COMP30027 Report

Anonymous

1. Introduction

Sentiment analysis is the 'interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques.' Sentiment Analysis is key in helping business have a better understanding of their customers' reviews. As time progresses, Sentiment analysis is becoming more and more useful, due to the many ways customers 'are able to express their thoughts and feelings', more openly than ever before thus other consumers are also able to 'make decisions on what to purchase'(Rayana and Akoglu, 2015). The ability sentiment analysis has to automatically analyse thousands of reviews will help businesses save a lot of time. However, clearly understanding the intention behind words and text, even for human beings, can be at times, difficult. Thus although sentiment analysis has been well-studied, the best sentiment analysis model remains yet undiscovered.

The main focus of this report is to understand some of the most popular sentiment analysis models, such as Naïve Bayes, Support Vector machines or DeepLearning the different ways it can be optimised and applying them to an open-ended problem (Pang, Lee and Vaithyanathan, 2016). The open-ended problem requires predicting star ratings (1, 3, or 5) for reviews on restaurants based on previous real reviews made on Yelp, courtesy of the data from (Mukherjee et al., 2013)

2. Data

The data acquired is from Yelp. This website connects restaurants together with reviews.

Data is manipulated through generating embeddings. Converting each of the words into a word embedding. Word embeddings, being the vector representations of their words. This will help capture any underlying relation between itself and other words in the sentence. For instance, words with similar meanings will be clustered closer together in the hyperplane compared rare accounts being positioned further away.

2.1 Count Vectorizer

One popular encoding method is the Count Vectorizer, which converts the text documents into a 'Bag of Words'. The information that it retains of the word is only its occurrences while ignoring its position information (Raoi and Devi, 2019).

2.2 TF-IDF Vectorizer

Term Frequency-inverse document frequency is a way to measure how relevant a word is to the review in the entirety of reviews. This, similar to count vectorizers, looks at the frequency of these words occurring however words common in every review, such as I, me, this, if, will rank low as they don't pertain to the reviews in particular.

3. Splitting

The entire data provided has already been split into training and testing. The testing data provided do not contain the response variable needed to cross-validate the models and as such, the training data will be splitted instead and used to cross-validate any models.

To avoid any overfitting of the data, the data will be split into 3, training, validation and testing. The training dataset will be used to fit the model, The validation dataset will be used to evaluate any given models and help fine-tune any model parameters without any chance of bias that may occur if the test dataset was evaluated instead. The test dataset will provide the final evaluation of the model once the model has been finetuned. Validation dataset was included over just splitting to training and testing as if you adjust your model based on your testing data, your evaluation of your model will be biased towards the testing data.

4. Feature Selection

Feature selection, also called variable or attribute selection is the automatic selection of features in your data that are most relevant to the model.

Feature selection is useful, even more for Sentiment analysis, as it acts mainly as a filter,

removing words that don't even appear in reviews or words that are too common with no relevance to any emotional connection, words such as; the, a, and… etc.

Chi squared filtering method assigns a statistical score to each feature and then ranked. To decide how many important features to keep, we tune it depending on the highest accuracy each different model produces with varying number of kept features. Plotting the graph which shows accuracy when fitting the training and validation data.

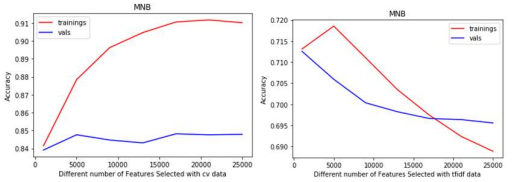

Figure 1: Multinomial Naïve Bayes accuracies on training vs validation data with varying amounts of Feature's selected.

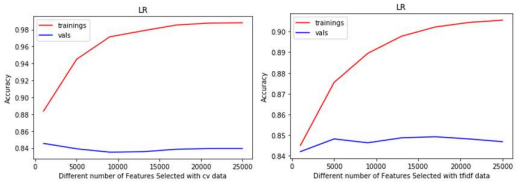

Figure 2: Logistic Regression accuracies on training vs validation data with varying amounts of Feature's selected.

The number of features that result in the highest accuracy. This is tested twice for each model, once on the data word embedded by Count Vectorizer the other, by TF-IDF Vectorizer.

Best Number of Feature Selections	Multinomial Naïve-Bayes	Logistic Regression
Count Vectorizer	5000	21000
TF-IDF Vectorizer	1000	17000

Figure 3: Table of best number of feature selections.

5. Models

The Models selected are Multinomial Naïve Bayes and Logistic Regression. These were chosen compared to the arguably better neural network models (Kolokas, Drosou and Tzovaras, 2019), due to its easier implementation and smoother runtime.

5.1 Multinomial Naïve Bayes

The multinomial Naïve Bayes estimates the conditional probability of a particular word given a class as the relative frequency of term t in reviews belonging to the same class. The variation takes into account the number of occurrences of the word in the training set, including multiple occurrences(Xu and Shuo, 2017).

This frequency-based probability may introduce zeros, if the word was not expected when calculating probabilities. Thus Laplace smoothing is adopted to counter such a problem. Laplace smoothing, alpha, is the smoothing parameter in where we add a number, alpha to every frequency whenever the model comes across a unique word when evaluating. This hyperparameter can also be finetuned to find the optimal alpha for Laplace smoothing.

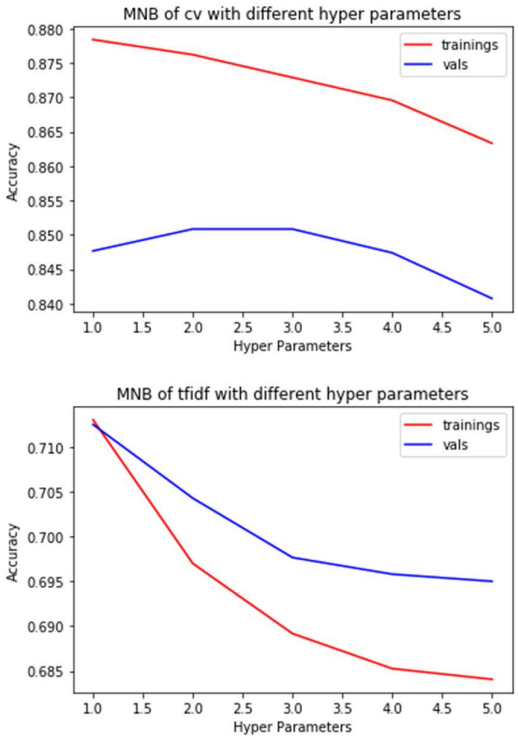

Figure 4: Learning Curves of Multinomial Naïve Bayes Accuracies on training vs validation with varying size of Laplace Smoothing on the dataset embedded by Count Vectorizer vs TF-IDF Vectorizer.

The figure above reveals higher overall accuracies when using count vectorizer on the text reviews, with the highest validation accuracy ('85%) occurring when Laplace smoothing variable is 3 (figure 4.).

5.2 Multinomial Logistic Regression

Logistic regression is the regression analysis usually conducted when the dependant variable is binary which is useful to explain a

relationship between one dependant variable with one or more ordinal independent variables (Sokal, 1995). In this case, we are given more than two predictor variables to analyse, them being the score ratings of 1, 3, 5 thus mores specifically, Multinomial regression is analysed. The penalty parameter for the model is either l2 or none. When validated, l2 shows a higher validation accuracy percentage with a lower training accuracy.

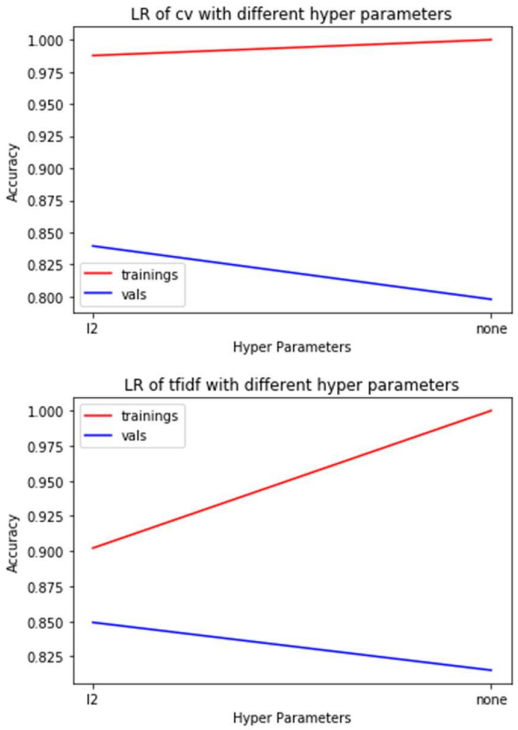

Figure 5: Learning Curves of Logistic Regression Accuracies on training vs validation with either a l2 or no penalty on the dataset embedded by Count Vectorizer vs TF-IDF Vectorizer.

The figure above reveals barely higher accuracy when using tf-idf on the text reviews, with the highest validation accuracy (‘84%) occurring when penalty is l2 (figure 4.).

6. Combining best parameters

The parameters tuned for the Multinomial Naïve Bayes.

Parameters	Optimal
Word Embedding	Count Vectorizer
Number of Features Selected	5000
Laplace Smoothing	3
Validation Accuracy	85%
Training Accuracy	84%

The parameters tuned for the Logistic Regression.

Parameters	Optimal
Word Embedding	TD-IDF Vectorizer
Number of Features Selected	17000
Penalty	L2
Validation Accuracy	84%
Training Accuracy	84.5%

When applying these two models on the testing set, the training accuracy of the LR is greater than MNB’s. Though very minimal, we vote our Logistic Regression the model to use.

7. Conclusions

In this work, we briefly tuned two models used in sentiment analysis as to better understand the importance of tuning. Whether it was to help with common issues such as overfitting or underfitting.

The main parameters tuned were the:

- Word Embedding Method
- Feature Selections (Chi-Squared)

And the hyper-parameters for their respective models:

- Laplace Smoothing Variable (for MNB)
- Penalty Method(for LR)

We found that the largest impact in tuning from order from largest to smallest change in accuracy could be:

Hyper-Paramater >> Word Embedding Method >> Feature selection

References

Pang, B., Lee, L. and Vaithyanathan, S., 2016. Sentiment Classification using Machine Learning Techniques. International Journal of Science and Research (IJSR), 5(4), pp.819-821.

Zebin, Y. and Aijun, Z., 2020. Hyperparameter tuning methods in automated machine learning. SCIENTIA SINICA Mathematica, 50(5), p.695.

Raoi, K. and Devi, M., 2019. Disquisition of Sentiment Inquiry with Hashing and Counting Vectorizer using Machine Learning Classification. International Journal of Innovative Technology and Exploring Engineering, 9(1), pp.737-743.

Sokal, R., 1995. LogXact: Logistic Regression Software Featuring Exact

Methods.LogXact-Turbo: Logistic Regression Software Featuring Exact Methods. The Quarterly Review of Biology, 70(1), pp.127-127.

Xu, Shuo & Li, Yan & Zheng, Wang. (2017). Bayesian Multinomial Naïve Bayes Classifier to Text Classification. 347-352. 10.1007/978-981-10-5041-1_57.

Kolokas, N., Drosou, A. and Tzovaras, D., 2019. Text synthesis from keywords: a comparison of recurrent-neural-network-based architectures and hybrid approaches. Neural Computing and Applications, 32(9), pp.4259-4274.

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.