Project 1 Report: Sentiment Analysis

Group 22

CAI, Shizhan CHEN, Zixin LIANG, Haoran ZHANG, Weiwen

Abstract

Abiding by the pipeline, we utilized several data pre-processing methods such as tokenize and n-grams, together with some data analysis approaches (e.g., Pearson correlation, ANOVA, etc.) to handle and analyze our data. After comparing different models (e.g., Logistic regression, LSTM with Word-2-Vec, CNN and BERT) and applying hyperparameter training, we finalize the best model based on the F1 score statistic. The results show that the BERT model has the best F1 score (0.672019) as well as the best accuracy (67.35%).

1. Introduction

Our task is to work on the text segmentation with the data of the reviews of restaurants. It is originally a classification task in machine learning fields. We firstly explored the dataset and then processed the data. During the training process, we implemented algorithms include but not limited to KNN, SVM, Logistic Regression, Decision Tree, Random Forest, Adaboost, GBDT, Xgboost, CNN, RNN, LSTM with Word2Vector and BERT (transfer learning), etc. Finally, the BERT outperformed other pipelines and surpassed the strong baseline on the validation dataset.

2. Data Exploration

Our dataset contains the information of the review of restaurants, 10,000 for training and 2,000 for validation. Information includes text, stars, business_id, cool, date, funny, review_id, useful, and user_id. In this segmentation task, the text is the main feature that we should focus on, and stars are the labels.

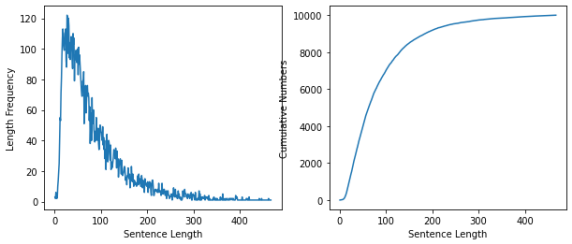

Figure 1: Length Distribution

Bottom left is the distribution of the length of the sentences. Except for BERT, all other models we tried were trained with max_length about 150~200.

On the analysis part, we first made the correlation coefficient analysis between the numerical columns. It shows that cool-funny has the strongest collinearity. Then we made a summary of the mean and variance of the numerical data. After that, we have made a one-way ANOVA test for the numerical columns of this dataset. The p-value is shown in the chart below. In the result, we only cannot reject the null hypothesis that variable cool has no influence on funny. Then, we made a calculation of the Wasserstein distance between those numerical variables. The results show that cool-funny has the shortest Wasserstein distance of 0.036. Whereas, during the training process, we found all these features except 'text' do not help in improving the accuracy. Thus, we kept the feature 'text' as the only feature for further processes.

ANOVA p	cool	funny	stars	useful
cool	-	0.41703220	0	1.11E-105
funny	-	-	0	8.84E-103
stars	-	-	-	1.53E-281
useful	-	-	-	-

wassers	cool	funny	stars	useful
cool	-	0.036	2.6576	0.9477
funny	-	-	2.6274	0.9273
stars	-	-	-	2.2009
useful	-	-	-	-

Correlation	cool	funny	stars	useful
cool	1	0.73580349	0.087558	0.5424997
funny	-	1	-0.061346	0.5077457
stars	-	-	1	-0.085296
useful	-	-	-	1

Describe	minmax	mean	variance	skewness	kurtosis
cool	(0, 53)	0.5184	3.2957910	11.88188	230.16302
funny	(0, 42)	0.5388	3.0221963	9.714569	155.40779
stars	(1, 5)	2.9916	2.2057500	-0.06253	-1.4015607
useful	(0, 247)	1.4661	15.319982	28.05905	1583.92217

Table 1: Charts of Analysis

3. Our Approach

We have tried many machine learning and deep learning approaches to satisfy this 5-class classification task. We firstly tried classical machine learning algorithms like KNN, SVC (SVM Classification) and Logistical Regression. Then we tried the tree method like Decision Tree, Random Forest, Adaboost, GBDT, Xgboost. But these methods didn't perform well. Then we tried deep learning methods like CNN, RNN, LSTM and the results were much better. Finally, we proposed to use BERT, a transfer learning approach and the result was the most satisfactory. (Table 2)

3.1. BERT Approach

BERT, Bidirectional Encoder Representations from Transformers, known as BERT, is a pretrained model that can perform state-of-art results on many natural language processing tasks.[1] BERT is built based on the

Transformer model proposed by the work of Vaswani et al. [2] The architecture of the transformer is shown below. (Figure 3)

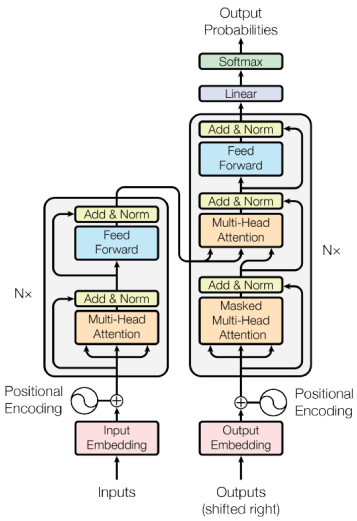

Figure 3: Transformer

In our work, we fine tune the pre-trained BERT base model, "bert-base-uncased", to realize transfer learning in our task. Our code is referred to the work of Li. [3]

Firstly, we use the BERT tokenizer, together with the pre-trained model, to process our data frame which contains the training data, validation data and test data. And then encode them to get prepared for the classification. Note that our task is to classify texts into 5 classes; however, the BERT pre-trained model should be fed with label 0~4. So, we need a preprocessor and a postprocessor to convert the data between 0~4 and 1~5. As for optimizers, we chose the AdamW. Finally, the training procedure commences, f1 score and accuracy are the performance metrics. After hyperparameters tuning (e.g., learning rate, batch size, epochs, etc.), we figured out that it outperforms other models that we have tried. The final prediction for testing set is based on one of the best models (Figure 6) we trained.

4. Result

For Logistic Regression model, we found the best model is to fit the full training data set and change the parameter 'multi_class' to 'multinomial' since our aim is predicting a multi-classes label. The accuracy on the validation set is 0.6025 for this model.

Here in the Figure 4--6, we demonstrate some results of CNN and BERT. As all other models did not perform well, we just ignore them in this report. (Results are reserved in 'Supplementary' folder) We simply keep the accuracy of each model on validation set in Table 2.

We can conclude that the BERT model has the best performance for this sentiment analysis task. The model structure, the optimizer as well as the hyperparameter choices have significant effect on improving accuracy. Further works may focus on changing the pre-trained model like "bert-large-uncased" for further improvement.

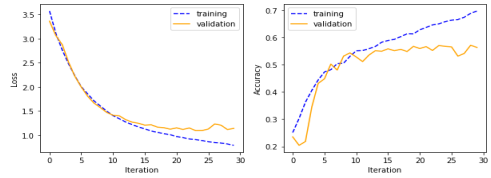

Figure 4: Result of CNN:
Best validation accuracy:0.612 0.6119999885559082

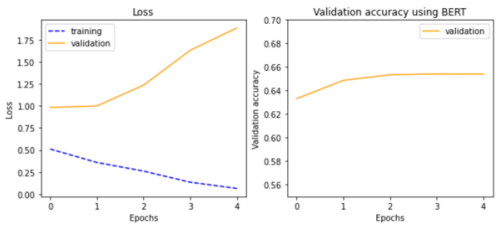

Figure 5: Result for one of the BERT models
(P.S. Not the best result, we found 2 epochs is the best by further exploring)

```
Epoch 2
Training loss: 0.6780793862342834
Validation loss: 0.7726137273311615
F1 Score (Weighted): 0.6720187888023456
```

k/2000
0.6735

Figure 6: Best model

Model	Val Acc	Model	Val Acc
SVC	25.85%	RF	52.40%
KNN	38.65%	Xgboost	55.60%
Decision Tree	39.45%	CNN	61.20%
Adaboost	50.75%	LSTM	55.68%
GBDT	55.10%	BERT	67.35%

Table 2: Accuracy on Validation Set

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv.org*, 2018. https://arxiv.org/abs/1810.04805.

[2] A. Vaswani et al., "Attention is All you Need," Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017, [Online]. Available: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[3] S. Li, "Multi Class Text Classification With Deep Learning Using BERT," Medium, Aug.02,2020. (accessed Apr. 06, 2021). https://towardsdatascience.com/multi-class-text-classification-with-deep-learning-using-bert-b59ca2f5c613