Prediction of Rating Review
Based on the Review Text Content

Boonrit Boonmarueng
Computer Engineering
King Mongkut's University of Technology Thonburi
boonrit.b@mail.kmutt.ac.th

Abstract

The reviews of products and services on the online platform have been increasing. As a result, a large number of reviews makes it difficult for the businesses to automatically classify them into different semantic orientations. However, text analysis is essential in nowadays to help reflect the opinions and needs of customers to the company. In this paper, we demonstrate the 5-classes text classification of rating review based on the review text content from the first Starbucks review dataset. We construct and compare the performance of the models among the traditional machine learning model including Multi Linear Regression (MLR), Naive Bayes, Random Forest (RF), XGBoost and K-Nearest Neighbors, and Bidirectional Recurrent Neural Networks (BRNN) based with combination of Long Short-Term Memory (LSTM) and Gate Recurrent Units (GRUs) schemes. The result shows that RF can achieve the highest accuracy above 95% by using only 1 ngram. For RNN based, 3 consecutive layers of GRUs provide the best result at 95%. Finally, we find out that by using some words from customer's review can lead to distinguish the level of rating on this dataset.

Keywords: Bidirectional Recurrent Neural Network, Text Analysis, Text Classification, Traditional Machine Learning

1. INTRODUCTION

E-commerce is developing fast, as a result of product or service reviews have grown immediately on the online website. The review can be an article review, a blog, comment, or a rating. If customers write or give a positive review, it will create reliability for the product and service of that company. On the other hand, if customers write a negative review, it will cause problems, create a bad reflection on the company that will result in a long-term period. Nowadays, many platforms allow customers to write reviews and give ratings to businesses such as Yelp in the USA, Pantip, Wongni in Thailand, etc.

A large number of reviews makes it difficult for businesses to automatically classify them into different semantic orientations such as positive, negative, neutral or rating the customer's reviews. Moreover, the textual review data is highly complicated. It is necessary to rely on an expert who has a high understanding and experience in text analysis. However, many companies still value online customer reviews because it helps to reflect customer feedback come back to the business to improve their defect.

The classification of the customer reviews that have been popular for a long time. There are many research articles on this filed. Grabner et al.[1] focused on the customer review about hotels using sentiment analysis and reported that the classification is correct with a probability of about 90% when the system has 2 classes. While Shaziya and Humera[2] analyzed the movie's review in a dataset that obtained from Cornell University. For an algorithm used in this field, Naive Bayes Classification and Support Vector Machines are experimented in these studies [2,3], K-mean clustering [4]. Moreover, Nadali et al.[5] applied a fuzzy logic model to perform semantic classifications of customers review.

There are different types of text classification: content based, sentence based and aspect based. Additionally, the classification can be supervised or unsupervised learning. Since the data that used in this paper has labels with the level of rating star, as a result of we mainly focused on the text content based with supervised learning.

The rest of this paper is distinguished into 8 main parts. Section 2 presents the basic idea of word representation and reviews related text classification models. Section 3 demonstrates the detail of dataset. Section 4 shows the preparing data process. Section 5 descries the pipeline structure and parameter setting. While Section 6 descries the designed structure of the recurrent neural network. Section 7 reports the results from each classification model. Finally, section 8 concludes the paper.

2. BACKGROUND

To understand the context of this paper, this section provides related literature of word representation techniques and text classification models.

2.1. Word Representation

Word Representation is the technique that used to represent the words in matrix form, which is easy to apply as input of the classification model. There are numerous methods to represent the words, but this paper use 2 techniques including TF-IDF and Global Vectors.

2.1.1. Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF consists of 2 terms: First, Term Frequency (TF) term reports how often that a word appears in all document. Secondly, the Inverse Document Frequency (IDF) used to descale the important of a word when it appears many times in all document. Finally,when we combine these two terms, it results in TF-IDF score or weight for a word overall document.

2.1.2. Global Vectors (GloVe)

The Global Vector (GloVe) is an unsupervised learning algorithm for obtaining vector representations for words. The process to transform a word to GloVe representation is similar to mapping the key and value in the dictionary by searching a word as key to get the matrix of feature as value.

2.2. Traditional Machine Learning

This topic provides the detail of 5 supervised traditional machine learning models that used in this paper consisting of Multi-Linear Regression, Multi-Class Naive Bayes, Random Forest, XGBoost and K-Nearest Neighbors.

2.2.1. Multi-Linear Regression (MLR)

This model is a classification method that performs logistic regression to multiclass/more than 2 classed by using the Softmax function to classify.

2.2.2. Multi-Class Naive Bayes

Multinomial Naive Bayes classifier is a unique instance of Naive Bayes classifier which uses multi-channel distribution for each feature.

2.2.3. Random Forest (RF)

Random forest, one type of classification, consists of a large number of individual decision trees that operate as an ensemble. The reason that why the RF model can provide higher accuracy because each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. While some decision trees may be predict incorrect class, many other decision trees will be correct, that make a group the decision trees are able to converge in the correct prediction.

2.2.4. XGBoost

XGBoost is one of the classification that also used ensemble and gradient boosted technique. Nowadays, this model has been used extensively in the competition because of its processing speed and excellent performance. There are 3 forms of gradient boosting that can be used including: 1).Gradient Boosting, 2).Stochastic Gradient Boosting, and 3). Regularized Gradient Boosting with both L1 and L2 regularization.

2.2.5. K-Nearest Neighbors (KNN)

KNN is one of the supervised learning algorithms that uses the concept of distance calculations from unknown instances to other K known instances.

2.3. Bidirectional Recurrent Neural Networks (BRNN)

BRNN is a one type of the recurrent neural networks (RNN). In the BRNN's structure, It connects adjacent layers in two directions, consists of feeding forward and backward to create relationship patterns or word patterns in both directions. In this paper, we implement 2 types of layer: 1). Long Short-Term Memory (LSTMs). and 2). Gate Recurrent Units (GRUs). In fact, the GRUs are simpler, train faster and perform better than LSTMs on less training data. While LSTMs outperform in tasks requiring modeling long-distance relations.

2.4. Model Evaluation

In this paper, we mainly focused on the text classification based on text content with supervised learning. To evaluating the effectiveness of each model, the same training data were used to train the model,and same testing data were used every time to reduce the bias that can occur. The performance metric used to measure model's performance in this paper is the accuracy, which is the percentage that the predicted class matches the actual class.

3. DESCRIPTION OF DATASET

The review data that used in this paper were obtained from the Yale website during the years 2006-2020 by using the ParseHub software, which is an open-source. The amount of data is 2,759 rows with attributes in the data consist of the reviewer's name, date, rating points in the range of 1-5 stars and text review. For classes distribution as shown in figure 1, the most of distribution at 5 stars (44.95%), 3 stars (22.17%), 4 stars (13.81%), 2 stars (22.17%) and 1 star (8.73%).

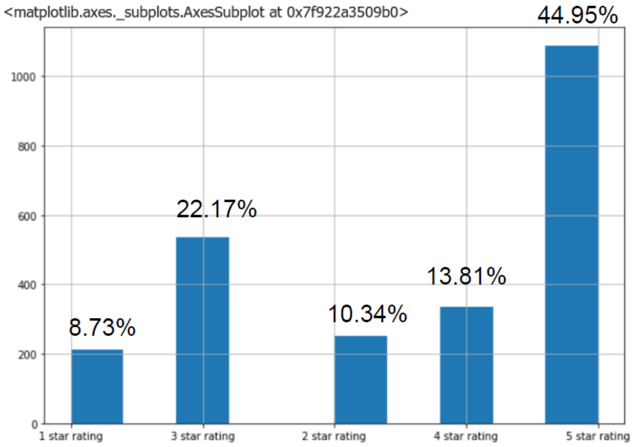


Figure 1: The classes distribution.

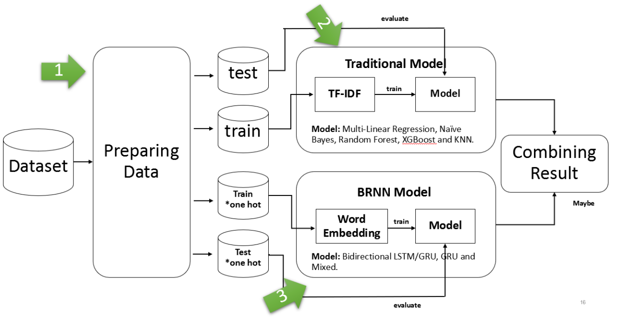


Figure 2: The overview of system.

4. PREPARING DATA

This section is the first part of Figure2 that shows the overview of our system. From the initial inspection, we found out that some reviews that does not have a review text written, but they still gave the rating. To solve this issue, we filtered out these type of reviews that make the number of rows reduced to 2,418 rows.

To avoid the bias from random splitting data, we always set the random state at 38 and 99 with 50% training set and 50% testing set.

For BRNN models that need the numeric input, we transformed that 5 rating from the category type into 5-classes Binary with the one hot encoding. Then, mapping all words with GloVe.6B100d corpus to create the word representation.

5. TRADITIONAL MACHINE LEARNING

This section show the detail of the second part of the system. In this paper, the sklearn library from Python was chosen to processing the raw text data and creating the model. We constructed the sklearn pipeline consists of 2 parts: first, creating the TFIDF matrix, which has ngrame parameters and basic text processing function such as cut the word, lowercase characters and removing the special characters. Second part, model classifier. In the experiment, we will change the ngrame parameter from range 1-3 ngrame and the type of the classification model as previously mentioned (not include KNN). Finally, the number of neighbor in KNN model was set at k=9.

6. BIDIRECTIONAL RECURRENT NEURAL NETWORK

The third part of the system, we chose the BRNN because using the concept of bidirectional could create the relationship from run inputs in 2 ways, both one from past word to future word and one from future word to past word. It could result in capturing the relationship of each word and give high chance to get the best result. We designed the structure of BRNN as the figure 3, by feeding the inputs with dimension 2300x100 Through 3 consecutive layers of LSTM/GRU and using SoftMax activation function to classify into 5 class from 1-5 stars. In Figure 4 showing examples of the structure and number of parameters used in BRNN learning. There are 2 pooling layers used to reduce the number of nodes before entering the Fully Connected layer. Finally, the result is a 5-bit binary.

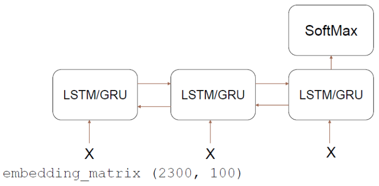
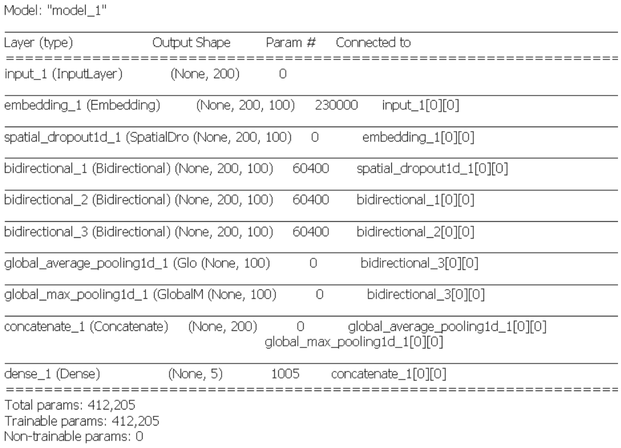


Figure 3: The overview of BRNN model.

7. RESULTS

In our experiment,we spitted dataset 2 times with random state (rc) at 38 and 99 to training and testing the model. The results for each text classification model show as the table 1. In the traditional machine model aspect, we found out that adding the number of ngram does not affect the increase in accuracy in all models.The RF has an average accuracy above 96%, which is more than other models. However, the model with the least accuracy is KNN. For in BRNN aspect, the 3 consecutive layers of GRUs achieve the average accuracy at 95%, while another BRNN based models give only 94.5%.



```
Model: "model_1"
________________________________________________________________________________
Layer (type)                   Output Shape        Param #     Connected to
================================================================================
input_1 (InputLayer)           (None, 200)         0
________________________________________________________________________________
embedding_1 (Embedding)        (None, 200, 100)    230000      input_1[0][0]
________________________________________________________________________________
spatial_dropout1d_1 (SpatialDro (None, 200, 100)   0           embedding_1[0][0]
________________________________________________________________________________
bidirectional_1 (Bidirectional) (None, 200, 100)   60400       spatial_dropout1d_1[0][0]
________________________________________________________________________________
bidirectional_2 (Bidirectional) (None, 200, 100)   60400       bidirectional_1[0][0]
________________________________________________________________________________
bidirectional_3 (Bidirectional) (None, 200, 100)   60400       bidirectional_2[0][0]
________________________________________________________________________________
global_average_pooling1d_1 (Glo (None, 100)        0           bidirectional_3[0][0]
________________________________________________________________________________
global_max_pooling1d_1 (GlobalM (None, 100)        0           bidirectional_3[0][0]
________________________________________________________________________________
concatenate_1 (Concatenate)    (None, 200)         0           global_average_pooling1d_1[0][0]
                                                                global_max_pooling1d_1[0][0]
________________________________________________________________________________
dense_1 (Dense)                (None, 5)           1005        concatenate_1[0][0]
================================================================================
Total params: 412,205
Trainable params: 412,205
Non-trainable params: 0
```

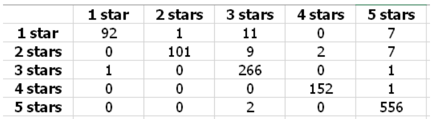
Figure 4: The architecture of BRNN model.

Model	Acc.(%) rs=38	Acc.(%) rs=99
MLR+TFIDF+(1,1)ngrame	96.00	96.00
MLR+TFIDF+(1,2)ngrame	96.00	96.00
MLR+TFIDF+(1,3)ngrame	96.00	96.00
NB+TFIDF+(1,1)ngrame	94.00	96.00
NB+TFIDF+(1,2)ngrame	94.00	96.00
NB+TFIDF+(1,3)ngrame	94.00	96.00
RF+TFIDF+(1,1)ngrame ntree=99	96.00	97.00
RF+TFIDF+(1,2)ngrame ntree=99	96.00	96.00
RF+TFIDF+(1,3)ngrame ntree=99	96.00	96.00
XGBoost+TFIDF+(1,1)ngrame	96.00	96.00
XGBoost+TFIDF+(1,2)ngrame	96.00	96.00
XGBoost+TFIDF+(1,3)ngrame	96.00	96.00
KNN K=9	93.71	94.46
BRNN 3LSTM	94.00	94.00
BRNN 3GRU	95.00	96.00
BRNN LSTM-GRU-LSTM	95.00	94.00

Table 1: The result of text classification models.

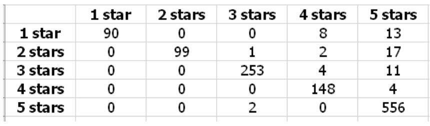
Next, we consider the false predictions or errors that occur in the RF and 3 consecutive layers GRUs of BRNN model. In the picture 5, which is the confusion matrix of the RF model, it shows that most errors occur at class 1 and 2 stars, only slightly found at class 3 and 4 stars. As in Figure 6, which is a confusion matrix of the 3 consecutive layers GRUs of BRNN model, it demonstrates that most errors are found at Class 1-3 stars. Moreover If considering the training time, the result shown that the RF model consumes only 3 minutes, while GRUs takes approximately 15 minutes.

However, the biggest problem encountered with BRNN



	1 star	2 stars	3 stars	4 stars	5 stars
1 star	92	1	11	0	7
2 stars	0	101	9	2	7
3 stars	1	0	266	0	1
4 stars	0	0	0	152	1
5 stars	0	0	2	0	556

Figure 5: The confusion matrix of RF model.



	1 star	2 stars	3 stars	4 stars	5 stars
1 star	90	0	0	8	13
2 stars	0	99	1	2	17
3 stars	0	0	253	4	11
4 stars	0	0	0	148	4
5 stars	0	0	2	0	556

Figure 6: The confusion matrix of 3 consecutive layers GRU of BRNN model.

is the overfitting. We found that the accuracy of training and validation is usually higher than 99%,while the accuracy of the test is only 95%. Also found that at the random state equal to 38 in training and testing of 3 consecutive layers GRUs, revealed that the accuracy of validation is higher than both training and testing which could imply that the dataset may not be divided appropriately.

To understand the context of the RF model why it gives relatively high accuracy and low dispersion errors. One important factor is using the ensemble of multiple decision trees (n=99). Nevertheless, it is difficult to show the important features that produced by every decision tree. We only extracted the important feature from the first tree as table 2 to show the top 10 words, which have high score in TF-IDF matrix.

Word	TF-IDF score
place	0.0178
line	0.0161
even	0.0135
starbucks	0.0109
long	0.0101
around	0.0097
every	0.0094
see	0.0089
experience	0.0085
coffee	0.0084

Table 2: The top 10 words with high TF-IDF score.

Lastly, to dig deeper, we created simple visualization to see customer's feedback, which is the words could for each class (1-5 stars). In figure 7 and 8, which are the words cloud from the reviews of customer in class of 2 and 5 stars, show that the most problems from is the long queue and long waiting time that customers have to experience in this coffee shop.

8. CONCLUSION AND FUTURE WORK

this paper shows the text classification models to predict the rating from customer's review of first Starbucks between the traditional machine learning models and BRNN based. It is evident that the combination of TF-IDF technique and random forest model can overcome other models

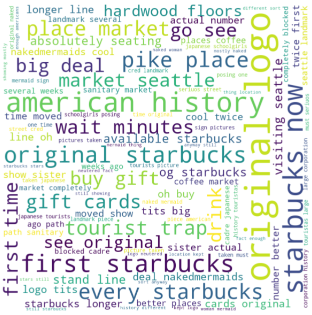


Figure 7: The words cloud from class 2 stars.

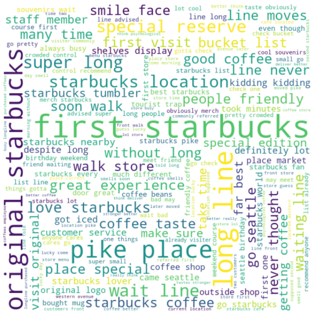


Figure 8: The words cloud from class 5 stars.

with using 1 ngrame. Additionally, RF model gives the average accuracy more than 96% with small errors that occur at class 1 and 2 stars, only slightly found at class 3 and 4 stars.

For future work, the error analysis is an important point to pay attention to, by studying what are the factors that cause the error. Moreover, do the experiment to find out the best parameters for each model. Finally, gathering results from each model to decide the final result can help improve system performance.

9. REFERENCE

[1] Gräbner D., Zanker M., Fliedl G., Fuchs M. (2012) Classification of Customer Reviews based on Sentiment Analysis. In: Fuchs M., Ricci F., Cantoni L. (eds) Information and Communication Technologies in Tourism 2012. Springer, Vienna

[2] Shaziya, Humera. (2018). Text Categorization of Movie Reviews for Sentiment Analysis. 4. 11255-11262. 10.15680/IJIRSET.2015.0411065.

[3] Channapragada., Shivaswamy., (2015)., Prediction of rating based on review text of Yelp reviews, Sementic-Scholar.

[4] Sulthana, A. Razia, and Ramasamy Subburaj. "An improvised ontology based K-means clustering approach for classification of customer reviews." Indian Journal of Science and Technology 9.15 (2016): 1-6.

[5] S. Nadali, M. A. A. Murad and R. A. Kadir, "Sentiment classification of customer reviews based on fuzzy logic," 2010 International Symposium on Information Technology, Kuala Lumpur, 2010, pp. 1037-1044.