Sentiment Analysis

Kody Sanchez

CS 5600 - Intelligent Systems
Vladimir Kulyukin

Introduction

Sentiment Analysis is determining the opinion sentiment of a given text. This is very helpful to businesses that want to know their customers opinions of their products. The following is an analysis of various machine learning classifiers trained to determine the sentiment of varying Amazon product reviews.

Data

The data for this project consists of a few million reviews and is found on Kaggle[1]. For this project, 100,000 samples from the training dataset were pulled. The data is separated into positive and negative sentences. Each review is scrubbed of special characters, changed to lowercase, de-hyphenated, and checked to see if the word is a valid english word. The data is then split into training, testing, and validation sets. The training data contains 75% of the data. Each input set is converted to a term frequency-inverse document frequency matrix that can be fed into the classification models. The answer set is one-hot encoded.

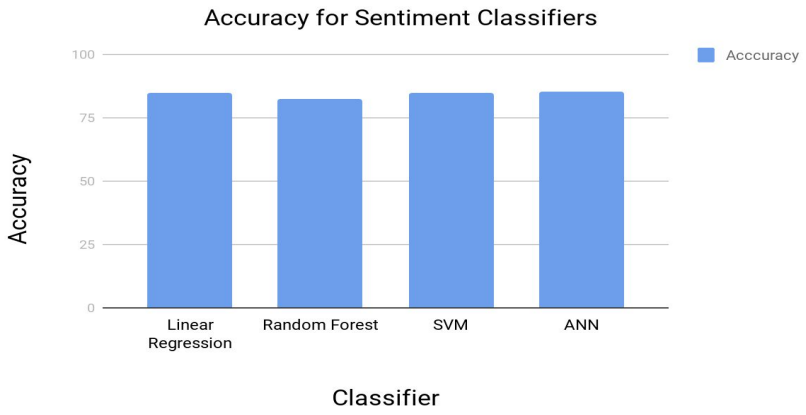
Methods

Six classifiers are created to be compared in terms of accuracy. The base model is a linear regression model. Other non-neural network models are a random forest model consisting of 100 trees and a SVM with a linear kernel. Three neural networks were created. First, an ANN with one hidden layer of 64 nodes activated with relu and trained for 100 epochs with a learning rate of 0.0001. Second, a CNN with 2 convolution layers the size of the input and 64 activated by relu and trained with a training rate of 0.00001 for 100 epochs. Finally, a RNN with one LSTM layer with 128 nodes trained with a learning rate of 0.000001 for 100 epochs.

[1] https://www.kaggle.com/bittlingmayer/amazonreviews.

Results

During the training of the CNN and RNN, the computer's GPU ran out of memory resulting in the tensorflow module becoming corrupted on the test computer. Fortunately, the results of the other models were saved. The base comparison is linear regression. With linear regression, an accuracy of 84.928% was achieved and is the benchmark. The random forest scored 82.588%, which is significantly lower than the benchmark compared to the other models. The SVM scored 84.884% , which is comparable to linear regression. It took significantly longer to train and test the SVM model than the linear regression, the random forest, and the ANN. The ANN scored 85.084% on the test data, and about 90% on the training and validation data.



Conclusion

Since there can be a disagreement on the sentiment of a given sentence, it is hard to gauge accuracy. Based on others contributions to this data set, as seen at on Kaggle[2], the classifiers under perform. Using a convolution neural network or a recurrent neural network would likely produce more accurate results, although the hardware to train such models was unfortunately not available for this project. The models tested perform well enough to be able to get a general picture of customers opinions on a product, which is very valuable to any marketing department.

[2] https://www.kaggle.com/bittlingmayer/amazonreviews/kernels