Classifying Soccer Tweets

Stats 133 Final Project

Nathan Kim[1] and Jun Ryu[1]

[1]Statistics & Data Science, University of California, Los Angeles (UCLA)

24th March 2024

Abstract

The goal of this project was to analyze the collection of tweets from 7 European Soccer Clubs (AFC Ajax, AC Milan, FC Barcelona, FC Bayern, Juventus FC, Liverpool FC, Real Madrid) between January 2019 and July 2019. Specifically, we wanted to see the various tones/sentiments of the clubs' tweets and determine how accurately we can predict the origin of the tweet based on its textual content. To conduct the analysis, we cleaned the data, found the term frequencies, ran sentiment analysis, and applied both unsupervised and supervised machine learning.

1. Introduction

Soccer clubs have utilized social media to communicate news with fans worldwide. Some of the most prestigious clubs have a multitude of international fan bases that they can run Twitter accounts in different languages for their fans. Clubs want to share a variety of news with their fans from club announcements to training session media, game highlights, and post-game reflections and conferences. Clubs will post similar content across different platforms but in different languages.

As clubs want to connect with their international fans, they attempt to maintain a positive tone as the season progresses, regardless of how successful the club is whether at winning its domestic league competition or earning prestigious continental trophies. Clubs want to motivate their fans and display that the players and the club are working hard towards winning trophies. During a season, we can analyze the meaning behind a club's tweet to determine how positive a club's tweet is and what terminologies the club has utilized frequently, whether it's the team's nickname or mascot, coach or player names, or competition that clubs are competing in.

With our soccer club Kaggle project, we were given tweets from 7 clubs, AFC Ajax, AC Milan, FC Bayern, FC Barcelona, Juventus FC, and Real Madrid, from 2018 to 2019. The dataset also included Twitter feeds from clubs' youth teams, women's teams,

or international accounts in different languages. With the clubs' tweets, we were given other characteristics, such as the number of followers a Twitter handle has or follows, the hashtags, source, and official descriptions. After we filter our data and select the only necessary columns and handles we need for our analysis, our goal is to analyze each club's Twitter messages to see what terms are frequently used, the sentiments behind a club's message, and determine whether we can predict the source of a tweet with a Random Forest model.

2. Data Cleaning

2.1 Initial Filtering

The unmodified version of the data was pulled from Kaggle[1]. With this raw data, we first wanted to grab only the data of interest: the tweets from the 7 official European clubs within the date range of January 2019 and July 2019 (the end of the data collection). The original dataset contained 131234 rows and 82 columns; however, after conducting the initial filtering, the dataset was reduced to 20883 rows and 9 columns. The 9 columns included information about the date the tweet was posted, the full text of the tweet, the official username of the club, and other metadata related to the tweet itself.

2.2 Transformations in Quanteda

After the proper data was selected, we needed to conduct further cleaning and apply various transformations on the tweets themselves to keep only the relevant information that would not affect our analysis. For example, having stop words such as "and" and "is" will not contribute much to sentiment analysis and will only hurt the computation complexity when running machine learning models. Thus, to prepare the text for transformations, we first created a corpus out of all the tweets and then created a list of tokens for each document (tweet) by splitting up each text into individual "words". Then, we proceeded with the following steps to extract only the necessary tokens:

1. Remove Punctuation, Numbers, Symbols (emojis)
2. Remove Links, Mentions, and Hashtags using Regular Expressions
3. Turn all Tokens into Lowercase
4. Remove Stopwords in 5 different languages (English, Spanish, German, Italian, Dutch)
5. Clear Whitespace using Regular Expressions

[1] https://www.kaggle.com/datasets/eliasdabbas/european-football-soccer-clubs-tweets?select=clubs_tweets.csv

6. Remove Emojis and other Non-ASCII Characters

7. Remove Own Stopword: “rt” (meaning retweets) and “gol”

8. Apply Stemming

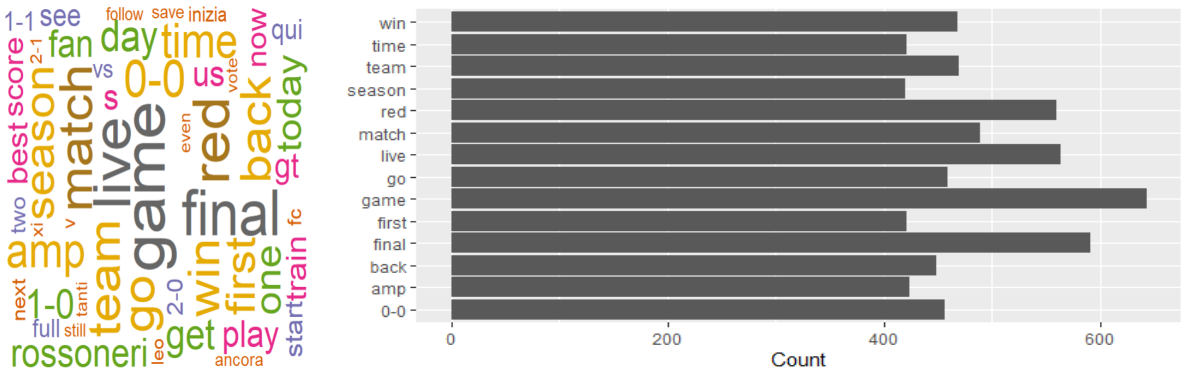
Before applying any transformations, the number of tokens across the 20883 documents was 414238. After applying the 8 steps as listed above, the number was cut significantly; the token count was 124723, only about 30% of the original amount.

Now that we completed the cleaning process above, we can analyze the Word Clouds and term frequencies per clubs to analyze which terms were frequently used per club.

3. Exploratory Data Analysis

3.1 Word Clouds/Term Frequencies

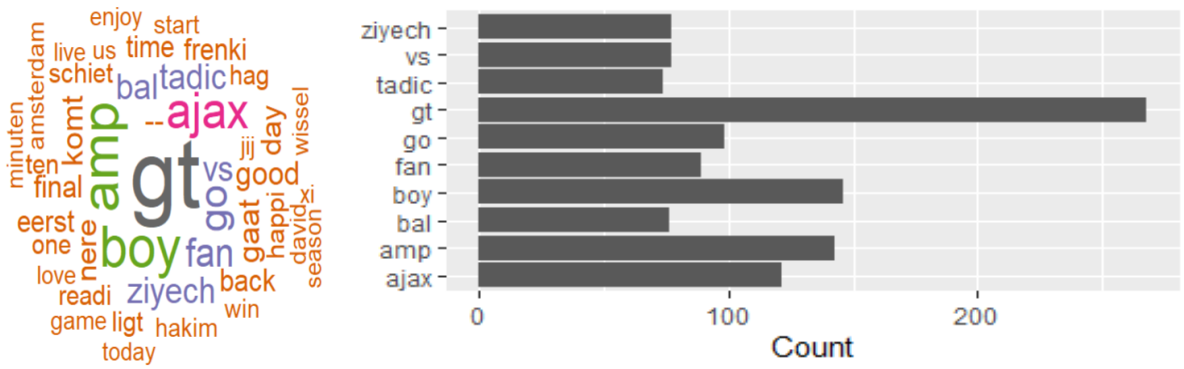
When analyzing the Word Clouds and term frequency charts for all clubs, some of the most common words were “live”, “game”, and “final”. These are the types of terms that clubs want to utilize to share information regarding their match training, performance, and results with their fans. Scores were also mentioned such as “1-0” or “1-1”. While these terms are common terms shared by all clubs, analyzing frequent terms by each club would display certain unique terms not observed in our charts below.



(a) (b)

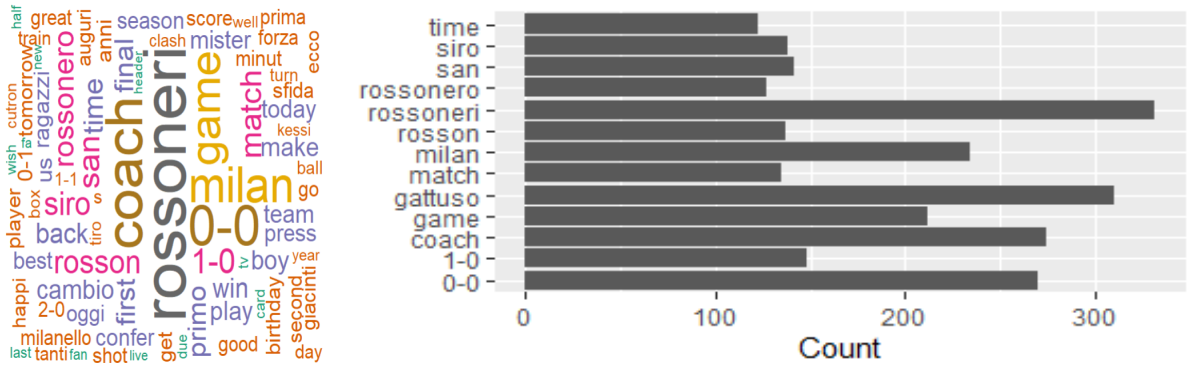
Figure 1: Plots for All Clubs: (a) Word Cloud, (b) Term Frequency

We see that for clubs such as AFC Ajax and AC Milan, clubs tend to utilize their team nicknames and player names frequently. For Ajax’s Word Cloud and term frequency chart, Ajax tweets mentioned terms such as “gt”, “ajax”, “boy”, and player names such as “ziyech” and “tadic”. For AC Milan, we see how the team’s nickname, “rossoneri”, was mentioned along with its coach “gattuso” and scores like “0-0” or “1-0”. From the two clubs’ tweet messages, AC Milan tends to highly mention its coach’s name, nickname, and team name in its tweet content, while Ajax frequently mention these two players’ names as they can be popular and the club’s star players.



(a) (b)

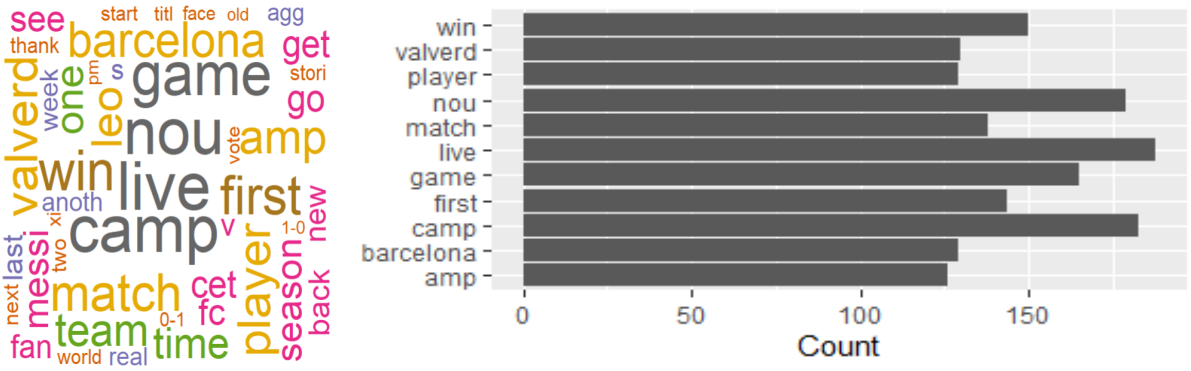
Figure 2: Plots for AFC Ajax: (a) Word Cloud, (b) Term Frequency



(a) (b)

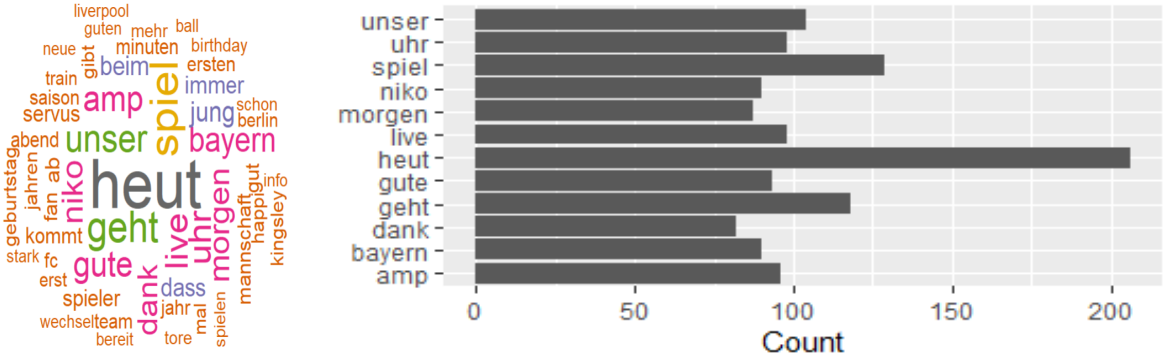
Figure 3: Plots for AC Milan: (a) Word Cloud, (b) Term Frequency

For FC Barcelona and FC Bayern Word Clouds and term frequency bar charts, Barcelona's home stadium name , "camp" and "nou" were mentioned along with player names such as "messi" and the coach's name, "valverd". Meanwhile, for FC Bayern, the club utilized German terms such as "heut", "geht", "spiel", and the word "live" as well. FC Barcelona seemed to frequently utilize English terms in its tweets, but Bayern's texts mostly involved German terms in its posts.



(a) (b)

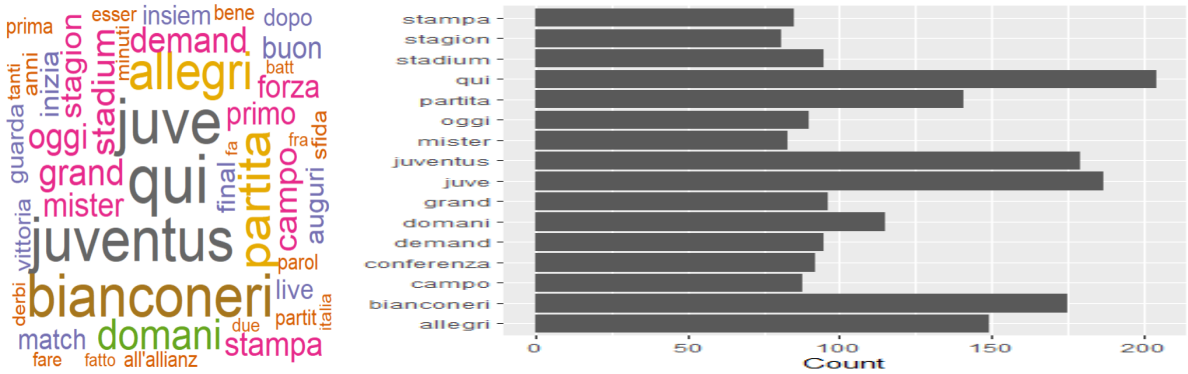
Figure 4: Plots for FC Barcelona: (a) Word Cloud, (b) Term Frequency



(a) (b)

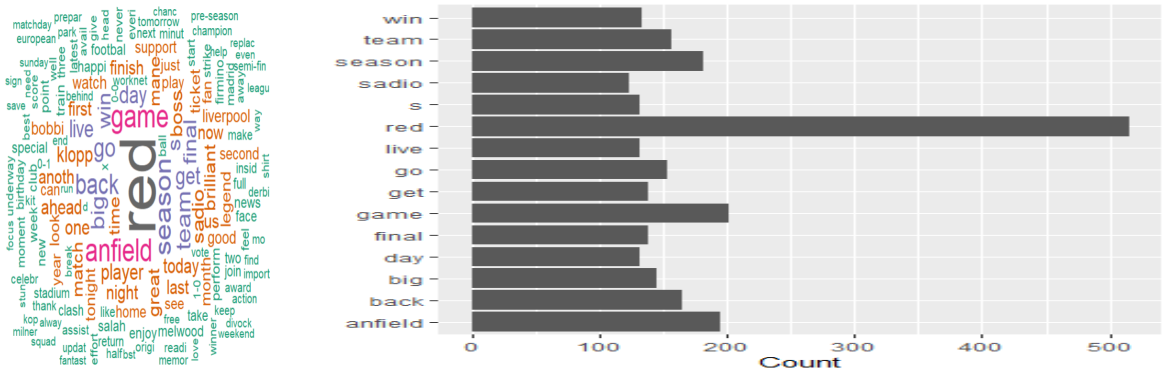
Figure 5: Plots for FC Bayern: (a) Word Cloud, (b) Term Frequency

With Juventus and Liverpool, Juventus often utilized many Italian terms in its tweets. Terms such as "bianconeri", their team nickname, "qui", "allegri", and their coach's names were mentioned frequently. Meanwhile, Liverpool mentions their team nickname "red" often in its tweets along with "game" and home ground, "anfield".



(a) (b)

Figure 6: Plots for Juventus FC: (a) Word Cloud, (b) Term Frequency

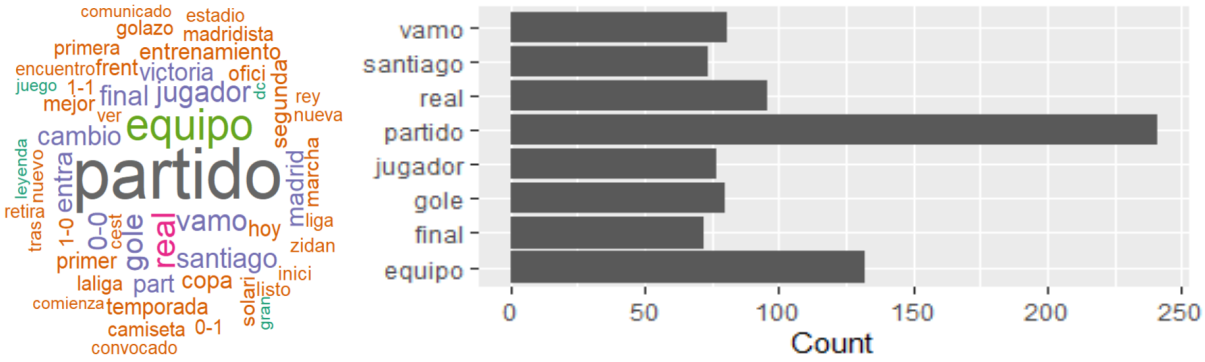


(a) (b)

Figure 7: Plots for Liverpool FC: (a) Word Cloud, (b) Term Frequency

Lastly, Real Madrid's frequent terms were mostly Spanish terms like "partido", match, and "equipo", team. Real Madrid tends to write its tweets with Spanish phrases, which

differs from its arch-rivals FC Barcelona who often use English words in its tweets.



(a) (b)

Figure 8: Plots for Real Madrid: (a) Word Cloud, (b) Term Frequency

Some of the most frequent types of terms that were mentioned in tweets were game-content, team name, and its home stadium names. Hence, when analyzing the tf-idf for clubs, certain unique terminologies were mentioned that did not appear under the term frequency charts. While some of the frequent terms did crossover with tf-idf and term frequency charts, the tf-idf table displays terms by term frequency and idf scores per club tweets. After analyzing the club's word clouds and term frequency bar charts, the next step would be to analyze the bigrams or the two terms that followed after each other from club tweet messages.

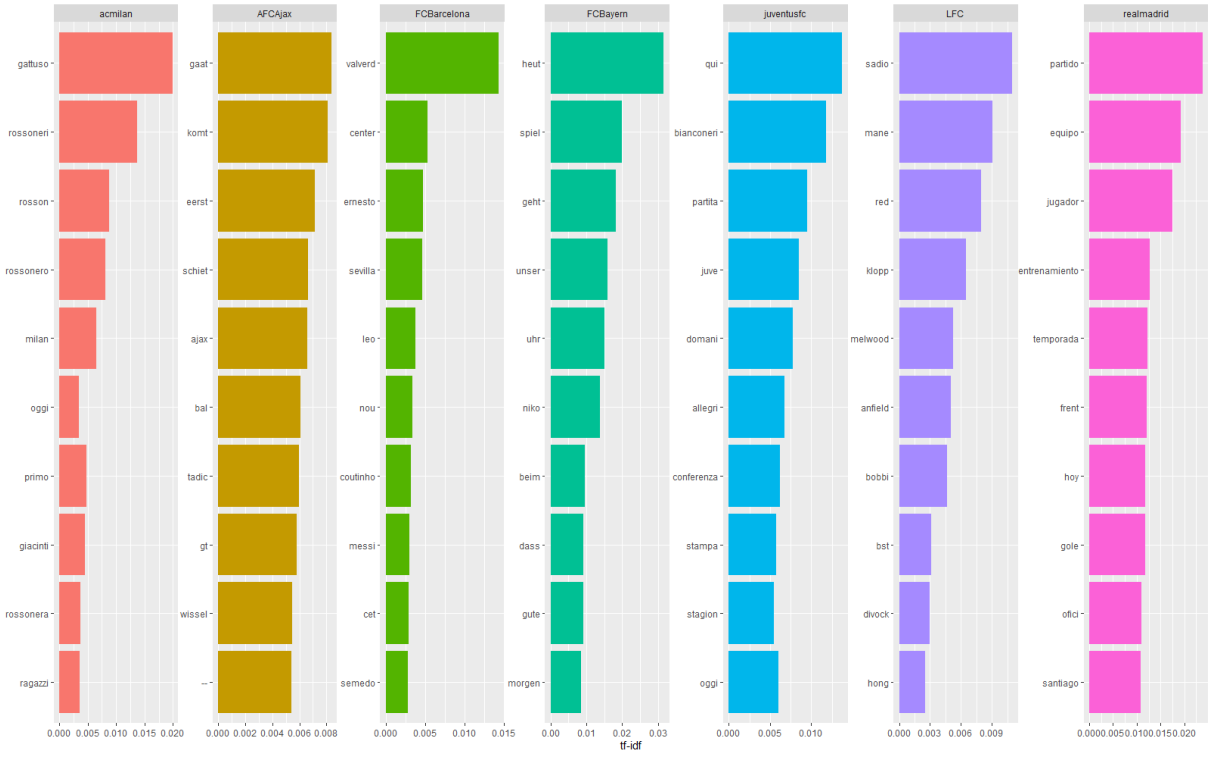


Figure 9: TF-IDF Plot for All Clubs

3.2 Bigrams

When analyzing clubs' bigrams, club names such as "real madrid" and "fc barcelona" were frequent, along with home grounds such as "san siro" and "camp nou". Italian terms were also quite frequent bigrams, like "tanti auguri" and "conferenza stampa".

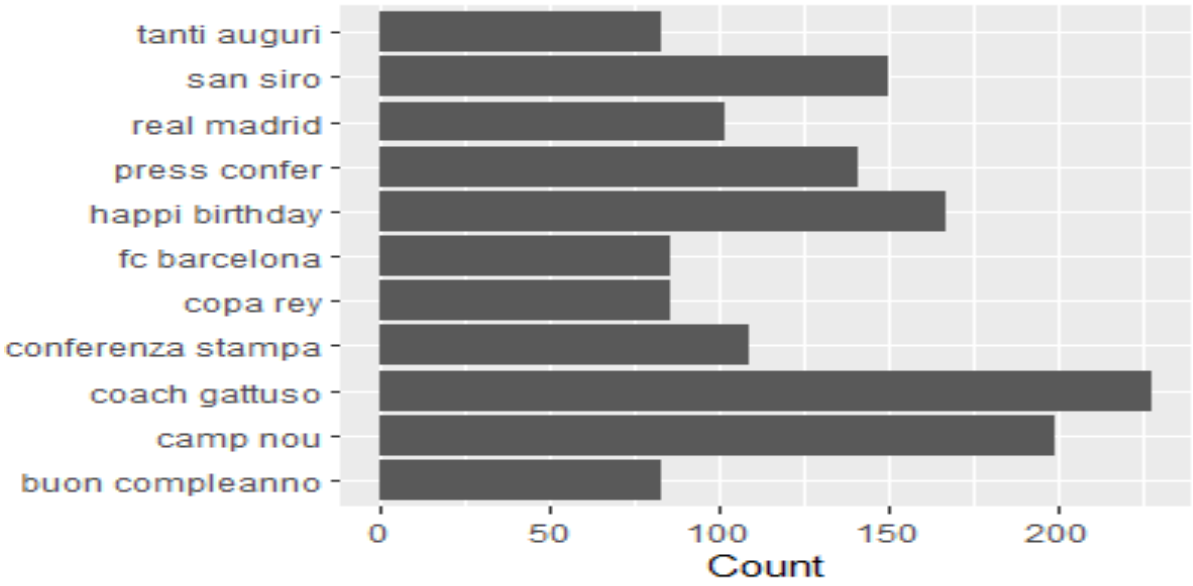
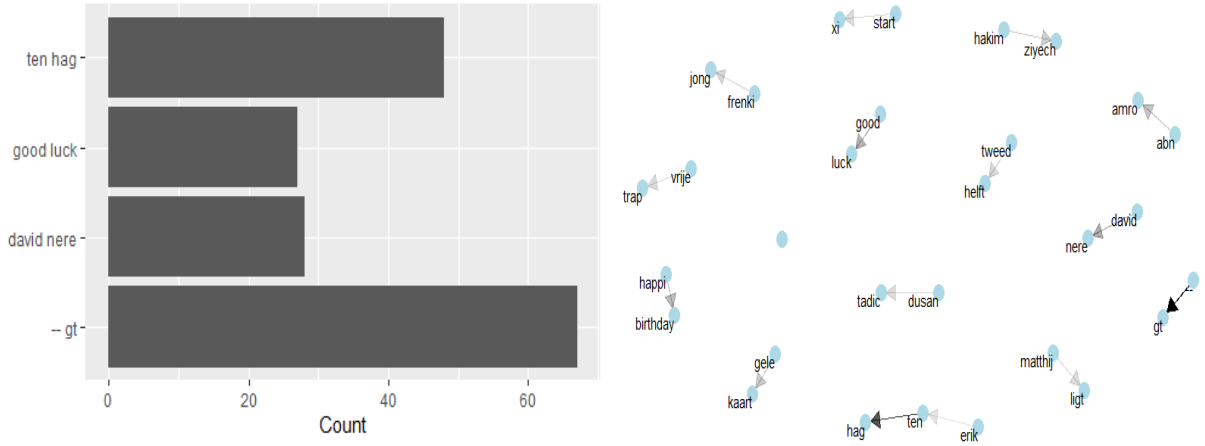


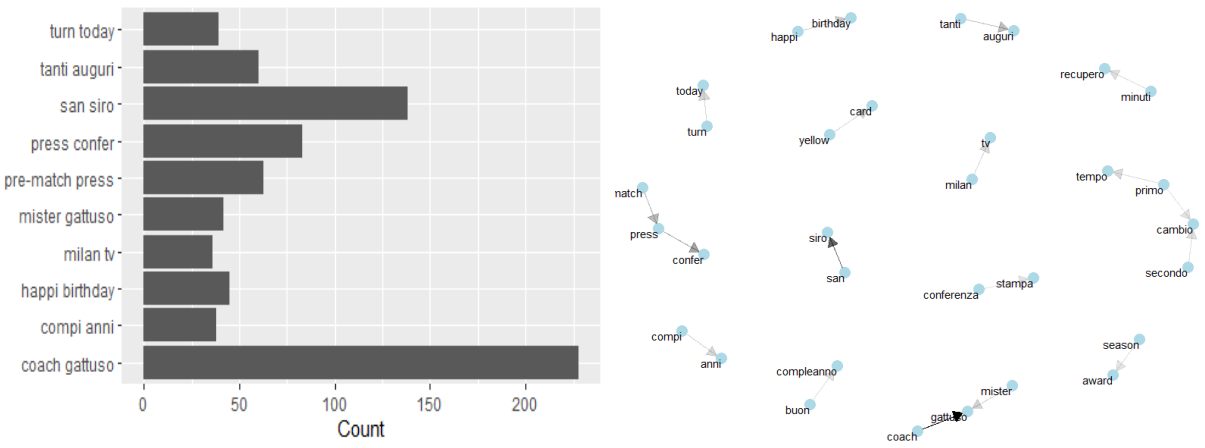
Figure 10: Plots for All Clubs: Bigram Count

For AFC Ajax, player and coach names were bigrams that were often mentioned. The coach's name "ten hag" and his players "david neres", "frenki jong", and "tadic dusan" are examples of the most frequent AFC Ajax bigram terms. In AC Milan tweets, we notice that the coach's name "coach gattuso" was quite popular and the club's home ground "san siro". English terms were also quite prevalent such as "press confer" and "pre-match press", terms that describe meeting with the press before or after a match.



(a) (b)

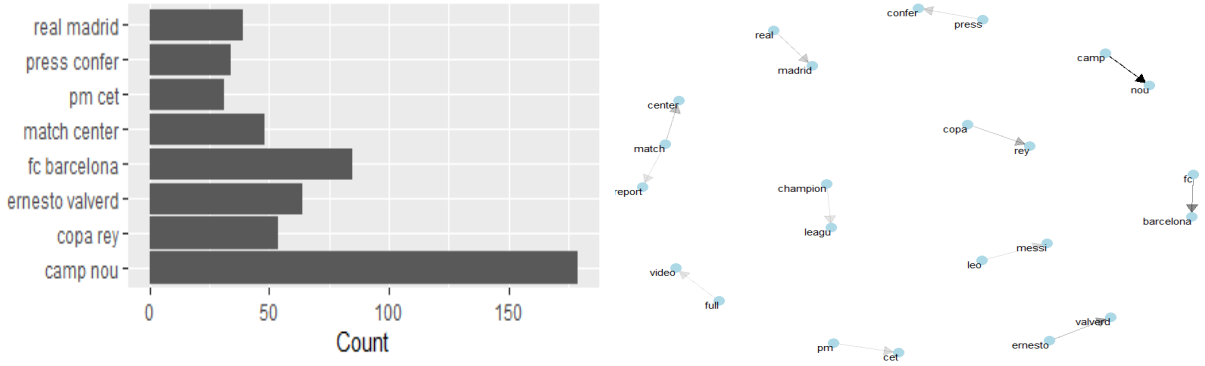
Figure 11: Plots for AFC Ajax: (a) Bigram Count, (b) Directed Graph



(a) (b)

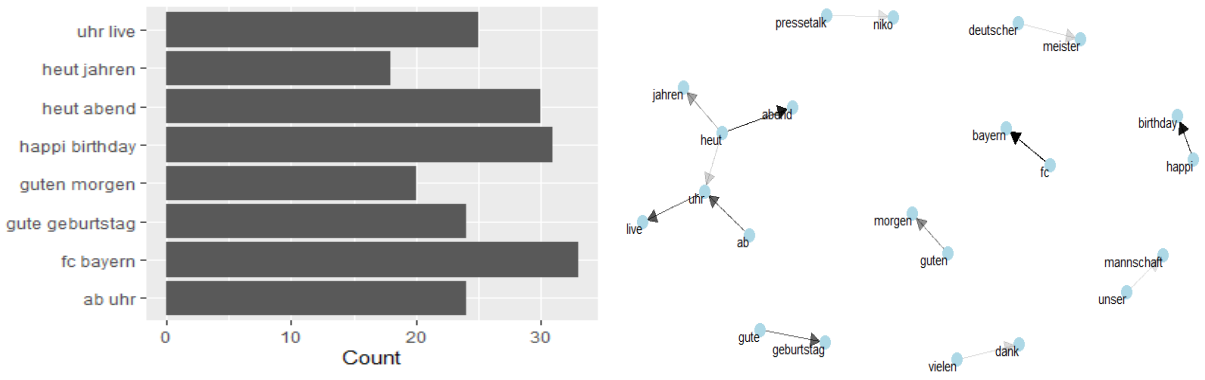
Figure 12: Plots for AC Milan: (a) Bigram Count, (b) Directed Graph

With FC Barcelona and FC Bayern, Barcelona included the name of its arch-rivals "real madrid" frequently in its tweets. This behavior displays how strongly they think about "real madrid" as their rivals. Other English terms were included such as "pm cet", "match center", and "press confer" for media-related matters. Bayern, on the other name, tends to display many German word combinations in its tweets.



(a) (b)

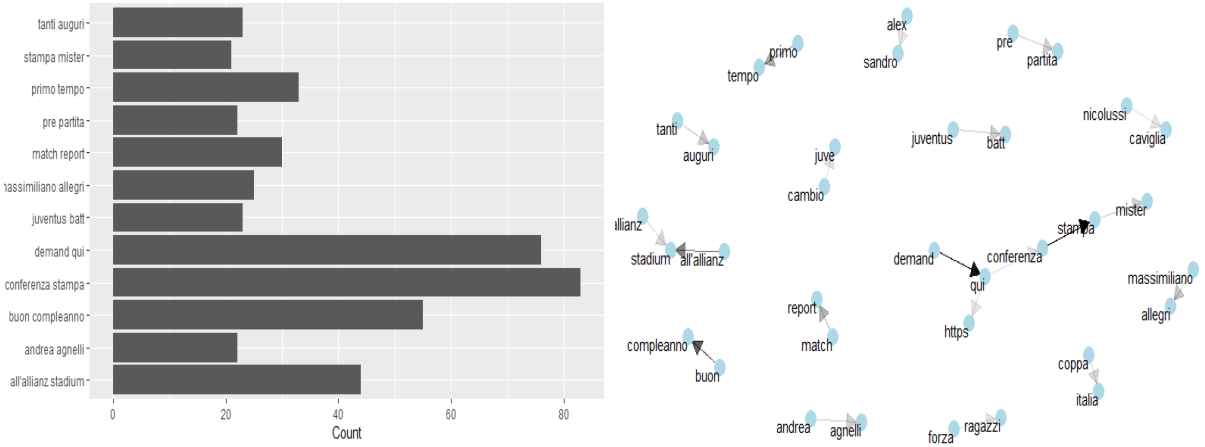
Figure 13: Plots for FC Barcelona: (a) Bigram Count, (b) Directed Graph



(a) (b)

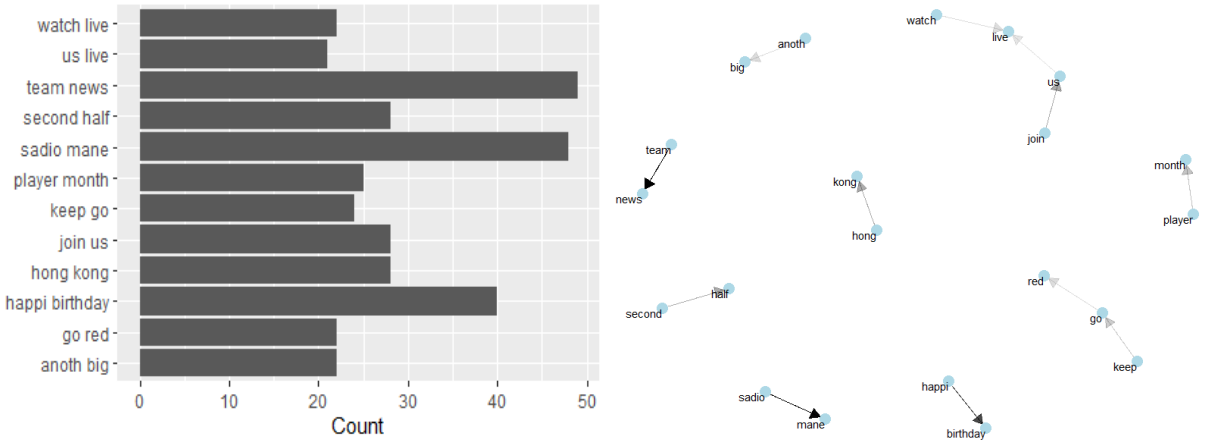
Figure 14: Plots for FC Bayern: (a) Bigram Count, (b) Directed Graph

With Juventus and Liverpool, Italian Bigram terms were highly frequent in its tweet messages, such as “primo temp” and “conferenza stampa”. For Liverpool, terms regarding the team and match such as “watch live” and “join us” were mentioned along with club players such as “sadio mane”.



(a) (b)

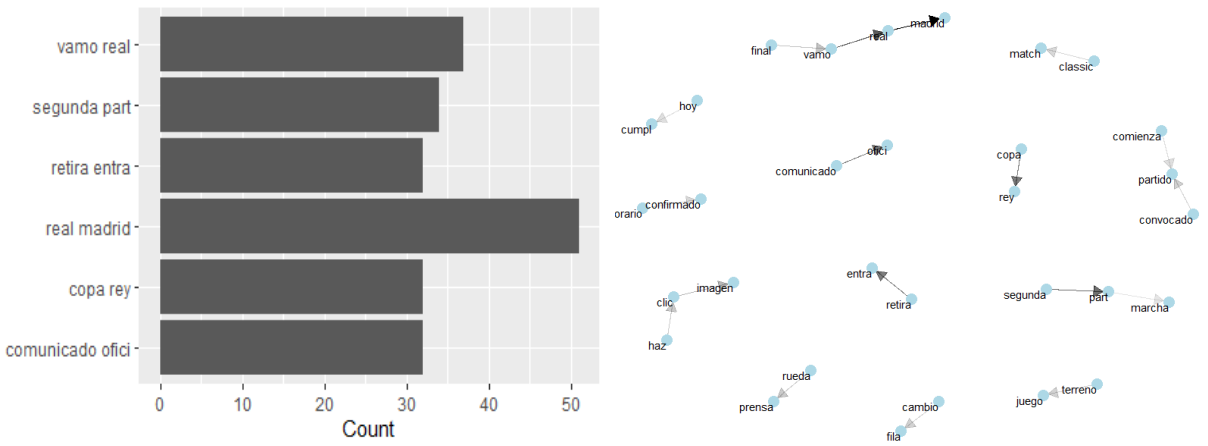
Figure 15: Plots for Juventus FC: (a) Bigram Count, (b) Directed Graph



(a) (b)

Figure 16: Plots for Liverpool FC: (a) Bigram Count, (b) Directed Graph

Real Madrid displayed mostly Spanish Bigrams in its tweets such as “copa rey” and “comunicado ofici”. Most of its frequent bigrams tend to be Spanish Bigram terms with English bigrams mixed in as well. Lastly, we also analyzed Bigram tf-idf terms by term frequency and idf scores and observed that teams displayed similar categories such as match terms, but there were other unique terms such as “scopri sorpr” and other player names that were not mentioned before.



(a) (b)

Figure 17: Plots for Real Madrid: (a) Bigram Count, (b) Directed Graph

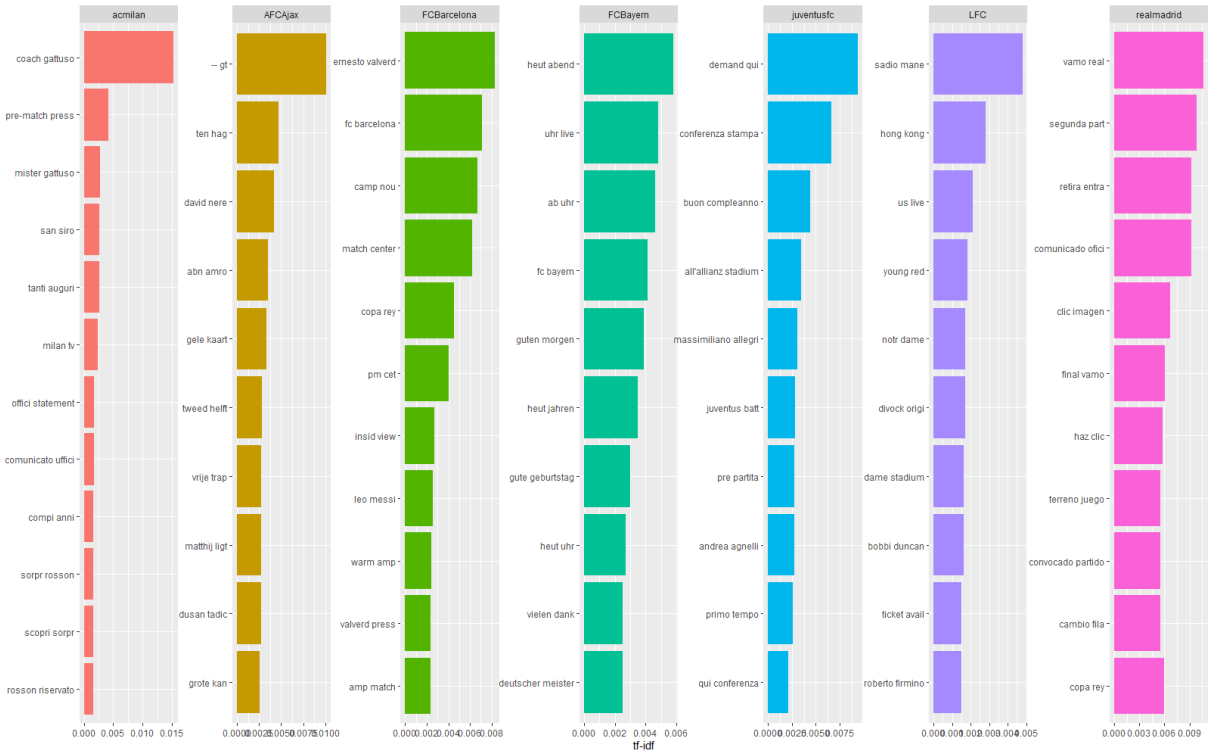


Figure 18: Bigram TF-IDF Plot for All Clubs

After analyzing bigrams, we want to analyze the correlations between words to determine which words were heavily correlated with one another.

3.3 Correlations

When analyzing word correlations for all clubs, some of the strongest word pairings began with terms such as: “back”, “day”, “go”, and “season”. These words were strongly correlated with words such as “win”, “go”, and “get”. These terms are utilized to describe seasonal matches and clubs’ will to win games.

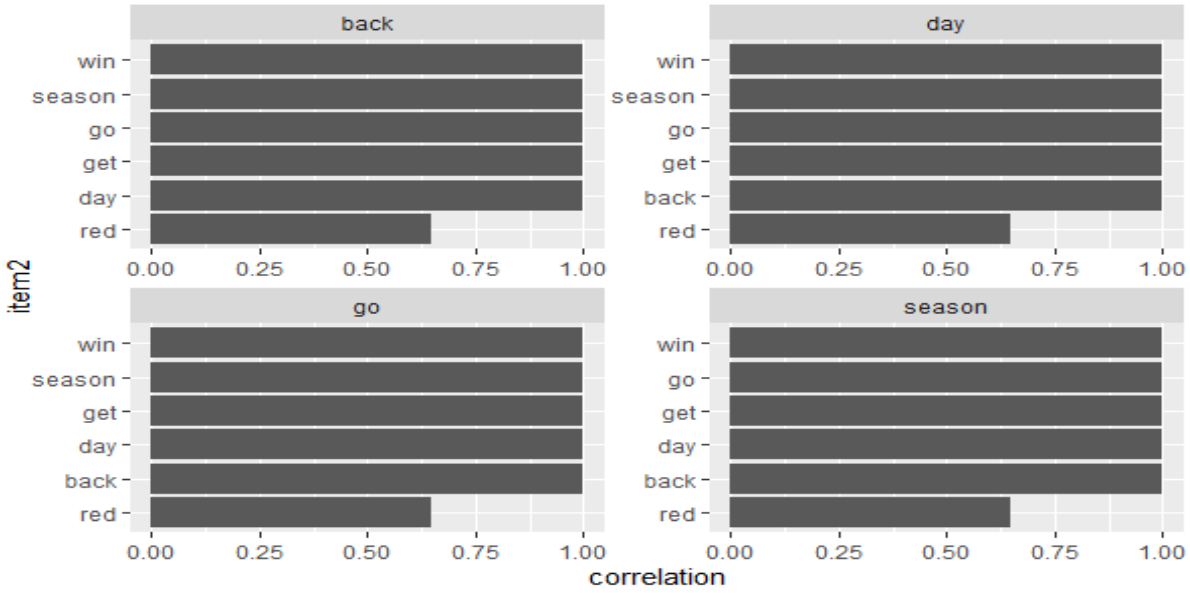


Figure 19: Word Correlation Plot for All Clubs

For AFC Ajax, Dutch terms such as gaan, julli, minuten, and komt are often followed by player names such as "donni", "jong", and "brobbery". These terms translate to go, minute, and come, and they have a strong correlation with player names to cheer on players or images of the players' performances in tweet images.

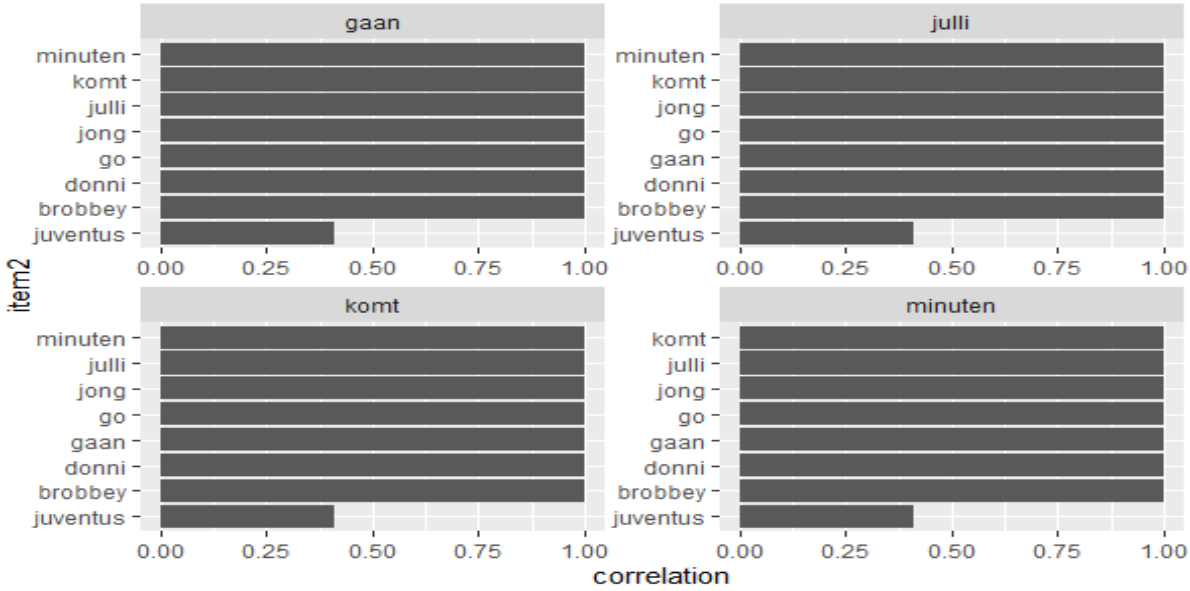


Figure 20: Word Correlation Plot for AFC Ajax

With AC Milan tweets, the score "0-0" was quite common, along with terms such as "cambio", or match, gattuso, and "tomorrow". These were scores that were followed after such words to describe results or scores to describe the second leg results from matches. These scores display Milan's final scores per match whether it is the second leg or single game match.

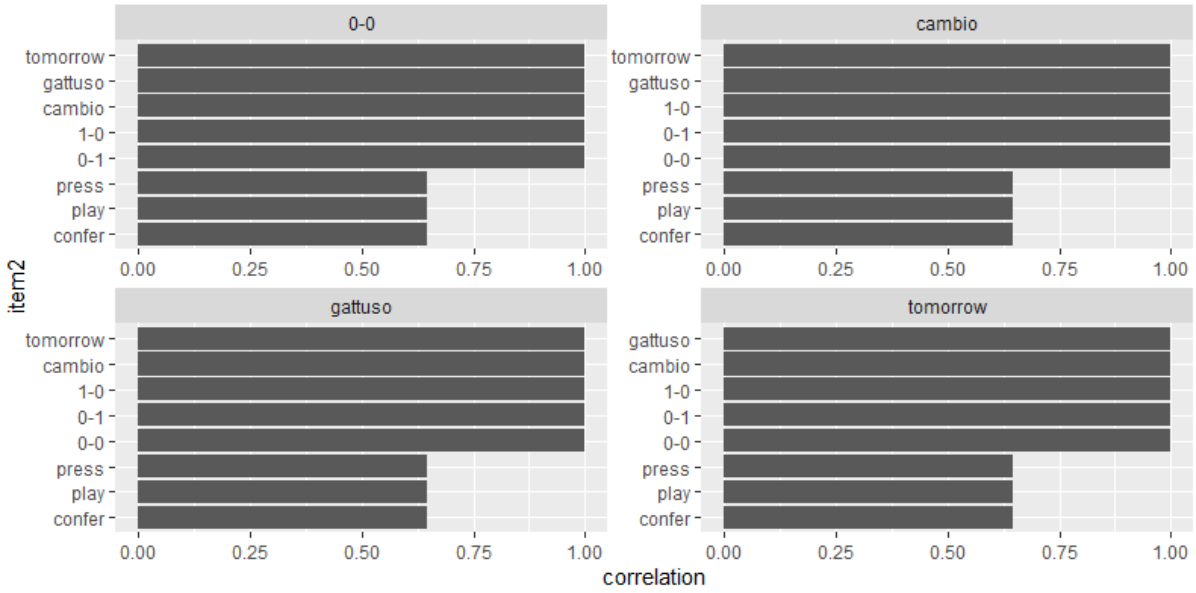


Figure 21: Word Correlation Plot for AC Milan

With Barcelona, "ernesto", "match", "sevilla", and "train" were terms which were followed by "live", "last", "valverd", and "cet". However, we noticed that for "match" and "train", the terms "live", "cet", "train", and "copa" were terms that tend to appear separately as their correlation terms were negative.

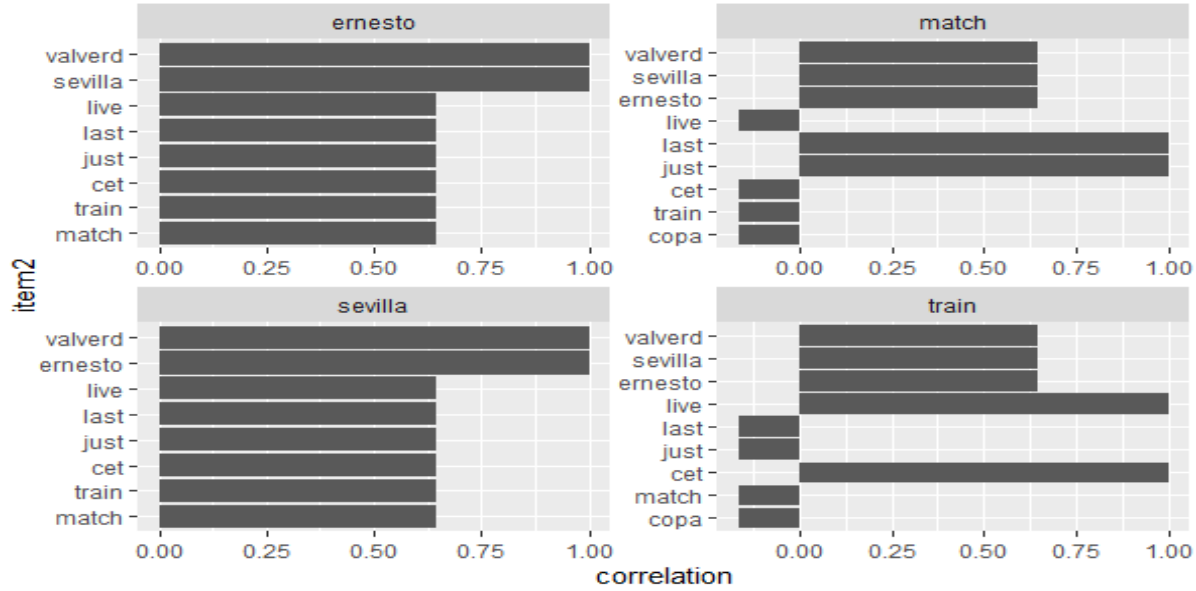


Figure 22: Word Correlation Plot for FC Barcelona

With Bayern tweets, we noticed the terms "dass", "fc", "geburtstag", and "gibt" were terms that were strongly followed by terms such as "saison', "immer", "happi", and "abend". With Bayern tweets, these terms had strong correlations with the terms that followed afterwards.

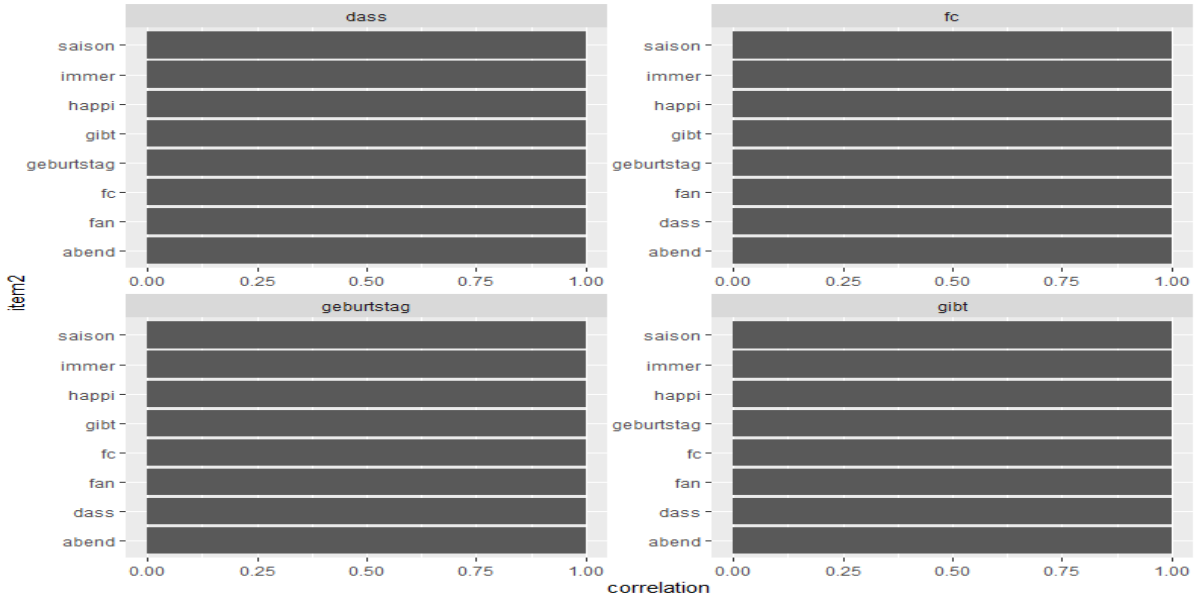


Figure 23: Word Correlation Plot for FC Bayern

In Juventus tweets, terms such as “allegri”, “domani”, “mister”, and “scudetto” were terms that were followed by “stadium”, “stampa”, “live”, and “squadra”. With “domani” and “mister” these terms were highly correlated with the terms that followed after it. These terms display matches that the coach Allegri coached in and the competitions that Juventus competed in.

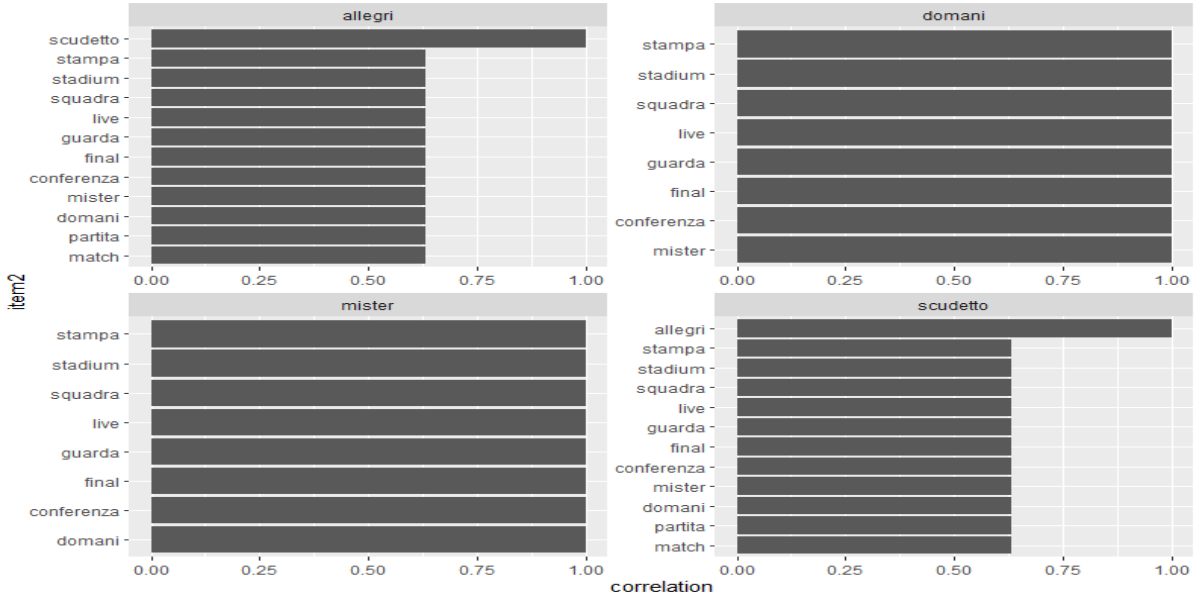


Figure 24: Word Correlation Plot for Juventus FC

For Liverpool, terms like “big”, “day”, “ klopp”, and “season” were highly correlated with terms like “win”, “today”, “time”, and “season”. These terms were also highly correlated with one another as they display how often these words followed one after another. The coach’s name “Klopp” was mentioned when he managed their games and helped the team win its matches.

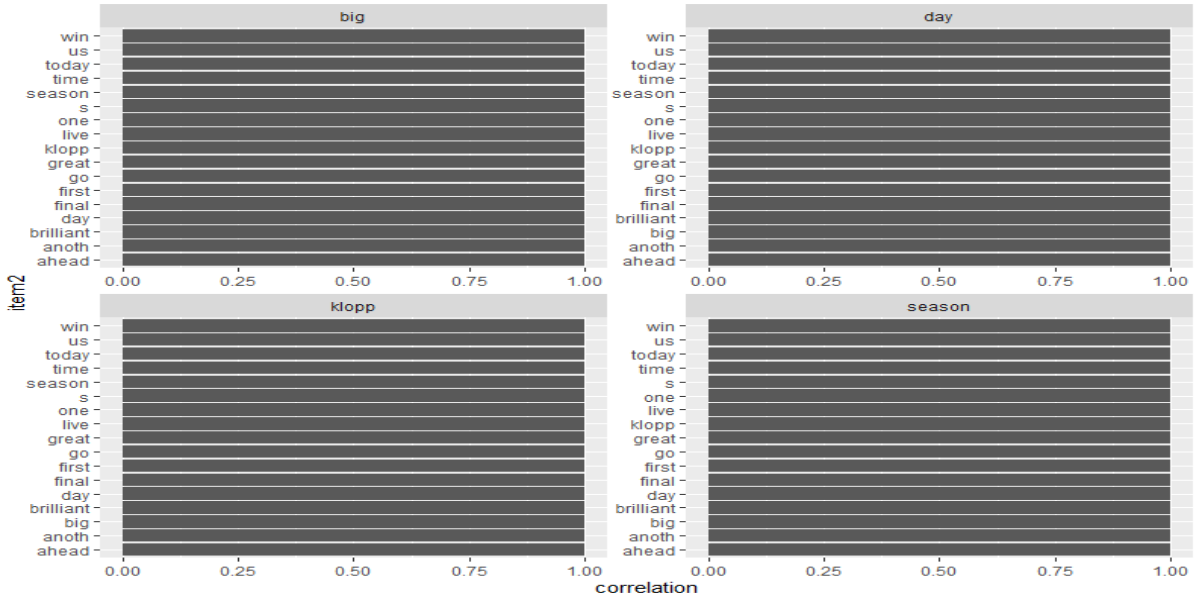


Figure 25: Word Correlation Plot for Liverpool FC

Lastly, Real Madrid terms such as “cambio”, “entra”, “vamo”, and “victoria” and these terms were highly correlated with terms such as “golazo”, “1-0”, and “inici”. These word correlations display match scores following each match with “cambio” and scores that followed after games Real Madrid won with the word “victoria”.

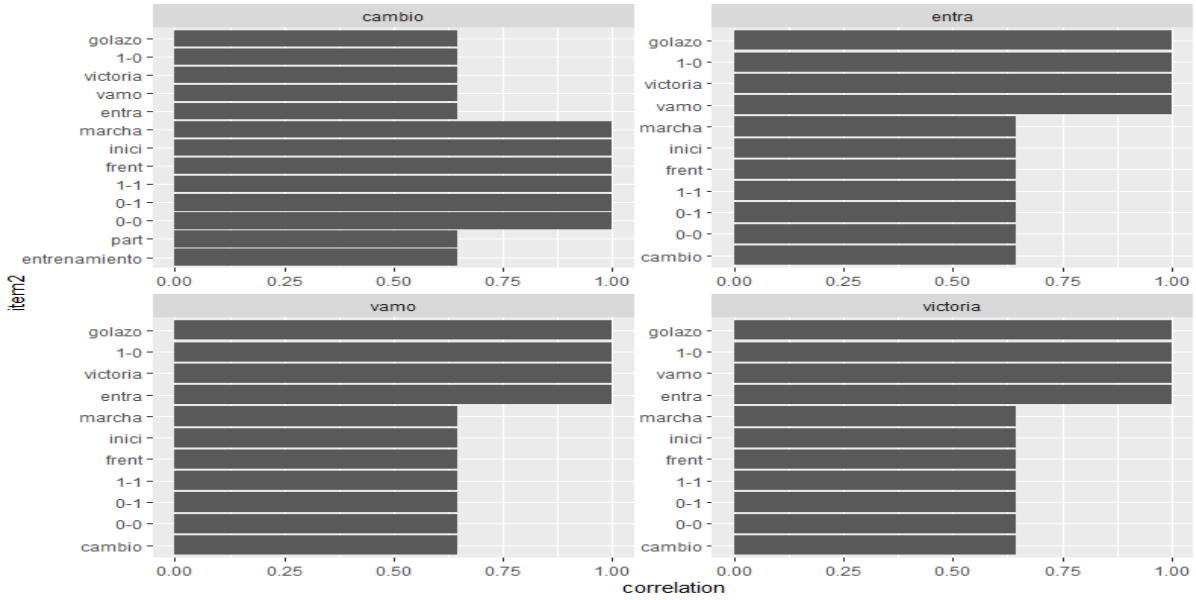


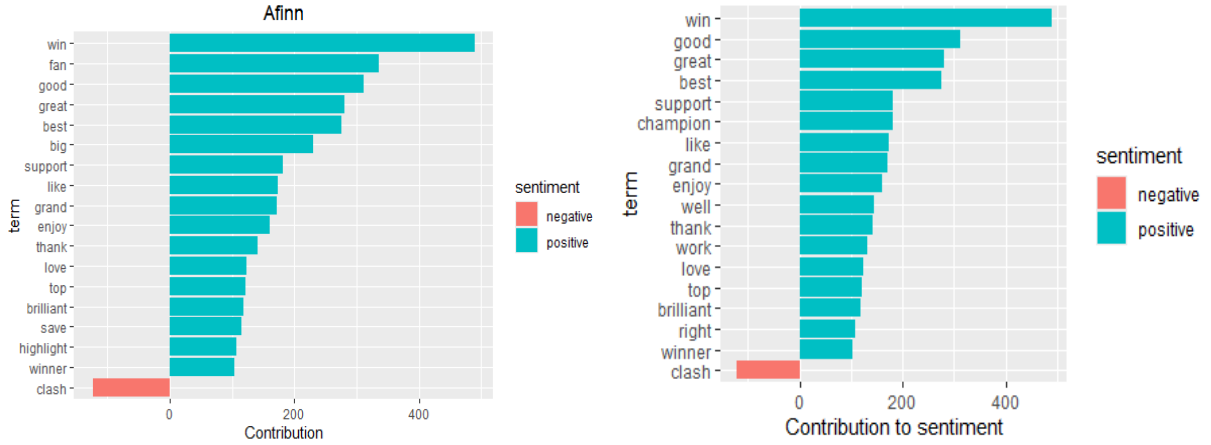
Figure 26: Word Correlation Plot for Real Madrid

After analyzing the correlation between words, the last step would be to conduct sentiment analysis on club tweets.

4. Sentiment Analysis

4.1 Afinn/Bing Sentiment Analysis

When analyzing the total clubs' tweet sentiments, we noticed that the negative term for Afinn and Bing sentiment analysis was the word "clash". Although that term is utilized as a synonym for match, it is quantified as a negative sentiment. There are similar terms in both categories such as "win", "enjoy", and "winner".



(a) (b)

Figure 27: Term Sentiments for All Clubs: (a) Using Afinn, (b) Using Bing

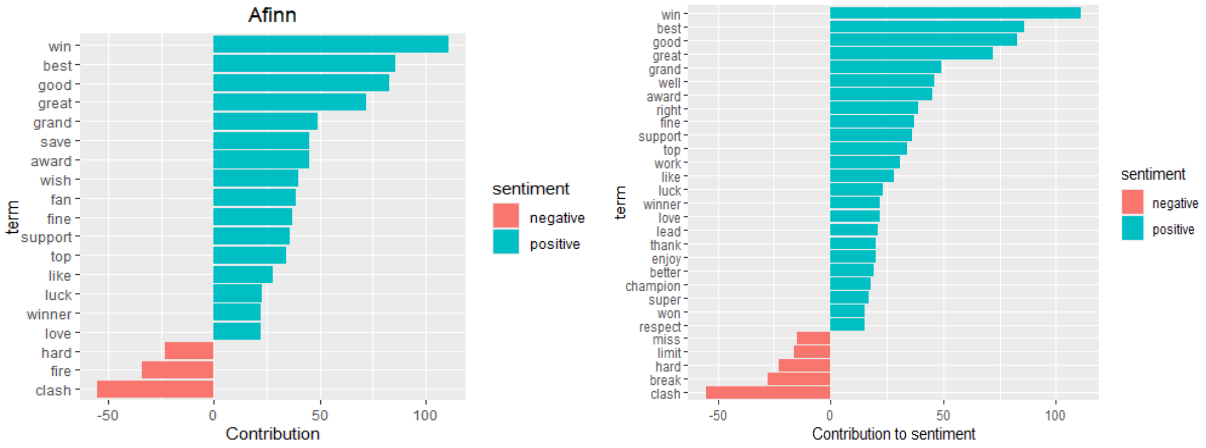
For Ajax's sentiment analysis, we noticed that for Afinn, there was one term that gave negative sentiment with "hard", but for Bing sentiment analysis, the words "trap" and "rust" were included. There are similar terms for Afinn and Bing, but we see unique terms such as "fan" and "highlight" in Afinn and "thank" and "well" terms with Bing.

(a) (b)

Figure 28: Term Sentiments for AFC Ajax: (a) Using Afinn, (b) Using Bing

In AC Milan's tweets, there were 3 negative terms from Afinn while there were 5 negative words found under Bing. There are similar terms such as "clash" and "hard" but

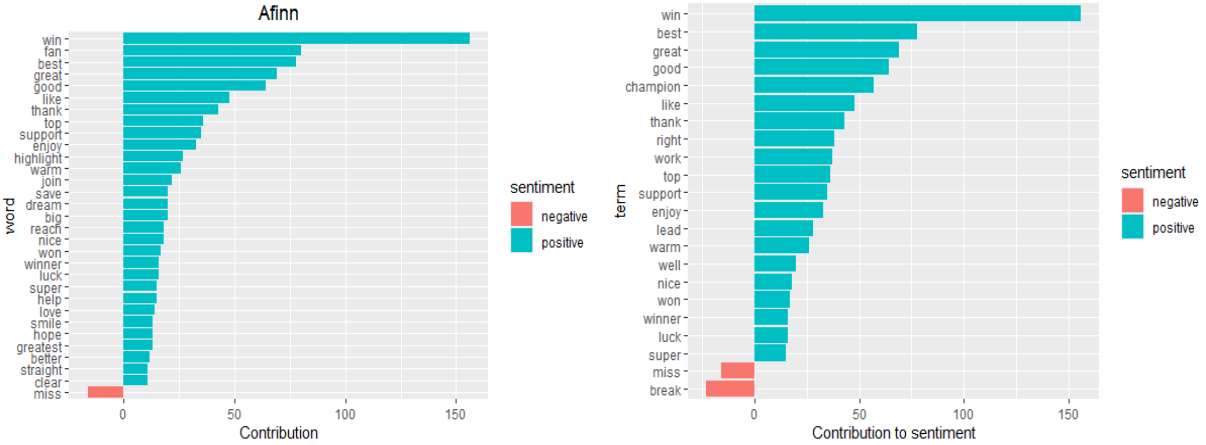
there are unique terms such as “fire” under Afinn and “miss” or “limit” for Bing. The top 5 positive terms were the same for Afinn and Bing sentiment analysis.



(a) (b)

Figure 29: Term Sentiments for AC Milan: (a) Using Afinn, (b) Using Bing

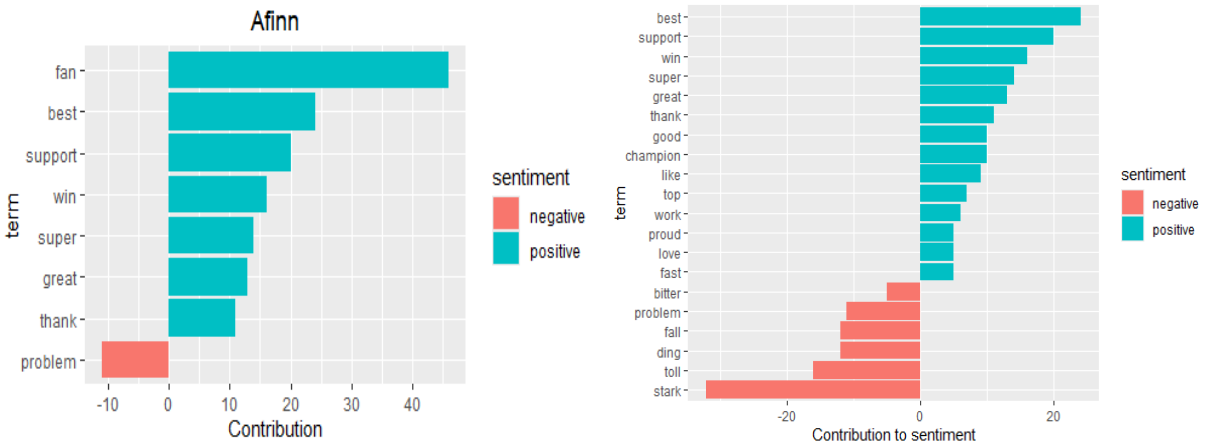
With FC Barcelona tweets, the common negative terms for Afinn and Bing were “miss” while Bing terms also included “break” as well. For positive sentiments, there were similarities such as “win”, “best”, and “thank”, with differences such as “fan” and “work” for Afinn and Bing.



(a) (b)

Figure 30: Term Sentiments for FC Barcelona: (a) Using Afinn, (b) Using Bing

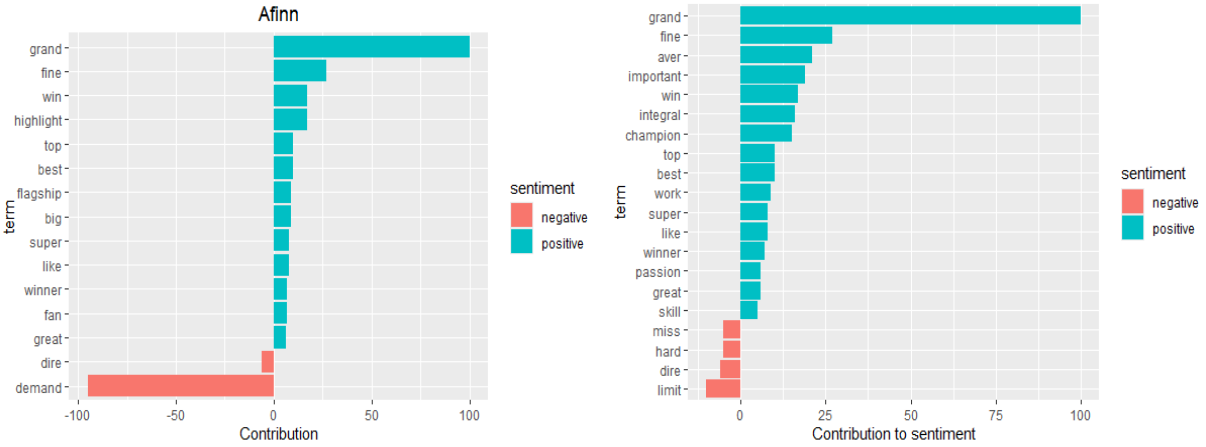
For FC Bayern tweets, the common negative sentiment word for both was “problem”, although with Bing sentiment analysis, there were other terms such as “bitter” and “stark” terms. For positive sentiments, similar terms were “best”, “support”, and “win”, with different positive sentiment terms being “super”, “fan”, and “good”.



(a) (b)

Figure 31: Term Sentiments for FC Bayern: (a) Using Afinn, (b) Using Bing

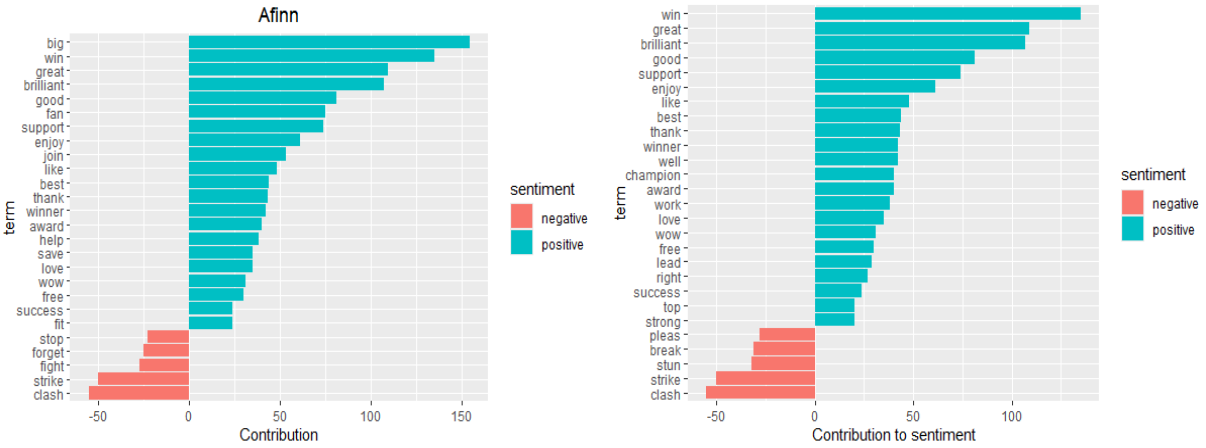
In Juventus tweets, the common negative term is “dire”, while the differing negative terms are “demand” for Afinn and “miss”, “hard”, and “limit” for Bing. With positive terms, there are different terms noted for both such as “aver”, “champion”, and “integral”.



(a) (b)

Figure 32: Term Sentiments for Juventus FC: (a) Using Afinn, (b) Using Bing

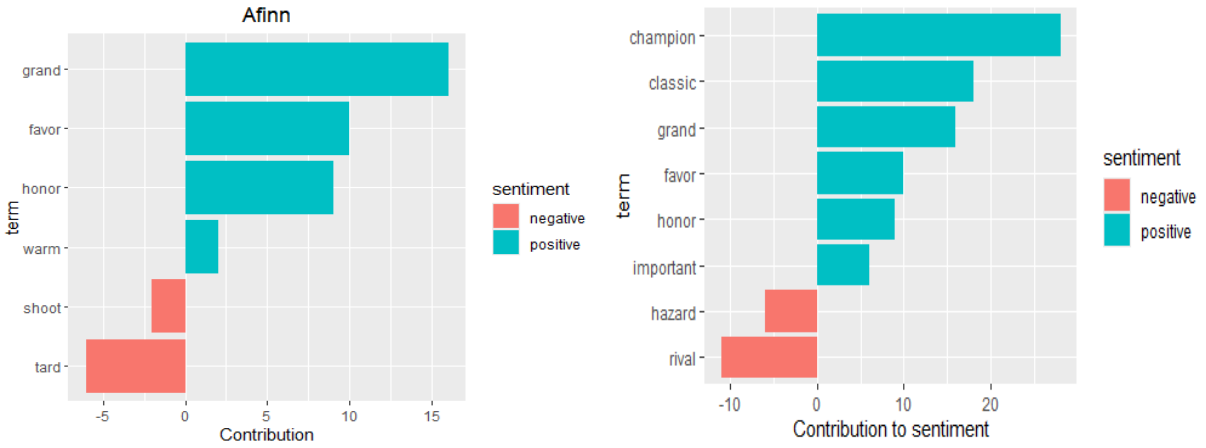
With Liverpool tweets, there were 5 negative terms found for both Afinn and Bing. There were similar terms such as “clash” and “strike”, with differences such as Afinn terms like “stop”, “forget”, and “fight’ and Bing terms such as “pleas”, “break”, and “stun”. With positive terms some differences found were “big”, “fan”, and “lead”, with similarities such as “grand” and “fine”.



(a) (b)

Figure 33: Term Sentiments for Liverpool FC: (a) Using Afinn, (b) Using Bing

Lastly, for Real Madrid, there were not many positive and negative terms noted as the club tends to utilize Spanish more than English terms. For negative terms, Afinn negative terms included "shoot" and "tard" while Bing terms were "hazard" and "rival". However, Hazard is the last name of a Real Madrid athlete named Eden Hazard, which the sentiment analysis did not differentiate. We notice similar positive terms such as "honor" and "grand', but differences such as "warm" and "champion".



(a) (b)

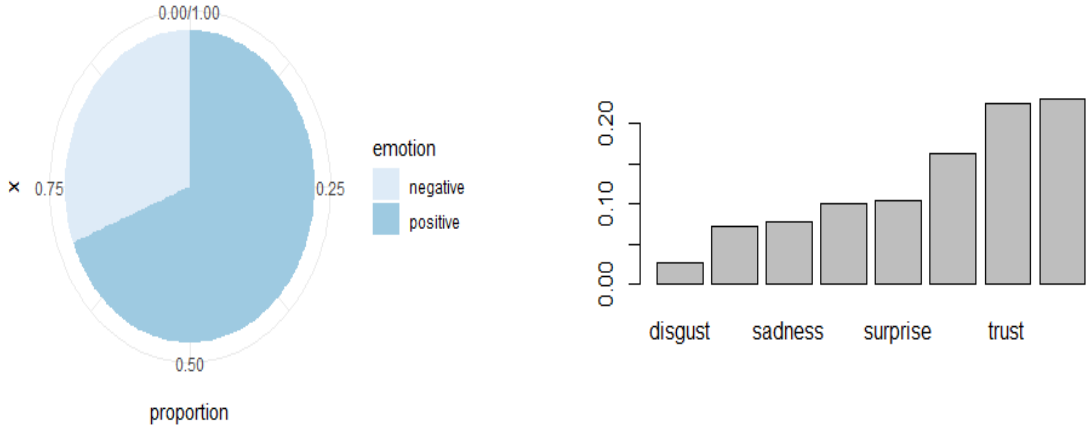
Figure 34: Term Sentiments for Real Madrid: (a) Using Afinn, (b) Using Bing

The next analysis to look at would be the NRC sentiment analysis, understanding the positive and negative sentiment proportions and what sentiment proportions were prevalent other than positive and negative in the clubs' tweets.

4.2 NRC Sentiment Analysis

With the pie chart, we notice that the overall club tweet messages tend to contain positive sentiments. When analyzing other NRC themes, some of the prevalent themes were disgust, sadness, surprise, and trust. Trust had one of the highest proportions since clubs

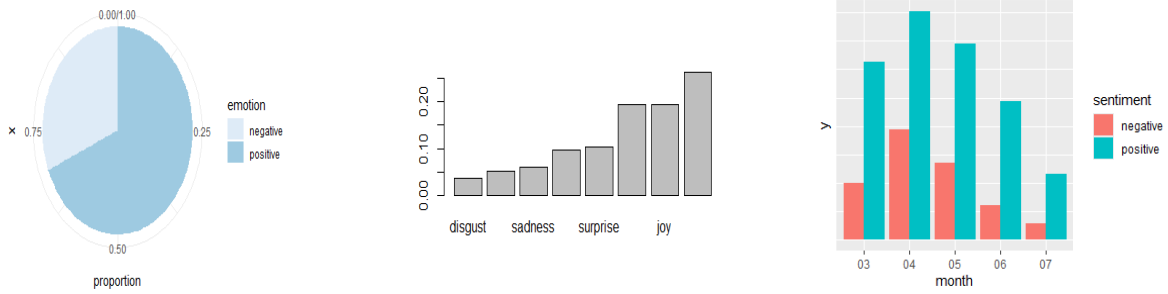
want to maintain their positive relationships with their fans. Overall, club sentiments are mostly positive rather than negative.



(a) (b)

Figure 35: Sentiments for All Clubs: (a) Positive/Negative Plot, (b) NRC Theme Plot

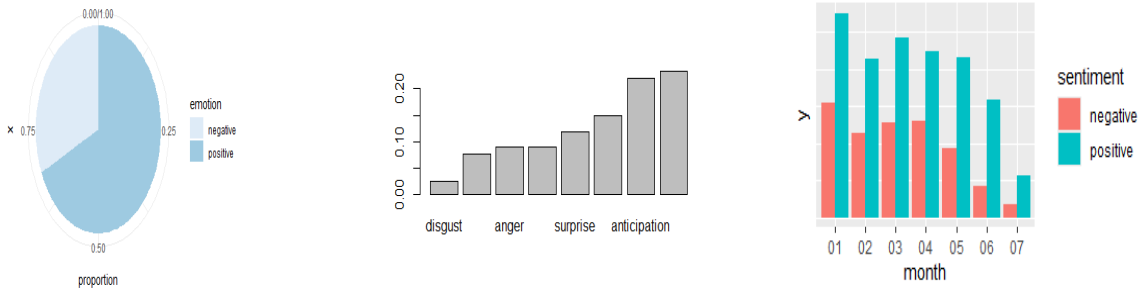
For AFC Ajax, the club displayed a mostly positive sentiment along with prominent positive tones such as surprise and joy. Ajax had an incredible run in 2019, as there was a jump from March to April and consistently higher positive sentiment counts than negative counts from March to May. With high positive NRC proportions and staggering differences between positive and negative sentiment counts, Ajax's messages tend to include more positive than negative sentiments.



(a) (b) (c)

Figure 36: Sentiments for AFC Ajax: (a) Positive/Negative Plot, (b) NRC Theme Plot, (c) Time Series Bar Plot

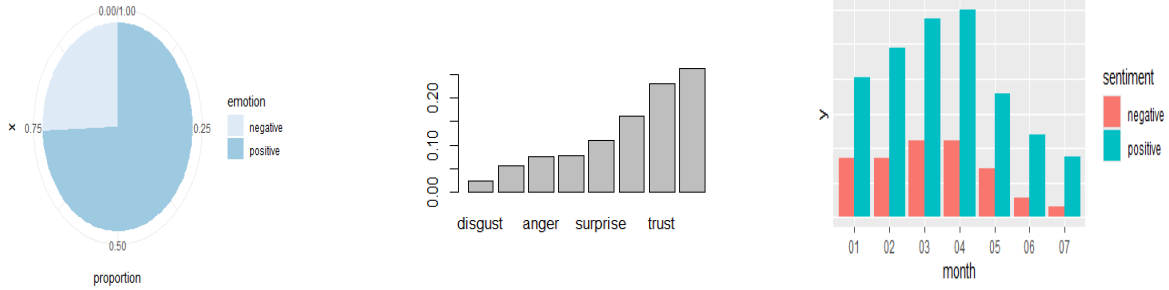
For AC Milan, the club had similar positive proportions as Ajax, but we notice there was anger and anticipation sentiment in its message along with the same NRC sentiments found in Ajax tweets. When analyzing AC Milan's positive vs negative proportion, AC Milan had its highest positive sentiment count in January, but we noticed how that count decreased across the season. AC Milan was able to start 2019 strongly, but its gradual positive sentiment decline to June indicates that the club had much anticipation and strong positive sentiment counts from January to May, but it never reached its heights in January.



(a) (b) (c)

Figure 37: Sentiments for AC Milan: (a) Positive/Negative Plot, (b) NRC Theme Plot, (c) Time Series Bar Plot

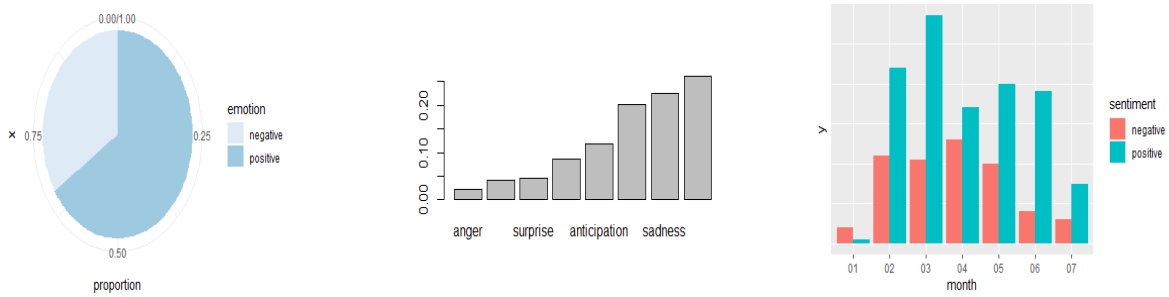
With FC Barcelona, the club was nearly 75% positive and its sentiments included disgust, anger, surprise, and trust. With positive sentiment counts, the club reached its peak in April and dropped drastically in May. The club clearly had a strong run until the drop in positive sentiments towards the end of the season. However, even after the drop in positive sentiments, the negative sentiment counts also decreased as well, displaying how much more positive than negative sentiments there were in its tweets.



(a) (b) (c)

Figure 38: Sentiments for FC Barcelona: (a) Positive/Negative Plot, (b) NRC Theme Plot, (c) Time Series Bar Plot

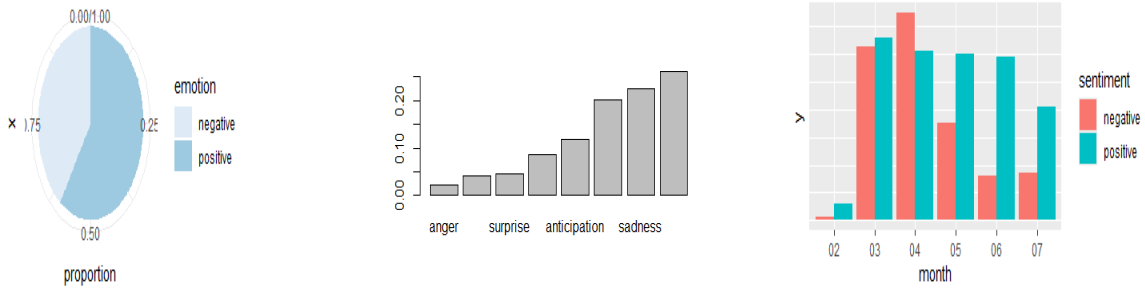
In Bayern's tweets, we noticed that tweets tend to be positive, but in the NRC sentiments, the sadness category was one of the highest proportions found in its tweets with anticipation followed afterward. Bayern's tweets reached their highest positive sentiment in March, but it drastically dropped in April and stayed consistent until June. It would not reach the same heights that the club had reached in March.



(a) (b) (c)

Figure 39: Sentiments for FC Bayern: (a) Positive/Negative Plot, (b) NRC Theme Plot, (c) Time Series Bar Plot

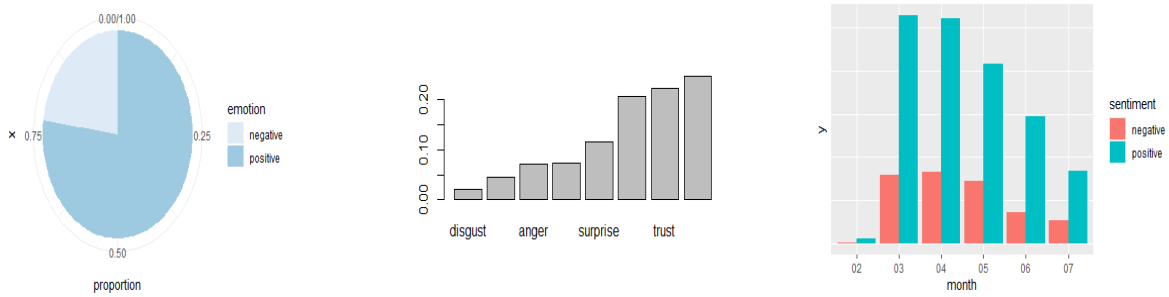
When analyzing Juventus's tweets, while the majority of Juventus tweets were positive, they seem to be less positive than other clubs' positive sentiment proportions. With the NRC sentiments, one of the largest sentiment proportions is the sadness sentiment. When comparing positive and negative sentiment counts by month, the negative sentiment was the highest in April but it decreased from then onwards. The club was able to end on a positive note, but the negative sentiment count indicated that the club had a rough start from February to April.



(a) (b) (c)

Figure 40: Sentiments for Juventus FC: (a) Positive/Negative Plot, (b) NRC Theme Plot, (c) Time Series Bar Plot

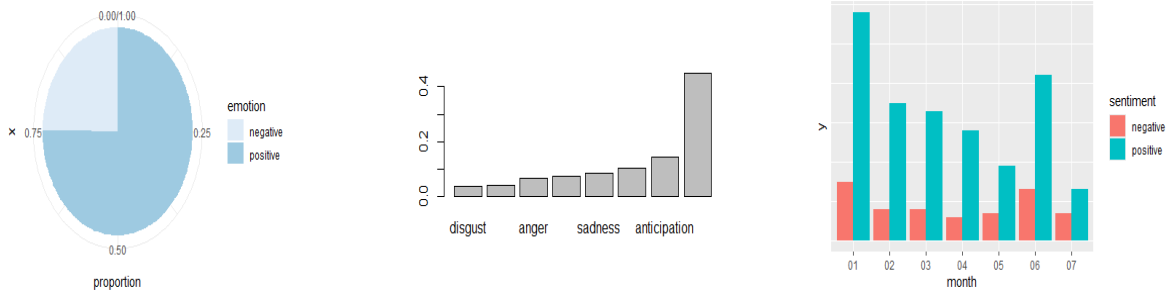
With Liverpool's tweets, the club displayed that the proportion of positive tweets was greater than 75%. Additionally, some of the highest NRC proportions are trust and surprise. Even when comparing positive and negative sentiment counts, Liverpool had the highest positive sentiment counts between March and April. Its negative sentiments were quite low throughout the season. Liverpool also had a sudden increase in positive sentiments from February to March, indicating that the club was able to maintain its positive sentiments throughout the second half of the season.



(a) (b) (c)

Figure 41: Sentiments for Liverpool FC: (a) Positive/Negative Plot, (b) NRC Theme Plot, (c) Time Series Bar Plot

Lastly, with Real Madrid tweets, the positive sentiment proportion was at 75% and trust was noted to be the highest proportion compared to other NRC sentiments. However, when analyzing the positive sentiment counts between January and July, the positive sentiment counts dropped heavily from January to May, then gradually increased again in June. While the club seems to display higher positive than negative sentiment counts in its tweets, the sudden decrease in positive sentiments from its height proves club sentiments would not reach these heights again in January and the positivity level altered in its tweets.



(a) (b) (c)

Figure 42: Sentiments for Real Madrid: (a) Positive/Negative Plot, (b) NRC Theme Plot, (c) Time Series Bar Plot

All the clubs tend to end on a positive note with larger positive sentiment counts than negative sentiment counts. There is a higher positive sentiment proportion than negative sentiments for all clubs with only 3 clubs above 75%. After analyzing the club's term frequencies and sentiment analysis, the next step is to run machine learning models to predict tweets by club.

5. Machine Learning

5.1 Unsupervised Learning Using LDA

LDA, or Latent Dirichlet Allocation, is a powerful topic modeling technique that utilizes unsupervised learning to divide text into natural groups. LDA particularly works well with textual data due to its ability to classify each document as a mixture of topics, rather than simply assigning one topic. Another strong advantage of topic modeling over other common clustering algorithms is that any given word inside a topic is not limited to just one topic and can appear in multiple (and possibly all) topics.

Since there are already groups (the soccer clubs) identified within our data, it might be worth mentioning why we are using unsupervised learning. The main objective of implementing topic modeling was for the following:

1. Visualize whether or not a particular club stays consistent with one topic across the seven months (the time interval of our data)

2. To spot any misclassified months

3. To compile a list of the most commonly mistaken words

To address the first objective, we needed to get our tokens in a tidy format where they are split by each club and also each month. We were essentially treating each combination of clubs and months as an individual "document"; for example, a document "AFCAjax_03" will contain tokens from the club AFC Ajax in March 2019.

	club_date	term	total_count
1	AFCAjax_03	–	3.00
2	AFCAjax_03	—	2.00
3	AFCAjax_03	0-0	1.00
4	AFCAjax_03	0-1	3.00
5	AFCAjax_03	0-2	1.00
6	AFCAjax_03	05u	1.00

Table 1: First 6 Rows of Tidied Tokens

Organizing the tokens in a tidy format, we created a seven-topic model. Since there are 7 clubs of interest, it was reasonable for us to set the number of topics as 7. Then, we plotted the list of the top seven words from each topic using the beta values, which is the probability of a word being generated from one specific topic. Since each topic supposedly models one soccer club, this plot looked quite similar to the TF-IDF plot. This was a good indication as we understood that each topic fully encapsulates one club and, thus, provides outstanding interpretability if one club were to fluctuate between multiple topics across the different months.

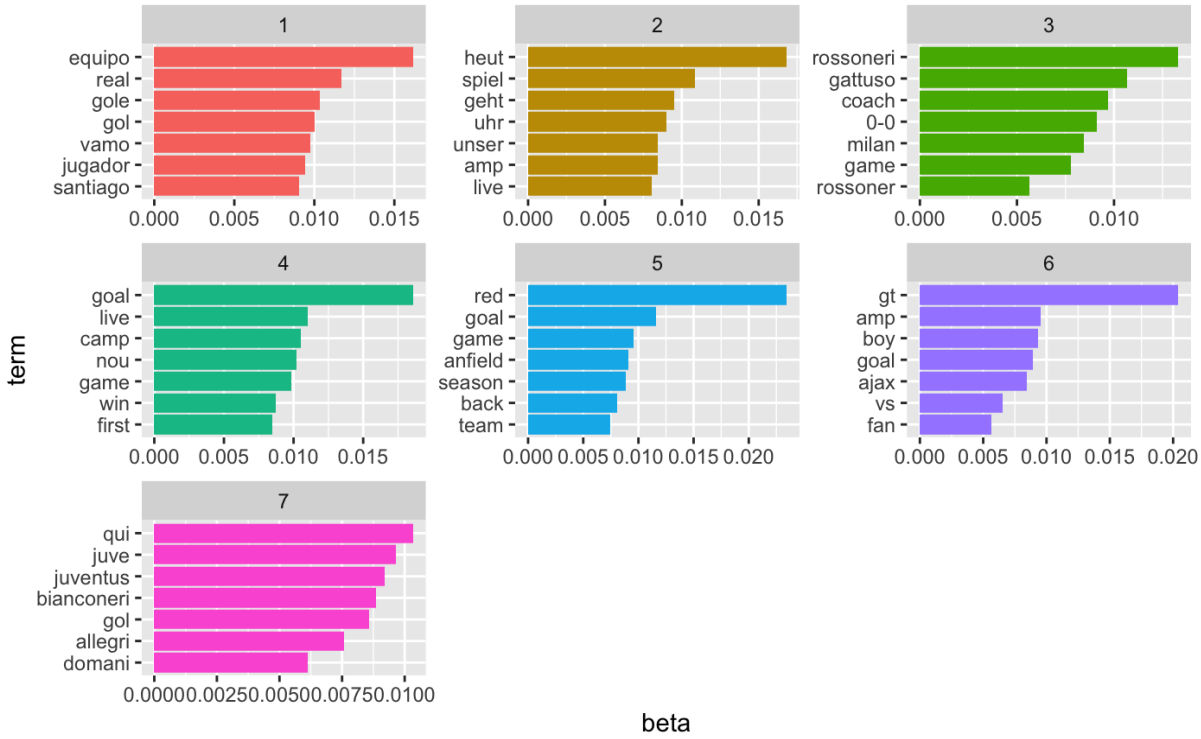


Figure 43: LDA Topics Plot by Beta Values

However, when we proceeded to plot the distribution of topics per club over the seven months, the results did not reflect many fluctuations. The following figure shows the per-document plot, where each document still represents one club and one particular month:

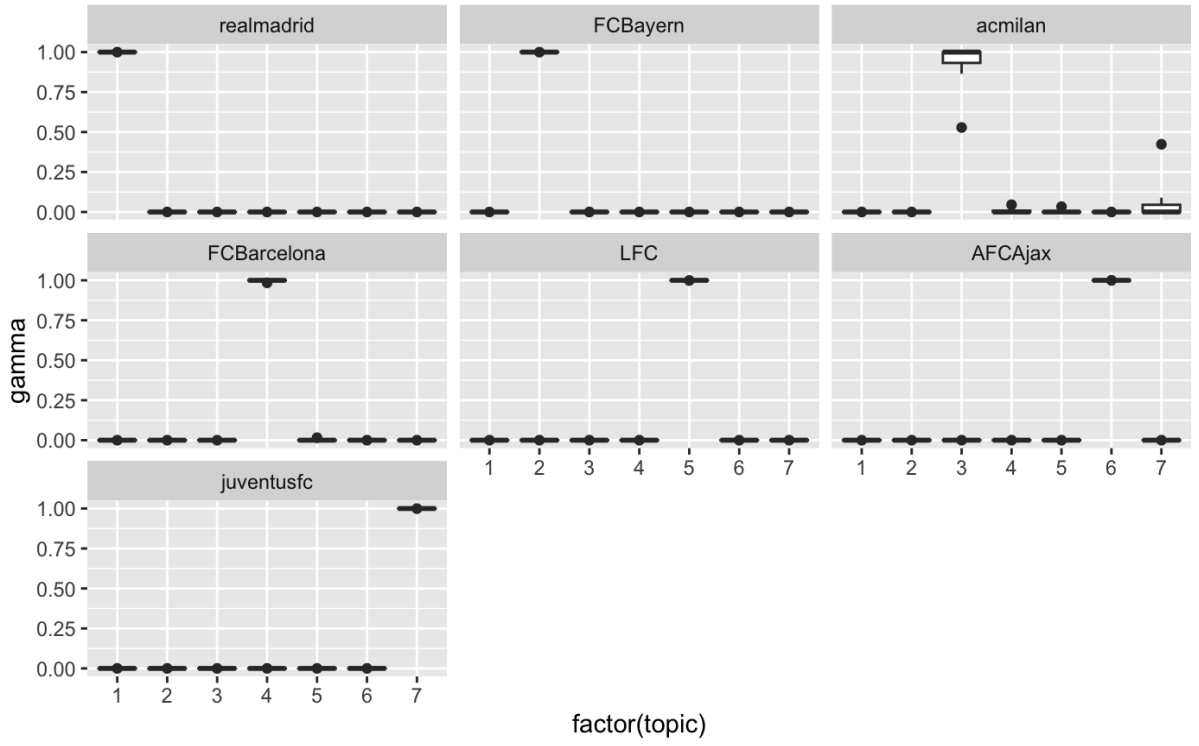


Figure 44: LDA Topics Plot by Gamma Values

All clubs with the exception of AC Milan displayed consistency across the seven months, meaning that each club stayed true to its own topic and did not share a lot of commonalities with any other club. However, in the case of AC Milan, we do see a slight variance in topics across the seven months, including a month where it was classified as 50% Topic 3 (belonging to AC Milan) and 50% Topic 7 (belonging to Juventus FC). It seems like AC Milan and Juventus FC definitely have some correlation based on the above plot, but why is such the case? AC Milan and Juventus FC are both Italian professional clubs, known for their long rivalry as two of the top competitors from the same country. Thus, we can predict that the Italian rhetoric used by the two clubs might be quite similar.

The second objective in this section was to pinpoint any misclassified months in terms of topics. The way we approached this was per each club and month, we picked out the "majority topic", one that contributes the most to the distribution of topics. If this "majority topic" was different from the real topic of that specific club, then we would classify such an instance as a misclassification. After going through this process, we found that there were no misclassified months, further corroborating the claim that each club stayed true to its topic.

The last objective of compiling a list of misclassified words required a bit more attention as it puts the emphasis on the individual words, rather than documents. The results from this investigation were reflective of the previous findings; there were little to no misclassified words as a whole. There were, however, some misclassified words that were categorized as coming from Juventus FC when in reality they are from AC Milan, which again confirms our findings from the first objective.

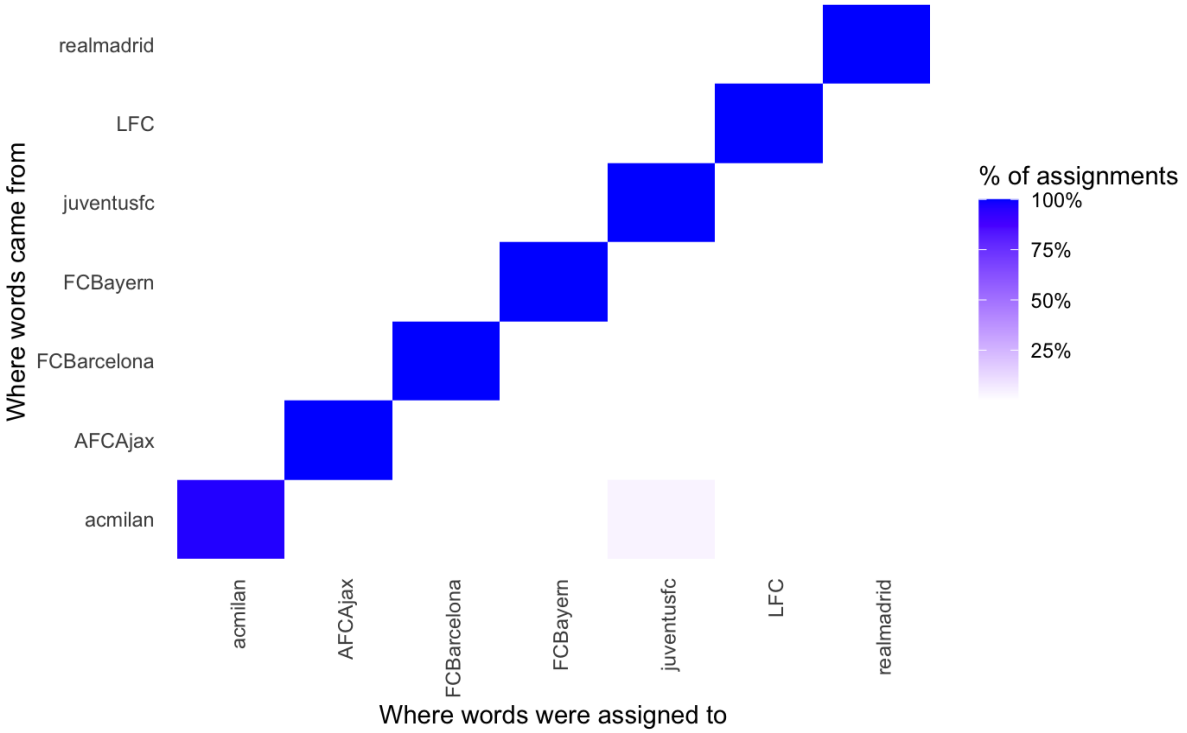


Figure 45: LDA Misclassification Matrix

	club	consensus	term	n
1	acmilan	juventusfc	award	38.00
2	acmilan	juventusfc	miglior	25.00
3	acmilan	juventusfc	gol	22.00
4	acmilan	juventusfc	stagion	17.00
5	acmilan	juventusfc	anni	16.00
6	acmilan	juventusfc	oggi	12.00
7	acmilan	juventusfc	parata	10.00
8	acmilan	juventusfc	azzurr	9.00
9	acmilan	juventusfc	forza	9.00
10	acmilan	juventusfc	bocca	8.00
11	acmilan	juventusfc	lupo	8.00
12	acmilan	juventusfc	merita	8.00
13	acmilan	juventusfc	scudetto	8.00

Table 2: Top Misclassified Words

The results from the three separate investigations all support that each club has its unique rhetoric. As there is not a lot of overlap between the clubs, this will be particularly helpful for the next section where supervised learning will be employed to answer the question of how well we can classify each tweet.

5.2 Supervised Learning Using Random Forest

After topic modeling has been used, we wanted to now determine how accurately we can predict the source of the tweet. To do so, we decided to apply supervised learning using a random forest model. A random forest model for classification comprises a number of decision trees, which are then combined using a majority vote to predict a single output. To prepare our tokens for random forest, we needed to cut the majority of sparse terms from our DFM (document-feature matrix).

The reason why sparse terms were removed was largely due to computational complexity. As the computational complexity of random forest (in training) scales with the number of features, having about 16000 features did not make much sense. Thus, we kept only the features that appeared in more than 50 documents (tweets), which came about to be 430 features. Note that here, we are using the cleaned version of the text.

On the topic of computational complexity, another important issue to address was the number of decision trees in the model. The random forest function in R defaults to 500 trees, which initially took a very long time to execute. With the number of trees affecting the computational complexity for both training and testing instances, we decided to scale this hyperparameter down to 50. With 50 trees, the algorithm ran in a much more reasonable time frame.

For the actual training and testing, we split up the preprocessed data into 70% training and 30% testing. This came about to be about 14618 documents for training and 6265 documents for testing. Now, using the training data, we trained the random forest model

and then evaluated the model on the testing data.

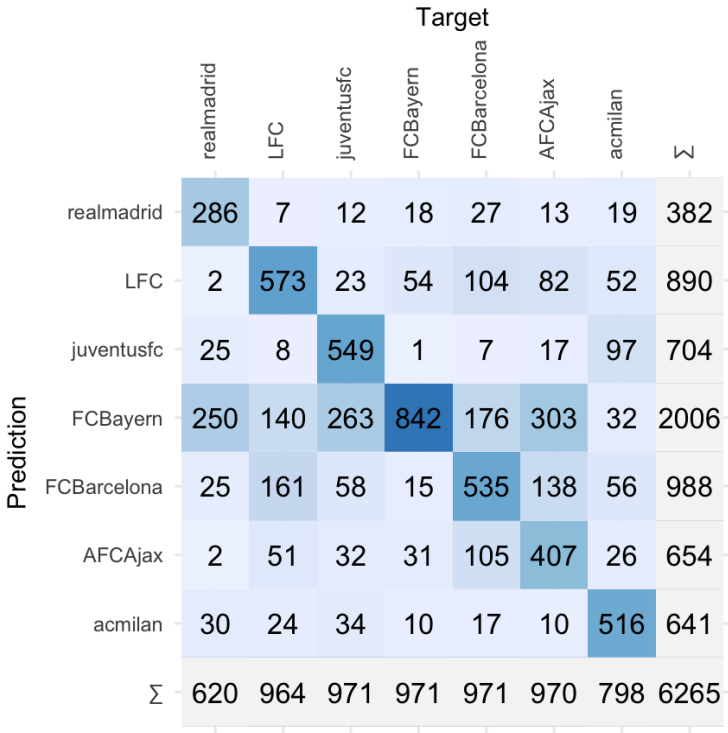


Figure 46: Confusion Matrix with Cleaned Text

As the above confusion matrix suggests, there are a significant number of misclassified instances. With a relatively balanced test set (a similar number of documents for each classification label), we determined that the accuracy was a fair metric to evaluate the model. The accuracy of the model (number of correctly identified instances divided by the total training set size) was a mere 59.2%, which was not ideal. In particular, we noticed that for clubs AFC Ajax and Real Madrid, the model performed poorly, with 407 correct classifications out of 970 for the former and 286 correct out of 620 for the latter. So, why was the model performing so poorly?

Running it with the cleaned text left us questioning whether the uncleaned text would perform better. Often times, an uncleaned text may outperform the cleaned text and we wanted to test if this was true in this particular scenario. Thus, we grabbed the unfiltered tokens, removed sparse terms, and split the data using the same proportions.

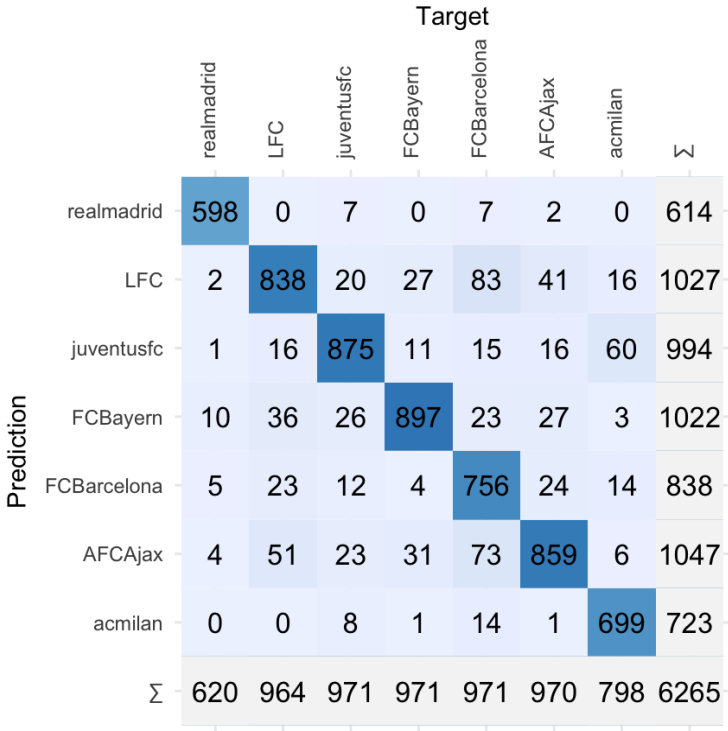


Figure 47: Confusion Matrix with Uncleaned Text

With the uncleaned text data, the model achieved a high accuracy of 88.1%, signifying an extreme jump from the previous iteration. We notice that most classifications are identified correctly, including that of AFC Ajax and Real Madrid. To explain this difference in accuracies between the cleaned and the uncleaned text, it is imperative to take a deeper look at some of the features (tokens) that were used in each run.

	Tokens
1	train
2	today
3	like
4	alway
5	winner
6	one
7	friend
8	love
9	day
10	back

Table 3: 10 Features Pulled from the Cleaned Text

	Tokens
1	@juventusfcyouth
2	#under23
3	#juventuswomen
4	#allegri
5	#getready
6	#ajaxjuve
7	#juveatleti
8	@sergioramos
9	#halamadrid
10	#realmadrid

Table 4: 10 Features Pulled from the Uncleaned Text

Comparing the two sets of features, we notice that the set of features for the cleaned text contains a lot of plain English words, while the set of features for the uncleaned text contains mentions, hashtags, and special symbols like emojis. With the former, the machine learning model had a harder time classifying each tweet as common English words like those can be prevalent across all 7 clubs. However, with the use of specific mentions and keywords in hashtags that may be more unique to each club, the model was able to easily pick up on the subtle nuances of each club, making it much easier to identify each tweet.

6. Conclusion/Recommendations

When analyzing the term frequency results for all clubs, we noticed that match terminologies such as "final", "game", and "live" were commonly frequent terms as clubs want to display their match game data and information to their fans. On the other hand, with analyzing frequent terms by clubs, certain themes such as team name and nicknames, player names, and unique soccer terms were frequent types of terms that clubs included in their tweets. When analyzing tf-idf, we notice certain frequent terms reappear in our tf-idf table along with other unique terms from clubs' tweets. With bigrams, we notice that celebratory terms such as "happy birthday" were frequent along with team names, home stadiums, player names, and media-related terms such as "watch live". Analyzing word correlations, scores were highly correlated with other scores and with match terms such as "cambio" and "press" and names as well. There were also terms that were slightly correlated and terms that were not correlated with each other at all.

In sentiment analysis, clubs tend to utilize higher positive sentiment than negative sentiment terms. There were few negative sentiment terminologies clubs ever utilized, although certain negative sentiment terms are soccer synonym terms that sentiment analysis deemed as negative connotations. With NRC sentiment analysis of positive and negative sentiment counts by months, all clubs ended their seasons on a positive note. Only 3 clubs were able to maintain positivity sentiment proportions at or above 75%. Certain NRC

positive sentiment proportions such as "trust", "joy", and "anticipation" were frequent, along with negative sentiments such as "disgust" and "anger". Positive and negative sentiment counts fluctuated across the season, but whenever negative sentiment counts were higher than positive sentiments, positive sentiment counts dropped abruptly, or positive sentiments reached their maximum and stayed consistent, club sentiments seemed to shift based on the club's performance and success.

As far as machine learning, the results we observed were quite satisfactory. Using topic modeling with LDA, we assigned a topic to each club, and from that, we were able to track each club's lexicon over the seven months of 2019. From this, we deduced that clubs did not show a lot of fluctuations in topics. We did, however, find that the club AC Milan tended to get confused with Juventus FC in certain months and we suspected that this was due to their similar vocabulary used as the two Italian clubs in the sample. We also found some of the most commonly misclassified words and saw that most of these words were Italian, which supports the fact that AC Milan and Juventus FC do share a lot of the same vocabulary. Moving on to supervised learning with random forest, a major finding for us was that training with the uncleaned text significantly outperformed training with the cleaned text. This was due to the fact that the uncleaned version had a lot more unique identifiers like hashtags and certain emojis that can be attributed to each club. With an accuracy difference of about 30%, the uncleaned version proved to be useful and also exemplified that, sometimes, in text mining, cleaning and transformations can be detrimental to the machine learning process.

To discuss some limitations and potential next steps, we must first talk about the data itself. Although the data we filtered was extensive and proved to be of great size, it only included data from the first seven months of 2019. Not only does this not perfectly encompass each club's Twitter history, but it fails to cover one full year as well. Thus, incorporating data from 2018 would help us visualize the club better as it completes a full year's worth of data. Moreover, adding other sub-accounts other than the official club accounts might also diversify our analysis; we could even do a deeper exploration into each club and see the similarities and differences across the sub-accounts of each club. Besides the data itself, more classification models, such as k-nearest neighbors or support vector machines, could be employed and we could come up with a summary table containing the results of each model. With this, we would be able to optimize the accuracies and truly recommend an "ideal" way to go about predicting these tweets.

References

Package cvms title cross-validation for model selection. (2024). https://cran.r-project.org/web/packages/cvms/cvms.pdf

Package quanteda title quantitative analysis of textual data. (2023). https://cran.r-project.org/web/packages/quanteda/quanteda.pdf

Silge, J., & Robinson, D. (n.d.). *Text mining with r: A tidy approach*. O'Reilly. https://www.tidytextmining.com/

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (n.d.). *R for data science* (2nd). O'Reilly. https://r4ds.hadley.nz/