IMDB Movie Review Sentiment Prediction

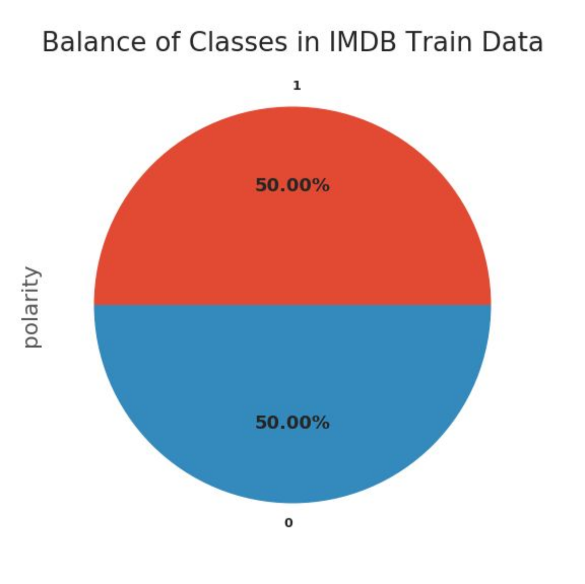
Chelsea, Helen, Jackson, Min, Temesgen

Background

- 25,000 reviews from IMDB
- Positive scores: >=7   Negative scores: <= 4
- AWS

GOAL:

- Sentiment analysis using NLP (Natural Language Processing) to predict positive or negative reviews on movies

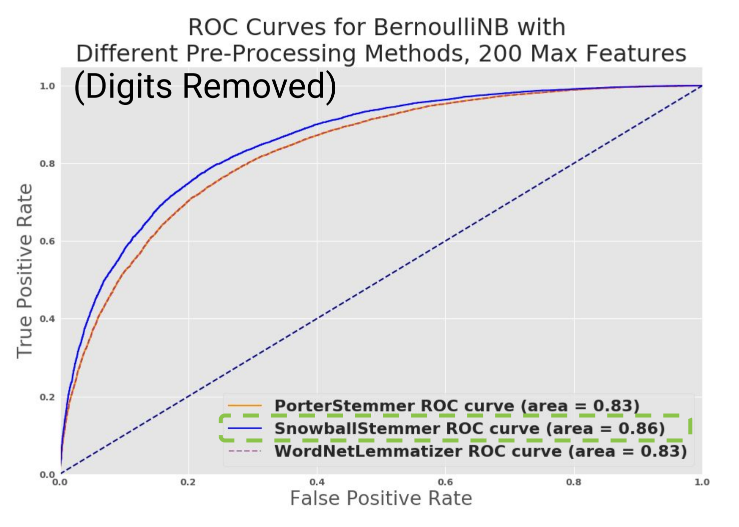


Initial Findings On Training Data

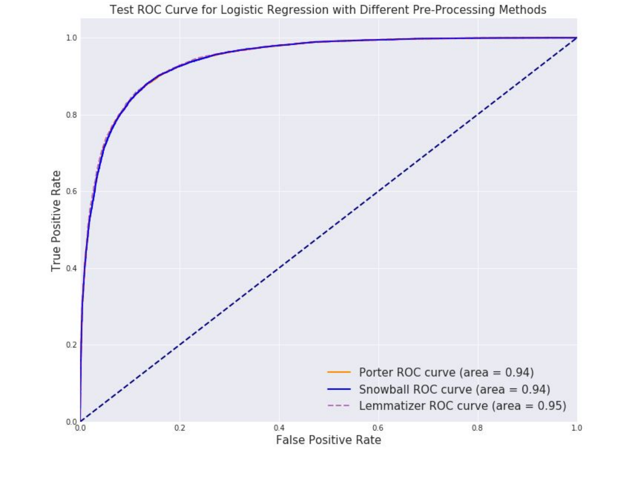
Porter	**Snowball**	**Lemma**
Logistic Regression: 0.95	Logistic Regression: 0.95	Logistic Regression: 0.95
Random Forest: 0.82	Random Forest: 0.84	Random Forest: 0.83
Gradient Boost: 0.89	Gradient Boost: 0.89	Gradient Boost: 0.89

Model Comparisons with ROC Curves

- SnowballStemmer tended to produce higher AUC scores, but not too much difference
  - All TF-IDF with varied max_features and stemmers/lemmatizers
- Overall, Logistic Regression performed better than Bernoulli Naive Bayes
- We decided to move forward with Logistic Regression on all features, digits included



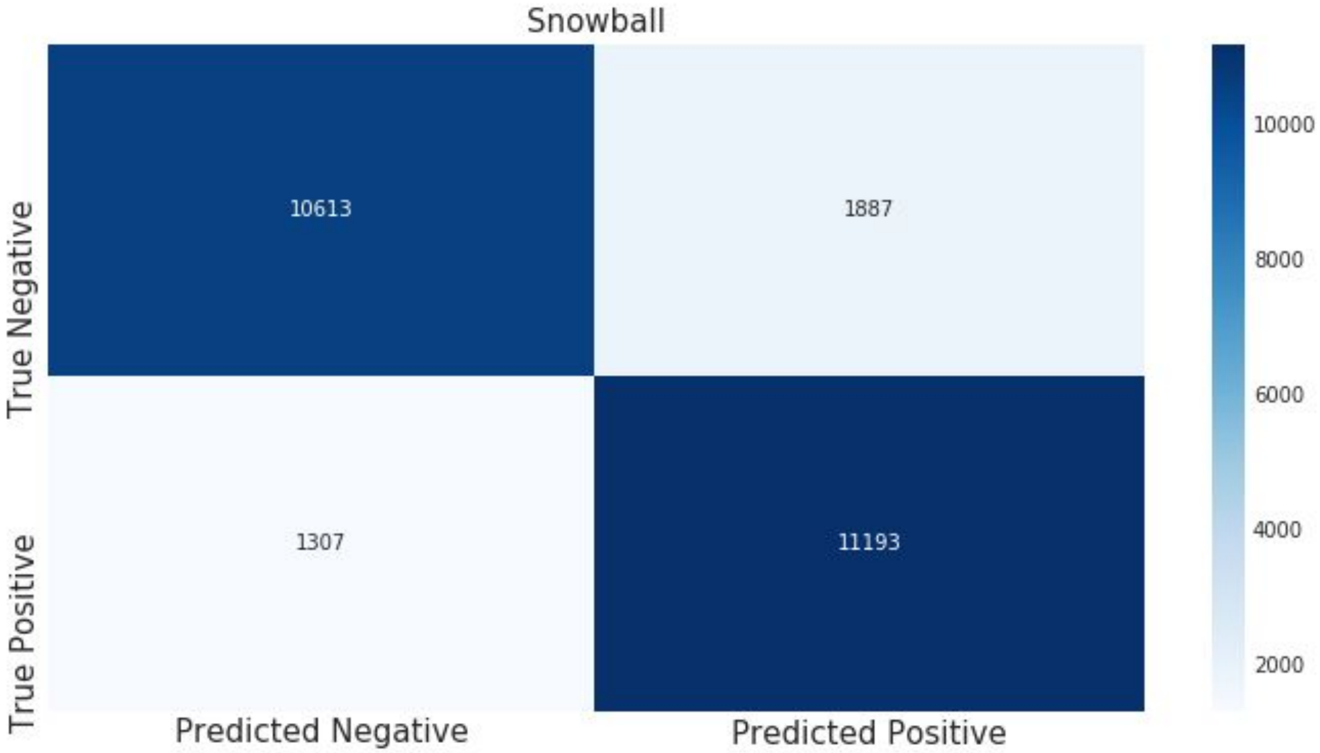
All Features (50k+), Digits Included

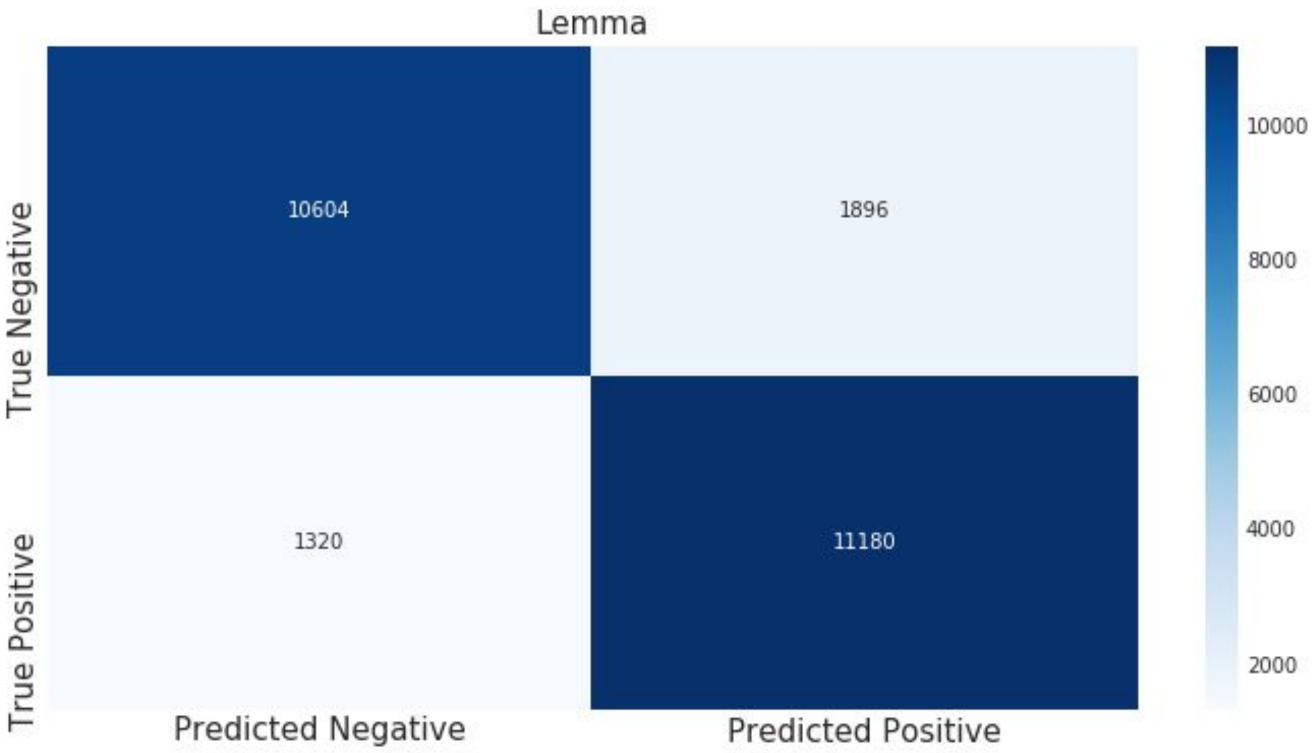


Model Selection

- Normalize data
- Include Numbers
- Tokenized stop words and reviews
- Stemming and lemmatization
- Logistic Regression
    - Porter: 0.8707
    - Snowball: 0.8722
    - Lemma: 0.8713

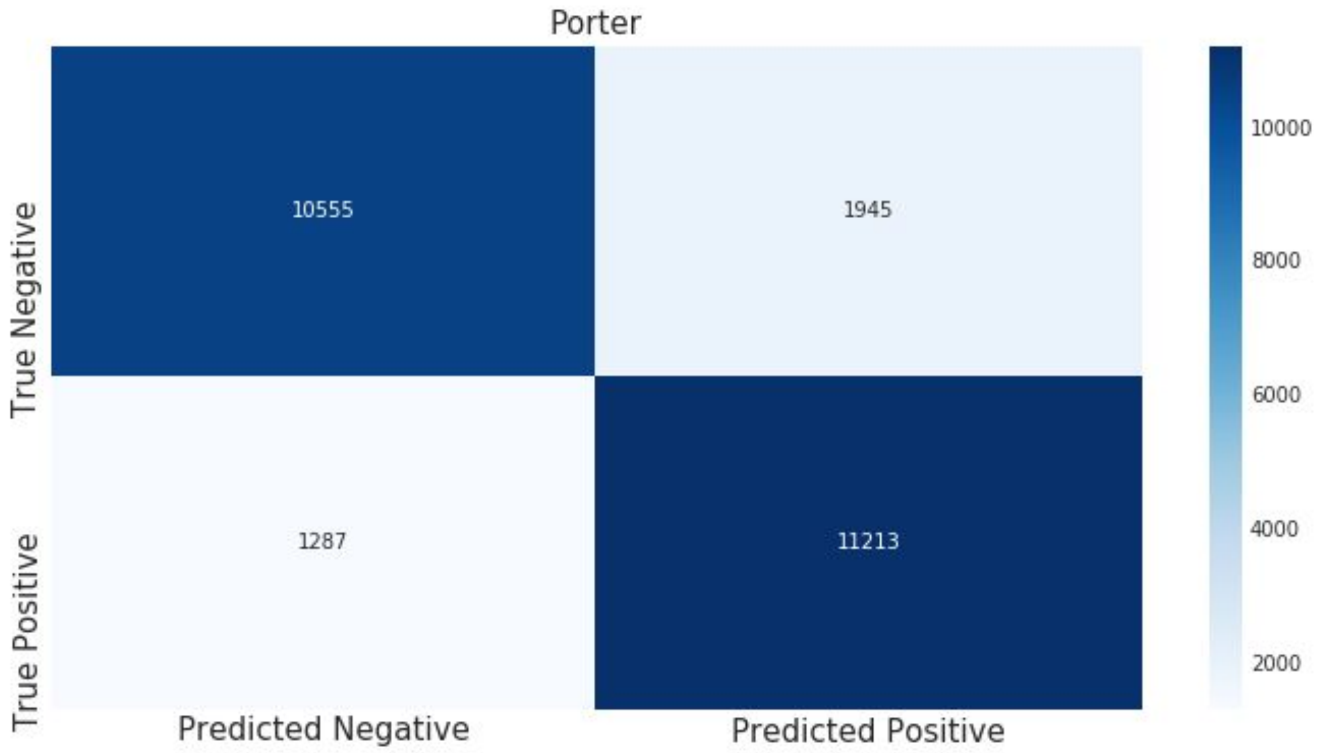
Confusion Matrix Model Comparison


Snowball
True Negative
True Positive
Predicted Negative
Predicted Positive
10613
1887
1307
11193
10000
8000
6000
4000
2000



Lemma

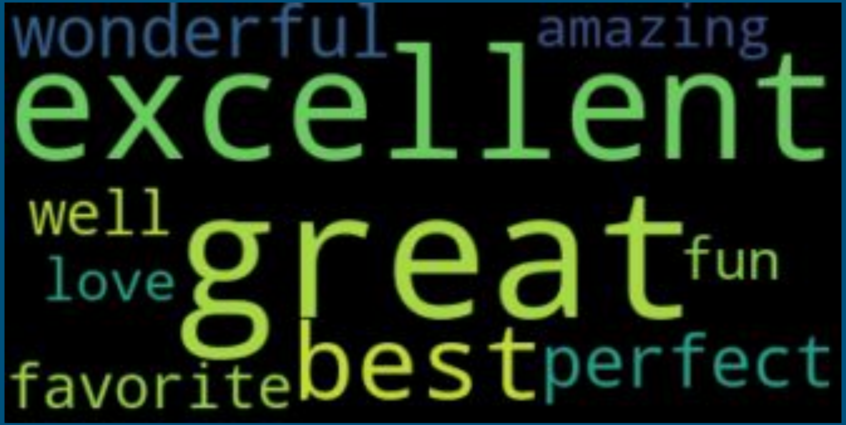
True Negative
True Positive
Predicted Negative
Predicted Positive
10604
1896
1320
11180
10000
8000
6000
4000
2000

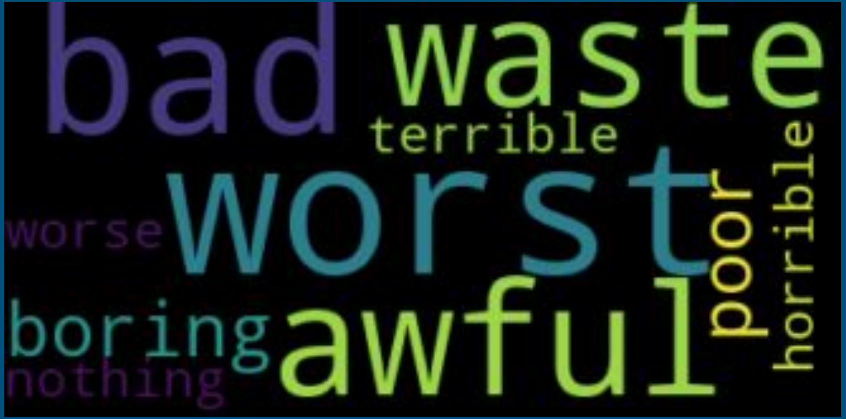


Porter

	Predicted Negative	Predicted Positive
True Negative	10555	1945
True Positive	1287	11213

Top 10
Positive Words vs. Negative Words


wonderful
amazing
excellent
well
great
fun
love
favorite
best
perfect


bad
waste
terrible
worse
worst
poor
horrible
boring
awful
nothing