COMP30027 Report

1. Introduction

The aim of this project is to predict the rating of a particular restaurant as either 1, 3 or 5 given a Yelp review.

The popular "bag of words" approach has a major weakness. It simply vectorizes each word and does not take into account the context of a word in a document or the words around it.

A better approach could be the doc2vec representation, an improvement of the word2vec method with the addition of document id parameter.

In this project, we will further investigate this hypothesis on various machine learning models. The main focus for this project is tailored to understanding the performance of various models rather than finding the best performing model.

2. Model Building

2.1 Preprocessing

The provided meta and doc2vec datasets were stored as Pandas Dataframes and the "bag of words" approach was stored as a scipy sparse matrix. There were no missing values in the training set, so no imputation was required.

The distribution of class labels is as following:

Rating	Count
1	2336
3	6444
5	19288

Table 1- Distribution of class labels.

We will use the 0R model as a baseline, which has an accuracy of 68.7%.

Features such as reviewer_id, distinct_id, review_id, business_id and date will all be removed.

2.2 Feature Selection/Dimension Reduction

A very naïve model

We will use the remaining attributes (vote_useful, vote_funny, vote_cool) in meta_train to build a test model. 5-fold cross validation will be used.

Model	5-Fold Accuracy
GaussianNB	67.4%
MultinomialNB	68.9%
DecisionTree	68.6%
Logistic Regression	69.2%

Table 2- Model Accuracies for a very naïve model.

We can observe from Table 2, that the model is terrible and barely outperforms the 0R baseline. As expected, the vote features by themselves are not good enough to explain the ratings of the reviews. We will now investigate the "bag of words" and doc2vec approaches.

"Bag of Words"

Dimension reduction will be performed by selecting the k-most important words from a dictionary of words on the training set. These will be selected according to mutual information and chi-squared residual analysis methods.

The top 100 features using the chi-squared residual analysis will be a starting point. Some of the words selected are food related or adjectives relating to food such as "cakey" or "unappetizing" which could have a predictive influence. There were also numbers such as 15 or 20, but for the purpose of the assignment we will not remove them.

The mutual information residual analysis

method yielded very similar results for the top words selected. However, the computation time was much longer compared to the chi-squared method.

doc2vec

Before diving into the doc2vec embeddings, we will run test models for an understanding of its performance. The results for 100 and 200 features were about the same as 50 if not worse.

Model	5-Fold Accuracy
GaussianNB	72.5%
DecisionTree	65.5%
Logistic Regression	81.5%

Table 3- Model Accuracies for doc2vec for 50 features.

Compared to Table 4, there is not a big improvement compared to the "bag of words" approach. We will not use the doc2vec embeddings for this project.

3. Optimal Model Results

The chosen evaluation metric was a 5-fold cross validation method. As cross validation should summarise model performance fairly well, other metrics such as precision and recall will not be used.

GaussianNB

Using 55 features yielded an accuracy of 74.3%. No smoothing is required for GaussianNB (Figure 1).

MultinomialNB

Using 300 features and a laplace smoothing parameter of 1 yielded an accuracy of 81.7% which is a big improvement from GaussianNB (Figure 2).

Decision Tree

Using 8 features yielded an accuracy of 71.6% (Figure 3), which just slightly outperforms the 0R model. The selection criterion was entropy and gini returned very similar results.

Logistic Regression

Using 125 features, with C = 10 using a true multinomial logistic regression yielded an accuracy of 80.1% (Figure 3).

3.1 Error Analysis

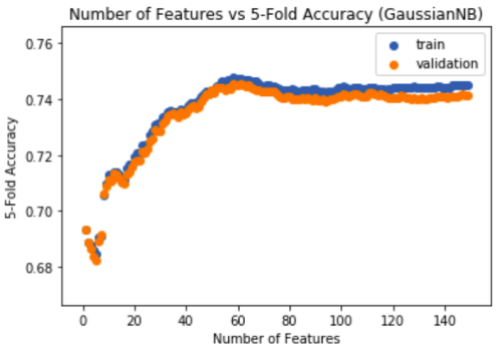

Figure 1- 5-fold prediction accuracies for GaussianNB.

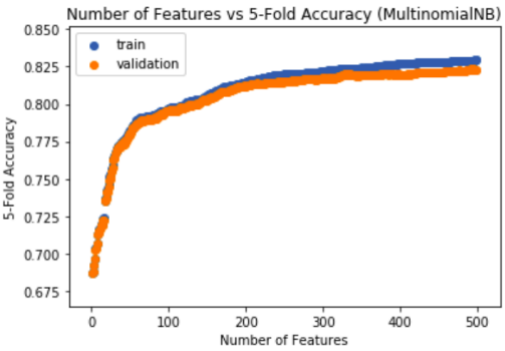

Figure 2- 5-fold prediction accuracies for MultinomialNB.

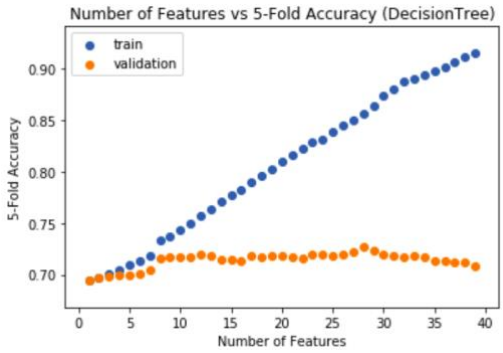

Figure 3- 5-fold prediction accuracies for Decision Tree.

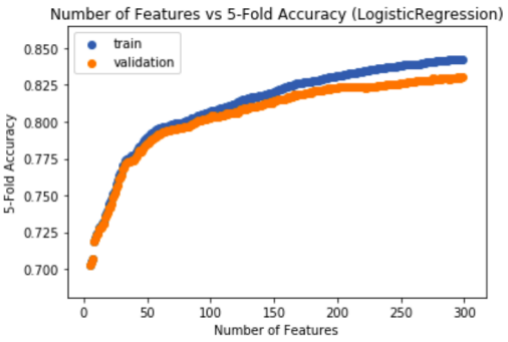

Figure 4- 5-fold prediction accuracies for Logistic Regression

4. Evaluation

GaussianNB was an average performer in comparison to the other models. One of the key reasons could be that some features are not normally distributed. Many words would not appear in every document so the density graphs would be highly right skewed (Figure 5).

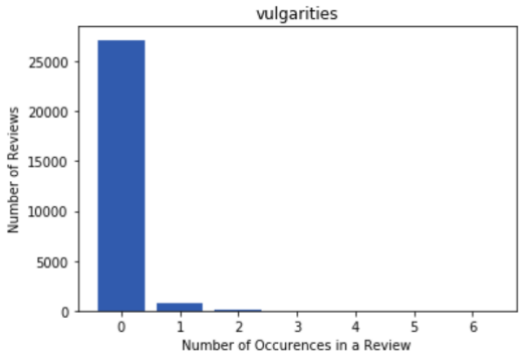

Figure 5- Bar Chart for the word "vulgarities"

Even without the '0' bar, the data is clearly still right skewed.

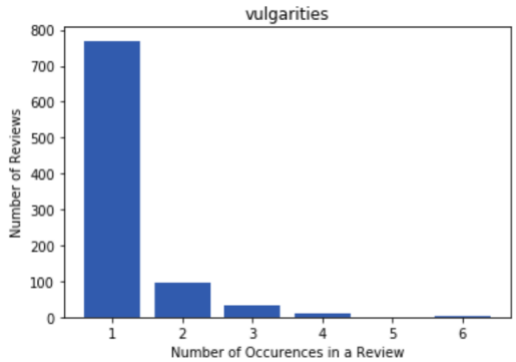

Figure 6- Bar Chart for the word "vulgarities" without the '0' bar.

In addition, the independence assumption of words given a particular rating would not be logical. Words such as "terrible" and "fantastic" when given a particular rating would not be independent of each other. The GaussianNB does not seem to be overfitted.

The MultinomialNB classifier performed fairly well, which is probably because the "Bag of Words" dataset is capturing the number of counts per word in each document, which can be modelled a multinomial distribution. Figure 2 also suggests that the improvement in accuracy starts to decrease but there does not seem to be a clear indication the model is overfitted.

In addition, the MultinomialNB is able to take in more features and still outperform the GaussianNB and decision tree models. The performance of both GaussianNB and decision tree models began to decrease after a certain number of features, whereas the MultinomialNB model still slightly improves. Similar results occurred in McCallum and Nigam's research article "A Comparison of Event Models for Naïve Bayes Text Classification" where the MultinomialNB had higher performance for larger vocabulary sizes.

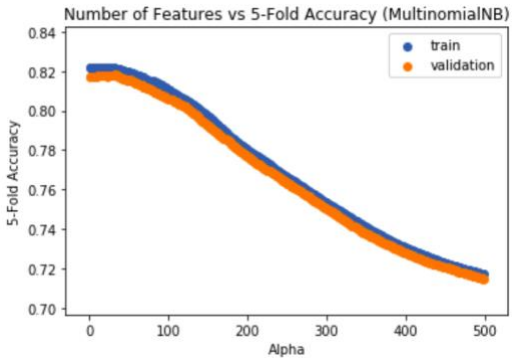

Figure 7- Graph for 5-fold accuracies for varying alpha in a MultinomialNB model with 300 features.

The other hyperparameter to tune for MultinomialNB is the smoothing

parameter alpha. Clearly, using the default alpha of 1 will yield a strong performance. The performance of the model decreases as alpha increases. This occurs as the denominator value of probability the smoothed parameter becomes very large so the overall probability will become very small. Eventually the model would predict every instance as the majority class, which is the same behaviour as 0R.

The decision tree classifier was the worst performing classifier from the four classifiers investigated. One of the trees used in 5-fold cross validation was visualised with max depth at 4 for simplicity (Image was too large to fit in the report). Intuitively, the decision tree makes decisions at each branch by deciding if a word in a review meets a certain threshold. This is repeated until a leaf node is reached. One of the reasons for its poor performance is the possibility of a poor selection of words from the chi-squared residual analysis. Some words used in the branches includes "romans", "preceded" and "oc" which do not make much sense in deciding the rating of a review. In addition, the decision tree cannot handle a large number of features as it quickly overfits (Figure 3). This suggests that the tree is unstable and additional features provided leads to large changes in the structure of the optimal tree.

The final model investigated was the logistic regression model. To tackle the multi-class classification problem, a true multinomial logistic regression model was implemented for better performance rather than the usual "one-vs-rest".

The first step was to tune the C hyperparameter. Higher C values focus on minimising the training errors and thus allows less regularisation.

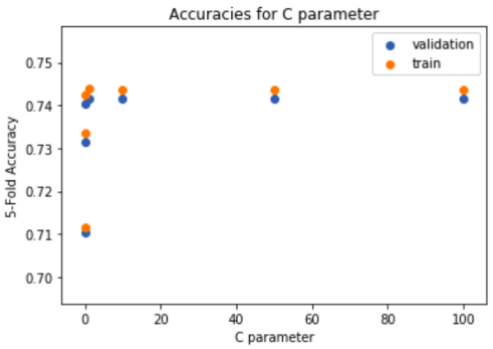

Figure 8- Graph for 5-fold accuracies for varying C parameters for a logistic regression model with 300 features.

Figure 8 displays the accuracies for varying C values. The results were very similar for other logistic regression models with a different number of features. The model performs worse with a smaller C value. This means that less regularisation is required, which suggests that it linearly separates review ratings efficiently without the need of allowing training errors.

Logistic Regression performed very similar to MultinomialNB in the sense that the prediction accuracy improves as the number of features increase. However, it tends to overfit a lot more when the features increase past 150. The computational time for logistic regression was also worse as the number of features increased.

Model	5-Fold Accuracy
GaussianNB	74.3%
MultinomialNB	81.7%
DecisionTree	71.6%
Logistic Regression	80.1%

Table 4- Model Performance using "Bag of Words".

5. Conclusion

Overall, the MultinomialNB was the best performing model as it was computationally efficient and could take in a large number of features without overfitting the model.

Model performance can be greatly improved by stricter data cleaning, further investigation of other text representations, more complex models or some form of stacking, boosting or ensemble learning.

In conclusion, we have achieved the purpose of this project, which was to understand the performance of various models using the "Bag of Words" representation.

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.

Shperber, G. (2017). *A gentle introduction to Doc2Vec*. [online] https://medium.com/wisio/a-gentle-introduction-to-doc2vec-db3e8c0cce5e. Available at: https://www.google.com/search?q=doc2vec&oq=&sourceid=chrome&ie=UTF-8

McCallum, A., Nigam, K. A Comparison of event models for Naïve Bayes text classification, 1998.