BT5153 Applied Machine Learning For Business Analytics 2023: Kaggle Project: Sentiment Classification on Movie Reviews

Muthukumaran Samiayyan (A0027474A)

Background

As part of the Kaggle Competition, a Movie Reviews Dataset was used to build a Machine Learning model to predict the sentiments of movie reviews. It is a binary classification problem (positive or negative sentiment). Multiple models were trained on a training dataset and accessed to derive the best model. The best model was used to predict on a test dataset to be submitted in Kaggle for assessment. The following sections detail the Data preprocessing, Feature Engineering and model validation performed on the multiple models before concluding on the final model to be used for submission in the Kaggle competition.

Dataset

The training dataset has 24995 records with equal number of positive and negative sentiments (balanced dataset). The Test dataset has 25865 records with no labels on which we need to use our model to predict the labels. The maximum length of the length of the reviews is ~1000 words and the mean length is ~400 words.

Data Preprocessing

For data preprocessing, TextHero package was used to preprocess the text to remove URLS, HTML tags, digits, punctuations, whitespace, mentions (@), hashtags and emails in addition lowercasing the text.

Feature Engineering

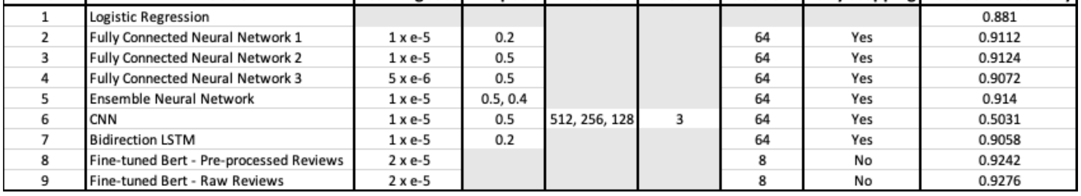
The Roberta-Large model is used to extract features with dimensions of 1024 from the pre-processed dataset. It is a pre-trained Mask-Language -Model that was pre-trained on a huge number of raw-texts and has learnt bidirectional representation of sentences. Therefore, the model can produce features that represent an inner representation of the English language and context in a sentence. Using these reviews pre-processed text as input, features of dimensions 1024 were extracted using this Roberta-Large model to be used input for our models. (for models 1-7 in Table 1)

Model Validation

The train dataset was split into 80% for training and 20% for validation. Accuracy was used to check the trained model's performance on the validation dataset.

Models

The extracted features from the Roberta-Large model were used as input to various models that were tested. As a baseline model, the logistic regression model was used and the validation accuracy was 0.88. Using this model as a baseline, multiple neural network models were assessed to deduce the best model to be submitted in the Kaggle Competition. The results of the multiple models accessed is shown in Table 1. For the Neural Network models, the loss function used was cross entropy and the optimiser used was momentum based Adam. The learning rate and dropout rate were assessed to test the model on accuracy to ensure that the model does not overfit. 250 epochs for training were used for the models 1-7 with early stopping when the validation loss stabilized for 5 epochs. 2 epochs were used for model 8 and model 9.



No	Models	Learning rate	Dropout	Filters	Filter size	Batch size	Early stopping	Validation accuracy
1	Logistic Regression							0.881
2	Fully Connected Neural Network 1	1 x e-5	0.2			64	Yes	0.9112
3	Fully Connected Neural Network 2	1 x e-5	0.5			64	Yes	0.9124
4	Fully Connected Neural Network 3	5 x e-6	0.5			64	Yes	0.9072
5	Ensemble Neural Network	1 x e-5	0.5, 0.4			64	Yes	0.914
6	CNN	1 x e-5	0.5	512, 256, 128	3	64	Yes	0.5031
7	Bidirection LSTM	1 x e-5	0.2			64	Yes	0.9058
8	Fine-tuned Bert - Pre-processed Reviews	2 x e-5				8	No	0.9242
9	Fine-tuned Bert - Raw Reviews	2 x e-5				8	No	0.9276

Table 1: Performance summary of models

1. Models 2-4 (Fully Connected Neural Networks)

For models 2 to 4, a fully connected neural network was used to train. With the features of dimension 1024 as input, 6 hidden layers of neurons, 2048, 1024, 512, 256, 128 and 64 respectively were built, the output layer was

1 neuron as it is a binary classifier. The activation function used for every hidden layer is RELU and dropout rate was used at every hidden layer. For the output layer, Sigmoid function was used for the binary classification. Learning rates of 1 x e-5 and 5 x e-6 and dropout rates of 0.2 and 0.5 were assessed in these models. Based on the validation accuracy, a learning rate of 1 x e-5 and dropout rate of 0.5 gave the best accuracy (Model 3 - 0.9124).

2. Model 5 (Ensemble)

An ensemble model of two neural networks of different architecture was also assessed. The first neural network was with the features extracted (dimensions 1024) as input and two hidden layers (2048 neurons and 256 neurons respectively) with RELU function as activation functions for the hidden layers and dropout rate of 0.5 for each hidden layer output. The output layer was 1 neuron with sigmoid function. The 2nd neural network was with the same input layer and output layer but with 3 hidden layers (1024, 512 and 128 neurons respectively). The activation functions were RELU for the hidden layers and sigmoid function for the output layer. Both these models were trained using the training dataset. The models were then predicted using the validation dataset and the mean predictions were used to assess the validation accuracy. The model performed better than the Models 2-4 with a validation accuracy of 0.914.

3. Model 6 (CNN)

A convolutional neural network was also assessed to deduce if using filters to learn different context in the reviews could help in the sentiments classification. 3 convolution layers of 512, 256 and 128 filters respectively with filter size of 3 was used followed by a fully connected network that is described in Model 2-4. However, the model did not perform well and only a validation accuracy of 0.5031 was achieved.

4. Model 7 (Bi-Directional LSTM)

A bi-directional LSTM model was also assessed to deduce if the sequence of the sentences in the reviews could be captured to better enhance the prediction of the sentiments. A bidirectional LSTM layer of 512 neurons was built where the output was then flattened to be passed to an output layer with a sigmoid activation function. A dropout rate of 0.5 was used. The model performed relatively well compared to CNN with a validation accuracy of 0.9058. However, it was still lower in performance compared to the fully connected networks and ensemble model.

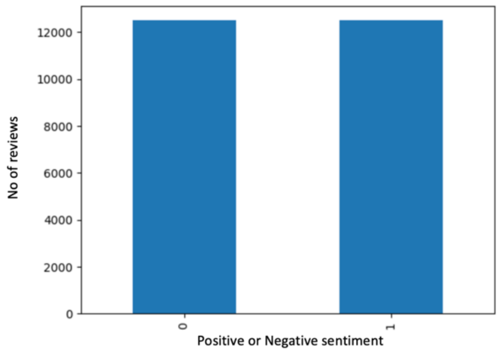
5. Model 8 – 9 (Fine-Tune Bert Model)

The last two models assessed was using the BERT model where the model was fine-tuned using our dataset. The preprocessed reviews text were tokenized using the BERT tokenizer with a maximum length of 512. The output of the BERT encoder model of dimensions 768 are pooled to a linear head output layer of 2 for our binary classification. A Softmax function was used in the output head layer for our binary classification task. Since the model is very large and pre-trained, an epoch size of 2 was sufficient to train the model. A learning rate of 2 x e-5 was utilized for the training. A validation accuracy of 0.9242 was achieved with this model which was the best model among all the models assessed thus far. This model (model 8) was used to predict the sentiments for the test dataset and submitted in Kaggle. The test accuracy from the public leaderboard was 0.92737 which is the best test accuracy that has been achieved thus far.

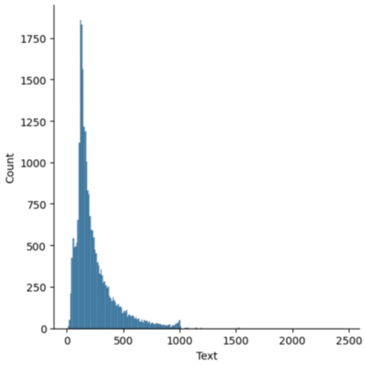
As a test to check if pre-processing the reviews text made reviews lose some context that could help in predicting the sentiments, the same model was trained using the raw text of the reviews which were tokenized as input (Model 9). The fine-tuned model with these tokenized raw reviews as input gave a better validation accuracy of 0.9276. This fine-tuned model was also used to make predictions on the test dataset and submitted in Kaggle. Interestingly, it shows that for the fine-tuning of the BERT model, pre-processing of the reviews removes some of the context of the reviews that may help in better predicting the sentiments of these reviews. This is the best model that predicts the sentiments of the movie reviews with a test accuracy of 0.9356 after submission in Kaggle.

Appendix

EDA of Dataset



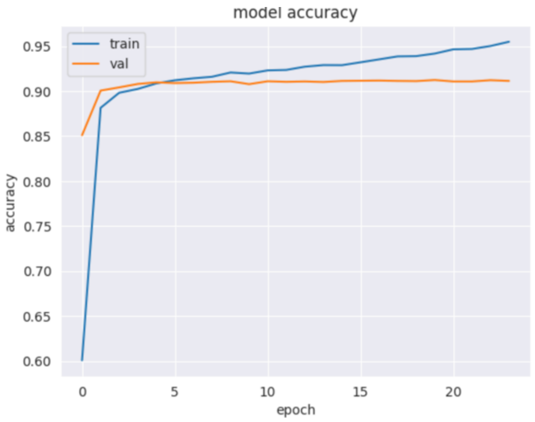
Number of positive and negative sentiments in Training dataset



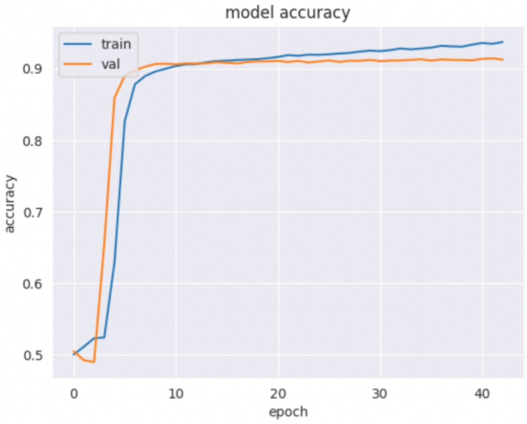
Word length distribution of reviews in Training Dataset

Model Accuracy Comparison

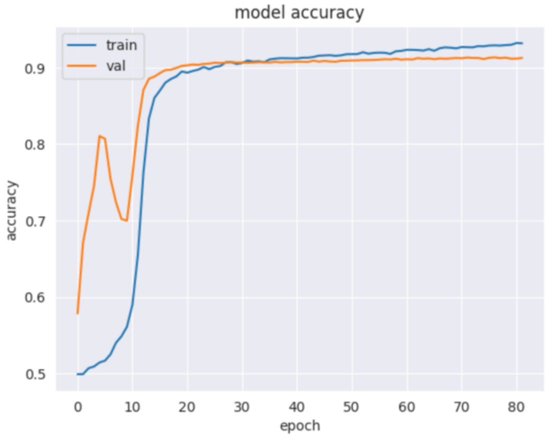
Model accuracy plots not available for model 5, 8 and 9 as model 5 is an esnsemble model of two neural networks and model 8 and 9 is only trained for 2 epochs.



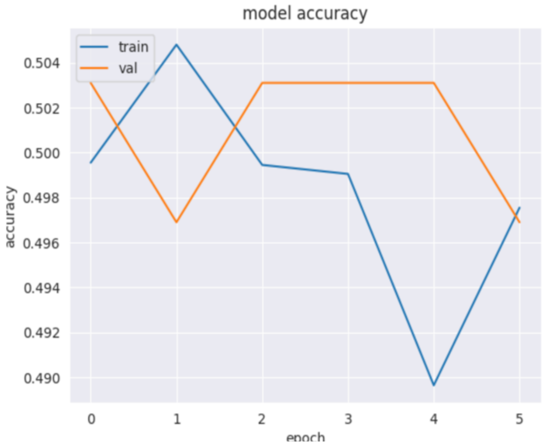
Model 2 training accuracy vs validation accuracy over no of epochs



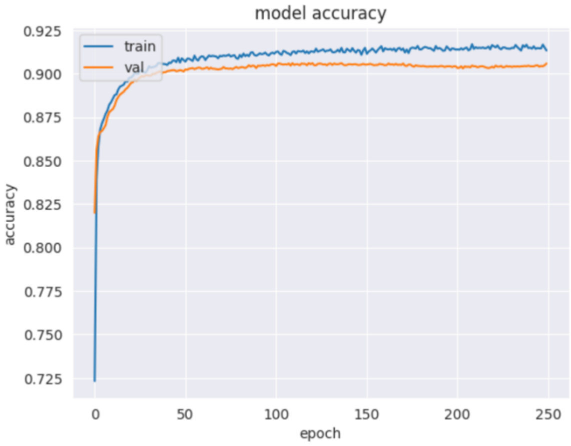
Model 3 training accuracy vs validation accuracy over no of epochs



Model 4 training accuracy vs validation accuracy over no of epochs



Model 6 training accuracy vs validation accuracy over no of epochs



Model 7 training accuracy vs validation accuracy over no of epochs

References

1. https://huggingface.co/docs/transformers/v4.27.2/en/model_doc/bert#transformers.TFBertForSequenceClassification
2. https://huggingface.co/roberta-large