Talk about using linear regression model

Describe about features used

Talk about the result

A linear regression model with l2 regularization is trained to predict the ‘intrinsic funniness’ of words. The data used to train the model are 4997 words with (manually labelled?) humor rating from 27.31 to 100 (cite?).

To train the model, each word in the dataset is represented by 19 different features: cosine distance to category defining vector (CDV) of six categories (i.e. sex, party, insult, profanity, body function and animals), the Boolean value of whether the word contains letter ‘k’, the log frequency of word, log average letter probability, log average phoneme probability of the word, ratio of log average letter probability and log average phoneme probabilty, the Boolean value of whether the word contains the phoneme /u/, cosine distance to CDV of eight-le words, valence \* arousal of word, valence \* arousal \* dominance, arousal \* dominance, valence, arousal, concreteness of the word. Those are also features verified in westbury to have some significant linear relation with humor.

First, the category defining vectors (CDV) of six categories (i.e. sex, party, insult, profanity, body function and animals) that are related to funniness are calculated (cite the westbury paper). Those vectors are calculating by first getting 19 words most related to those categories as described in the Appendix A of westbury and find their average word2vec embedding; then finding cosine similarity between words within the dataset and the average embedding and find the average embedding of first 100 words with greatest similarity (cite westbury).

The CDV of the le words are average of word2vec embeddings of gaggle, jiggle, tinkle, waddle, wiggle, wriggle, gobble and nibble. The reasoning for using that is described in westbury as those are Cons+le words with humor rating above 2 SD and do not very an evident semantic relation to humor and is tested to have significant r value (cite westbury).

The frequency of word is gathered from <https://www.kaggle.com/datasets/rtatman/english-word-frequency>

The letter frequency of word is gathered from [Practical Cryptography](http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/english-letter-frequencies/)

The pronunciation of a word is gathered from Carnegie Mellon University (CMU) Pronouncing Dictionary [CMU in IPA](https://people.umass.edu/nconstan/CMU-IPA/) and the frequency of each phoneme is gathered through [Relative Frequencies of English Phonemes | cmloegcmluin](https://cmloegcmluin.wordpress.com/2012/11/10/relative-frequencies-of-english-phonemes/)

Define each term

The Valence, arousal, dominance, and concreteness of a word are gathered from estimated of Hollis (cite Hollis).

After training linear regression model via cross validation method, the calculated root mean squared test error is 7.6717 and R2 to 0.37022. The distribution of original data and tested data are shown in below.