COMP30027 Report – Yelp Rating Prediction with Naïve Bayes and SVM

1. Introduction

The advancement of technology has led to the rise of online review sites such as Yelp. Customers today rely heavily on online reviews to make purchasing decisions. However, the quality of reviews often fluctuates as many of the reviews may be biased or contradictory.

To analyse online reviews, sentimental analysis can be conducted by extracting information from texts to understand its underlying opinion. For example, an online review can be classified as negative, neutral, or positive. Some related work in this field includes Hu and Liu (2004) who had mined product features commented by customers and identifying the sentiment in each review. Sasikala (2018) has also implemented various classifiers to predict the sentiments of online reviews with empty ratings.

In this paper, the Naïve Bayes and SVM models are built to test the effectiveness of these models in predicting the star rating of restaurants in Yelp according to the review texts.

2. Dataset

The Yelp reviews used in this experiment were provided by Mukherjee, Venkataraman, Liu and Glance (2014) and Rayana and Akoglu (2015).

2.1 Feature Extraction

After cleaning the reviews, Doc2Vec features with 50, 100 and 200 dimensions have been extracted for use in the classifiers built (Figure 1).

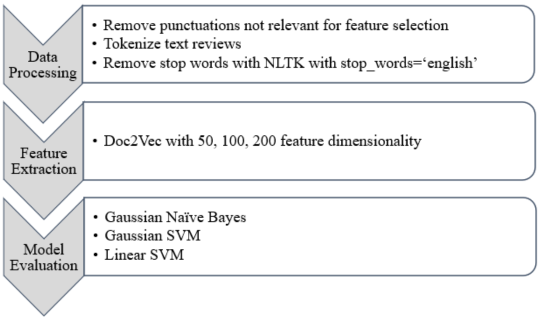

Data Processing
• Remove punctuations not relevant for feature selection
• Tokenize text reviews
• Remove stop words with NLTK with stop_words='english'
Feature Extraction
• Doc2Vec with 50, 100, 200 feature dimensionality
Model Evaluation
• Gaussian Naïve Bayes
• Gaussian SVM
• Linear SVM

Figure 1- Experiment process

2.1.1 Doc2Vec

Doc2Vec is an extension to Word2Vec where the entire document is encoded instead of individual words. This text feature is used as each restaurant review can be viewed as a separate paragraph. Doc2Vec provides sematic information of texts as words that have similar context will be closer to each other in the vector space (Le & Mikolov, 2014) .

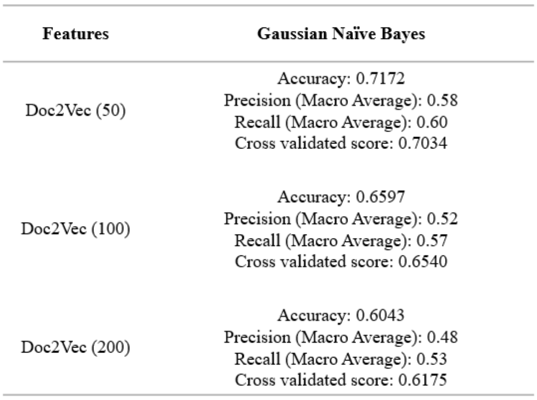
3. Model Evaluation

3.1 Gaussian Naïve Bayes (GNB)

Naïve Bayes is a probabilistic learning method based on the Bayes' theorem with the assumption of independence between each pair of features.

The input for multinomial NB model must be non-negative, hence the GNB model is used for this experiment instead as the Doc2Vec vector is normally distributed.

As seen in Table 1, the performance of the GNB model deteriorates as the number of dimensions of the Doc2Vec increases.



Features	Gaussian Naïve Bayes
Doc2Vec (50)	Accuracy: 0.7172 Precision (Macro Average): 0.58 Recall (Macro Average): 0.60 Cross validated score: 0.7034
Doc2Vec (100)	Accuracy: 0.6597 Precision (Macro Average): 0.52 Recall (Macro Average): 0.57 Cross validated score: 0.6540
Doc2Vec (200)	Accuracy: 0.6043 Precision (Macro Average): 0.48 Recall (Macro Average): 0.53 Cross validated score: 0.6175

Table 1- GNB performance

The learning curve of GNB (Figure 3) shows us that from approximately 450 training data onwards, the validation mean squared error (MSE) would almost be the same, which explains why adding more dimensionality to the Doc2Vec feature did not improve the overall performance of the model (Perlich C., 2009). The training score decreases as the number of training example increases, indicating the underfitting of data and a high bias.

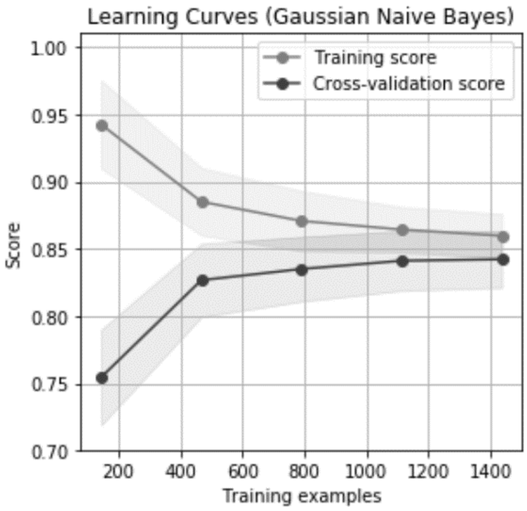


Figure 2- GNB learning curve

The bias is caused by the skewed distribution of rating labels, where 68.72% of ratings were 5-star ratings (Figure 2).

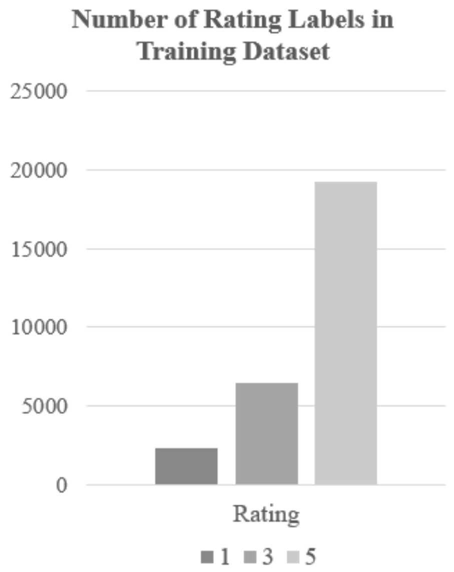


Figure 3- Number of Rating Labels in Training Dataset

The high frequency of the 5-star rating label would cause the model to be biased in predicting that particular label (Frank E. & Bouckaert. R.R., 2006). The asymmetric confusion matrix (Figure 3) also shows us that the model is biased as it has correctly classified many of the 5-star rating reviews whereas performed poorly in correctly classifying other ratings.

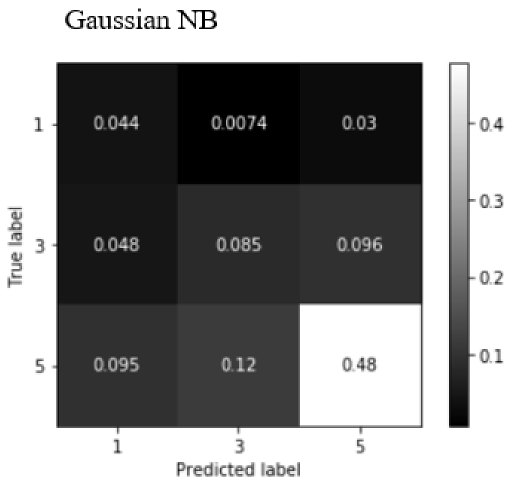


Figure 4- GNB normalised confusion matrix

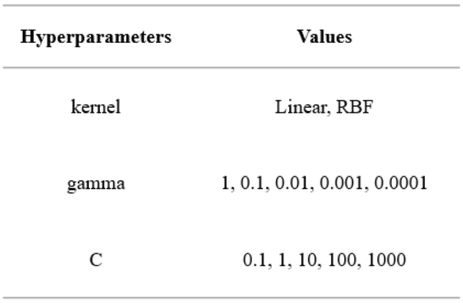
While simple to implement, the GNB model suffers from high bias which resulted in inaccuracies in classification of reviews. Therefore, instead of using more data for this model, a more complex model should be built for the given problem.

3.2 Support Vector Machine (SVM)

SVM is a classifier that separates points with a hyperplane. This model has been chosen for this experiment as it works well with the sparse Doc2Vec matrix and can also generalise well in high dimensional feature spaces (Joachims, 1998).

3.2.1 Hyperparameter Tuning

Hyperparameter tuning was performed to explore the best combination for the given experiment. The grid search method is implemented to build a model for every combination of hyperparameters (Liu et al. 2006, p.713). The hyperparameters tuned and the range of values tested are presented in Table 2.



Hyperparameters	Values
kernel	Linear, RBF
gamma	1, 0.1, 0.01, 0.001, 0.0001
C	0.1, 1, 10, 100, 1000

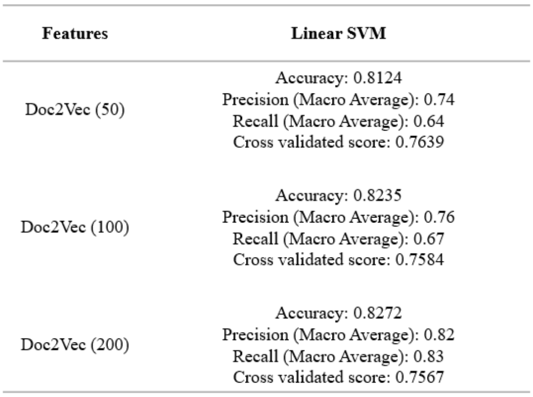
Table 2- SVM hyperparameters tuned

The optimal hyperparameter combination found for this experiment was a radial basis function (RBF) kernel with gamma=0.001 and C=10. Gamma influences the hyper-line flexibility where a higher value may cause overfitting. C determines the penalty of error term, where a lower value indicates a smoother decision boundary and leniency on incorrect classifications.

The linear and gaussian SVM models are further examined below.

3.2.2 Linear SVM

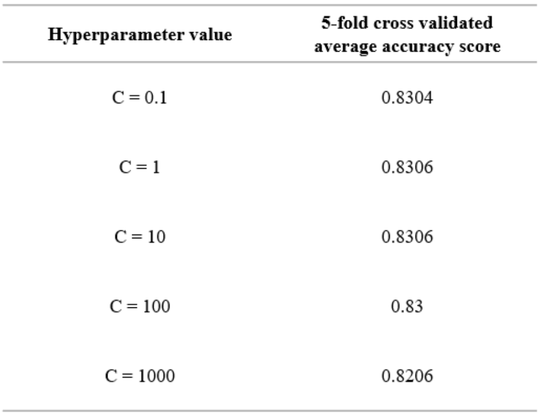
Table 3 shows that the improvement of performance of the linear SVM model as the number of dimensions of the Doc2Vec increases.



Features	Linear SVM
Doc2Vec (50)	Accuracy: 0.8124 Precision (Macro Average): 0.74 Recall (Macro Average): 0.64 Cross validated score: 0.7639
Doc2Vec (100)	Accuracy: 0.8235 Precision (Macro Average): 0.76 Recall (Macro Average): 0.67 Cross validated score: 0.7584
Doc2Vec (200)	Accuracy: 0.8272 Precision (Macro Average): 0.82 Recall (Macro Average): 0.83 Cross validated score: 0.7567

Table 3- Linear SVM performance

With a linear kernel, the accuracy score of the SVM model would decrease if the C hyperparameter is too low or too high, hence a moderate value would optimise the performance of the model (Table 4).



Hyperparameter value	5-fold cross validated average accuracy score
C = 0.1	0.8304
C = 1	0.8306
C = 10	0.8306
C = 100	0.83
C = 1000	0.8206

Table 4- Accuracy score of linear SVM with different C values

The learning curve of this model (Figure 5) shows a large gap between the training score and cross-validation score, which indicates a high variance. This suggests that a more complex model should be implemented or more data should be gathered for the given model. Additionally, the training score is almost at its maximum regardless of the number of training examples, which may imply the overfitting of data.

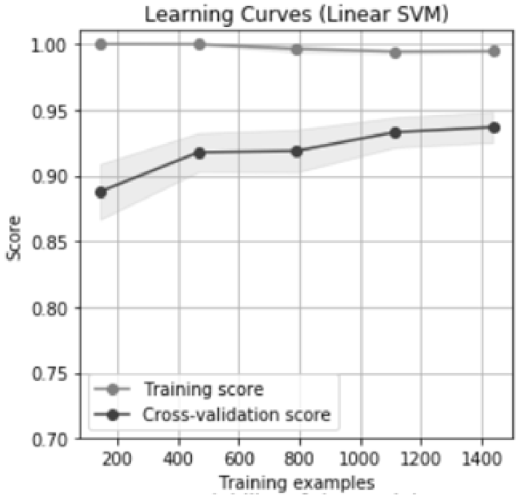


Figure 5- Linear SVM learning curve

Similar to the GNB model, SVM is also biased in correctly classified many of the 5-star rating reviews due to the unbalanced rating labels (Figure 6).

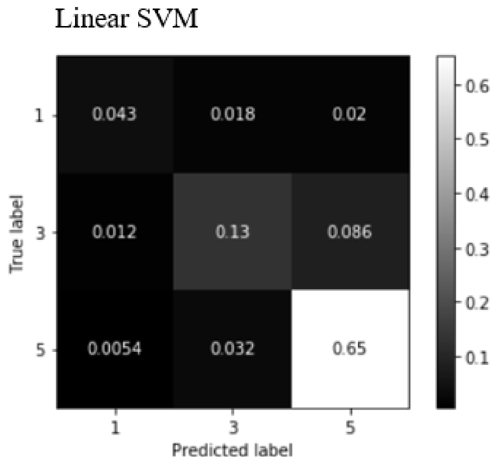
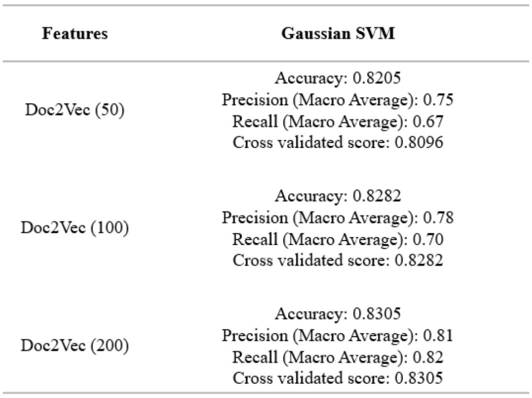


Figure 6- Linear SVM normalised confusion matrix

3.2.3 Gaussian SVM

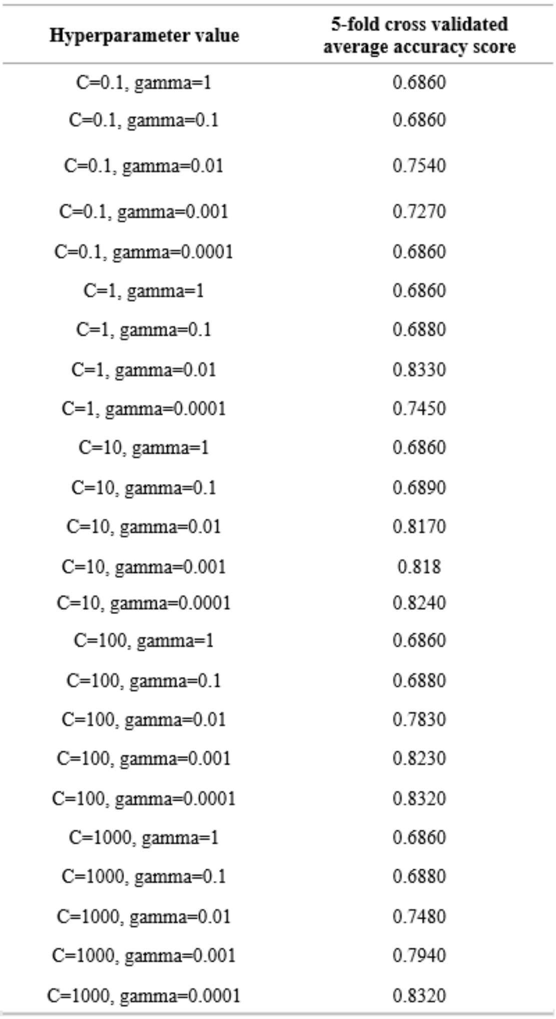
The performance of the SVM model with an RBF kernel has a slightly better performance compared with the linear kernel. Similarly, the model performance also improves when the dimension of Doc2Vec is higher.



Features	Gaussian SVM
Doc2Vec (50)	Accuracy: 0.8205 Precision (Macro Average): 0.75 Recall (Macro Average): 0.67 Cross validated score: 0.8096
Doc2Vec (100)	Accuracy: 0.8282 Precision (Macro Average): 0.78 Recall (Macro Average): 0.70 Cross validated score: 0.8282
Doc2Vec (200)	Accuracy: 0.8305 Precision (Macro Average): 0.81 Recall (Macro Average): 0.82 Cross validated score: 0.8305

Table 5- Linear SVM performance

While performing hyperparameter tuning, it is evident that the accuracy score of the model is higher when the gamma value is lower whereas the value of C has minimal impact on the accuracy scores (Table 6).



Hyperparameter value	5-fold cross validated average accuracy score
C=0.1, gamma=1	0.6860
C=0.1, gamma=0.1	0.6860
C=0.1, gamma=0.01	0.7540
C=0.1, gamma=0.001	0.7270
C=0.1, gamma=0.0001	0.6860
C=1, gamma=1	0.6860
C=1, gamma=0.1	0.6880
C=1, gamma=0.01	0.8330
C=1, gamma=0.0001	0.7450
C=10, gamma=1	0.6860
C=10, gamma=0.1	0.6890
C=10, gamma=0.01	0.8170
C=10, gamma=0.001	0.818
C=10, gamma=0.0001	0.8240
C=100, gamma=1	0.6860
C=100, gamma=0.1	0.6880
C=100, gamma=0.01	0.7830
C=100, gamma=0.001	0.8230
C=100, gamma=0.0001	0.8320
C=1000, gamma=1	0.6860
C=1000, gamma=0.1	0.6880
C=1000, gamma=0.01	0.7480
C=1000, gamma=0.001	0.7940
C=1000, gamma=0.0001	0.8320

Table 6- Accuracy score of Gaussian SVM with different C and gamma values

It is shown in Figure 7 that the gap between the training score and cross-validation score has been narrowed after using an RBF kernel. However, the training score of learning curve of this model is still high regardless of the number of training examples, meaning that there is overfitting of data. This may explain why the optimal value of the hyperparameter gamma is rather low, as the hyperplane is made less flexible to avoid overfitting.

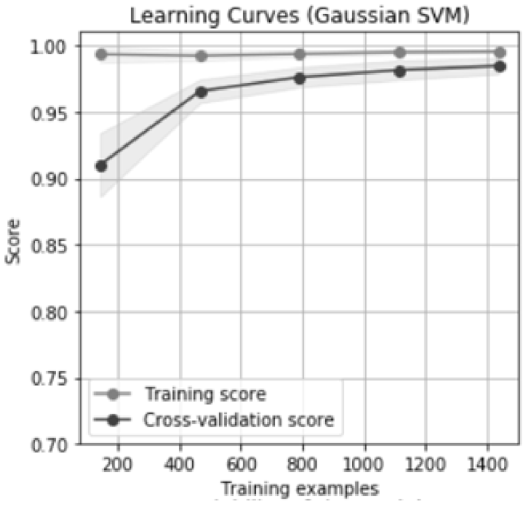


Figure 7- Gaussian SVM learning curve

The Gaussian SVM model also suffers from the bias for correctly classifying 5-star ratings due to the unbalanced rating labels.

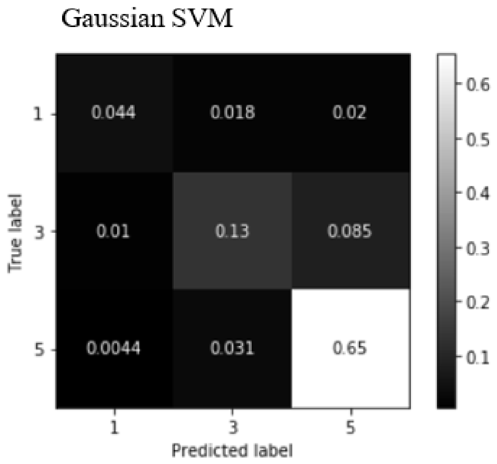


Figure 8- Gaussian SVM normalised confusion matrix

4. Model Comparison

Figure 9 summarises the performance of the 3 models implemented for the experiment. Although GNB has faster implementation compared to SVM, the SVM models have clearly outperformed GNB in this experiment. This may be due to the independence assumption of Naïve Bayes in contrast of SVM which considers the interactions between the features. Wang and Manning (2012) also argued that SVM works better than Naïve Bayes for full-length reviews compared to text snippets.

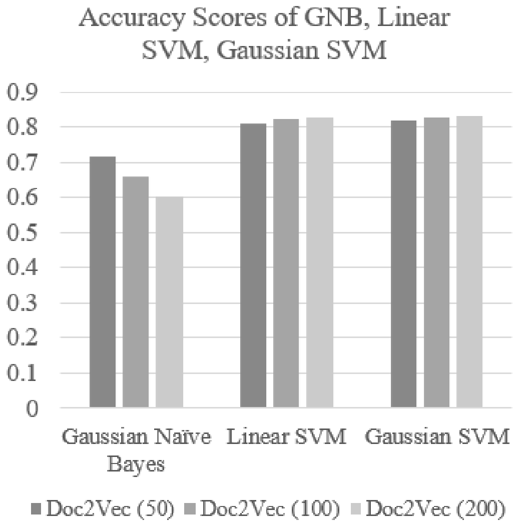


Figure 9- Model accuracy comparison

Furthermore, the GNB model has a high bias and low variance while the SVM model has a high variance and low bias. This bias-variance trade-off exists due to low variance models' tendency to be less complicated whereas low bias models are more complex with a more flexible underlying structure.

5. Conclusions

In this paper, the GNB and SVM models built to predict star ratings for restaurants with Yelp reviews have been contrasted. The performance of these models has been examined and SVM has performed better than GNB with the Doc2Vec text feature in the given problem. It is also evident that having a training dataset with a balanced label can vastly affect the performance of a model.

6. References

Frank E., Bouckaert R.R., Naive Bayes for Text Classification with Unbalanced Classes, 2006.

Hu M., Liu B., Mining and Summarizing Customer Reviews, 2004.

Joachims, J., Text Categorization with Support Vector Machines: Learning with Many Relevant Features, 1998.

Liu R., Liu E., Yang J., Li M., Wang F., 2006, *Optimizing the Hyper-parameters for SVM by Combining Evolution Strategies with a Grid Search*. In: Huang DS., Li K., Irwin G.W. (eds) Intelligent Control and Automation, pp. 712-721.

Mukherjee, A., Venkataraman, V., Liu, B. & Glance, N. What Yelp fake review filter might be doing? 7th International AAAI Conference on Weblogs and Social Media, 2013.

Perlich C., Learning Curves in Machine Learning, pp. 2 – 4, 2009.

Q.V. Le, & T. Mikolov, Distributed Representations of Sentences and Documents, pp. 4, 2014.

Rayana, S. & Akoglu, L. Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015. 985-994

Sasikala, P., Sentiment Analysis and Prediction of Online Reviews with Empty Ratings, 2018.

Wang. S., Manning. C., Baselines and Bigrams: Simple, Good Sentiment and Topic Classification, 2012.