BDS 516 Homework #9
Galapagos Penguins

Introduction

Although they achieved fame in different ways, Kanye West and Barack Obama are two of the most influential people of our time. Kanye West has amassed over 30 million followers, mostly from his incredibly successful entertainment career and relationship with mega-socialite family, the Kardashians, but has also received recent attention for his most recent foray into the world of politics as a 2020 candidate running on a conservative platform. On the other hand, Barack Obama established his prominence as an American politician and lawyer. He served as the 44th president of the United States, serving from 2008-2016, and was the first African-American president of the U.S. He is a member of the Democratic party and has a relatively progressive political platform. Since his presidency, Obama continues to remain an influential figure, both in politics and pop culture.

One factor that makes this comparison interesting is that both Kanye and Obama are politicians (to differing degrees). However, they have very different political ideologies, are in different phases of their careers, differ in personality, and differ in demographic characteristics (such as age). These characteristics influence the characteristics of their followers which may, in turn, influence the content of their tweets and how their followers engage with their tweets. Therefore, this may affect factors such as the timing, source, content, sentiment, and engagement level of their tweets. For this reason, our analysis aims to address the following question: can an algorithm based on tweet source, use of quotes, use of photos, and sentiment scores successfully predict whether a tweet was made by Barack Obama or Kanye West?

Methods

Twitter data from both Kanye West and Barack Obama was collected through a Twitter Developer account. After receiving a twitter API, tweets from both @kanyewest and @BarackObama were extracted from twitter in R Studio using the *rtweet* package. Other packages used to run models, manage and clean data, create visuals, text mine, run text analyses, and calculate statistics include: *httpuv, tidytext*, *tidyverse*, *dplyr*, *lubridate*, *ggplot*, *scales*, *readr*, *syuzhet*, *mlbench*, *caret*, *vtreat*, and *InformationValue*.

Of all tweets posted, 3,200 tweets from both @BarackObama and @kanyewest were sampled randomly on April 21st, 2021. Dataframes containing the two sets of tweets were cleaned to only contain the source (which device the tweet was made on), status_id (ID of the tweet), text (text content of the tweet), created_at (time when tweet was created), retweet_count (number of retweets), favorite_count (number of favorites), is_retweet (whether or not the tweet is a retweet), and screen_name (screen name of the individual who tweeted it).

Prior to modeling, a number of summary statistics were calculated. To analyze differences in source of tweets, the count function was used to examine from which device both Kanye and Obama tweeted from most often. For time of day, the percent of total tweets sampled tweeted at each hour of the day was calculated for both Kanye and Obama and graphed as a line graph. Time tweeted was standardized using Eastern Standard Time (EST). Furthermore, the percentage of total tweets sampled starting with a quotation mark was calculated. Tendencies to include pictures or links in tweets was captured by calculating the total percentage of tweets containing either a picture, a link, or both among the tweets sampled. Furthermore, retweets and favorites were analyzed by calculating the percentage of tweets that are retweets out of the total sample, the average number of retweets on each of their tweets

among the tweets sampled, and the average number of favorites on each of their tweets among the tweets sampled. Lastly, average sentiment for tweets were calculated for both Kanye and Obama. This was performed by removing all non-alphabetic characters using the *str_replace_all* command. Sentiment scores were calculated using the NRC Sentiment Dictionary. By running the command get_nrc_sentiment, sentiment scores were calculated for tweet texts. The average for both Kanye and Obama was then calculated for each of the following sentiments and valence: anticipation, fear, disgust, joy, sadness, surprise, trust, negative valence, and positive valence. Notably, percentages were used for comparisons because Kanye had fewer than 3,200 tweets while Obama had enough tweets to reach 3,200 sample tweets. Therefore, this was done to account for the difference in total number of tweets in the dataset for each person.

Multiple logistic regression was used to develop the predictive algorithm, using the *glm* function in R. The following predictors were included: whether the tweet uses a quotation mark, whether the tweet features a picture or link, and sentiment scores for the sentiments anticipation, fear, joy, trust, and positive valence. These factors were chosen due to a notable difference in behaviors observed between Kanye and Obama based on the summary statistics. The outcome variable was captured as being tweeted by Obama or not tweeted by Obama (i.e. tweeted by Kanye). The algorithms predictive power was then tested on a new set of tweets made by both Kanye and Obama. A train-test split was created with 70% of tweets from both Kanye and Obama included in the train set and 30% from each included in the test set.

Model diagnostics were calculated, including the optimal prediction probability cutoff, the misclassification error, the sensitivity level, and the specificity level. An AUC-ROC curve was created to demonstrate the ability to distinguish between the positive (i.e. Obama's tweets) and negative class (i.e. Kanye's tweet).

Lastly, the model was used on a set of tweets from an unrelated user. A random sample of 3,200 of Drake's tweets (@Drake) run through the model. Examples of tweets by Drake that were classified as Obama and Kanye were then analyzed by examining how the characteristics of the tweets related to each component included in the model.

Results

Summary Statistics

Primary Tweet Source

When analyzing the source of Kanye and Obama's tweets, we see that Kanye primarily tweets from an iPhone (n=1830 or ~98%) while Obama primarily tweets from a desktop/laptop computer (n=2446 or ~76.5%).

Time of Day

The majority of Obama's tweets were made between 10am and 4pm with very few made before 7 am or after 9 pm. This is potentially due to staffers making many of his tweets on his behalf, especially for tweets made while he was in office. On the contrary, Kanye's time of tweets is much more variable, with many tweets being made in the late morning/early afternoon (~11 am to 3 pm) but a considerable number being made late at night (~9 pm to 1 am). This may indicate fewer individuals making tweets on Kanye's

behalf/a more personal use of twitter since these tweets are occurring outside the typical workday.

Percentage of Tweets Starting with Quotation Marks
While neither Kanye nor Obama begin tweets with quotation marks particularly often, Obama does appear to use them significantly more since Kanye uses them so infrequently. While 17% of Obama's tweets use quotes at the beginning, fewer than 1% of Kanye's tweets use quotation marks.

Percentage of Tweets Containing Pictures or Links
Whether or not a tweet contains pictures may also be useful information for predicting whether a tweet was made by Obama or Kanye. Out of this sample, 92% of Obama's tweets contain a picture or link whereas only 55% of Kanye's tweets do.

Percentage of Tweets that are Re-tweets
Obama and Kanye have nearly the same percentage of tweets that are retweets. 11% of Obama's tweets are retweets and 10.6% of Kanye's tweets are retweets.

Retweet/Favorite Counts on Original Tweets
Overall, Obama gets 11,400 retweets on average and 58,068 favorites. Meanwhile, Kanye gets 9,314 retweets on average and 48,920 favorites. However, it is important to note that Obama has 130 million followers and Kanye only has 30 million followers. This may indicate a higher engagement rate with Kanye's tweets compared to Obama's. Considering we are looking at average retweets/favorites regardless of follower count, the similarities in retweet and favorite counts for Kanye and Obama may indicate that these characteristics may not be particularly predictive.

Sentiment
On average, it seems that Obama expressed more emotion through his tweets than Kanye. Obama's tweets had a higher sentiment score across all sentiments included in this analysis (i.e. anticipation, fear, disgust, joy, sadness, surprise, trust, negative valence, positive valence) compared to Kanye. Obama's tweets scored particularly higher for "anticipation", "trust", and "positive valence".

Predictive Algorithm

All predictors in our logistic regression model were significant at the $p<.001$ level when holding the other predictors constant. Whether or not the tweet included a quotation mark had an odds ratio of 797. This means that tweets containing a quotation mark have a 797x greater odds of being Obama's tweets than Kanye's tweets when all other predictors in the model are held constant. Furthermore, tweets with a picture or link included have a 30x greater odds of being Obama's tweet than being Kanye's tweet when all other predictors are held constant. Sentiments do not demonstrate such a strong relationship with the author of the tweet, but do demonstrate a meaningful relationship. The odds ratios for anticipation, fear, trust, joy, and positive valence are

1.74, 1.78, 2.08, 0.316, and 2.27, respectively. This means that tweets expressing more anticipation, fear, trust, and positive valence have a higher odds of being tweeted by Obama. Nonetheless, tweets with higher joy scores have a lower odds of being authored by Obama.

Unfortunately, Kanye's total number of tweets did not sum up to 3,200. Therefore, there were no new tweets to run the algorithm on. Therefore, the model was evaluated on a 70-30 train-test split. Using the test data, the optimal prediction probability cutoff was found to be 0.52. Furthermore, the misclassification error was 0.15, indicating that the percentage of incorrectly classified instances is 15% and the AUROC was 0.8916, indicating that there is a 90% chance that the model will be able to distinguish between the positive class (i.e. Obama's tweets) and the negative class (i.e. Kanye's tweets). Furthermore, the sensitivity and specificity were 0.88 and 0.79, meaning that the model correctly identified Obama's tweets as being tweeted by Obama 88% of the time and the model correctly identified tweets by Kanye as being tweeted by Kanye 79% of the time. Overall, these diagnostics indicate that our model has strong predictive power in distinguishing between tweets by Kanye and tweets by Obama.

Running the Algorithm on @Drake's Tweets

When the original prediction algorithm was used on a data set composed of tweets posted by just Drake, the algorithm classified 91% of those tweets as being Kanye West's tweets and 9% being Obama's.

When pulling a tweet from Drake that was classified as an Obama tweet, we found that the tweet did not begin with a quote or link but included a picture. It had an anticipation score of 4; fear score of 1; joy score of 2; trust score of 3; and positive score of 5. Given that our algorithm found that pictures and sentiments of anticipation, fear, trust, and positive *all* increase the odds of a tweet belonging to Obama instead of Kanye, it is unsurprising that a tweet by Drake consisting of an image and with such sentiment scores was coded as an Obama tweet.

We then pulled a Drake tweet that was classified as a Kanye tweet. This tweet also did not begin with a quote, and included a picture. However the sentiment scores were different. It had an anticipation score of 0; fear score of 0; joy score of 1; trust score of 0; and positive score of 0. While our algorithm found that a one unit increase in the sentiment score for joy *decreases* the odds of the tweet being authored by Obama by 68%, it also found that tweets with pictures and links have a 30 times odd of being an Obama tweet. As such, this classification seems to be rather perplexing given the differing magnitude of these variables.

Conclusions

We found that the tweets from Obama's account were more likely to be created from a desktop or laptop computer compared to the tweets from Kanye's account which were

from an iPhone. This may, in part, explain why we also found that Obama was more likely to tweet between 11am and 3pm - within standard work hours - compared to the greater variability in the timing of tweets from Kanye. It is also important to note that, as a performer, Kanye may be working late at night (i.e. when concerts tend to occur) which may also explain why his tweet behavior looks differently than someone like Obama who may, along with his staffers, work closer to the conventional work day. Finally, it is important to note that both Kanye and Obama likely travel often, leading to variability in the time zones in which they make tweets. This is a weakness of looking at when tweets are made in EST since the increase in Kanye's tweet time variability may reflect more frequent travel rather than a meaningful difference in which tweets are made. Nonetheless, although this may not directly measure the hour at which each of them tweet, this may still be a useful input in our model to predict which tweets belong to whom.

We also found that Obama's tweets conveyed more emotion compared to Kanye's. This may not be particularly surprising as Obama and Kanye may be tweeting to different audiences and have different pressures when crafting their tweets. It is also possible that Obama's tweets are registered as being more emotive as it may be considered inappropriate or risky for him to use more complex language, like sarcasm. As someone in office, he likely had to be careful that his tweets were literal so that they were less likely to be misconstrued by the public which would then make it easier to associate his tweets with specific emotions. Kanye may not face quite as much public scrutiny as an entertainer, so he may use language that is considered to be more neutral by the NRC algorithm or use more sarcastic language. There may also be less pressure on Kanye to relate or empathize with his twitter audience whereas there may be more of an expectation for a president to do so. This raises an interesting question if Kanye's sentiment levels changed following his announcement that he would run for president (although, not every politician conforms to society's expectations of presidential conduct).

As noted above, our model has an AUC of .9, indicating that there is a 90% chance that any given tweet will be correctly diagnosed as either Obama's or Kanye's. Further, the model's true positive rate and true negative rate are both about .85, indicating that our model has strong predictive power. When we tried our prediction algorithm on a third user's tweets (Drake's), it overwhelmingly classified his tweets as having originated from Kanye's account. This is not very surprising, given that Kanye and Drake are entertainers of a similar (43 vs. 34) age.

Overall, we feel that our predictive model does an impressive job of classifying tweet origin. Barack Obama and Kanye West are both African American men with a significant cultural impact. Yet, it is unsurprising that they tweet in different ways, considering their difference in personal characteristics, backgrounds, and primary audiences. Nonetheless, it is noteworthy that those differences can be ascertained by our model and put to use on test data.