Erdos Institute Data Science Bootcamp Spring 2025

Project title: Predicting review sentiment and star rating from text
Group members: Duncan Clark, Samantha Jarvis, Brennan Register, Andrew Silva

Executive summary: Using Google review data, we predict the sentiment and star rating based on review text. Our model is trained on review data from restaurants in the US, roughly ~2 million samples.

Our main model is a recurrent neural network with long short term memory. This model uses a pre-trained word embedding layer from the GloVe vector representation of words model. Specifically, the corpus of review text is limited to the 20,000 most common words, and the overlap between those and the GloVe words are assembled to build an embedding matrix. The LSTM is bidirectional and can process reviews with a maximum length of 128 words (reviews past that length are truncated).

Two variants were trained on this data set:

- A model for predicting sentiment of a review as positive (4, 5 star), negative (1, 2 star), neutral/mixed (3 star)
- A model for predicting the star rating of the review text.

The sentiment analysis model achieves an accuracy of over 98% (F1 0.98132) on the testing set, and the star rating model achieves an accuracy of over 94% (F1 0.94016).

Both LSTM models demonstrated comparable or superior performance to transformer-based models such as BERT or ROBERTa, with faster training and prediction times, and a smaller storage footprint. Confusion matrices can be found in the github repository.

Additional datasets for vegetarian and Thai restaurants in the US Midwest were used to further test the models. The project concluded that LSTM is sufficient for sentiment analysis of specific establishment types, while ROBERTa may offer advantages in handling text with spelling errors. Future work may include additional text processing, model ensembling, robustness testing across different establishment types, and using sentiment analysis to refine restaurant recommendations.