Capstone Project

Case Study 3: Document Classification

Spencer K’Burg

September 2019

**Overview**

The goal of this project is predict whether a review is positive or negative based on the text of the review. This is a classic sentiment analysis problem. I was able to build a predictor that correctly predicted the sentiment of the review 87% of the time.

**Problem Statement**

Given a set of previously unseen reviews, correctly predict whether they are positive or negative reviews.

**Metrics Used**

I primarily used the accuracy score in evaluating the model. I also used the confusion matrix to help understand where I was missing with my predictions.

**Analysis**

**Dataset and Features**

The dataset consists of 50,000 reviews. Half of the reviews are positive and half are negative. I split the dataset into a train and test set (75%/25%) and didn’t look at the test set until it was time to review the results. There was one string of text for each review and that is the only feature given. I was able to use the text in that field to create word vectors that I used in my predictions.

**Data Exploration**

I looked to see if the dataset was balanced between positive and negative reviews (it is). This helped to set a baseline to compare my predictions against.

**Algorithms and Techniques Used**

I used word vectorization to convert the text strings into something that I could build a model on. I used the Word2VecKeyedVectors in the gensim library and the

word vectorization available at <https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M.vec.zip> as a basis for the word vectorization. I did some basic tokenization before applying the vectorization to get better results. A logistic regression model was used to fit the model.

**Results and Conclusion**

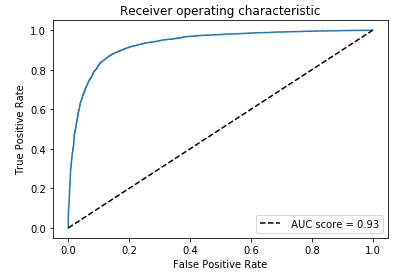
**Graph 1.**

Confusion Matrix with results on the test dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted Label | |
|  |  | Positive | Negative |
| True Label | Positive | 5,277 | 826 |
| Negative | 814 | 5,583 |

**Graph 2.**

ROC Curve for the test dataset



I was able to correctly predict the sentiment of the reviews in the test set 87% of the time. This seems to be a pretty good result given that 50% of the reviews are positive and 50% of the reviews are negative. This also seems reasonable compared to other classification tasks that I’ve seen in other settings. The confusion matrix also shows that the misclassification is pretty evenly split between false positives and false negatives. The model also has the benefit of not taking too long to fit and make predictions. Overall, the model is pretty consistently able to correctly identify the sentiment of a review based on its text.