Jared Balkman

DS710 – Final Project

16 December 2020

EXECUTIVE SUMMARY: TWEET POPULARITY AND SENTIMENT

Introduction

It's not hard to find opinions from all sides on any given divisive issue on social media. Often, the more divisive the issue, the stronger the opinion. It was this observation in regard to the recent presidential election that made me wonder if the strength or, say, subjective nature of someone's opinion, and how positive or negative that opinion is, posted in a tweet, would have anything to do with how popular that tweet would be. My project aimed to answer this question on the basis of polarity and subjectivity scores, taken from sentiment dictionaries, and their comparison with the number of likes and retweets of a given tweet. The information from this analysis could prove useful to twitter users or brands looking to grow their social media presence as well as market research analysts and scientists interested in online human behavior.

Data Collection and Analysis

No particular topic or trend concerned me for this project. I was only interested in collecting tweets from as far back as possible, thus the REST API was used to collect 10,037 tweets on December 12, 2020, using search criteria consisting of the 10 most common words used on Twitter ('the', 'I', 'to', 'a', 'and', 'is', 'in', 'it', 'you', and 'of') and a date criterion to limit tweets to those posted before December 6 (though in the case of retweets, the original tweets were posted prior to December 5 or earlier on December 5). The reason for getting the earliest tweets possible was the belief that a tweet would have accumulated most if not all of its likes and retweets by this point.

An initial, status-quo hypothesis was that there was no relationship between polarity and subjectivity, and the number of likes and retweets. To test this hypothesis, the data were fitted to separate multiple linear regression models – one with likes as the dependent variable, and the other with retweets. The independent variables were polarity, calculated using vaderSentiment, a Python sentiment analysis package; subjectivity, calculated using a different package TextBlob; and the number of followers of the author of a tweet. This was included as a control variable as while it was not interesting per se to the research question, it was deemed very likely to influence the number of likes and retweets.

Results and Conclusion

Both models were significant ($p < 2.2e-16$ or very near zero for both likes and retweets) and returned significant variables (see Fig. 1), giving us evidence to reject the status-quo hypothesis of no relationship and claim that sentiment does affect a tweet's popularity. There was some question, though, as to the validity of the model, owing in part to the abundance of neutral tweets , or those with polarity and/or subjectivity scores of zero (Fig. 2). A separate analysis was done after removing these values (shown in Fig. 3) and also came back significant ($p < 2.2e-16$). However, visual diagnostic plots examining the assumptions of the linear model were still difficult to interpret, and a low overall correlation showed that little of the variation in popularity was explained by the sentiment.

In sum, we have evidence of a link between polarity and popularity, but this link requires further research to draw stronger conclusions about its nature. Improvements to the analysis could include training a novel sentiment analyzer specifically on Twitter data, or an exploration of possible non-linear associations in the dataset.

Fig. 1: Significant Coefficients for Likes and Retweets Models

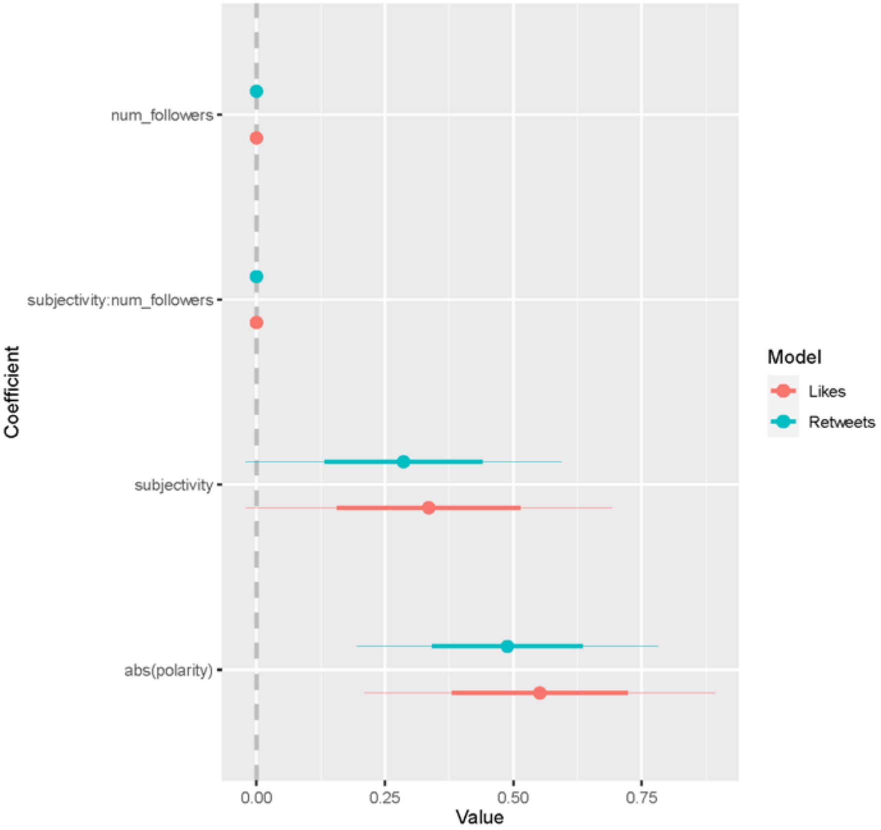


Fig. 2: Histograms for Polarity (blue) and Subjectivity (red) with Zero Values, n=9149

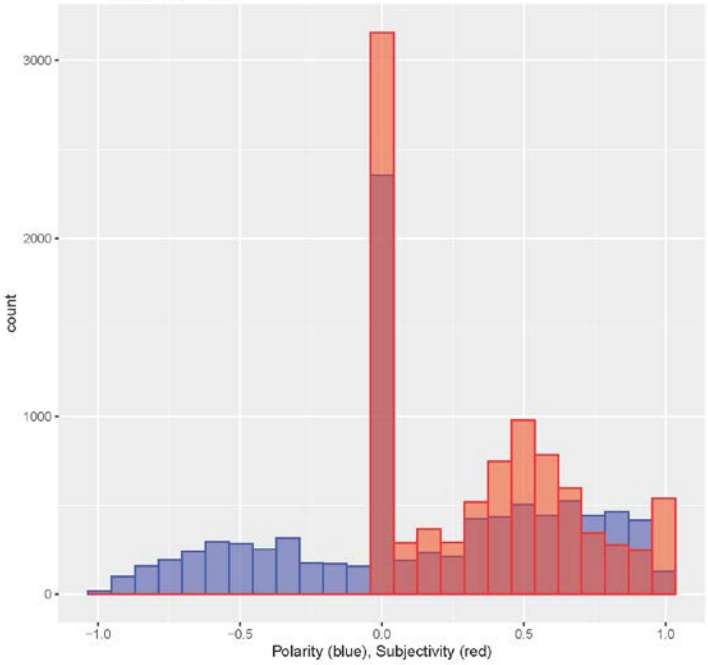
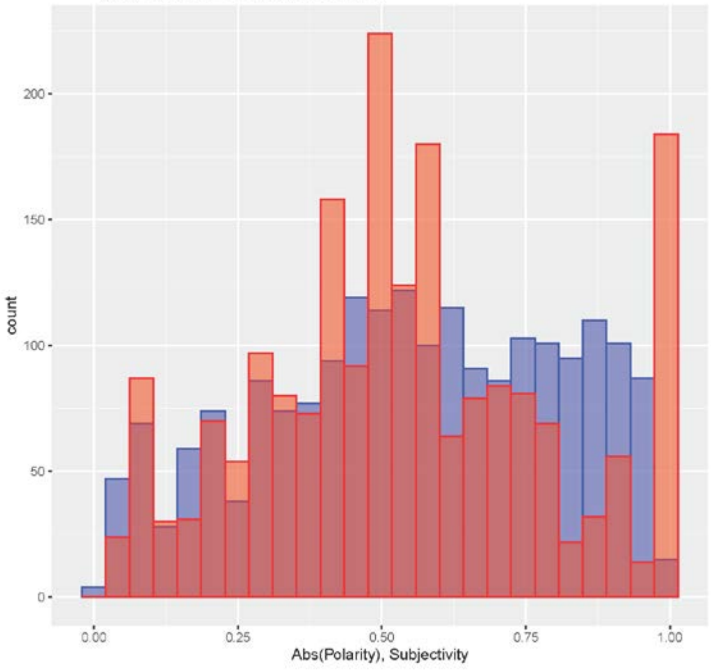


Fig. 3: Histograms for Absolute Value of Polarity (blue) and Subjectivity (red) without Zero Values, n=2021


count
Abs(Polarity), Subjectivity