Methodology:

Data used to generate the features:

- IMDB Data: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews?resource=download
    - This is a dataset containing 50,000 movie reviews from IMDB, as well as a label that shows whether the review is positive or negative

- Positive Words: https://gist.github.com/mkulakowski2/4289437
    - This is a text file that contains ~2000 positive adjectives

- Negative Words: https://gist.github.com/mkulakowski2/4289441
    - This is a text file that contains ~5000 negative adjectives

Features used to train the model:

- Positive word count
- Negative word count
- log(word count of review)
- 1 if “no” ∈ review, 0 if otherwise
- 1 if “!” ∈ review, 0 if otherwise

Models used for sentiment analysis:

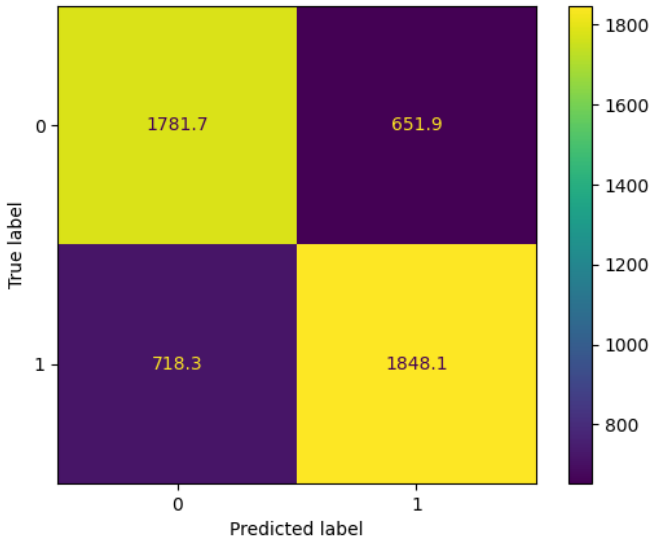
- Logistic regression
- Multinomial Naive Bayes

Evaluation:

- Used 10-fold cross validation to calculate accuracy, precision, recall and a confusion matrix for each model

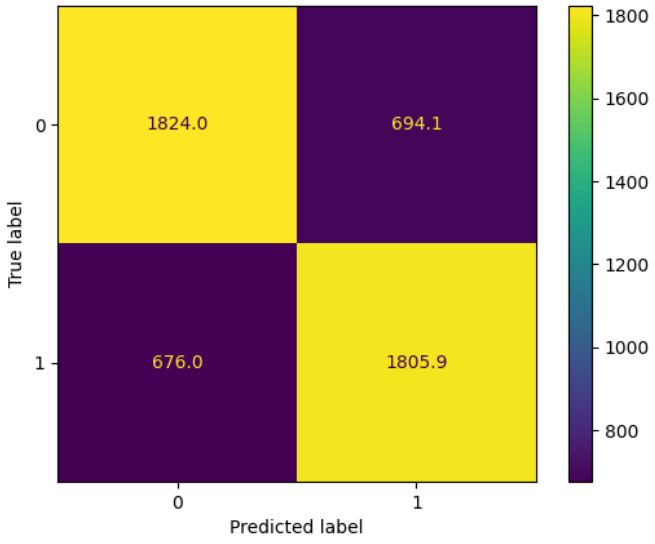
Results:

Logistic Regression:



Accuracy: 0.72596
Precision: 0.7322062360401083
Recall: 0.71268

Multinomial Naive Bayes:



Accuracy: 0.72598
Precision: 0.7244552063837617
Recall: 0.7296

Comparing the Models:

Based on the results from each model, we cannot conclude that one is superior to the other. The differences between the accuracy, precision, and recall of each model are negligible.

When comparing the confusion matrices, we can see that Multinomial Naive Bayes produced on average 42.3 more true positives and 42.2 more false positives than Logistic Regression. Logistic Regression produced on average 42.3 more false negatives and 42.2 more true negatives than Multinomial Naive Bayes.

Both models are ~22% more accurate than the baseline of randomly guessing whether a review is positive or negative (would result in 50% accuracy on average).

Comparing Features:

The features used to train the models and produce the results above are:

- Positive word count
- Negative word count
- log(word count of review)
- 1 if “no” ∈ review, 0 if otherwise
- 1 if “!” ∈ review, 0 if otherwise

When comparing these features we found that Positive word count and Negative word count are the most important features. When training and evaluating the models using just these two features we get the evaluation metrics:

Logistic Regression:

- Accuracy: 0.7196199999999999
- Precision: 0.7121819353036504
- Recall: 0.73732

Multinomial Naive Bayes:

- Accuracy: 0.71956
- Precision: 0.71670352008824
- Recall: 0.7263200000000001

When not using the Positive word count and Negative word count features, and training and evaluating the model using only the other 3 features we get the evaluation metrics:

Logistic Regression:

- Accuracy: 0.56658

- Precision: 0.5497395573062347
- Recall: 0.7362000000000001

Multinomial Naive Bayes:

- Accuracy: 0.5681200000000001
- Precision: 0.5504595314748688
- Recall: 0.74344