Dalya Adams
MSDS 422 section 59
Sentiment Analysis

One of the greatest benefits of machine learning for companies is the increase in day to day efficiency. These benefits can be seen when customers are correctly categorized, and thus targeted with the best marketing campaign, and more recently in identifying the sentiment inherent in text. When a comment box, chat log or email can be analyzed to determining the general sentiment, or feelings, of a customer, the customers case can be efficiently prioritized and assigned to a customer service representative to ensure the customer receives the attention the need in order to solve the problem efficiently while also maintain strong customer relations.

In translating the idea of sentiment analysis for customer service and relations into an actual system that can correctly categorize and prioritize customer issues, we must first address the computing power and the amount of training data. Utilizing recurrent neural networks (RNN) ensures increased accuracy but computations require more processing power. We will assume that the company we are providing this analysis for is willing and able to provide or pay for the necessary processing power. The next issue is the data on which to train the model. For sentiment analysis we need a large dataset which contains strings of text and a general indicator of whether the text was positive or negative. Ensuring that we understand what we are most concerned with, in this case negative feedback would require more immediate attention, allows us to ensure we have a robust sample of negative text strings with which to train the model.

Outlined above were issues we need to ensure we account for, specifically a dataset with which to train our RNN. To ensure we have a large vocabulary with which to train our model, we turn to pre-trained word vectors and subsets of pre-trained word vectors. We can utilize technologies such as word2vec, GloVe and FastText. We test the accuracy of these pre-trained word vectors on movie reviews, collecting both the training set accuracy as well as the test set accuracy. The model with the highest accuracy will be investigated further for use in the customer service sentiment analysis. We compare the accuracy between pre-trained word vectors by toggling between different dimensions and vocabulary size.

We chose the GloVe pre-trained word vectors for this analysis and tested the accuracy of the word vectors with dimensions of 50d and 200d.We also tested the impact of a vocabulary size of 10,000 and 100,000 on the accuracy of the model. The vocabulary size is a list of the top n words in the English language. By comparing the impact of the top 10,000 words in the English Language versus the top 100,000 words in the English language, we can hone in on the optimal vocabulary size for the analysis. The combinations and results of this 2x2 experiment can be found in the appendix of this report. We also took the most accurate model and adjusted the learning rate of 0.001 to 1.0. A higher learning rate corresponds with faster analysis but can also result in decreased accuracy.

In analyzing the effectiveness of word vectors in classifying sentiment, we assess accuracy on movie reviews. By providing the RNN with the text from both positive and negative movie reviews, we can assess the general accuracy of the model in capturing the negative or positive experience the customer had when watching that movie. The GloVe 50d pre-trained word vector with the top 100,000 words of the English language proved to be the most accurate model, with a training accuracy of 0.82 and a test accuracy of 0.685. The 50d word vector with a vocabulary size of 10,000 was the second most accurate with a training accuracy of 0.81 and a test accuracy of 0.68. In order to increase the accuracy of this model, removing stop words could be beneficial, also testing the model on a more relevant dataset may ensure that the accuracy of the model on movie reviews translates to customer communication. Finally, testing different word vector technologies could prove beneficial in accurate sentiment analysis, as well as performing a grid search to tune the model to its optimal hyperparameters.

Appendix

Embedding Source	**EVOCAB Size**	**Learning Rate**	**Train Accuracy**	**Test Accuracy**
glove.6B.50d	10,000	0.001	0.81	0.68
glove.6B.50d	100,000	0.001	0.82	0.685
glove.6B.200d	10,000	0.001	0.98	0.63
glove.6B.200d	100,000	0.001	1.0	0.62
glove.6B.50d	10,000	1.0	0.5	0.47