Predicting Yelp Review Ratings

Andy Chen

1 Introduction

Online reviews are a valuable data source for businesses to get feedback. This data is easily accessible, but the task of extracting useful information is prohibitively difficult. Hence, automated processing of review data has become a topic of interest in machine learning. Previous research has shown that Naive Bayes, support vector classifiers and logistic regression are suitable for sentiment analysis tasks (Vinodhini et al. 2012). In this paper, we apply some of these machine learning models on review text and metadata to generate predictions for ratings.

2 Data & Features

This review rating prediction task was performed on data from Yelp (Mukherjee et al., 2013) and (Rayana and Akoglu, 2015). Each review was associated with a rating of 1, 3 or 5, and various features were constructed from the review text and the vote counts (useful, funny, cool).

2.1 Uni-grams (Bag-of-Words)

In the uni-gram (Bag-of-Words) approach, each unique word in the text is considered as its own feature. We opt to include stop-words, as the incorrect exclusion of relevant words (which may exist in pre-compiled lists of stopwords) has negatively impacted past sentiment analyses (Saif et al. 2014). We also consider as features: 1) words with different capitalisation and 2) exclamation marks with their preceding word. The motivating reason is that capitalisation and punctuation can convey emotion in written language, and past sentiment analyses have shown these as potentially useful features (Koto and Adriani, 2015).

2.2 Bi-grams

The uni-gram approach fails to capture the meaning which arises from interacting words. This limitation is particularly significant in this task, since we expect there to be many relevant word pairs involving a modifier (eg. "tasty burger") or negations ("not good"). Thus, two-grams are included to overcome this limitation. Hundreds of thousands of features are generated, and we consider the $30,000$ features most correlated with rating in the training data. Correlation is assessed by χ^2 distance.
In addition, explicit mentions of rating (strings of the form "x star(s)") are included as a feature.

2.3 Paragraphs

Paragraphs, represented as feature vectors, have been shown to outperform bag-of-word features in the past (Le and Mikolov, 2014). To avoid missing relevant patterns, 200 features are used for the task.

2.4 Voting data

Finally, the number of "useful", "cool" and "funny" votes are used as 3 features.

3 Model Construction

We constructed three classes of models. The models were trained on 80% of the data and validated on the remaining 20%. The most accurate model on the validation set was taken as the final model. We then refitted this model to ensure that the variance was low and the performance was consistent.

3.1 Decision Tree

We first established a baseline using 0R. This achieved 69.20% accuracy.

We then constructed a decision tree, with Gini impurity as the splitting criterion. The decision tree model was assessed across a range of values for maximum depth.

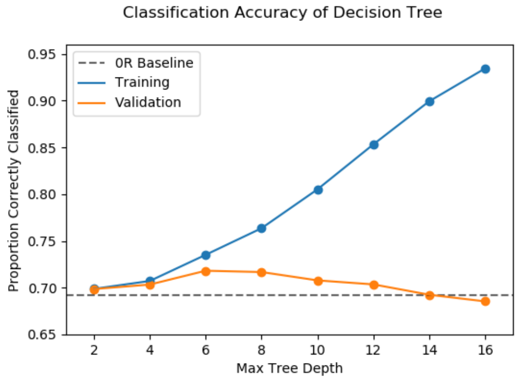

Figure 1: Maximum Decision Tree Depth vs. Accuracy

Figure 1 shows that the decision tree classifier did not generalise well. Deeper trees were overfitted, and the best validation accuracy of 71.28% was achieved when depth was restricted to 6.

Theoretically, a restriction on tree depth prevents overfitting and reduces model variance. We tested five more models.

Trial	Validation accuracy
1	72.53%
2	70.75%
3	71.21%
4	71.16%
5	72.03%
Mean	71.54%

Table 1: Decision Tree Model Variance

Table 1 shows that the accuracies were similar to the original best model, which suggests that model variance is low.

3.2 Multi-class Logistic Regression

We fitted a logistic regression model, using a one vs. rest strategy for multi-class prediction. $L2-$regularisation was used to shrink the coefficients of less influential features, thus preventing overfit on noise in the training data. A range of regularization parameters were tested.

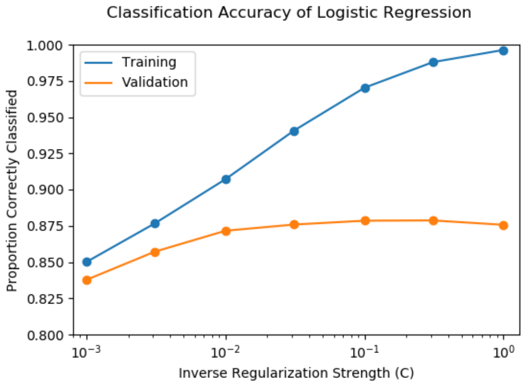

Figure 2: Regularisation Strength vs. Accuracy for Logistic Regression

Figure 2 shows that an 87.87% validation accuracy was achieved in the best model ($C = 0.31$). To test if this result replicates for different datasets, the model was reconstructed for 5 different splits.

Trial	Validation accuracy
1	88.30%
2	87.98%
3	87.71%
4	87.94%
5	87.96%
Mean	87.98%

Table 2: Logistic Regression Model Variance

The results in Table 2 indicates that the best model has a low variance.

3.3 Support Vector Classifier

We fitted a support vector classifier (SVC) using a one vs. rest strategy for multi-class prediction. We opted for a linear kernel, as other non-linear kernels were prohibitively slow in training. Overfitting was prevented with soft margin classification, which maximises the distance between the support vectors while allowing misclassification on the training data. This was controlled with a regularization parameter, a range of which were tested.

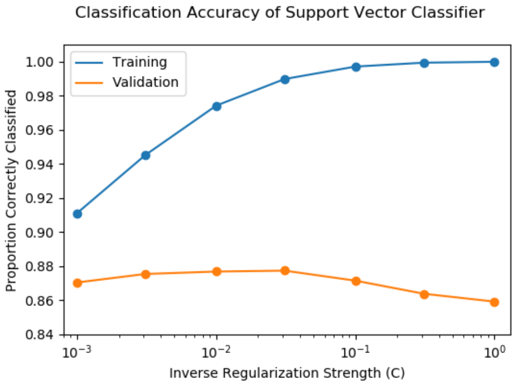

Figure 3: Regularisation Strength vs. Accuracy for Support Vector Classifiers

Figure 3 shows that the best model ($C = 0.031$) had a 87.74% accuracy. The inability to get 100% in validation indicates that the feature space is probably not linearly separable.

The SVC can be vulnerable to overfitting. Figure 3 shows how validation accuracy is poorer when the training accuracy is near 100%. The benefit of regularisation is evident, since the validation accuracy is better at around $C \approx 10^{-2}$, where the training accuracy isn't perfect.

Five models were constructed with different train/validation splits.

Trial	Validation accuracy
1	88.30%
2	87.98%
3	87.71%
4	87.94%
5	87.96%
Mean	87.98%

Table 3: Logistic Regression Model Variance

Table 3 shows that the model has low variance, so the results are consistent across datasets.

4 Comparison of Models

4.1 Best Performance

Classifier	Accuracy
0R Baseline	69.20%
Decision Tree	71.28%
Logistic Regression	87.87%
Support Vector Classifier	87.74%

Table 4: Comparison of Model Performance

The results in Table 4 suggests that SVC and logistic regression are far superior to the decision tree, which is only marginally better than the baseline. The SVC and logistic regression models are practically equivalent, as the difference in validation accuracy is small relative to the model variance.

4.2 Detailed Performance Metrics

Rating	Precision	Recall	F1 Score
1	57.7%	8.7%	15.2%
3	46.8%	35.0%	40.1%
5	77.2%	91.4%	83.7%

Table 5: Performance Metrics for the Decision Tree

Rating	Precision	Recall	F1 Score
1	84.6%	65.7%	73.9%
3	78.4%	71.6%	74.8%
5	90.8%	95.8%	93.3%

Table 6: Performance Metrics for Logistic Regression

Rating	Precision	Recall	F1 Score
1	82.5%	66.3%	73.5%
3	77.9%	71.8%	74.7%
5	91.0%	95.5%	93.2%

Table 7: Performance Metrics for SVC

Different performance metrics show that the decision tree is generally a poor model, and performs very badly on 1 and 3-star reviews. The logistic regression model and SVC were similar and generally performed well on all metrics. All 3 models performed better on 5-star reviews compared to 1 and 3-star reviews, which indicates model bias.

4.3 Error Analysis

		Predicted			
		1	3	5	Total
Actual	1	41	181	247	469
	3	15	441	804	1260
	5	15	320	3550	3885
	Total	71	942	4601	5614

Table 8: Confusion Matrix of the Decision Tree Classifier

		Predicted			
		1	3	5	Total
Actual	1	308	99	62	469
	3	44	902	314	1260
	5	12	150	3723	3885
	Total	364	1151	4099	5614

Table 9: Confusion Matrix of the Logistic Regression Classifier

		Predicted rating			
		1	3	5	Total
Actual	1	311	96	62	469
	3	52	905	303	1260
	5	14	161	3710	3885
	Total	377	1162	4075	5614

Table 10: Confusion Matrix of the SVC

The confusion matrix showed misclassifications amongst all 3 classes. It is not surprising for 3-star reviews to be misclassified, as 3-star reviews would share similarities with 1 and 5-star reviews. However, it is surprising that 1-star reviews were confused with 5-star reviews, as we'd expect them to be completely different. The three models were also affected by bias, as the models overpredicted 5-star ratings and underpredicted 1 and 3 star ratings. This bias was likely because of the unbalanced class labels, as there was far more 5-star ratings in the data set.

5 Conclusion

In this paper, we use decision trees, logistic regression and support vector classifiers to predict review ratings. We found that logistic regression and support vector classifiers were best suited to this task.

6 References

1. Vinodhini, G. and Chandrasekaran, R.M., 2012. Sentiment analysis and opinion mining: a survey. International Journal, 2(6), pp.282-292.

2. Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. 2013. What Yelp fake review filter might be doing? In *7th International AAAI Conference on Weblogs and Social Media.*

3. Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 985-994.

4. Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of twitter. In *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14).*

5. Fajri Koto and Mirna Adriani. 2015. A comparative study on twitter sentiment analysis: Which features are good?. *International Conference on Applications of Natural Language to Information Systems.*

6. Distributed Representations of Sentences and Documents. Q. Le & T. Mikolov, ICML, 2014.