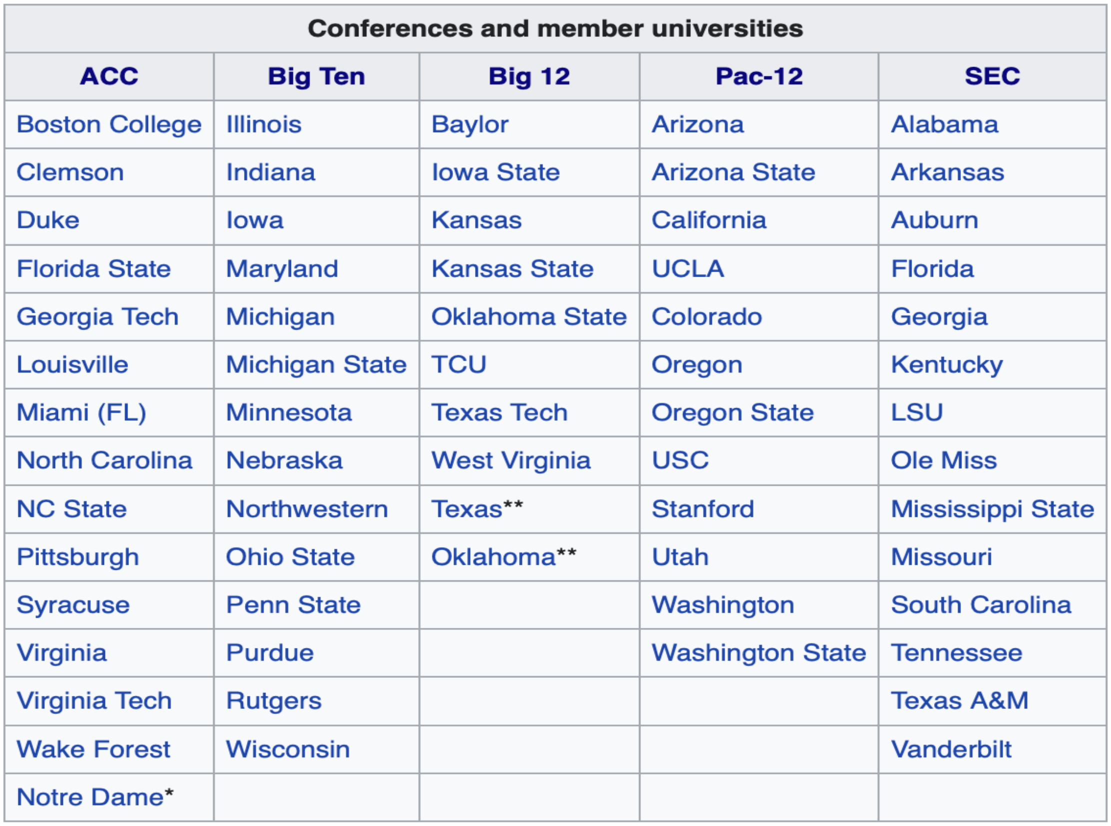
The World's Perception of College Football

By: Bryant Wise

Introduction:

College football is one of the most prominent topics discussed among sports fans, and the forum that many people use to express their opinions on the topic is Twitter. These fans use this forum to either praise their own team or bash another team, which is why I chose to take a deep dive into tweets posted pertaining to the Power 5 Conferences. These collections of college institutions from all over the United States are the 5 athletic conferences that are considered to be the most elite, which makes these teams the topic of most conversation about college football.

The five conferences that belong to the Power 5 are: the SEC, the ACC, the Big 10, the Big 12, and the PAC-12. The SEC, or Southeastern Conference, is an American college athletic conference founded in 1933, and the fourteen members are primarily located in the South Central and Southeast United States. The ACC, or Atlantic Coast Conference, holds 15 members who are all located in the eastern United States. The Big 10 conference is the oldest Division 1 collegiate athletic conference, and currently holds 14 members with 2 affiliate institutions. These members can be found in each of 11 states stretching from New Jersey to Nebraska. The Big 12 conference has 10 members located in the states Iowa, Kansas, Oklahoma, Texas, and West Virginia. Finally, the Pac-12 Conference has 12 members who are located in the states of Arizona, California, Colorado, Oregon, Utah, and Washington. All of these conferences consist of a mix of public and private institutions with different academic and social standing in the United States. These vast differences between each conference and institutions make it a very large melting pot of students and athletic programs. Below is a list of the teams that make up each conference:



Conferences and member universities

ACC	Big Ten	Big 12	Pac-12	SEC
Boston College	Illinois	Baylor	Arizona	Alabama
Clemson	Indiana	Iowa State	Arizona State	Arkansas
Duke	Iowa	Kansas	California	Auburn
Florida State	Maryland	Kansas State	UCLA	Florida
Georgia Tech	Michigan	Oklahoma State	Colorado	Georgia
Louisville	Michigan State	TCU	Oregon	Kentucky
Miami (FL)	Minnesota	Texas Tech	Oregon State	LSU
North Carolina	Nebraska	West Virginia	USC	Ole Miss
NC State	Northwestern	Texas**	Stanford	Mississippi State
Pittsburgh	Ohio State	Oklahoma**	Utah	Missouri
Syracuse	Penn State		Washington	South Carolina
Virginia	Purdue		Washington State	Tennessee
Virginia Tech	Rutgers			Texas A&M
Wake Forest	Wisconsin			Vanderbilt
Notre Dame*				

With the application of web scraping and sentiment analysis of the Twitter API, I was curious to see which Power 5 conference is the most negatively discussed compared to the others. Thus, the question I will be researching is:

Which Power 5 conference: the Big Ten Conference, the Big 12 Conference, the Atlantic Coast Conference (ACC), the Pac-12 Conference, or the Southeastern Conference (SEC), is the most negatively discussed in the United States?

I believe that looking into this topic will reveal how the fans of college football feel about each of the top tier conferences, and in the end, reveal which one is the most hated/negatively discussed. It is taking a look at qualitative data, which is normally overlooked when involving sports.

Related Research:

In the world of sports, there is always some type of statistic involved with players performance or the team as a whole, and with college football that fact does not waiver. There are an endless amount of analytical statistics available in sports that the area my question is based upon is regularly overlooked. The statistics of college football typically involve quantitative numbers based on a player's performance, but my research question is analyzing the accounts of everyday people and their thoughts on college football. In the field of my research question, I was able to find many sources pertaining to my topic; however, many of them come from sports media outlets, which are not academic sources unfortunately.

The first related research source is a list of the top 10 most hated football programs in the nation based off of a survey conducted by TheTopTen.com. From the data collected, it is shown that 9 of the top 10 teams on this list belong to a Power 5 Conference. Two of the teams come from the Big Ten, two come from the SEC, two from the Big 12, two from the ACC, and one from the Pac-12. This list displays an even spread of the top 10 most hated across the Power 5 conferences. This is of key importance because it will be interesting to see how the survey differs from the sentiment analysis of the tweets collected in my project ("Top 10 Most Hated College Football Programs.").

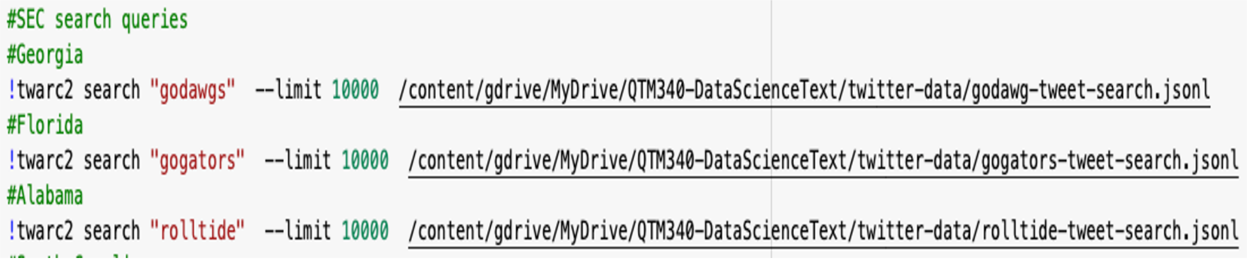
An analytical project done by a Data Analyst and 'college football stats expert' named Kyle Umlang was featured on Sports Illustrated. In his project, he analyzed many different aspects of college football in order to decipher the top college football programs ever. The data collected included the history of the AP Poll, NFL Draft selections, All-Americans, bowl games, national championships and Heisman winners. Now, this project takes a different approach and looks at quantitative data to make Kyle’s analysis, but it gives information on the top programs, which are normally in turn the programs of controversy. The reason this project is relevant to mine is because typically the most dominant teams historically get a lot of hate because of their success; therefore, with my analysis, this statement may be proven true (“Data Analyst Releases List of Top 35 College Football Programs All-Time.”).

Finally, an article by Bleacher Report, one of the most well known sporting news networks, put together a list of the best college football conferences. Bleacher Report’s analysis is taking into account many quantitative statistics to make claims on where a conference should land on this list. This creates a sense of where each conference falls quantitatively, but I collected qualitative data to generate a similar list. The SEC is ranked number 1, and from personal knowledge, they are one of the most discussed conferences in the nation; thus, it may receive a lot of praise or a lot of hatred (“Power Ranking Every FBS College Football Conference.”).

Corpus:

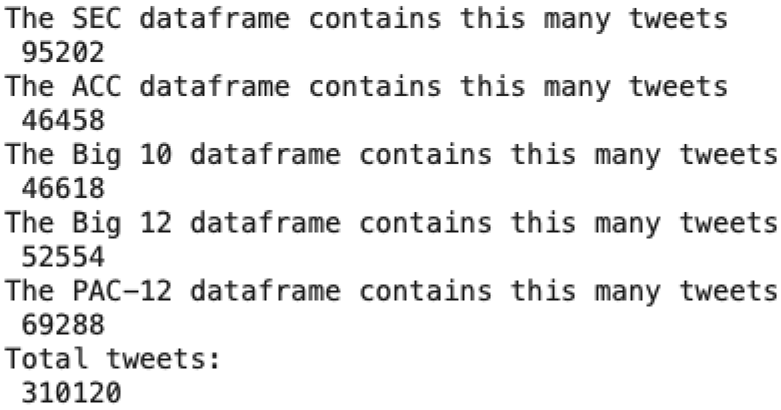
The corpus used for the analysis was a data set created using a python library called twarc2, which is a command line tool and library for gathering Twitter JSON data. In this corpus, the following variables are present: date, username, text, sentiment_ score, positive_score, negative_score, likes, and retweets. The text included in the corpus is all of the tweets found within the past 7 days that contained the search term associated with each team. These search terms implemented in the data collection process using twarc2 were found via an external source that created a list of all NCAA Division 1 Twitter handles and hashtags for each program (“NCAA Division I Football and Basketball Twitter Hashtags and Handles.”).

The data collection process was a main focal point of my project because every analysis that I made was from the data I scraped from Twitter myself. The beginning stages of the data collection process dealt with creating a list of all the teams that belonged to each Power 5 Conference, and the search terms I discussed earlier. This research was then implemented with twarc2 using the search function found in the twarc library, which would pull all of the tweets that used the search term associated and create a JSON file.



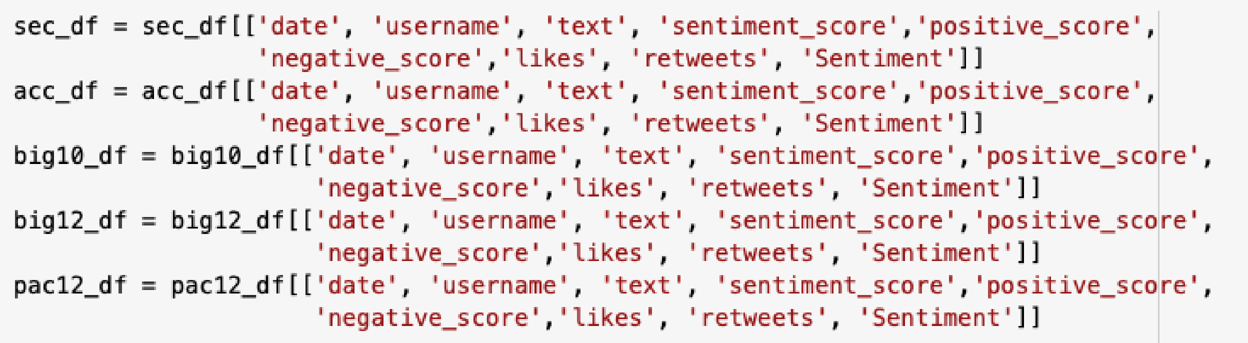
```
#SEC search queries
#Georgia
!twarc2 search "godawgs"  --limit 10000  /content/gdrive/MyDrive/QTM340-DataScienceText/twitter-data/godawg-tweet-search.jsonl
#Florida
!twarc2 search "gogators"  --limit 10000  /content/gdrive/MyDrive/QTM340-DataScienceText/twitter-data/gogators-tweet-search.jsonl
#Alabama
!twarc2 search "rolltide"  --limit 10000  /content/gdrive/MyDrive/QTM340-DataScienceText/twitter-data/rolltide-tweet-search.jsonl
```

The Twitter API only allowed access for tweets between the dates November 14th, 2021 to November 24th, 2021. This JSON data was then converted to a csv file in order to be loaded into my notebook easily using Pandas. After all the data frames were created for each team, I separated and joined each team's data into a larger set that contained all of the Twitter data collected for the teams within each conference. The total amount of tweets involved in my analysis totaled to 310,120 tweets, and a breakdown of the totals for each conference is displayed below:



```
The SEC dataframe contains this many tweets
 95202
The ACC dataframe contains this many tweets
 46458
The Big 10 dataframe contains this many tweets
 46618
The Big 12 dataframe contains this many tweets
 52554
The PAC-12 dataframe contains this many tweets
 69288
Total tweets:
 310120
```

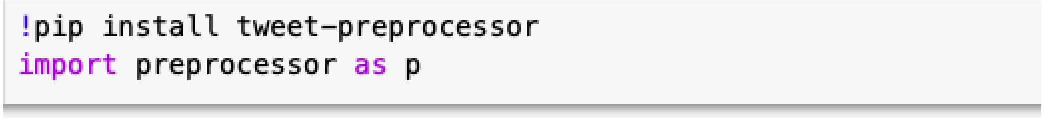
The data cleaning process was fairly simple because I only needed 5 of the 60 fields that the twarc programming pulled for each tweet; thus, all that was needed was renaming of the columns I wanted for my analysis and selecting those specific columns for each conference's data frame.



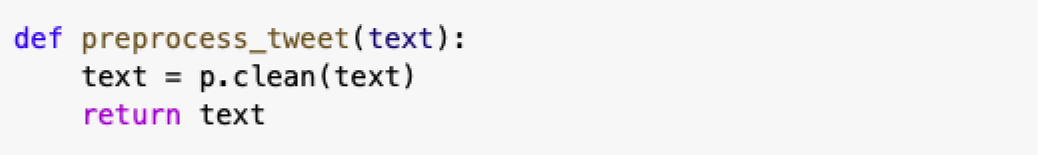
```
sec_df = sec_df[['date', 'username', 'text', 'sentiment_score','positive_score',
                 'negative_score','likes', 'retweets', 'Sentiment']]
acc_df = acc_df[['date', 'username', 'text', 'sentiment_score','positive_score',
                 'negative_score','likes', 'retweets', 'Sentiment']]
big10_df = big10_df[['date', 'username', 'text', 'sentiment_score','positive_score',
                     'negative_score','likes', 'retweets', 'Sentiment']]
big12_df = big12_df[['date', 'username', 'text', 'sentiment_score','positive_score',
                     'negative_score','likes', 'retweets', 'Sentiment']]
pac12_df = pac12_df[['date', 'username', 'text', 'sentiment_score','positive_score',
                     'negative_score','likes', 'retweets', 'Sentiment']]
```

Process and Methods:

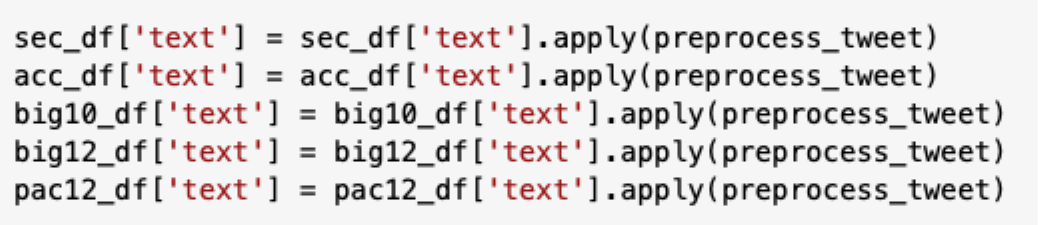
For the analysis of which college football conference is most negatively discussed, I chose to run the Sentiment Intensity Analyzer from the VADER library. This package is defined as a “natural language processing that analyzes people’s opinions, sentiments, evaluations, attitudes, and emotion via computational treatment of subjectivity in text” (“Introduction to Cultural Analytics & Python.”). The VADER SentimentIntensityAnalyzer is a program calculates a positive, negative, and compound score of the text that it is applied to, and it accomplishes this by combing through the text looking for a lexicon of words that have been assigned a predetermined sentiment score along with some other simple rules. This package is a very reliable and efficient way of gathering a sentiment analysis; however, there are certain punctuations, numbers, and emojis that deteriorate the programming. Thus, the first method I implemented was a simple function using the tweet-processor python library that removed all of these specific nuisances to the VADER program in order to obtain the most accurate scoring.



```
!pip install tweet-preprocessor
import preprocessor as p
```

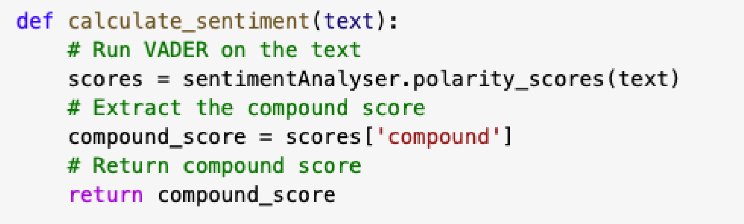


```
def preprocess_tweet(text):
    text = p.clean(text)
    return text
```

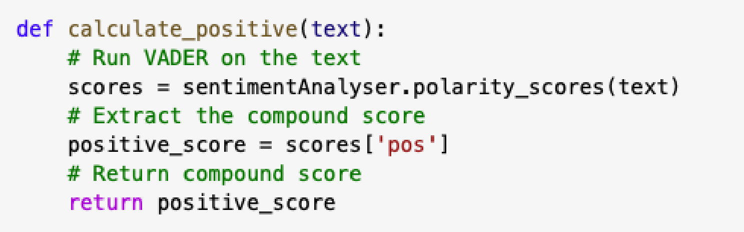


```
sec_df['text'] = sec_df['text'].apply(preprocess_tweet)
acc_df['text'] = acc_df['text'].apply(preprocess_tweet)
big10_df['text'] = big10_df['text'].apply(preprocess_tweet)
big12_df['text'] = big12_df['text'].apply(preprocess_tweet)
pac12_df['text'] = pac12_df['text'].apply(preprocess_tweet)
```

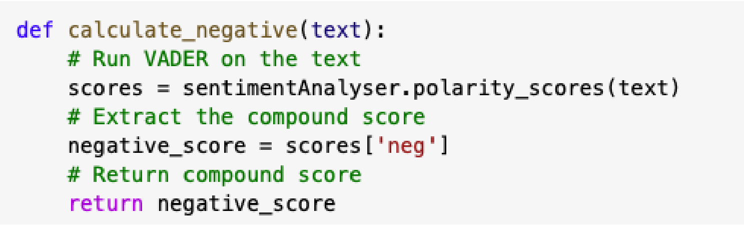
Once this was accomplished, I created functions that were to be run on the 'text' column of my data frames to obtain the sentiment scores for each tweet within the dataframes. Each function had the job of running the SentimentIntensityAnalyser, but each one was built to pull the compound score, positive score, and negative score.



```
def calculate_sentiment(text):
    # Run VADER on the text
    scores = sentimentAnalyser.polarity_scores(text)
    # Extract the compound score
    compound_score = scores['compound']
    # Return compound score
    return compound_score
```

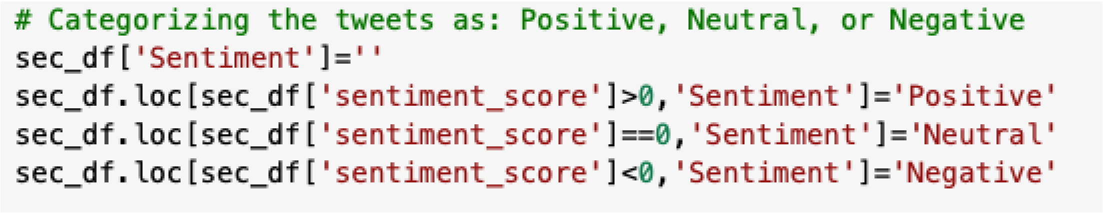


```
def calculate_positive(text):
    # Run VADER on the text
    scores = sentimentAnalyser.polarity_scores(text)
    # Extract the compound score
    positive_score = scores['pos']
    # Return compound score
    return positive_score
```



```
def calculate_negative(text):
    # Run VADER on the text
    scores = sentimentAnalyser.polarity_scores(text)
    # Extract the compound score
    negative_score = scores['neg']
    # Return compound score
    return negative_score
```

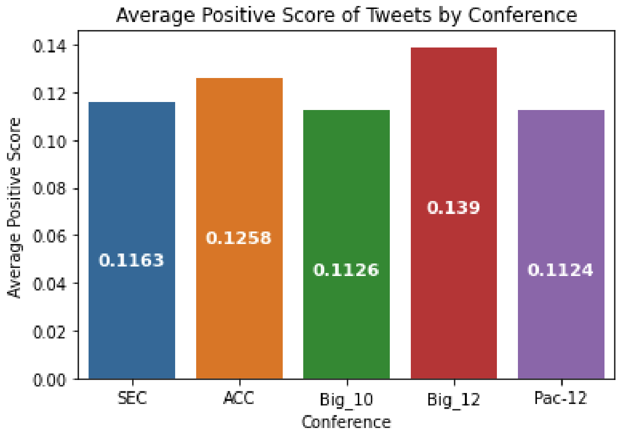
During the application of each of these three functions, a new column was created in each conference's dataframe to store the specific type of score. Furthermore, I then categorized these sentiment scores based on the compound scores, and created a new column that displayed whether a tweet was Positive, Neutral, or Negative.

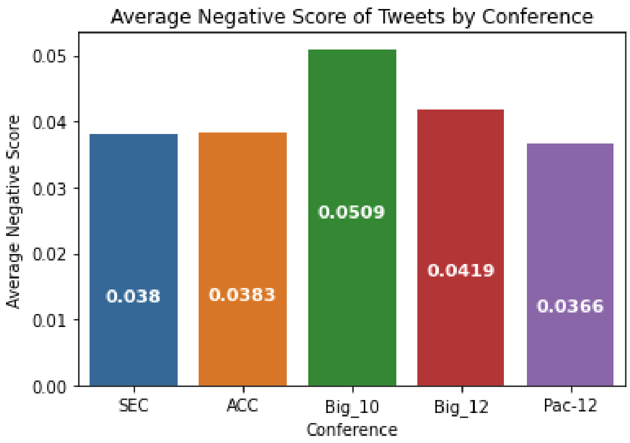


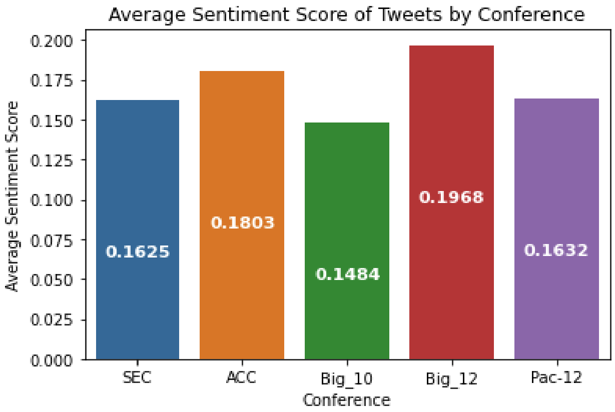
```
# Categorizing the tweets as: Positive, Neutral, or Negative
sec_df['Sentiment']=''
sec_df.loc[sec_df['sentiment_score']>0,'Sentiment']='Positive'
sec_df.loc[sec_df['sentiment_score']==0,'Sentiment']='Neutral'
sec_df.loc[sec_df['sentiment_score']<0,'Sentiment']='Negative'
```

Results and Discussion:

In order to make the Sentiment scores comparable based on each conference, I chose to apply the aggregate function of 'mean' to obtain the overall mean of the compound, positive, and negative scores. The visuals below display the results obtained from this process.





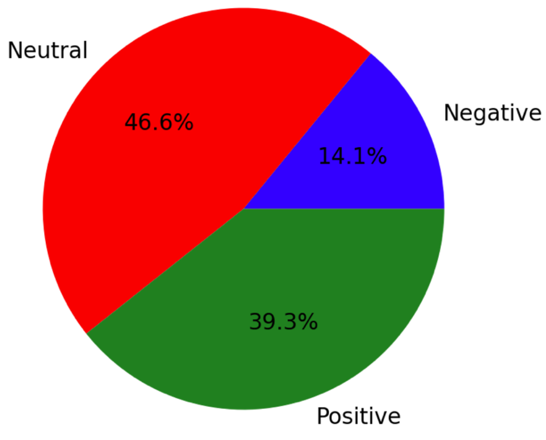


From these visuals, I was able to rank each of the Power 5 conferences from most negative to most positive for all three scores. The ranking is displayed here for all three categories.

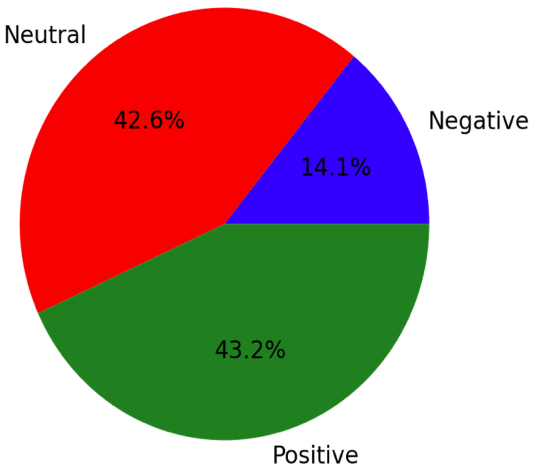
Conference	Overall Ranking (Most Negative to Most Positive
Big 10	1
SEC	2
Pac-12	3
ACC	4
Big 12	5

These rankings give a clear answer as to which Power 5 Conference is most negatively discussed and that is the Big 10 conference. This may be due to many factors of the fanbase, recent events, or overall record, but it was very interesting to see that the second best FBS College Football conference, based on the article by Bleacher Report, was at the bottom of these rankings.

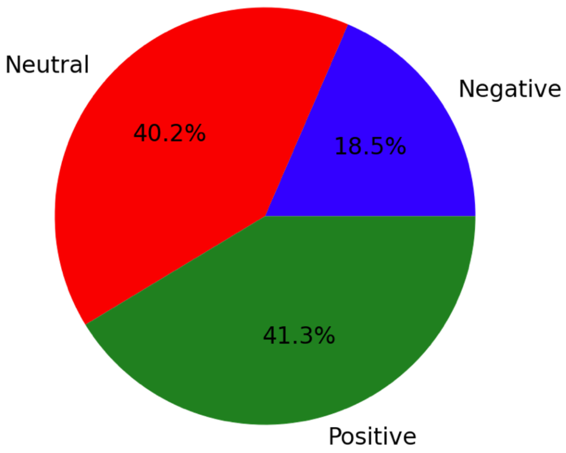
To validate the results from taking the averages of the sentiment scores, I chose to implement another analysis that dealt with the proportions of the tweets being positive, negative, or neutral. This was done by using the categorical data that was created at the end of my process, and the results are displayed below:



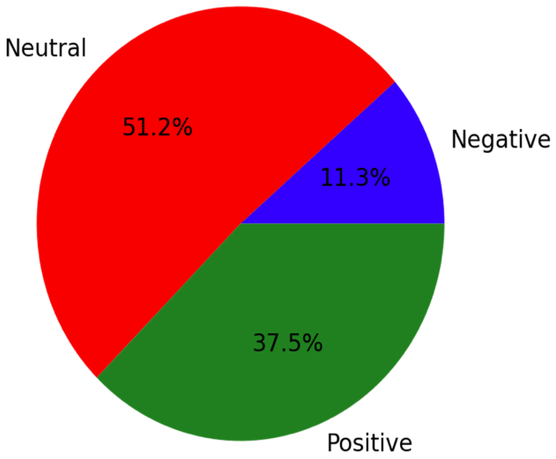
Proportion of Sentiment for SEC



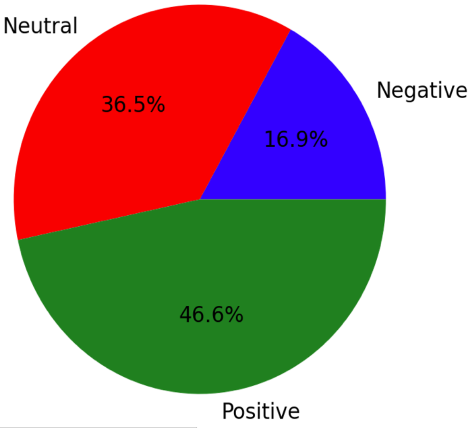
Proportion of Sentiment for ACC



Proportion of Sentiment for BIG 10



Proportion of Sentiment for Pac-12



Proportion of Sentiment for BIG 12

The result is portrayed the same when looking at the proportion of all the negative tweets pertaining to each conference. It is clear that the Big 10 had the highest proportion of negative tweets.

In all of these data visualizations, most of the conferences were around the same number when compared to the others. The survey discussed earlier done by TheTopTen.com, also, showed that there was an even spread of teams from each Power 5 conference in the Top 10 Most Hated College Football Programs. After analyzing the findings, I was able to find out from the Sports Illustrated article that 3 teams from the top 8 that Kyle Umlang selected were from the Big 10.

The final results came as a surprise to me based on my prior knowledge about college football. My initial guess as to the most negatively discussed conference was the SEC, but from these findings I was proven wrong. This, also, points out a form of bias in my initial thoughts because I was born and raised here in Georgia; therefore, the most prominent college football conferences are the ACC and the SEC which may have skewed my judgement.

Conclusion and Next Steps:

In conclusion, through the process of using twarc2 to obtain tweets relevant to the Power 5 Conferences I was able to give insight into the general public's opinion and thoughts of the beloved game that is college football. This project was just a glimpse into this topic because of the limitations of access to the Twitter API. I would like to take a look at tweets that date further back into the past to see if there is a tendency that occurs when comparing conference overall records to the sentiment scores of these times/seasons. Along with this, I would like to add further search terms to the initial search to obtain more tweets that pertain to each team. Overall, this text analysis manifested fantastic results that could be applied in a much broader way to the world of college sports.

Work Cited:

"Introduction to Cultural Analytics & Python." *Twitter API Setup - Introduction to Cultural Analytics & Python*, https://melaniewalsh.github.io/Intro-Cultural-Analytics/04-Data-Collection/11-Twitter-API-Setup.html.

"NCAA Division I Football and Basketball Twitter Hashtags and Handles." *All My Sports Teams Suck*, 31 Dec. 2019, https://www.allmysportsteamssuck.com/ncaa-division-i-football-and-basketball-twitter-hashtags-and-handles/.

The Spun. "Data Analyst Releases List of Top 35 College Football Programs All-Time." *NewsBreak*, The Spun, 3 July 2020, https://www.newsbreak.com/news/1594598073349/data-analyst-releases-list-of-top-35-college-football-programs-all-time.

"Top 10 Most Hated College Football Programs." *TheTopTens*, https://www.thetoptens.com/football/most-hated-college-football-programs/.

Wallace, Greg. "Power Ranking Every FBS College Football Conference." *Bleacher Report*, Bleacher Report, 6 Oct. 2017, https://bleacherreport.com/articles/2734949-power-ranking-every-fbs-college-football-conference.