Sentiment Analysis Using Google’s Word2Vec

Kaitlyn Carroll, Maddison Davenport, Alison Rust, Sina Sontowski

INTRODUCTION

- 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 be above 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-paragraph movie reviews, which gives the model a large vocabulary and allows it to be tested effectively. *Figure 1*— below— shows the total number of reviews plotted by sentiment (1 for positive; 2 for negative).

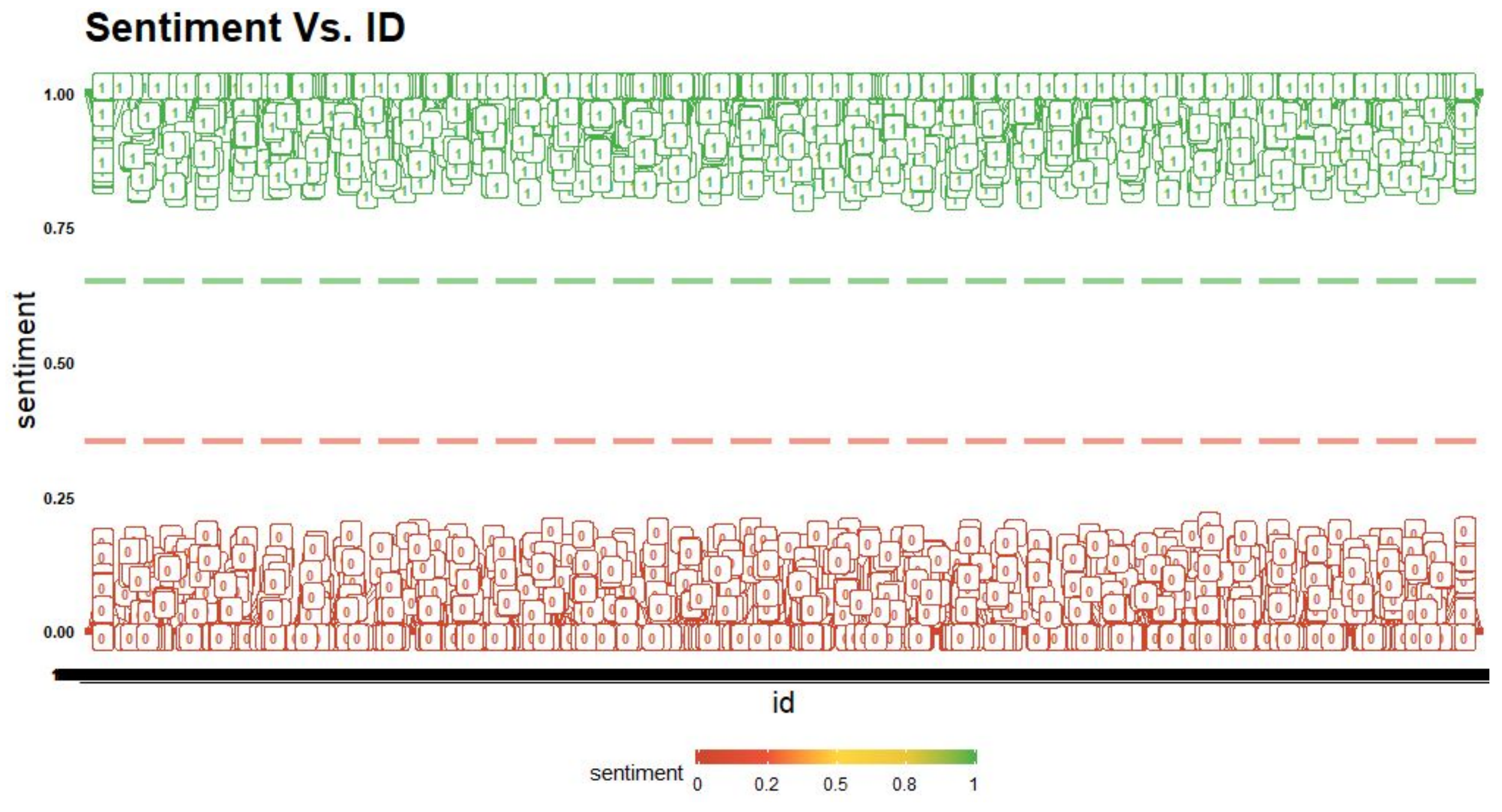

Figure 1: Scatter plot of sentiment

METHODS AND TOOLS

- **Cleaning the Data**
  In order to use Word2Vec, the data has to be cleaned. In order to create our sentiment analysis model, the reviews were stripped of any HTML encoding, the reviews were converted to lowercase, and non-alphabetic characters removed.
- **Word2Vec**
  Word2Vec is a modeling tool, and is the basis for this project. A training set was used to train the model to learn word associations.
- **Doc2Vec**
  Doc2Vec is an extension of Word2Vec. While the popular Bag-of-Words method for improving the accuracy of Word2Vec models uses vector averaging, Doc2Vec instead converts the reviews to word-vector-representation features, eliminating the need for such a method. This saves time and computing resources.
- **Random Forest**
  A Random Forest was used for classification of the reviews as either positive or negative. The Random Forest fits a number of decision trees classifiers on various sub-samples of the dataset and uses averaging.

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.

GRAPHICAL REPRESENTATION AND ANALYSIS

- *Figure 2* shows that the ROC curve falls under the diagonal line which represents perfect chance.
- *Figure 3,* shows five consecutive tests of the model wherein the accuracy of each of the trials was below 52%.
- *Figure 4* visualizes the five runs of the model showing the false positives and false negatives were always almost equivalent.
- It should be noted that while during testing, some trials showed values in the upper 50 percentile range, these can be determined to be outliers, as the five consecutive tests that were documented show the tests almost consistently under 50%, with only one recorded trial exceeding 50%.

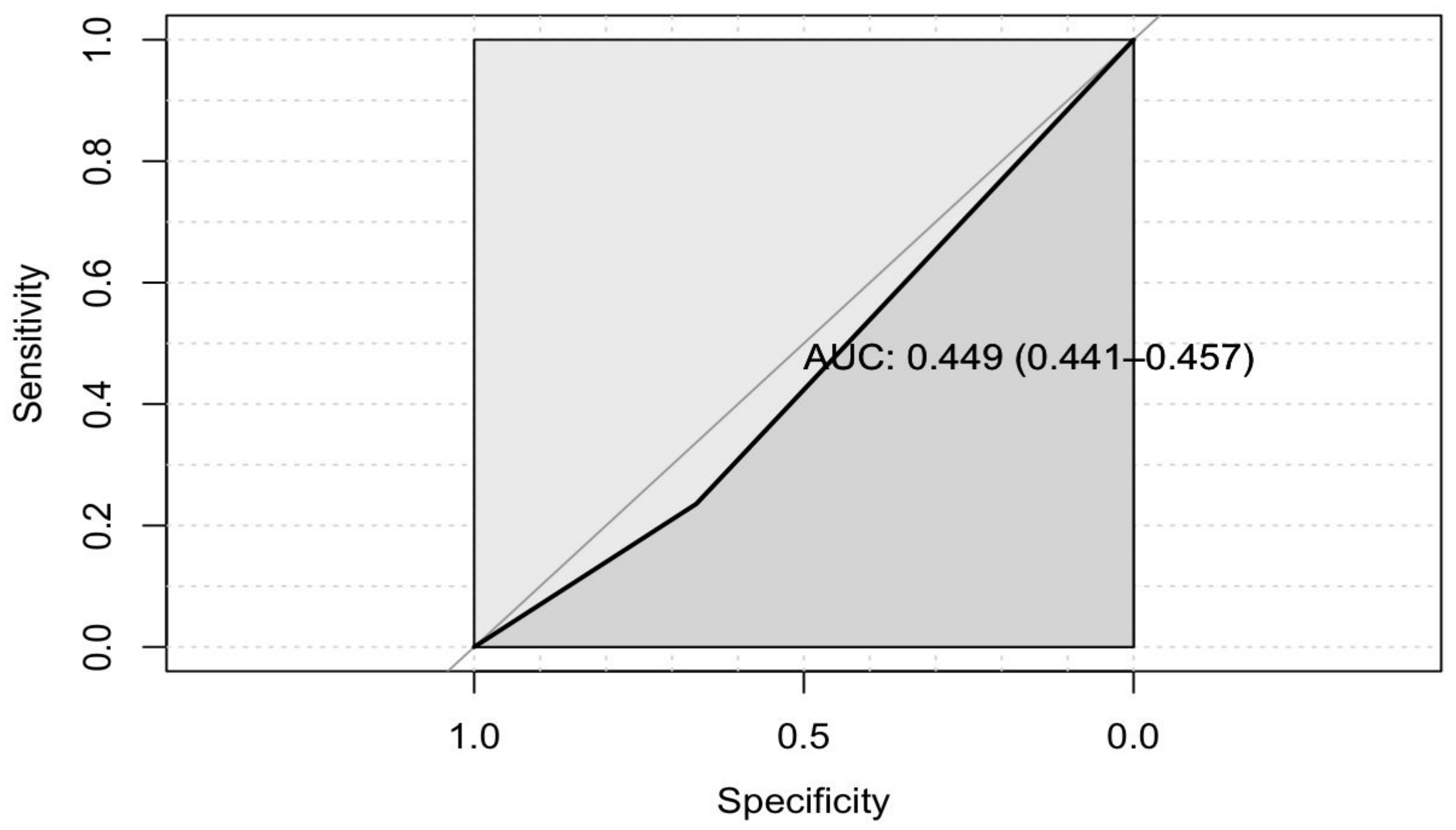

Figure 2: Results plotted on ROC curve

TimesModelRan	Accuracy(%)	True Positive	False Positive
1	48.78	5854	6146
2	51.14	6137	5863
3	47.25	5670	6330
4	45.23	5428	6572
5	49.20	5905	6095
Average	48.32	5798.80	6201.20

Figure 3: Table showing accuracy of five consecutive trials

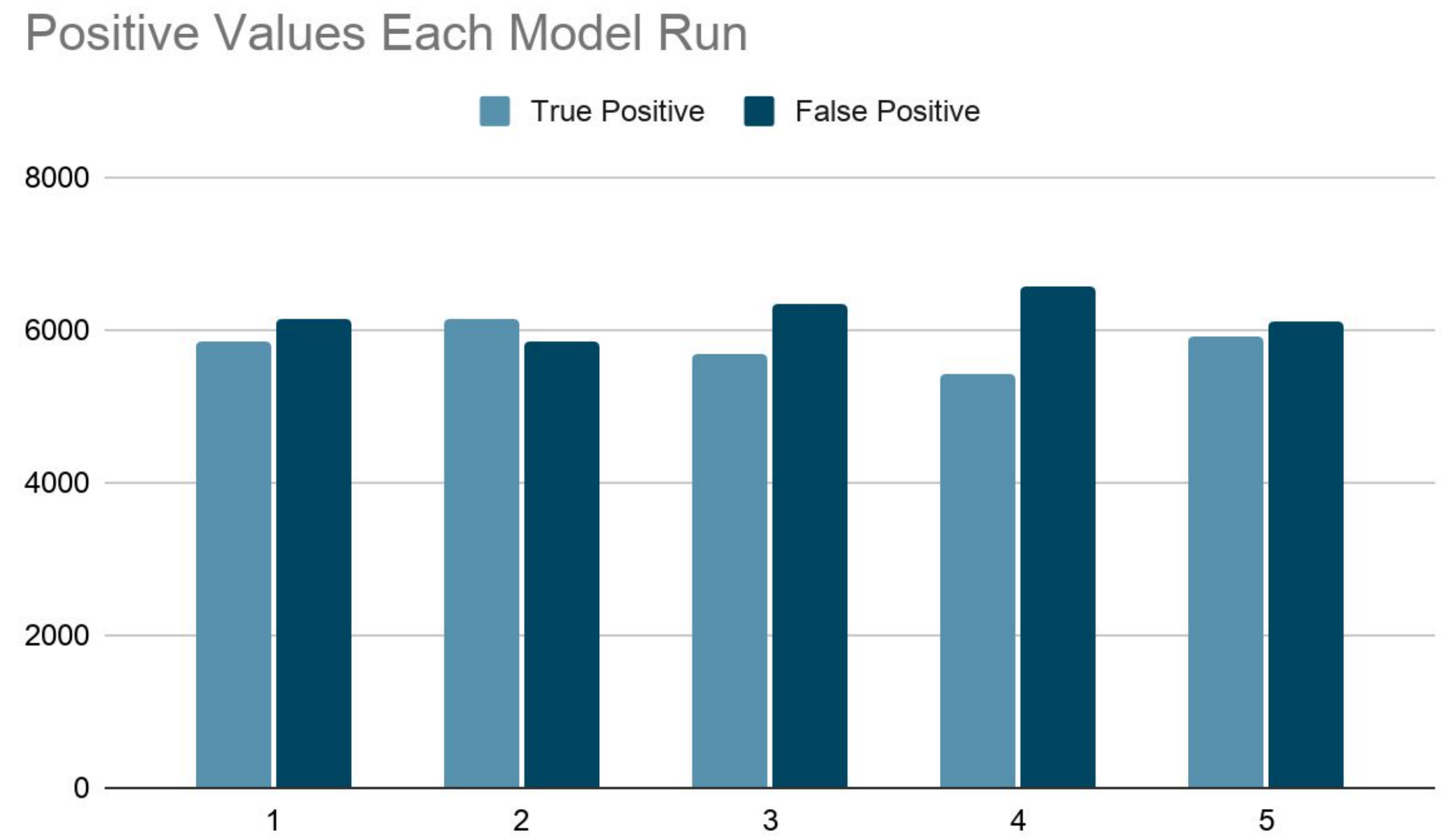

Figure 4: Bar Plot of True Positives and False Positives

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.

CONCLUSION

- 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 60% 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

I. Jang B, Kim I, Kim JW (2019) Word2vec convolutional neural networks for classification of news articles and tweets. PLoS ONE 14(8): e0220976. https://doi.org/10.1371/journal.pone.0220976

II. Shah, C. (2020). A hands-on introduction to data science. Cambridge, United Kingdom: Cambridge University Press.

III. Wijffels, J. (2021, March 28). Distributed representations of sentences, documents and topics [r package doc2vec version 0.2.0]. Retrieved April 01, 2021, from https://cran.r-project.org/web/packages/doc2vec/index.html

IV. Zou, Q., Xing, P., Wei, L., & Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. Rna, 25(2), 205-218.

V. Bag of Words Meets Bags of Popcorn. (n.d.). Retrieved from https://www.kaggle.com/c/word2vec-nlp-tutorial/

VI. Package 'word2vec'. (2020, November 26). Retrieved March 19, 2021, from https://cran.r-project.org/web/packages/word2vec/word2vec.pdf

VII. Wickham, H., Chang, W., Henry, L., Pederson, T. L., Takahashi, K., Wilke, C., . . . Dunnington, D. (n.d.). Create elegant data visualisations using the grammar of graphics. Retrieved April 2, 2021, from https://ggplot2.tidyverse.org/

Acknowledgments: Dr. William Eberle, Professor, Computer Science, Tennessee Technological University