



# Emotion Analysis of Elon Musk Tweets to Predict Audience Interaction

Skylar Kolisko, 2022 | Dr. Eni Mustafaraj, Advisor

## Background

I spent the last year working with Natural Language Processing, so for my Data Science Capstone, I wanted to create a new use for Language Models. This project aims to determine which emotions public figures should express to their followers to get maximum information spread (retweets)

## Research Question

The goal of this project is to see if an NLP tool's decisions can be used to predict the amount of engagement Elon Musk's tweet will get? I will use the RoBERTa Q&A model, fine-tuned RACE, from the University of São Paulo (2) to creatively to perform the task of emotion analysis to decide which emotions Elon should communicate in his tweets to get a high amount of retweets.

## Technical Background

Natural Language Processing (NLP) - Branch of AI dealing with making machines do to tasks related to human language as opposed to machine

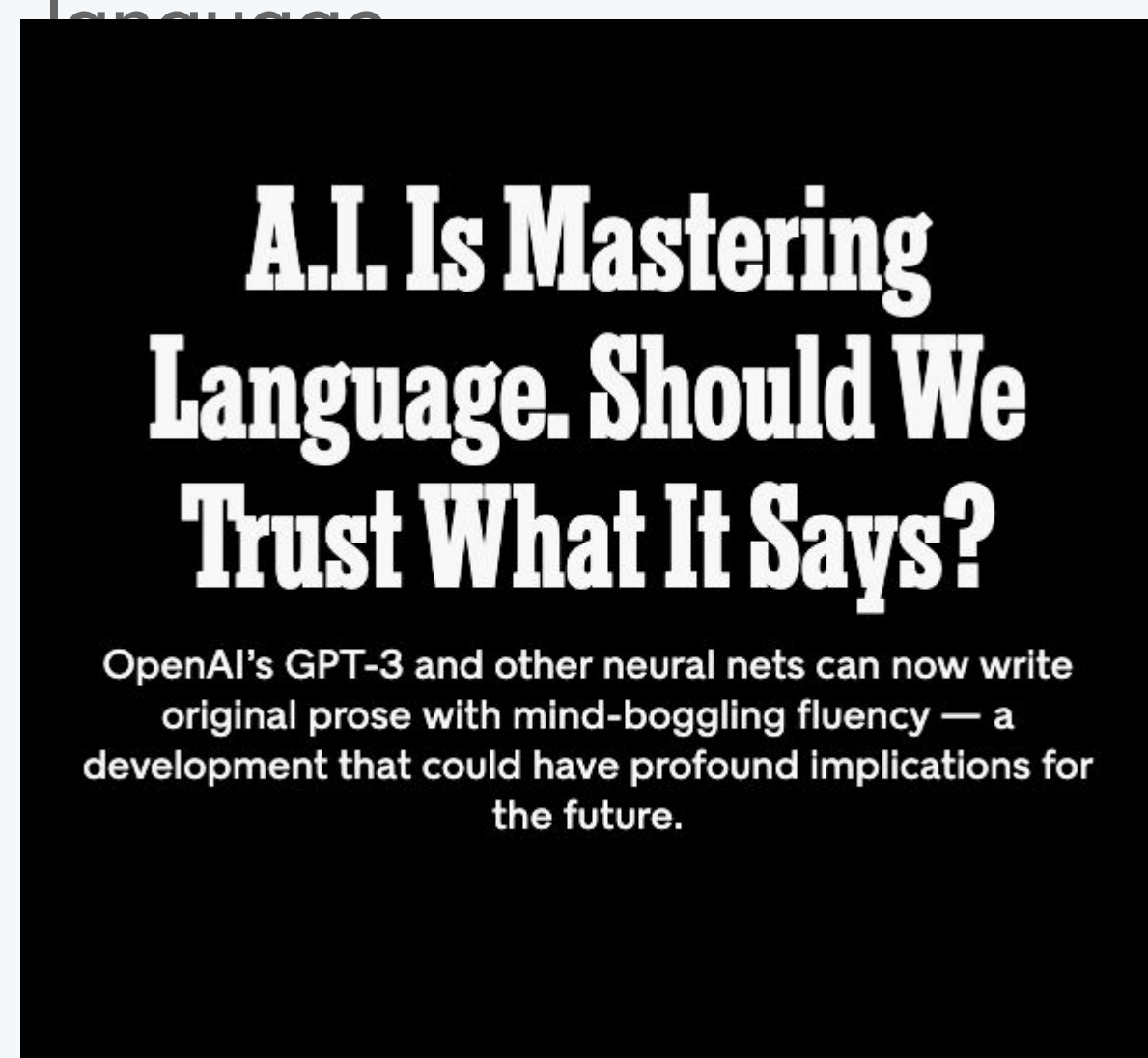
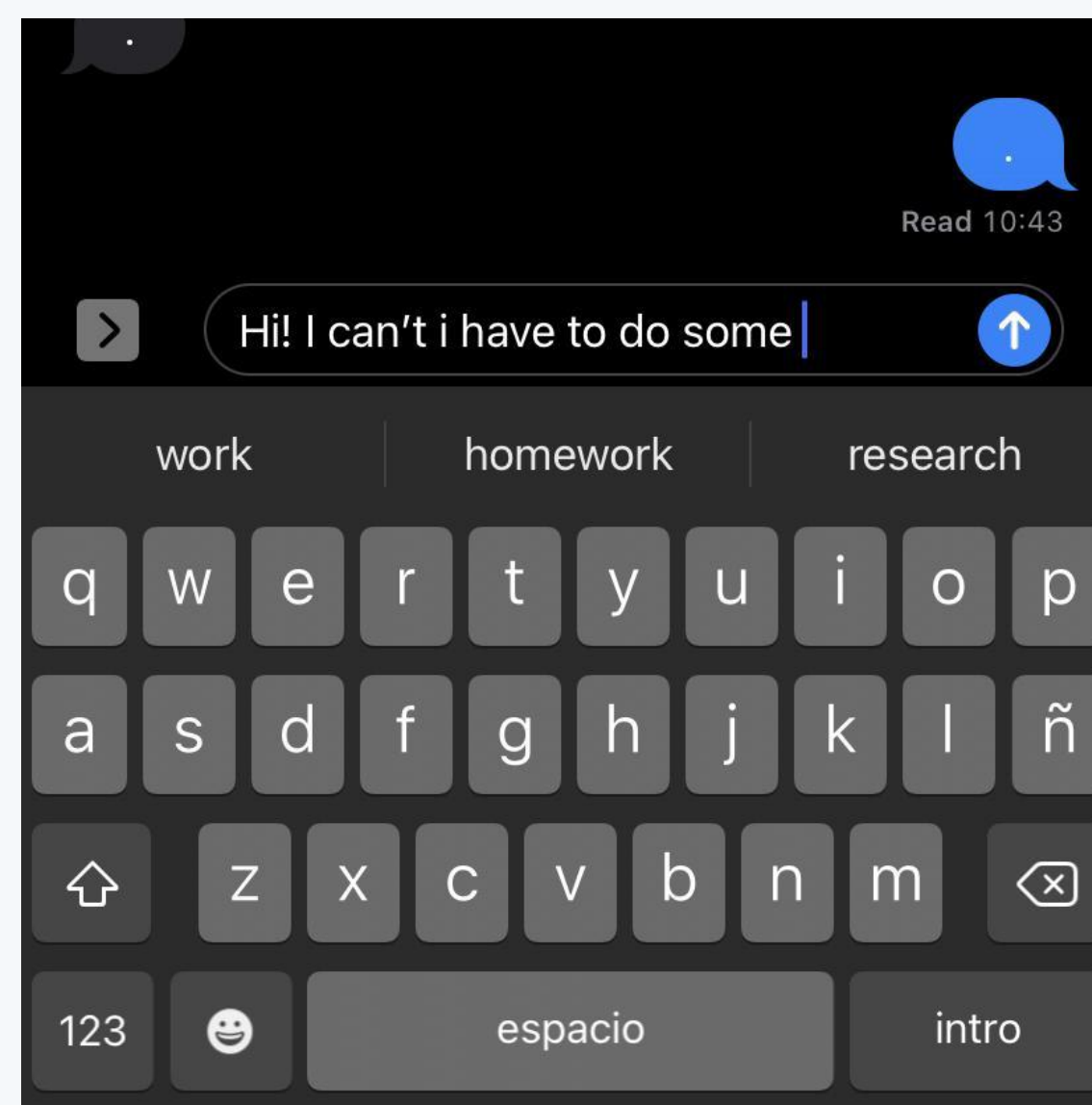


Fig 1. (left) NYT article explains the ubiquity & potential dangers of NLP.

Fig 2. (right) Apple's predictive text is a language model.



Language Model - probability distribution of which word is likely to occur next.

Emotion Analysis Task- Model determines how much it believes a piece of natural language to be aligned with a certain emotion.

Multiple Choice Q & A Model - takes as input context, question, & possible answers, outputs a weight vector, representing model's confidence in each answer.

## Original Data

The original data for this project comes from a Kaggle dataset which includes 635 of Musk's tweets from early July, 2022 to late September, 2022.



Examples of tweets in the dataset.

## Data Preprocessing &

1. Use Twitter API to get non truncated tweets

The amount of attention on me has gone supernova, which super sucks. Unfortunately, even trivial articles about me... <https://t.co/XcwDQMs4aZ>

1. Make input to Q&A model

tweet_text	tweet_id	retweets	favorites	date	adj
Sock Con, the conference for socks	1544377493263720450	1451	30753	2022-07-05	creative
Sock Con, the conference for socks	1544377493263720450	1451	30753	2022-07-05	generous
Sock Con, the conference for socks	1544377493263720450	1451	30753	2022-07-05	rude
Sock Con, the conference for socks	1544377493263720450	1451	30753	2022-07-05	mean

fav_to_follower_ratio	retweet_to_follower_ratio	question	answer_a	answer_b
-8.099269625438120	-11.153004202005400	How can this statement be described?	creative	not creative
-8.099269625438120	-11.153004202005400	How can this statement be described?	generous	not generous
-8.099269625438120	-11.153004202005400	How can this statement be described?	rude	not rude
-8.099269625438120	-11.153004202005400	How can this statement be described?	mean	not mean

1. Input to model to get the tweets' emotion scores. Thanks to Prof. Anderson for letting me use her server!

Note: because of limitations on Twitter's lowest API access, ran steps 1-3 twice.

1. Compile the "yes" scores as the final dataset.

Note: the "yes" keys are the weights the model put on the answer to the question, as in, the amount it believed the tweet to be best aligned with some

tweet_text	retweets	prediction_rude	prediction_sarcastic
Sock Con, the conference for socks	1451	0.5096931458	0.9819375277

## Model Selection

Goal: prune to achieve a parsimonious model to find a few emotions for Elon to emit or avoid.

As a start, from the original forty adjectives, a multiple linear regression model was made. It predicted the "retweet score" of a tweet, where

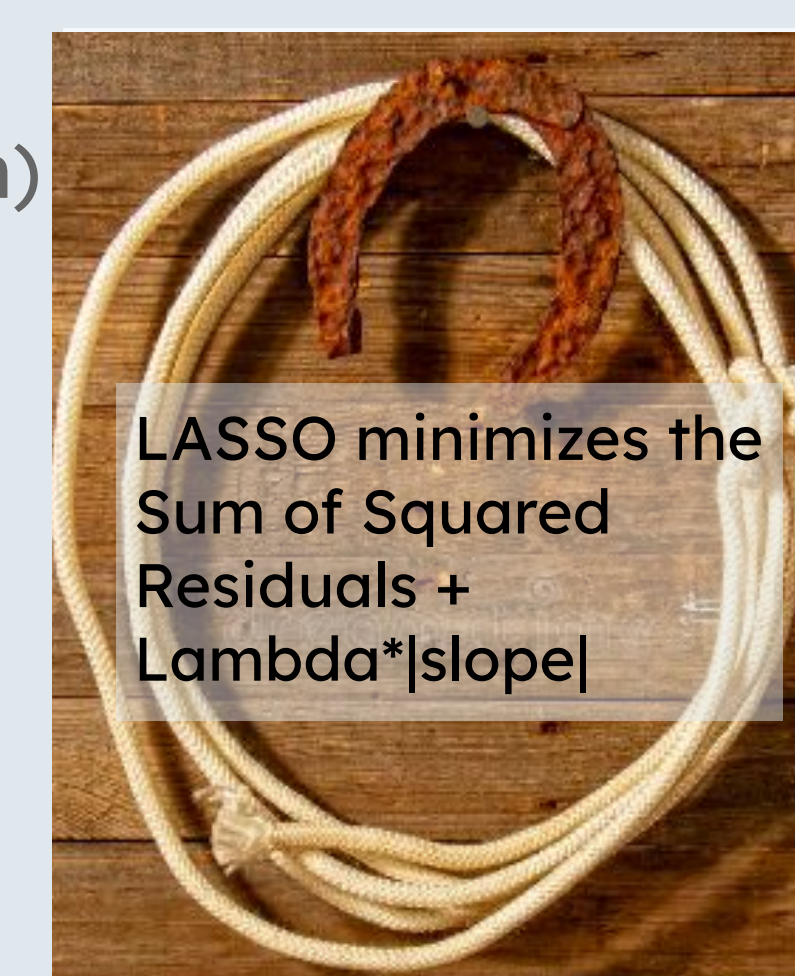
$$\text{retweet score} = \ln(\text{retweets}/\text{Musk followers})$$

Pruning this by hand demonstrated the model's multicollinearity problem.

Solution: LASSO model (Regression using L1 regularization)

Useful because:

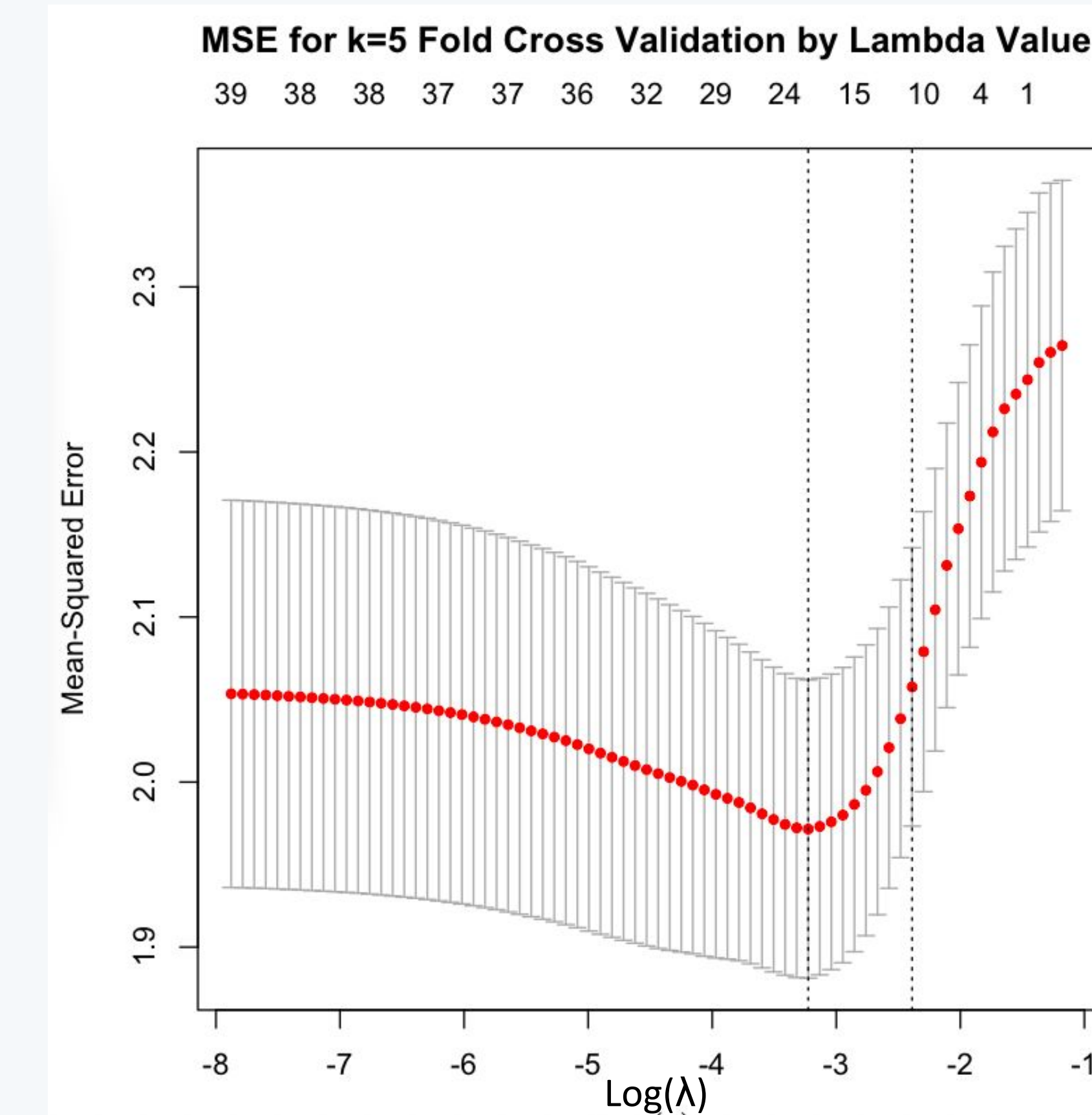
1. Can handle low to moderate levels of multicollinearity
1. Automatic feature selection
1. Creates a sparse model



LASSO minimizes the Sum of Squared Residuals +  $\lambda \sum |s_{\text{slope}}|$

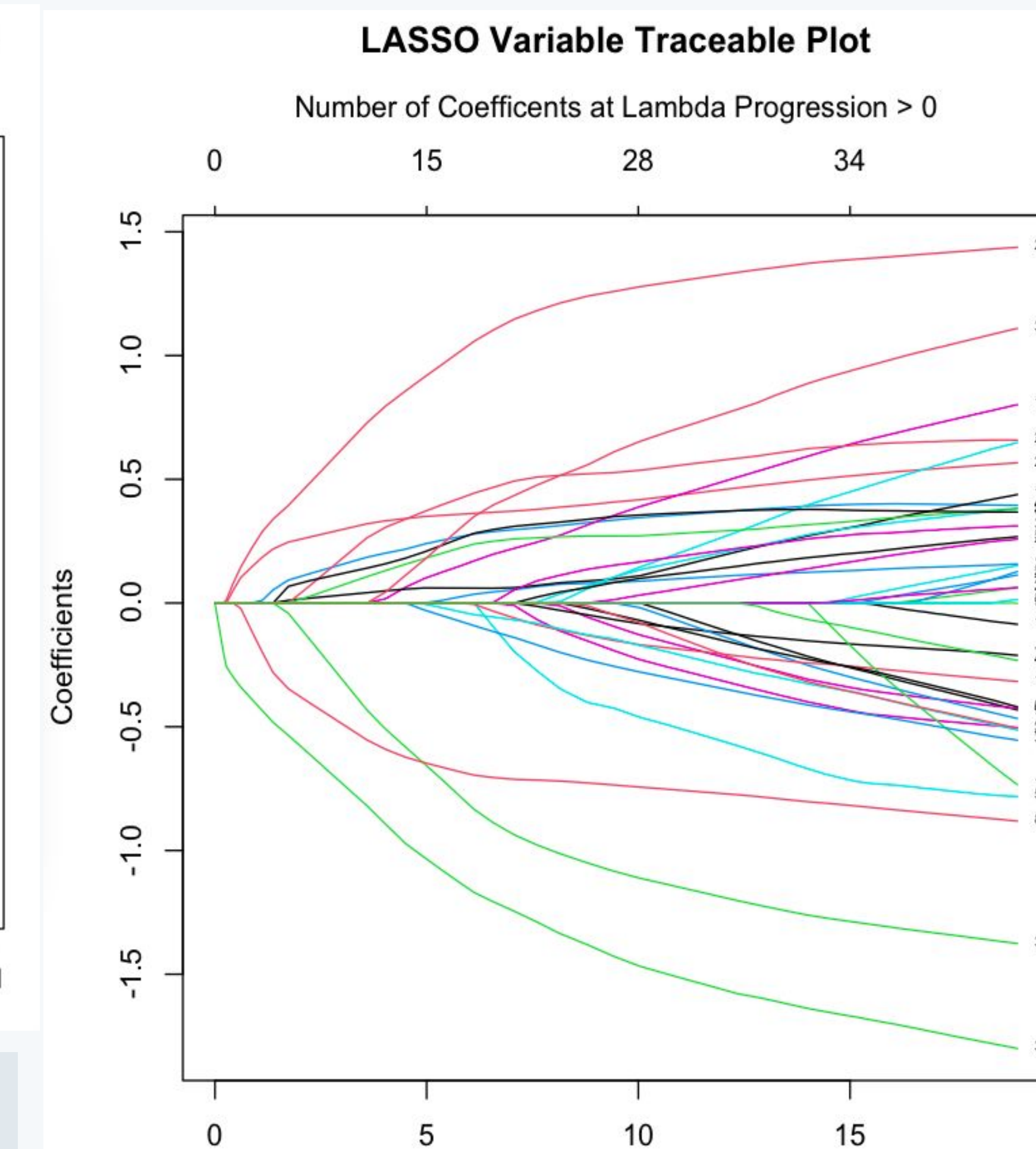
## Model Fitting & Validation

From the original 40 emotions, 11 were selected by LASSO as significant enough to be included in the model at that optimal  $\lambda$  of 0.03971, determined via 5-folds cross validation.



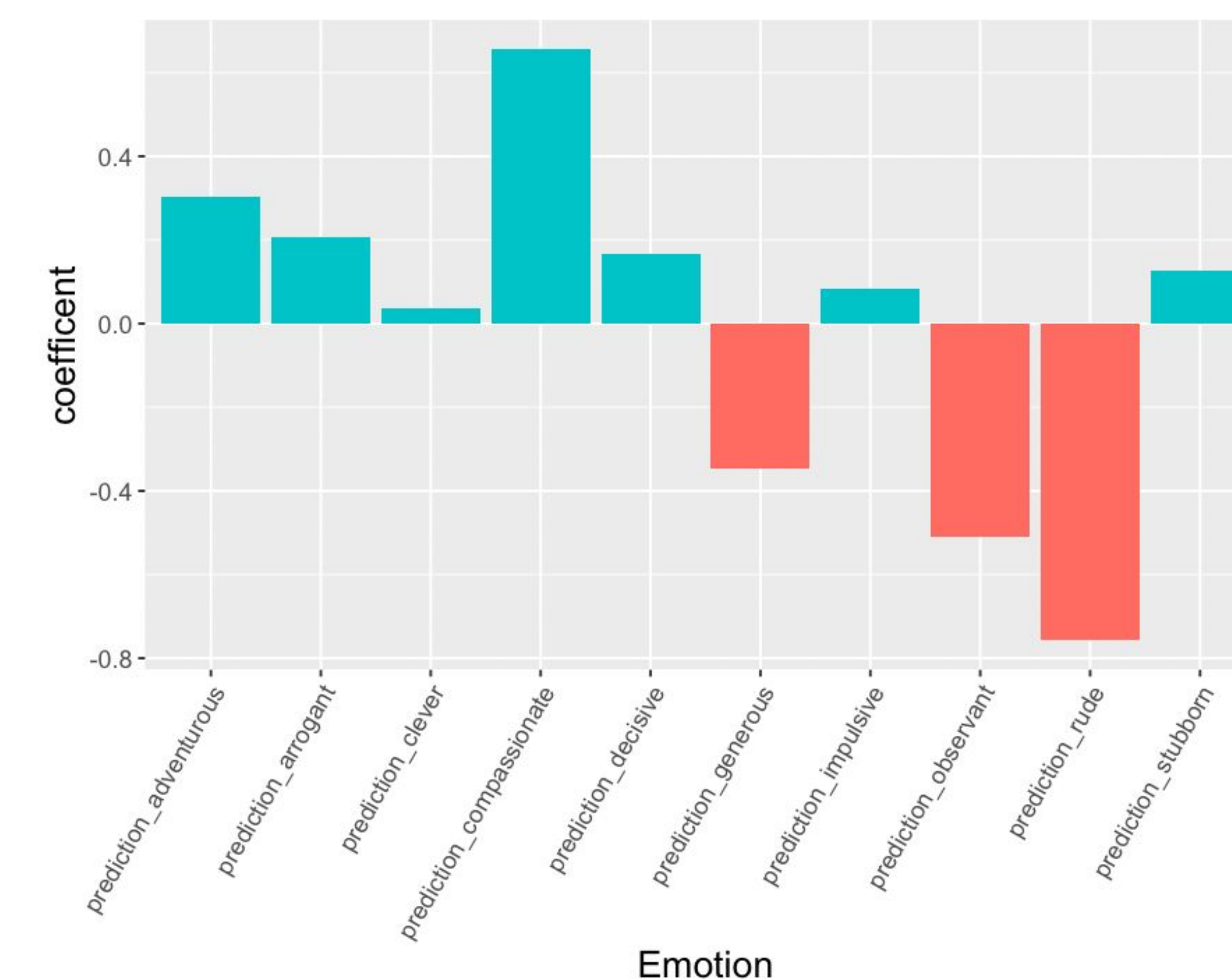
Top Left: MSE is achieved at  $\lambda_{\min}$ ,  $\lambda = 0.03971$ , and  $\lambda_{1se}$  is at  $\lambda = 0.00177$

Bottom Left: Variables (emotions) and their coefficient values at  $\lambda = 0.03971$ .

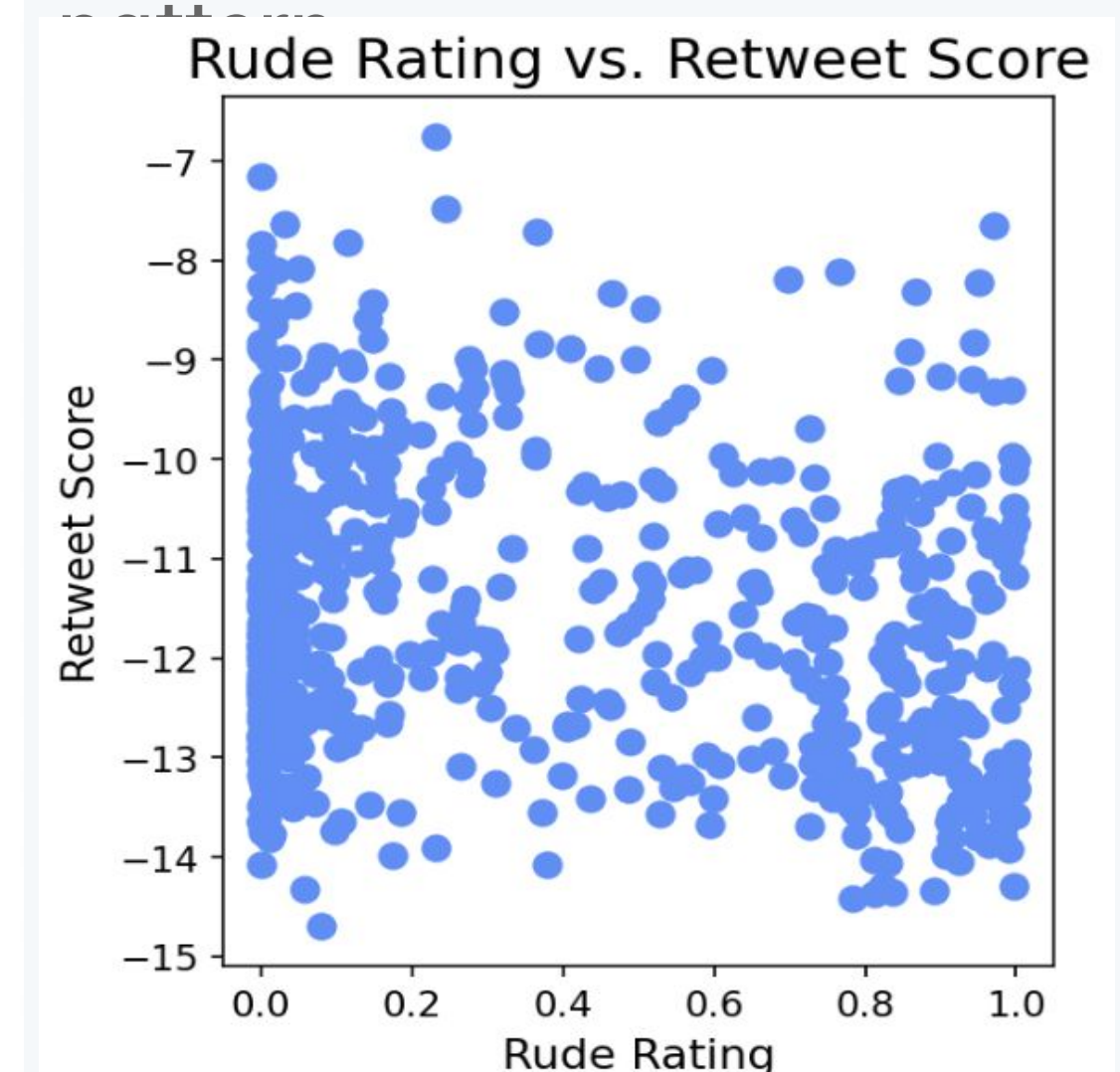


Top Right: As  $\lambda$  approaches 0, L1 norm grows, model approaches OLS model.

### Lasso Coefficients with $\lambda = 0.03971$



Below: Even most significant coef. shows little



Model's  $R^2$  is 18.8%.  
Paths for improvement:  
1.Small size of data set  
2.Q&A model's training data

## References

1. <https://www.npr.org/2022/10/08/1127689351/elon-musk-calls-himself-a-free-speech-absolutist-what-could-twitter-look-like-un>
2. <https://huggingface.co/LIAMF-USP/roberta-large-finetuned-race>