TrumpTweetArchive Project

Introduction

Trump's use of Twitter for promotion and communication is unprecedented for the United States Presidency. He is an active user of Twitter, and the trumptwitterarchive is a collection of his every tweet, retweet, and favorite.
"David S. Ferriero informed two Democratic senators last week that the Trump’ss tweets are being preserved for posterity, so future generations can revel in the president's peculiar and unprecedented use of Twitter as an art form and governing tool." -source theVerge

Problem Statement

Predict engagement levels based on text sentiment and other feature columns. The model analyzes the dataset starting after the Republic National Convention on July 18, 2016.

Hypothesis and Assumptions

- Build a model that forecasts the public engagement level from Trump’s tweets
- Target Engagement (y), which is defined as the combination of favorite_count + retweet_count
- Feature Columns: ['source', 'text', 'created_at', 'is_retweet']
- Additional Feature Columns from Text:

    Create a sentiment score on Text. Does negative sentiment increase engagement?

Natural language processing

Diving into sentiment analysis, scoring is based on TextBlog’s sentiment polarity function (TextBlob(text).sentiment.polarity), which returns a score between the range of -1 to 1. A Score of 1 has a very positive sentiment. Next, the model categorizes sentiment into 4 bins: Very Positive, Moderately Positive, Moderately Negative, and Very Negative sentiments.

In addition, the model uses the CountVectorizer function to understand Trump’s vocabulary on Twitter. We store the text column names and sum of counts to find the 25 most used words. Specifically, we used the CountVectorizer(min_df=5, max_features=5000 ,stop_words='english', ngram_range=(1,3)). Ngram set range 1-3 to capture phrases such as “make america great” or “fake news”.

LinearRegression and Analysis

LinearRegression model deployed for its ease of interpretation. Using Split/Train/Test, we derive the following scores. We can validate that the model performs better than a model of just mean values.

	Model	Null Model
MAE:	0.582941962	0.776758012
MSE:	0.814353678	1.182711508
RMSE	0.902415469	1.087525406

Here are the results from the Linear Regression model on Log(Engagement):

Factor	Value
Intercept	11.226
is_retweet	-1.972
is_afternoon	-0.260
is_evening	-0.245
is_latenight	-0.285
moderately_positive_sentiment	-0.082
moderately_negative_sentiment	0.100
very_negative_sentiment	0.227

The results are not surprising. A very negative Trump’s tweet leads to a .227% increase in engagement. More recent news uncovered Russian efforts to promote divisiveness, and it’s possible that their bots and trolls are involved in increasing overall engagement, especially around politically negative messages (source article from nytimes).