ISYE 6740 – Spring 2021
Final Project Report

Team Member Names: Alex Dreo, Evan Reed, and Skye Sheffield

Project Title: What produces high engagement rates in tweets?

Problem Statement

Over the past few years, Twitter has become a battleground to make or break companies and personalities. The famous "Chicken War" between Chick-fil-a and Popeyes on Twitter was so impactful that Popeyes sold out of three months' worth of chicken in just two weeks (Taylor, 2019). Since then, Popeyes has increased sales by $400K per restaurant (Klein, 2021). The final tweet on the Chadwick Boseman account announcing the "Black Panther" stars death holds the record for the most liked tweet ever followed by the tweet announcing the death of Kobe Bryant (Spangler, 2020). So, what does it take for a tweet to generate high engagement?

Based on 2019 statistics, there are roughly 250 million posting users and roughly 300 million tweets per day; however, roughly 50% of all tweets are retweets which means there are only about 150 million unique tweets a day (Leetaru, 2019). The most followed user on Twitter today is Former President Barack Obama with roughly 130 million followers; however, just having the most followers does not necessarily mean you always have the most engaging tweets just the largest audience. For the purposes of answering this question, the project focused on investigating the most followed Twitter users as they tend to have the most activity and what combination of factors produce a high engagement rate: interactions per number of followers.

Data Sources

The data for this project was gathered using Twitter's API. The data consisted of approximately 3,200 tweets, retweets, and replies from each of the 50 most popular accounts currently on the social network, determined by the number of followers of the accounts. We began our analysis by first preprocessing the data - transforming the raw data into a usable structure suitable to train the models. To reduce the scope of the model, retweets and direct messages were removed from our dataset, as the features that indicate a highly liked tweet may not extend to these specific types of tweets. Features for time and day were one-hot-encoded so that each of our features were either a count or binary indicator. The metadata features shown in Table 1 were derived from the tweet text and raw tweet metadata.

Table 1: Initial feature definitions

Feature	**Definition**
URL Count	Number of URLs included in the Tweet text
Hashtag Count	Number of Hashtags included in the Tweet text
Mention Count	Number of Mentions included in the Tweet text (A mention is tagging another user)
Contains URL	Binary indicator of URL (1) or not (0)
Contains Hashtag	Binary indicator of Hashtag (1) or not (0)
Contains Mention	Binary indicator of Mention (1) or not (0)
Time Buckets (One hot encoded)	Binary indicator for the UTC time of the tweet (6 columns in total) - Early Morning - Morning - Midday - Evening - Night - Late Night
Day Buckets (One hot encoded)	Binary indicator for the day of the week of the tweet (7 columns in total) - Column for each day of the week: Monday,…,Sunday
Is Weekend?	Binary indicator for if tweeted on Saturday or Sunday (1) or not (0)
Emoji Count	Number of emojis used in the tweet
Tweet Length	Number of characters in the tweet

Furthermore, non-English tweets were removed from the data set to leverage an English-based Natural Language Processing (NLP) model (*spaCyTextBlob*). Cleaned tweets, with emojis and URLs removed, were passed to this model. The model then tokenized the tweet to break it up into words. Then, these tokens are passed to the pre-trained sentiment model, which assigned polarity and subjectivity. Polarity and subjectivity, defined below, each captured some of the intent behind the tweet, while the derived feature, emotionality, represents polarity in a different way.

Table 2: NLP feature definitions

Feature	**Range**	**Definition**
Polarity	[-1,1]	Positivity or negativity of tweet, with -1 corresponding to a very negative tweet.
Subjectivity	[0,1]	Objectivity vs subjectivity where 0.0 is very objective and 1.0 is very subjective
Emotionality	[0,1]	The degree of positive or negative sentiment associated with a tweet. This field is defined as the absolute value of polarity.

Once all these fields were calculated, the historical follower counts of each of the users were collected for each year in the dataset. Since historical follower counts are not available to the free tier of the Twitter API, follower counts had to be manually scraped using the Wayback Machine (*Internet Archive: Wayback Machine*). A linear regression model was then created for each user to estimate follower counts at the time of each tweet, as it was not possible to measure the actual follower count for each time of tweet.

In the earlier years of twitter, a non-linear follower count trend was observed. This exponential growth trend, shown in appendix figure 1 eventually transitioned to linear growth in approximately 2012. For this reason, tweets prior to 2012 were excluded to increase the accuracy of the linear follower count model. This simple imputing technique achieved a mean absolute error of approximately 3 million followers, which is relatively low considering some users have more than a hundred million followers. The training data used to create these linear regressions is shown in appendix figure 2.

Using these follower count values, like ratio was calculated as the number of likes the tweet received divided by the follower count at the tweet post time. This like ratio was used to normalize the effect of followers on a tweet's success. Without this normalization, higher followed users would tend to receive many more likes than those with fewer followers. This ratio effectively de-coupled a user attribute, follower count, from the tweet attribute: number of likes.

Methodology

After completing preprocessing, a feature selection analysis was conducted to identify prominent features and eliminate insignificant features in the dataset. To explore the possibilities of the predictive abilities of the data for different outputs related to tweet engagement, the process was done over two different response variables (like count and like ratio).

Using the Elastic Net technique, a $k = 10$ cross fold validation model was constructed for each response. A range of alpha values were used to tune the models to minimize the sum of squares regression (SSR). After analyzing the results that produced the lowest SSR error for each of the two response variables, it was concluded that the Elastic Net approach was not able to eliminate any of the features from either of the models as none of the coefficients converged to zero at this optimal solution on minimizing error.

To further explore the relationship between the features, a correlation matrix of the features was created. As expected from the output from the Elastic Net analysis, the results did not indicate any significant level of correlation between the fields (except for the one-hot-encoded features and the fields directly derived from one another).

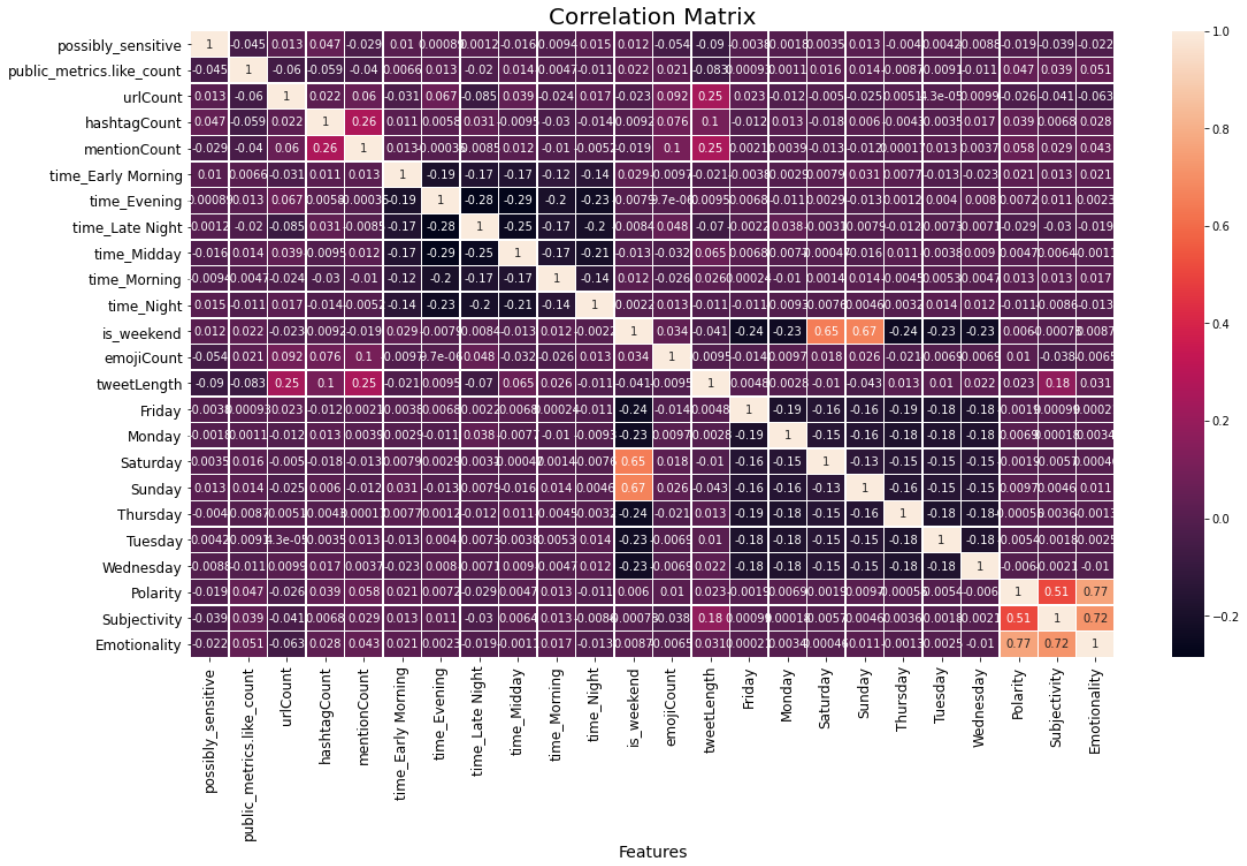

Figure 1: Correlation plot between features.

Next, using the Elastic Net alphas that produced the smallest SSR, the linear regression models were run against the respective testing data sets. The resulting mean absolute errors (MAE) were used to set a baseline measure of the predictive quality of the regression models with the goal of improving the accuracy.

Table 3: Model results from Linear Regression Model

Data	Response variable	MAE
Unfiltered Data	Like Count	24829.047051
Unfiltered Data	Like Ratio	0.0006132

Following the Elastic Net tuned linear regression models, a Random Forest regression model was built to predict the number of likes an influencer would have on their tweet based on the tweet features. The initial parameters of the random forest model were as follows: no max_leaf_nodes, no max_dept, min_sample_left of 1, min_sample_split of 2, and 100 n_estimators. The first response variable used was the number of likes. This model was developed with all the data after preprocessing - no data was filtered out and the response variable was the number of likes. This model predicted tweet length, subjectivity, polarity, and emotionality to be the four most significant features with relative importance values of 0.29,

0.13, 0.1 and 0.1 respectively. All the other features had importance values of .04 or smaller. This model was 1.86% better than random guessing with a MAE value of 30137.69 rather than 30710.12 from random guessing. The improvement of 1.86% was calculated by taking the difference between the MAEs for the random guess and random forest model and dividing the result by the random guess MAE. The random guess, or mean guess, model is determined by using the y predicted in the calculation as the mean of the response variable.

The distribution of likes was heavily skewed towards 0 likes per tweet, as can be seen by the overall like count histogram in appendix figure 3. To address this and improve the model performance, outliers were removed from the dataset to create a filtered dataset. Tweets with less than 100 likes, and tweets with a like ratio greater than 0.01, were excluded from our dataset. These cutoff thresholds were chosen through iteration over many possible cutoff values. The lower bound was necessary to remove the tweets with little to no engagement, as the goal of this model was to predict successful tweets. The upper bound of the like ratio of 0.01, was chosen to normalize to the heavy-tailed tendency of this distribution, which can be seen in the yearly-like histogram in appendix figure 4. The improvement of the model after these highly liked tweets are removed represent a gap in the current model. This means additional features would be required to improve performance at the extreme highs of the like-ratio. Even with this upper bound, the like-distribution was still heavily skewed towards 0, as can be seen from appendix figure 4.

This filtered random forest model also had tweet length, subjectivity, polarity, and emotionality to be the four most significant features with importance of 0.28, 0.13, 0.1 and 0.08 respectively with all the other features also having importance values of less than 0.04. This model was 3.14% better than random guessing with mean absolute error values (MAE) of 24,567 and 25,364 respectively. Furthermore, we choose to filter out all tweets that took place after March 31st, 2021 as users may not had enough time to interact with these newer tweets considering the data was pulled on March 15th, 2021. This resulted in an improvement of 4.84% in respect to random guessing although the overall MAE decreased in comparison.

Staying with the filtered data but changing the response variable to the like ratio with respect to the number of followers increased improvement by 7.72%. This model also had tweet length, subjectivity, polarity, and emotionality to be the four most significant features with importance of 0.29, 0.13, 0.09 and 0.08 respectively with all the other features also having importance values of .04 or smaller. All these results can be found in Table 4.

Table 4: Model Comparison

Random Forest Model	Data	Response variable	Random Forest MAE	Random Guessing MAE	Improvement over Random Guessing
Base	Unfiltered Data	Number of likes	30137.69	30710.12	1.86%
Base	Filtered Data	Number of likes	24566.64	25363.67	3.14%
Base	Filtered out Data after March 2021	Number of likes	25398.81	26689.65	4.84%
Base	Filtered Data	Like Ratio	0.000574	0.000622	7.72%
Grid Search	Filtered Data	Like Ratio	0.000543	0.000622	12.7%

Finally, hyper tunning of the random forest model was conducted to improve the results using a grid search method with k fold cross validation to find the best parameters. The best estimators from the grid search were 13 max_depth, 4 min_samples_leaf, 10 min_samples_split, and 157 n_estimators which resulted in a 5.40% improvement on the like ratio filtered data model with an MAE of 0.000543. Thus, resulting in an improvement of 12.7% compared to random guessing. This final model when applied the testing data illustrates effectiveness of the model on data outside of the training set. Ultimately, this was the best result of any model that was created, showing a notable improvement in the MAE compared the initial Elastic Net tuned regression model.

Evaluation and Final Results

Initial expectations for the model were to produce results for the selected set of users that would sizably outperform random guessing. However, only marginally better results were obtained, although the model did illustrate a quantifiable improvement. The best model performed 12.7% better than random guessing using random forest with grid search. We believe that additional features that were not included in our model, such as the effect of major events, disasters, election cycles, etc., would be crucial in decreasing error.

As stated in the initial proposal, if these models are used to predict engagement rates for other users, their accuracies will most likely suffer even more as this dataset of most followed users is biased towards the rich and famous. This is particularly true for users with a low number of followers, as a small change in number of followers will have a drastic effect on the like ratio. In order to avoid this pitfall, the models would need to be trained on a larger dataset with a varied group of twitter users; however, this was not in scope for our project given the timeframe.

Future Work

Besides including more tweets into the dataset, the simplest way to improve this model would be to include more relevant features. Additional analysis of tweet content could be done through further sentiment analysis. A bag of word analysis on the hashtags could be conducted to determine if there are trends between the use of a hashtag and the engagement rate. Additionally, adding in the effect of outside factors not captured in the tweet data would likely contribute significantly to the popularity and engagement of the users such as major events, disasters, election cycles, etc. Furthermore, the response variable influenced the model performance. Initially, we choose to look at the number of likes per tweet and then moved to a tweet ratio approach and further work could be done to examine the overall engagement rate if users comment or retweet a post without liking it. Additionally, investigation into which tweet attributes would cause one demographics to engage in higher rates could be an additional avenue to improve tweet like prediction.

Team Contribution

All team members contributed equally and collaborated on all sections to complete this project.

Appendix

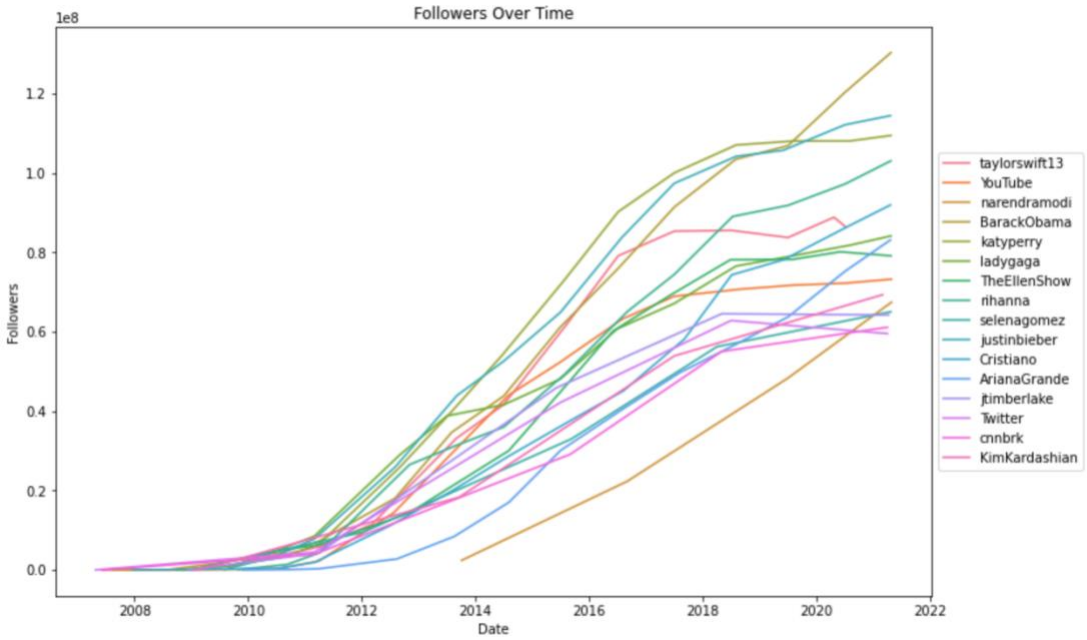

Figure 1: Follower Subset Over Time. This subset of users' follower counts over time shows the non-linearity in the follower trends during the earlier years of twitter.

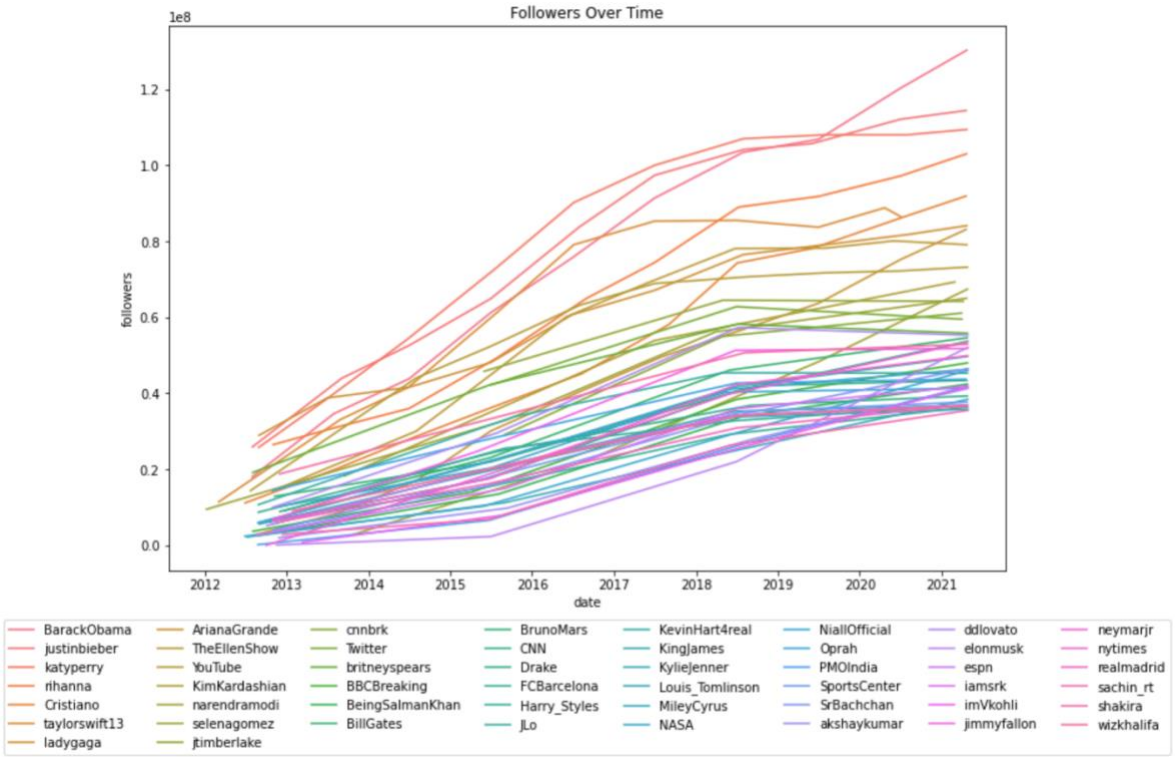

Figure 2: Followers Over Time. This plot starting at 2012, of all 50 users, shows that the model can be treated as a linear regression with respect to time.

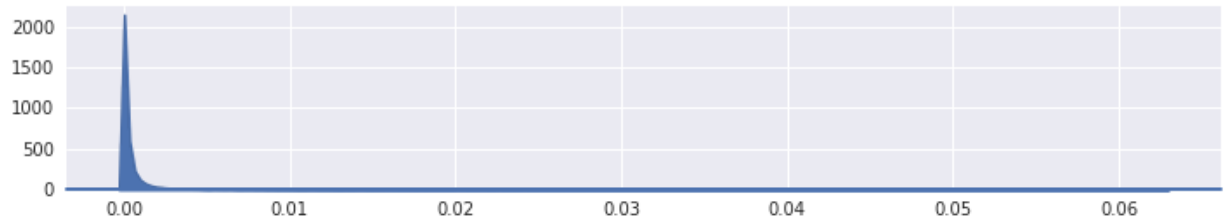

Figure 3: Untruncated Like-Ratio Histogram. This plot demonstrates the heavy-tailed-ness of the tweet distribution, with respect to the number of likes they receive.

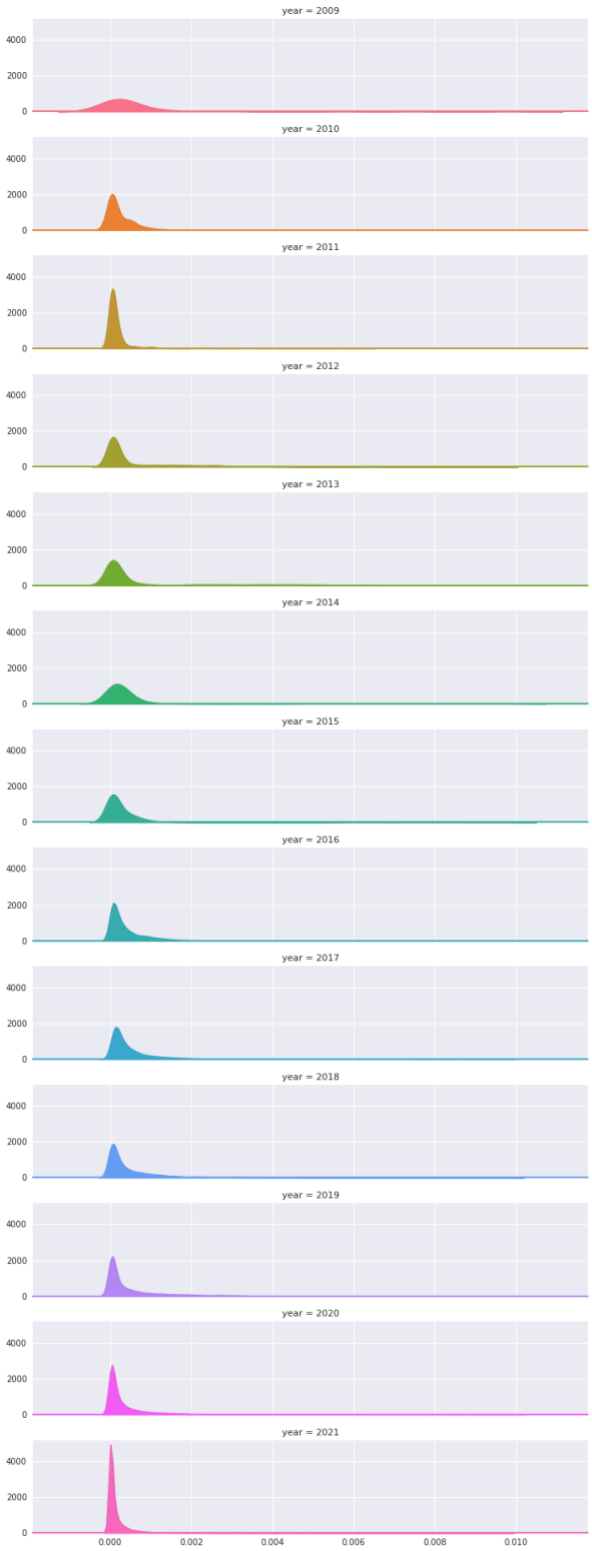

Figure 4: Truncated Like-Ratio Histogram. This plot shows the distribution of like-ratios.

Resources

Klein, D. (2021, February 21). *Thanks to Chicken Sandwich, Popeyes is Making $400K More Per Restaurant*. QSR Magazine. https://www.qsrmagazine.com/fast-food/thanks-chicken-sandwich-popeyes-making-400k-more-restaurant

Leetaru, K. (2019, April 24). *A Fading Twitter Changes Its User Metrics Once Again*. Forbes. https://www.forbes.com/sites/kalevleetaru/2019/04/23/a-fading-twitter-changes-its-user-metrics-once-again/?sh=68b758007a31

Spangler, T. (2020, December 7). *Chadwick Boseman Twitter Post Announcing His Death Is Most-Retweeted of 2020*. Variety. https://variety.com/2020/digital/news/chadwick-boseman-twitter-most-retweeted-most-liked-1234847935/

Taylor, K. (2019, December 12). *7 fast-food Twitter feuds that defined the decade*. Business Insider. https://www.businessinsider.com/fast-food-twitter-feuds-wendys-popeyes-chick-fil-a-decade-2019-12?international=true&r=US&IR=T#taco-bell-battles-old-spice-1

spaCyTextBlob · spaCy Universe. spaCyTextBlob. (n.d.). https://spacy.io/universe/project/spacy-textblob.

"Twitter API Documentation | Twitter Developer." *Twitter*. Twitter. Web. 15 Mar. 2021.

Internet Archive: Wayback Machine. (n.d.). https://archive.org/web/.