows that the model predicted a large number of neutral sentiment tweets during the season, followed by negative sentiment and finally positive sentiment. The results are as expected as many tweets about the use of VAR simply state that a decision was made without any additional opinions about the technology. The sentiment model, like most models, struggles to pick up sarcasm in a tweet, and may classify these tweets as neutral because they have positive text with a negative underlying message.

Figure 14: *Predicted sentiment for every game-specific tweet from the first half of the 2019-2020 English Premier League season*

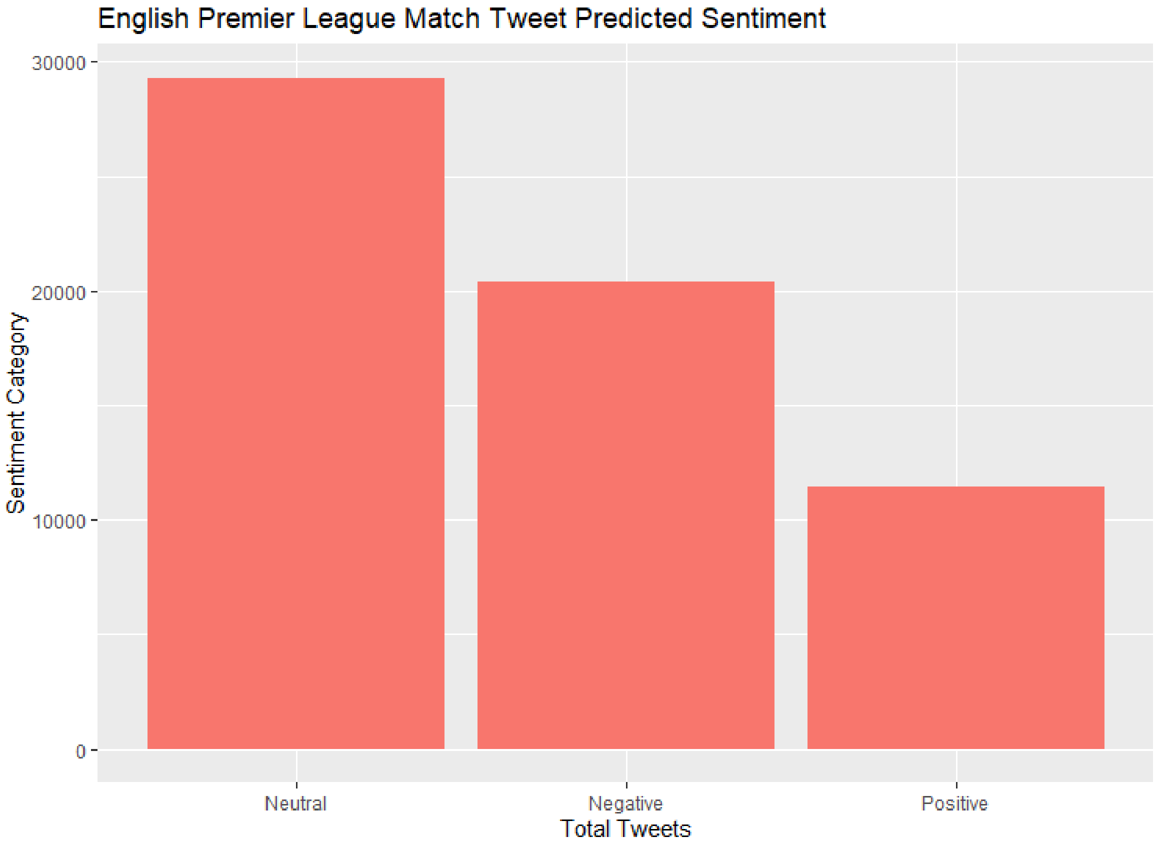


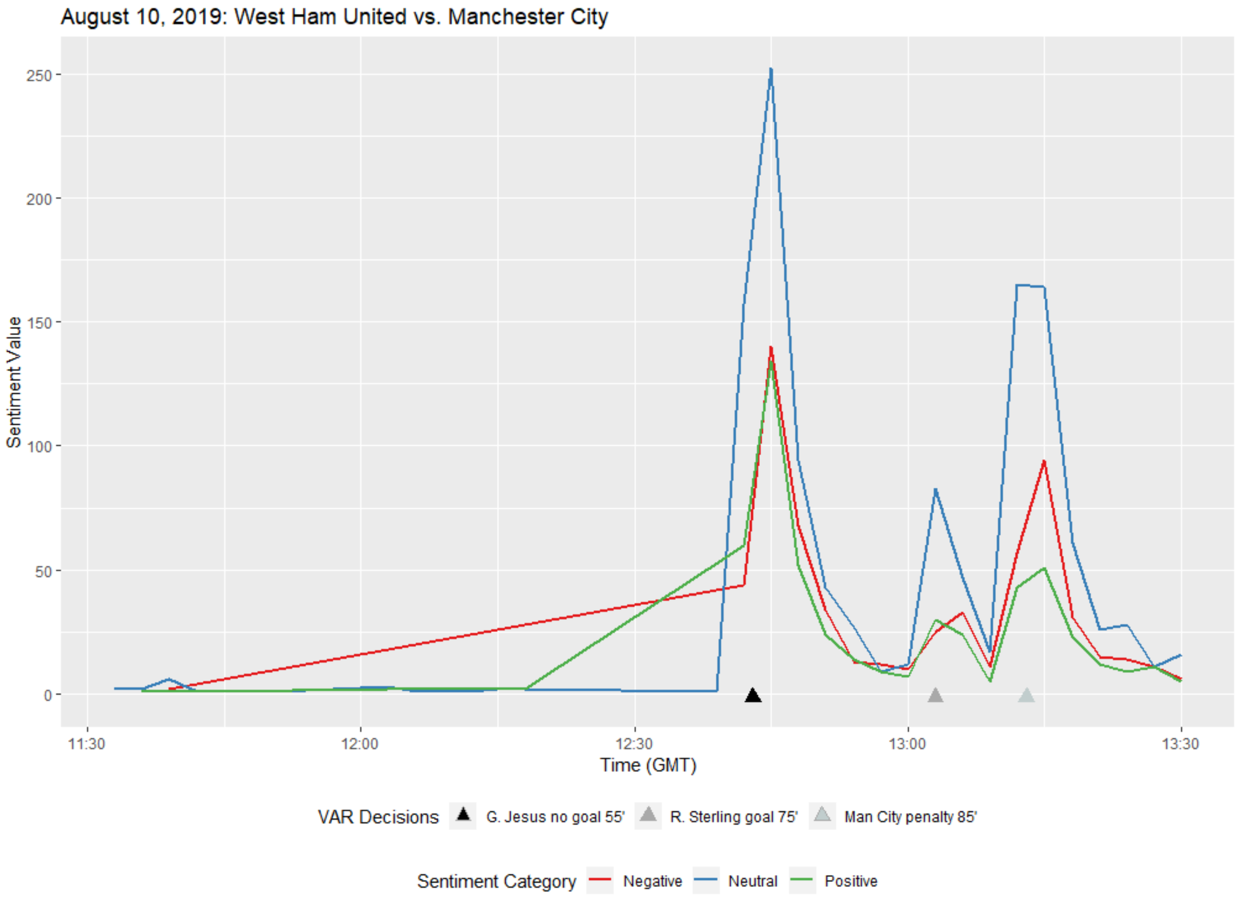
Table 5 shows several examples of tweets that were classified by the model. The positive, neutral, and negative tweet examples seem like they were correctly classified given the context of the tweets. The positive tweet says that VAR has been used correctly in the Premier League and shows belief in the technology. The neutral tweet simply states that a VAR decision has been made giving a goal to Arsenal and has no sentiment associated with it. The negative tweet states that the VAR decision was incorrect and that particular of the game does not need to be officiated so closely. The table also shows examples of team fans that may be biased in their sentiment towards VAR in given scenarios. The positive home fan praises the use of VAR which has ruled out a goal for Wolves and keeps the match at 0-0. The negative away fan is disgusted by the use of VAR in the Manchester City vs. Tottenham match that ended level after a Gabriel Jesus winner was ruled out by the technology. Both tweets show sentiment toward VAR in support of their favorite teams and does not tell us if the call on the field was right or wrong. Finally, the positive sarcastic tweet shows where the model can incorrectly classify a tweet. The tweet says VAR is brilliant, but that opinion may change several times throughout a match. The tweet seems to have positive sentiment towards VAR, but in reality, it would be better classified as negative or neutral because of the context surrounding the tweet.

Table 5: *Examples of classified tweets*

Sentiment	Tweet
Positive	"I actually think they've done VAR right in the PL. You can tell they've done it with speed in mind. Not perfect but a start. #TOTAVL"
Positive Home Fan	"We love you VAR we do We love you VAR we do We love you VAR we do Oh VAR we love you #LCFC #LEIWOL"
Positive Sarcasm	"#VAR is clearly brilliant *#LEIWOL*opinion may change several times a game"
Neutral	"VAR say goal, 1-1 #MUNARS"
Negative	"Ew. Not fond of the VAR involvement in the penalty. Doesn’t seem like that’s an area of the game crying for oversight. #WHUMCI"
Negative Away Fan	"…I said and I will say it again, VAR is the perfect weapon to stop Manchester City."

I also wanted to take a look at the breakdown of tweets during a match to see if there was more negative sentiment after VAR decisions. If we look at the Manchester City vs. West Ham game once more, we can see that there were large spikes in neutral sentiment after each VAR decision. This is expected as there would be many tweets reporting that a VAR call was made, but not as many criticizing or praising the technology. It can be seen that the first decision was split evenly in terms of positive and negative sentiment suggesting that fans were possibly undecided on whether or not the decision was correct which is a bit different than what the NRC Lexicon suggested. The decision was a very tight offsides call and was the first VAR decision in Premier League history, which may have led to the indecision among fans. The second decision was very similar to the first, but the third call was met with more negative sentiment than positive. Fans were not pleased with the third call because of the nature of the decision which led to the Manchester City penalty being retaken for a minor infringement. This information can be very useful to the Premier League as they continue to improve the use of VAR. If fans do not want to see VAR used to retake penalties for slight violations of the rules, the Premier League may want to rethink their use of VAR in such situations.

Figure 15: *Predicted sentiment throughout the Manchester City vs. West Ham United match*



Future Work

In future work with this project, I will continue to analyze the sentiment associated with VAR decisions in professional soccer. Through the use of my own sentiment model and the Bing and NRC lexicons, I hope to be able to classify VAR decisions into good and bad calls to evaluate how the Premier League can change the review system to improve the game of soccer. I will then be able to analyze specific VAR decisions that were deemed incorrect and investigate any similar traits among them. I will also analyze VAR decisions in other top professional leagues and competitions to see how the use of VAR differs. I will look into any features that may lead to more or less VAR calls such as the time remaining in the match and the score. I will also seek to find any biases in VAR decisions such as home-field advantage, favoring of the "better" or statistically favored team, and the referees themselves.

Conclusion

In conclusion, I have utilized Twitter data to analyze fan sentiment associated with Video Assistant Referee decisions in professional soccer. I first found that VAR decisions during professional soccer matches does in fact increase social media activity. As the number of VAR decisions in a match increases, we would generally expect the number of tweets about VAR to increase as well. I also found that the sentiment of tweets about VAR decisions is more negative rather than positive or neutral. Through the use of the Bing and NRC Lexicons it is clear that fans show more negativity towards the use of VAR than positivity or indifference. I was also successful in creating a model to predict the sentiment associated with a given tweet concerning VAR through the use of Natural Language Processing. The model helped to emphasize the negative nature of tweets about VAR in the English Premier League.

By looking at tweets at a game-level, patterns arose about the negative sentiment associated with VAR. Many of the tweets with high levels of negative sentiment occurred in three aspects of the game. First, fans were upset with some of the extremely tight offsides decisions such as the ones shown in *Figure 1*. Offsides has been classified by the Premier League as a "black and white" decision that will not be upheld due to a lack of "clear and obvious" evidence. As a result, attackers have been called offsides by less than three centimeters. Although the technology used by VAR is advanced and technically sound in terms of determining the body position of players in three dimensions, the frame rate of the cameras used by VAR will always lead to some margin of error. The determination of when the ball was kicked as well as the furthest point of the player is a bit subjective, and small differences in the offside line placement can rule out important goals. Second, fans were not pleased with decisions that involved the retake of a penalty due to early entry into the penalty box by a defender or the goalkeeper being off his line. Such decisions were rarely made before the introduction of VAR because of the difficulty referees had watching every player during a penalty kick. Fans believe this type of decision, much like a tight offsides call, is a case of "over-refereeing" the game. Soccer purists would much rather a call not be made just as it was before the introduction of VAR. Even though the call may be the correct decision, fans are discouraged by the extra time wasted to check VAR on these tight decisions that may not have a big impact on the play. Finally, fans have not been happy with the consistency of VAR on calling penalties. Some key penalties in big matches were not called during the first half of the season that led to supporters criticizing the technology. This subjectivity is part of the VAR technology and will continue to be a part of the game. However, referees must stay consistent with their use of VAR on penalties throughout the season to avoid any potential bias.

My main suggestion to the Premier League and other professional soccer leagues using VAR would be to put regulations in place to limit the number of tight decisions being made in matches. One suggestion would be to change offsides decisions only if a "clear and obvious" error has been made. Liverpool manager Jürgen Klopp has proposed making the offsides line thicker to create a greater margin of error for the attacker. If a set margin of error is listed for offsides decisions, then there would be less controversy surrounding the calls. Another suggestion would be to have more leniency on early intrusion into the penalty box. If the player clearly enters the box early or the goalkeeper is well off his line before the kick is taken, then the penalty should be retaken, but if the encroachment is minor then a call will not be made. Finally, a time limit should be set on VAR decisions to keep the pace of play fast in professional soccer. For example, if a decision cannot be made in 30-45 seconds, then the call on the field should stand. Some may argue for a longer time frame in key scenarios, but if such a call takes longer than 30-45 seconds to make, then it most likely was called correctly on the field the first time. Such implementations to the VAR technology could satisfy the demands of fans and players alike and lead to improved ratings for the Premier League. VAR technology has already improved the game of soccer, and by repairing the flaws in the system, it will continue to help the beautiful game.

References

1. Goff, S. (2019, June 26). Analysis | VAR is working at World Cup, FIFA says. That's not wrong, but it could work better. Retrieved September 10, 2019, from https://www.washingtonpost.com/sports/2019/06/26/var-is-working-world-cup-fifa-says-thats-not-wrong-it-could-work-better/?noredirect=on.
2. Johnson, D. (2019, August 10). VAR in the Premier League: The big decisions explained. Retrieved August 13, 2019, from https://www.espn.com/soccer/english-premier-league/story/3917260/var-in-the-premier-league-the-big-decisions-explained.
3. VAR in the Premier League: How did first weekend go for technology in top flight? (2019, August 11). Retrieved August 13, 2019, from https://www.bbc.com/sport/football/49307496.
4. Yu, Y., & Wang, X. (2015, February 20). World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets. Retrieved October 3, 2019, from https://www.sciencedirect.com/science/article/pii/S074756321500103X.
5. Corney, D., Martin, C., & Goker, A. (2014, April 13). Spot the Ball: Detecting Sports Events on Twitter. Retrieved October 3, 2019, from http://dcorney.com/papers/ECIR_SpotTheBall.pdf.
6. Gratch, J., Lucas, G., Malandrakis, N., Szablowski, E., Fessler, E., & Nichols, J. (2015, September 21). GOAALLL!: Using Sentiment in the World Cup to Explore Theories of Emotion. Retrieved October 5, 2019, from http://ict.usc.edu/pubs/GOAALLL! Using Sentiment in the World Cup to Explore Theories of Emotion.pdf.
7. Taspinar, A. (2019, November 4). TwitterScraper. Retrieved December 20, 2019, from https://github.com/taspinar/twitterscraper
8. Yurko, R. (2018, July 5). FCScrapR. Retrieved August 1, 2019, from https://github.com/ryurko/fcscrapR
9. Wasser, L., & Farmer, C. (2017, April 19). Sentiment Analysis of Colorado Flood Tweets in R. Retrieved January 8, 2020, from https://www.earthdatascience.org/courses/earth-analytics/get-data-using-apis/sentiment-analysis-of-twitter-data-r/
10. Mejova, Y., Srinivasan, P., & Boynton, B. (2013, February 1). GOP Primary Season on Twitter: "Popular" Political Sentiment in Social Media. Retrieved January 14, 2020, from https://dl.acm.org/doi/10.1145/2433396.2433463
11. Beal, V. (n.d.). Huge List of Texting and Online Chat Abbreviations. Retrieved September 15, 2019, from https://www.webopedia.com/quick_ref/textmessageabbreviations.asp
12. TweetTokenizer. (n.d.). Retrieved September 15, 2019, from https://kite.com/python/docs/nltk.TweetTokenizer
13. Hutto, C. J. (2019, September 6). VADER Sentiment Analysis. Retrieved September 15, 2019, from https://github.com/cjhutto/vaderSentiment