Restaurant Star Review Rating Prediction Using Stack Ensemble Model of Multinomial NB and Logistic Regression

1. Introduction

NLP – the ability for computers to comprehend human language - has been widely used in various fields, and one of them is text classification. The most common approach to tackle this problem is using machine learning classification algorithms, such as logistic regression and multinomial naïve bayes (MNB). However, these texts must be extracted in a form of vector words; hence, bag-of-words (BOW) may be a starting point, as it is simple and easy to interpret. (Harris, cited in Le and Mikolov, 2014). Another option is to implement paragraph vectors that may generalize the vector words dimension, as it captures the semantics of the words (Le and Mikolov, 2014).

This report will discuss the prediction result of the rating using BOW and paragraph vectors as input for MNB and logistic regression models. In addition, a stack ensemble model is used to aggregate the result for consistency. Moreover, the report will demonstrate how combining doc2vec and BOW may improve the model's accuracy.

2. Method

2.1 Data Preparation

The Yelp user review data that was given has been pre processed into BOW using count vectorizer feature selection, and paragraph vector using doc2vec method (Le and Mikolov, 2014). The BOW data's features are further reduced into 50, 100 and 200 features, using Chi and Multinomial feature selection method. Also, the user votes from the meta data are extracted and added to each BOW and doc2vec data.

2.1.1 Model

This experiment applied four models: 0R, MNB, logistic regression and stack ensemble. Firstly, 0R was chosen for the baseline performance due to its consistency in any dataset used, as it only takes the most popular label. Secondly, MNB is famous for a benchmark model in text classification, and it is fast compared to other complex models with reasonable accuracy. It is based on a probabilistic model that uses the word frequency likelihood to help classify the label (Rennie et al., 2017). Thirdly, logistic regression may be a sturdy text classifier, as it considers higher weight for words that may be effective in discriminating the labels (Jurafsky and Martin, as cited in Indra and Wikarsa, 2016). Lastly, to aggregate the result from both models, stack ensemble uses the output from base models probability prediction as input for the meta classifier; hence, this may aggregate the overall accuracy (Ma et al., 2018).

3. Experiment and Result

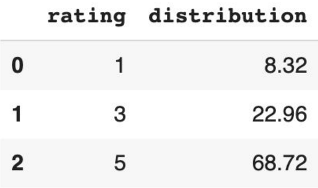
Model	Accuracy (%)
0R	68.72
MNB	85.01 (9 CV)
Logistic Regression	82.25 (6 CV)
Stack	87.03 (10 CV)

Table 1- Maximum average validation accuracy from each model.

3.1 0R

The 0R model was tested on all datasets using validation methods.

Results: The results from each dataset were identical, and it is expected that 0R has 68.72%, as rating 5 is the most popular label with 68.72% occurrences (Table 2).



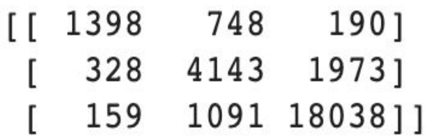
	rating	distribution
0	1	8.32
1	3	22.96
2	5	68.72

Table 2- Class rating distribution in the training dataset.

3.2 MNB

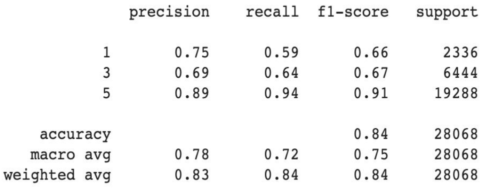
Since MNB only considers natural numbers for its attribute value, the doc2vec datasets were excluded. The model was tested on count vector datasets, and the reduced feature to 50, 100 and 200 using chi and mutual information feature selection. Out of these datasets, the count vector with the added user vote achieved the highest accuracy from the validation set. Then, the model was tuned on its alpha and resulted in an alpha of 1.0. After the tuning step, the model was cross validated with folds ranging from 3 to 10 using the training data to get the highest MNB fitted classifier. Finally, the classifier was used to make a prediction on the test data.

Results: MNB ranked in third with 85.01% accuracy (Table 1). In addition, it performed poorly in rating 1 and 3's recall with f-score below 64% (Table 4). Nonetheless, it has a robust recall for rating 5. Table 6 exhibits that MNB indeed showed a fast algorithmic performance, that is 0.278 over 28068 instances.



```
[[ 1398   748   190]
 [  328  4143  1973]
 [  159  1091 18038]]
```

Table 3- MNB confusion matrix on 9 fold cv



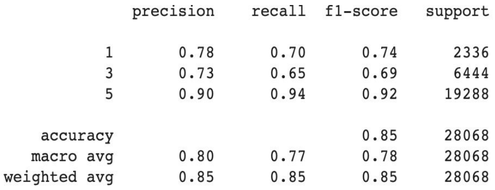
```
              precision    recall  f1-score   support

           1       0.75      0.59      0.66      2336
           3       0.69      0.64      0.67      6444
           5       0.89      0.94      0.91     19288

    accuracy                           0.84     28068
   macro avg       0.78      0.72      0.75     28068
weighted avg       0.83      0.84      0.84     28068
```

Table 4- Classification report table for MNB with 9 fold cv

3.3 Logistic Regression

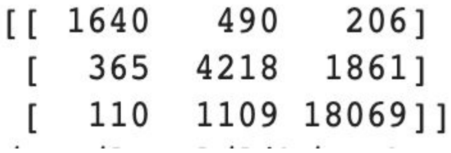
For logistic regression, it is susceptible for negative values; thus, doc2vec was included. However, original count vector data with user vote outperformed doc2vec dataset. The model underwent the same process as MNB. The results for the tuned hyperparameter were C of 10, multinomial saga solver. Subsequently, the model was trained and the maximum classifier was taken to make a prediction on the test data.

Results: Logistic regression surpassed MNB f1-score in every aspect. It achieved a higher accuracy by approximately 1%. Unfortunately, the logistic regression's algorithm complexity far outweighs its performance compared to MNB, as MNB is approximately 3000 times faster than logistic regression with slightly lower accuracy.



```
              precision    recall  f1-score   support

           1       0.78      0.70      0.74      2336
           3       0.73      0.65      0.69      6444
           5       0.90      0.94      0.92     19288

    accuracy                           0.85     28068
   macro avg       0.80      0.77      0.78     28068
weighted avg       0.85      0.85      0.85     28068
```

Table 5- Classification report table for logistic regression with 6 fold cv



```
[[ 1640   490   206]
 [  365  4218  1861]
 [  110  1109 18069]]
```

Table 6 - Logistic Regression confusion matrix on 6 fold cv

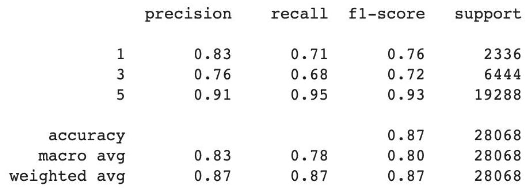
Model	Time (sec)
MNB	0.295
Logistic Regression	969.626
Stack Ensemble	1213.66

Table 7 - Time taken for each model to train and evaluate in the cross validation process of 10 fold.

3.3 Stack Ensemble

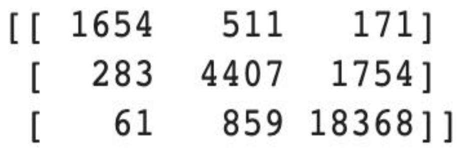
Additionally, for the stack ensemble, validation predictions probability from MNB and logistic regression top classifier were inserted as input for the meta logistic regression. Moreover, another logistic regression model was added with doc2vec plus user vote dataset with the intention to broaden the options for the meta classifier - logistic regression. The meta model was validated with 10 fold, and prediction was done on the most performed fitted meta classifier.

Results: The model increased the overall accuracy by 1%, and improved label 1 and 3 f1-score, which indicates that doc2vec dataset may help stabilize the prediction result. However, it took 1213.66 second to train the model.



```
              precision    recall  f1-score   support

           1       0.83      0.71      0.76      2336
           3       0.76      0.68      0.72      6444
           5       0.91      0.95      0.93     19288

    accuracy                           0.87     28068
   macro avg       0.83      0.78      0.80     28068
weighted avg       0.87      0.87      0.87     28068
```

Table 8 - Classification report table for stack ensemble using logistic regression with 10 fold cv



```
[[ 1654   511   171]
 [  283  4407  1754]
 [   61   859 18368]]
```

Table 9 - Stack Ensemble confusion matrix on 10 fold cv

4. Discussion

From the result, all of the model classifiers did make improvement on top of the plain 0R. Nonetheless, each of them has different behaviour. For MNB, its performance may be affected by the prior probability, as rating 5 dominates the class label; thus, it is likely that the model was heavily trained on label 5, as there are more words that may act as its likelihood or indicator. However, this is not the case for logistic regression. This model may be superior to MNB, as it does not rely on the attribute likelihood. It would approximate the posterior directly using gradient descent or ascent. Hence, this may explain the logistic regression's f1-score superiority over MNB. Unfortunately, the approximation increases the model's complexity. Moreover, introducing another logistic regression model with the doc2vec dataset for stack ensemble worked well. The most probable reason may be that doc2vec adds the semantic values for the words that were lost in the count vector. Stack also increased the true positive of every label; henceforth, this might indicate that the meta classifier managed to correct some of the base models' mistakes. Lastly, all of the confusion matrices show a similar pattern in the result, where the models are having difficulty in distinguishing closely related labels, such as 1-3, and 3-5. This may happen as the sentiment in the word may not be clear for label 3. In other words, both positive and negative sentiment words may be mixed up; thus, this may confuse the models. However, the result also explains that progressively, the chosen model manages to reduce the error of each class.

5. Conclusion

In conclusion, the report has shown the prediction result for each model, with 87.03% max average validation accuracy achieved by the stack ensemble model of MNB and two logistic regression, using logistic regression as the meta classifier. Furthermore, the report has demonstrated that combining the result of two logistic regression trained from doc2vec and BOW made an improvement in the accuracy. For future investigation, one could consider in depth text preprocessing, such as eliminating stop words and tune doc2vec dataset by expanding its feature quantity.

6. References

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.

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.

Le, Q. & Mikolov, T. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning, 2014.

Indra, S. & Wikarsa, Liza & Turang, Rinaldo. Using logistic regression method to classify tweets into the selected topics, 2016. 385-390.

Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.R. Tackling the poor assumption of naive bayes text classifiers. Proceedings of the 20th International Conference on Machine Learning, 2003.

Ma Z, Wang P, Gao Z, Wang R, Khalighi K. Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose. 2018. PLoS ONE 13(10): e0205872. https://doi.org/10.1371/journal.pone.0205872