Diego Valdes

IST 707

Home Work 8

6/1/2019

**Introduction**

Data is everywhere. While many worry about the carbon foot print of a human, it isn’t often that one’s data foot print is discussed. How much data does the average person leave behind, without even knowing about it? Search a destination before a morning drive, zooming under street light cameras, swiping credit card for morning coffee, then later for lunch, then later at the vending machine, asking Google a question, commenting on a Facebook post, buying a few items on Amazon, and finally, liking that twitter post that’s trending.

All the while, with every credit card swipe, and key stroke, a person’s data is being recorded. Most know that it’s used in algorithms to make product recommendation, suggest locations to visit, detect fraud, and all of these things are helpful in the world we’ve created that these abilities seem like a necessity. To companies, this is an opportunity, a chance to get to know customers, predict their needs and be able to better meet demands. And as the collection methods become broader, and the algorithms smarter, and the possibilities endless, the technology that just recommended a product, can now complete thoughts before they are written.

Earlier this year, Google introduced a feature in Gmail where the system would offer predictions for sentences. After just a few words, a prediction would appear for the user to accept. The system, at times, doesn’t even need a sentence to start; after typing a period a user can be given their next sentence in its entirety. Google has not released any data about the acceptance rate of recommendations, but it isn’t hard to imagine that all the data being generated from users is being used to improve the feature.

**Analysis**

**About the Data**

The deception dataset represents a series of reviews that are legitimate and fake, as well as the sentiment, positive or negative. Each row is one review. To prepare the data for analysis, all punctuation was removed. Words using apostrophes where treated as one word, ie: when’s as whens. This created a dataset with 3 columns: lie, sentiment, and review. A ‘title’ column was added to identify each review (review\_1, review\_2, …review\_n), but wasn’t necessary in analysis.

The dataset was converted from it’s form to a larger matrix where each word was a column, the review was a row, and the value was the word frequency. Stop words were removed (and, it, as, …). This left 1337 words among the 92 reviews. Further exploration of the data found that the reviews didn’t share many of the 1337 words. Only 3 words appeared in 30% of the reviews. Dropping the percentage of common words led to final results: 10% - 42 words, 5% - 121 words. (fig 1.1, fig1.2)

Fig 1.1 words appearing in 10% of reviews Fig 1.2 words appearing in 5% of reviews

A column was added to the final dataset with the total word length of the review and the frequency was transformed to a percentage based on the length. For analysis, 3 datasets were generated to be used in the models. These datasets where all scaled accordingly based on the length of the review and the differences are listed.

* Datasets
  + Set 1
    - Complete 1340 words
  + Set 2
    - 10% of shared words, 42 words
  + Set 3
    - %% of shared words, 121 words

**Models**

**Naïve Bayes (e1071::naiveBayes)**

All three sets were used to predict sentiment and legitimacy of the reviews. The model used the default values for each set, as experimenting with the parameters didn’t bring any improvements. Improvements in the model was seen based on the dataset used.

**SVM (e1071::svm)**

All three sets were used to predict sentiment and legitimacy of the reviews. The model went many alterations between the datasets and the kernels. None of the alterations within the parameters produced an increase in results and the default values were used, as these seemed to produce the best accuracy in testing.

**Results**

All three datasets were randomized and then separated into training and test sets, composed of 70% (65 observations) for training and 30% for testing (27 observations). Each set was used to create a model to test review legitimacy and sentiment, for a total of 12 models. Models using datasets 2 and 3, which used less words than the 1340 performed less accurately than dataset 1. For that reason, only models using dataset one will be discussed.

The model performing best for predicting review legitimacy was naïve bayes with 67% accuracy on the test data. Examining the other statistics (kappa, precision, and recall) are more stringent determinations of the performance of the model’s classification abilities. This model is only slightly performing better than the random selection possibility and may not be reliable for implementation, fig 2.1.

The SVM model performed poorly in every important statistic, eliminating it from possibility of use for predicting legitimacy, fig 2.2.

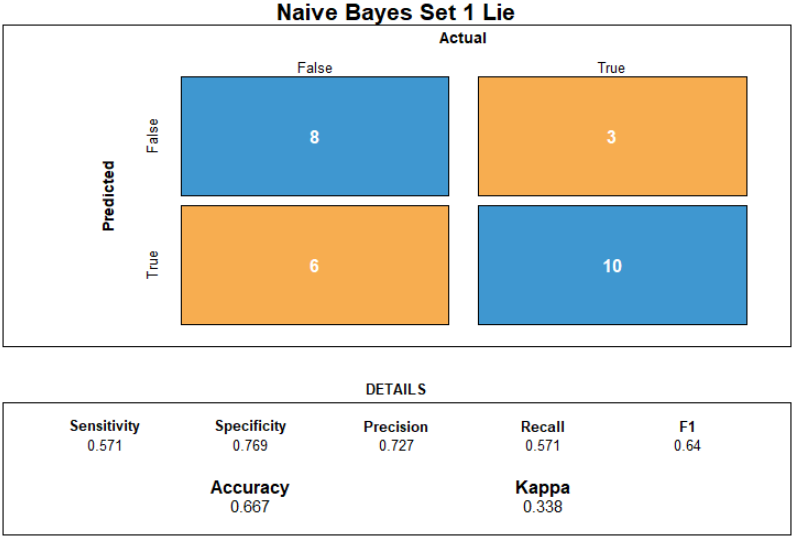
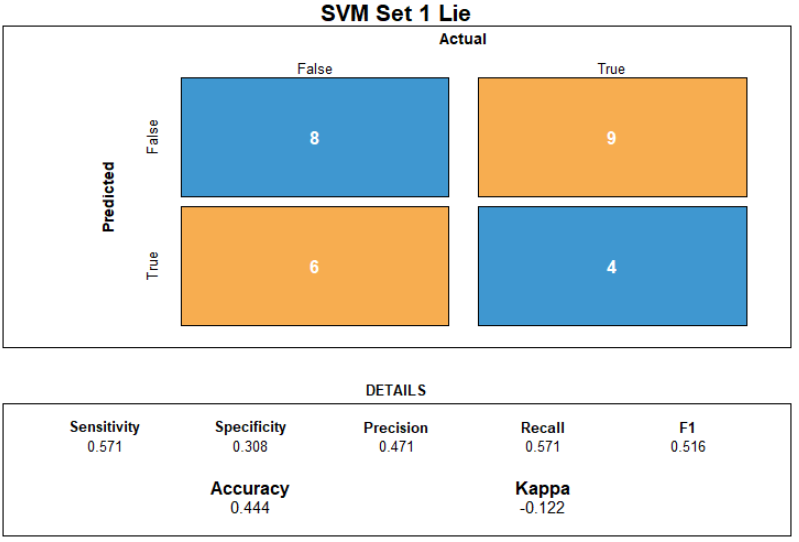
 

Fig 2.1 Confusion matrix: Naive Bayes - Lie Fig 2.2 Confusion matrix: SVM - Lie

When predicting sentiment, naïve bayes performed outstanding on the test data. Accuracy, precision, kappa, and recall are in good places to make it a good choice for usage in predicting review sentiments, fig 2.3.

SVM, as with legitimacy, performed poorly. Although precision is in a good range, accuracy is performing just as well as random selection, fig 2.4.

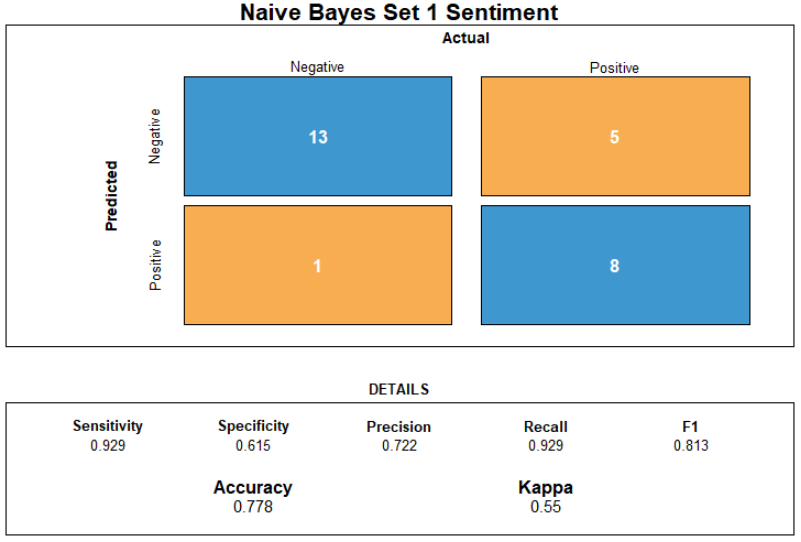
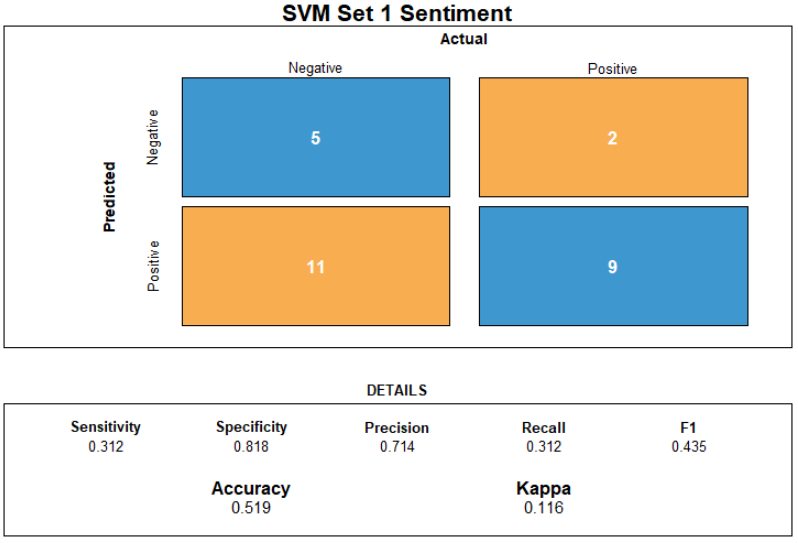
 

Fig 2.3 Confusion matrix: Naive Bayes - Sentiment Fig 2.4 Confusion matrix: SVM - Sentiment

**Conclusion**

SVM performed poorly in both predicting legitimacy and sentiment. In attempting to predict legitimacy, the model performed less accurately than random guessing. This model, in both situations, cannot be used for predicting either legitimacy or sentiment of the reviews.

The naïve bayes model appears to be adequate in predicting legitimacy based on it’s accuracy of the test data. However, many of the other measures of model performance weren’t promising. With the kappa score being at 0.338, this model is only performing slightly better than random guessing and may fluctuate below 50% if used against other data. Of the two measures determining positive predictions, precision, was good, being higher than 70%, but recall was not at around 50%. This model as well is not suitable for predicting legitimacy.

Naïve bayes did perform well in all measures in determining sentiment. The accuracy on the test set was above 70%. The other measures determining positive predictions and performance above random selection were at levels that bring reliability to using this model to predict sentiment.