This study examines the use of machine learning in performing sentiment analysis on 100,000 IMDB movie reviews to determine whether the sentiment behind a review is meant to be positive or negative. Word recognition is used to decide if the movie reviews are positive or negative. The model's identification system will use Word2Vec, and it will be able to detect sarcasm, ambiguity, and plays on words.

Goals
- The goal of this implementation and analysis is to measure the success of a Google Word2Vec model when it comes to sentiment analysis.
- The model is compared to a baseline which includes correctly-classified movie reviews and is distinct from the training dataset.
- The classification accuracy is calculated by comparing the number of correctly-identified movie reviews to the total.
- A correct prediction means the model correctly identified a movie review as positive or negative.
- As a validation to the classification accuracy score, an ROC curve is also plotted.
- The model will need to exceed the line of no-discrimination (better than random).
- Together a classification accuracy score of above 60% and the ROC curve being above the line of no-discrimination implies that the model is better than a random guess, and therefore the model is a success.

Data Source and Obtaining Data
- Data is obtained from the International Movie Database (IMDb) and accounts for a total of three data sets: two training sets and one test set.
- The data consists of a total of 100,000 multi-paragaph movie reviews, which gives the model a large vocabulary and allows it to be tested effectively.

Results
- Despite multiple trials, the accuracy of the Word2Vec model using Doc2Vec did not amount to the indication of success of 60%.
- Essentially, this model’s sentiment analysis is no better than a random guess.
- Figures 2, 3, and 4 (right) support these results, showing the ROC curve of the model’s results, a table documenting the results of five separate trials, and a bar graph comparing the amount of true positive to false positives, respectfully.

Discussion
- Word2Vec is most often used with Python due to its convenient functions and utilities. Therefore, not much documentation is provided for Word2Vec or Doc2Vec for R, making troubleshooting difficult.
- Word2Vec itself worked correctly and we were able to test it.
- There was not much classified data provided, meaning training was conducted on a limited dataset; this could have contributed to the low accuracy reported.
- Vector averaging and logistic regression are being planned in future work as next steps to improve the accuracy of the model.
- Bag-of-words is also being considered as an alternative to Doc2Vec in future work.

Conclusions
- Word2Vec and its extension, Doc2Vec, are not an effective model in performing sentiment analysis on the movie reviews derived from the International Movie Database (IMDb).
- The resulting average accuracy when using this model was approximately 48.32%; in order to be considered correct, the accuracy would need to be 80% or greater.
- In future iterations of this project, a new method should be implemented using the Bag of Words approach and vector averaging along with the bag-of-words method, as opposed to Doc2Vec.
- The Bag of Words model is a representation of texts used in natural language processing and information retrieval wherein provided text is broken down into fixed-length vectors.
- Vector averaging is a method that computes the average of data stored in a vector.
- Making the recommended changes to the project is likely to result in a model that is efficient, effective, and accurate.

References