Assignment 5.1: Sentiment Analysis with BERT on IMDb Movie Reviews

Patricia Enrique

Department of Engineering, University of San Diego

AAI-520: Natural Language Processing

Prof. Mokhtari

October 9, 2023

Data Loading and Preprocessing

The dataset used for the sentiment analysis with BERT is compiled by Standford (Maas et al., 2011) and is comprised of 50000 movie reviews with a binary sentiment. The distribution of positive and negative sentiment reviews is equal with 25000 for each classification. The reviews are of varying lengths and can contain emojis, breaks, and spelling errors. A random sample of the training dataset is shown in Table 1. The sentiments for the reviews are converted to a numerical scale with 1 indicating a positive sentiment and 0 indicating a negative sentiment. For this analysis, all the reviews were changed to lowercase before using the pre-trained BERT model.

Table 1

Movie review examples

ID	Review	Sentiment
415	A rating of "1" does not begin to express how dull, depressing and relentlessly bad this movie is.	negative
9437	Hated it with all my being. Worst movie ever. Mentally- scarred. Help me. It was that bad.TRUST ME!!!	negative
1419	For pure gothic vampire cheese nothing can compare to the Subspecies films. I highly recommend each and every one of them.	positive
9746	A great film in its genre, the direction, acting, most especially the casting of the film makes it even more powerful. A must see.	positive
21044	Brilliant. Ranks along with Citizen Kane, The Matrix and Godfathers. Must see, at least for basset in her early days. Watch it.	positive

Text Tokenization and Conversion to BERT Input Features

The BERT tokenizer is used to tokenize the reviews and generate IDs. An example of a sentence after tokenization and the corresponding IDs is shown in figure 1. During the tokenization process, a special token [SEP] is added to mark the end of a sentence,[CLS] is added to the start of the sentence to indicate that classification is being performed, and [PAD] is added to the end of sentences to pad them to a uniform length. The BERT pre-trained model is limited by a maximum length of 512 and any sentences

longer are truncated. Attention masks allow for differentiation between tokens that are used for padding and ones that are not by setting a 0 for padding and a 1 for all real tokens (Singh, 2021).

Figure 1

Tokenization and corresponding IDs from a review

```
Original:  A rating of "1" does not begin to express how dull, depressing and relentlessly bad this movie is.
Tokenized:  ['a', 'rating', 'of', '"', '1', '"', 'does', 'not', 'begin', 'to', 'express', 'how', 'dull', ',', 'de', '##pressing', 'and', 'relentless', '##ly', 'bad', 'this', 'movie', 'is', '.']
Token IDs:  [1037, 5790, 1997, 1000, 1015, 1000, 2515, 2025, 4088, 2000, 4671, 2129, 10634, 1010, 2139, 24128, 1998, 21660, 2135, 2919, 2023, 3185, 2003, 1012]
```

Once the data preprocessing is completed, the IDs and masks are split into training, testing, and validation. Half of the data is saved for testing with the other half used for training and validation. The number of reviews used for training is 20000, validation is 5000, and for testing is 25000. The data is then converted to a tensor to be fed into the pre-trained BERT model.

Model Definition, Training, and Evaluation

The pre-trained BERT model for sequence classification is loaded from the Transformers library and AdamW is set as the optimizer. To fine-tune the BERT model on the preprocessed IMDb dataset, the loss and accuracy is measured for each epoch using the training and validation data (Dubey, 2020). Table 2 summarizes the accuracy and loss for a run with 4 epochs. As seen from the table, the accuracy does not significantly increase with the addition of epochs; however, the loss is reduced. Due to the time cost to run with higher epochs, the model is limited to 2.

Table 2

Accuracy and loss to fine-tune BERT model on IMDB dataset

Performance Metric	Epoch			
	1	2	3	4
Accuracy	0.93	0.93	0.94	0.94
Loss	0.26	0.13	0.07	0.03

Once the fine-tuning is completed, the trained model is evaluated on the testing set using accuracy, precision, recall, and F1-score metrics. The final performance can be seen in Table 3. The model obtained a 93% performance on all the metrics with minimal fine-tuning. This indicates that 93% of the positive sentiment predictions were correct and similarly, 93% of actual positive sentiments were identified correctly.

Table 3

Performance metrics for the trained model on the testing set

Model	Performance Metric			
	Accuracy	Precision	Recall	F1 Score
Testing Set	0.93	0.93	0.93	0.93

Sample Movie Review Predictions and Explanations

The movie review samples introduced in Table 1 are individually fed into the pre-trained model for their sentiment prediction and the results are shown in Table 4 (Tiwari, 2020). It can be seen that the model is able to predict the correct sentiment for each of the review samples. This result is expected from the performance metrics obtained during the testing process.

A major disadvantage to the BERT pre-trained model is the model size and computational effort required to run it, and the training process is slow due to the number of weights that need to be updated. However, the BERT model is able to obtain high accuracy and can be fine-tuned and used immediately for various tasks (Tunstall et al., 2020).

Table 4

Sentiment predictions from sample reviews

ID	Review	Sentiment Prediction	True Sentiment

415	A rating of "1" does not begin to express how dull, depressing and relentlessly bad this movie is.	negative	negative
1419	Hated it with all my being. Worst movie ever. Mentally- scarred. Help me. It was that bad.TRUST ME!!!	negative	negative
9437	For pure gothic vampire cheese nothing can compare to the Subspecies films. I highly recommend each and every one of them.	positive	positive
9746	A great film in its genre, the direction, acting, most especially the casting of the film makes it even more powerful. A must see.	positive	positive
21044	Brilliant. Ranks along with Citizen Kane, The Matrix and Godfathers. Must see, at least for basset in her early days. Watch it.	positive	positive

References

BERT. (n.d) Hugging Face. Retrieved October 5, 2023, from https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertForSequenceClassification

Classification: Precision and Recall. (2022, July 18) Google for Developers. Retrieved October 8, 2023, from https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Dubey, A. (2020). BERT for sequence classification. *Kaggle*. https://www.kaggle.com/code/akshat0007/bert-for-sequence-classification

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Large Movie Review Dataset* (Version 1.0) [Data set]. Stanford University. http://ai.stanford.edu/~amaas/data/sentiment/

Singh, Y. V. (2021). Classification using Pre-trained Bert Model (Transfer Learning). *Medium.* https://medium.com/@yashvardhanvs/classification-using-pre-trained-bert-model-transfer-learning-2d50f404ed4c

Tiwari, S. (2020). IMDB Sentiment Analysis using BERT(w/ Huggingface). *Kaggle.* https://www.kaggle.com/code/satyampd/imdb-sentiment-analysis-using-bert-w-huggingface

Tokenizer. (n.d) Hugging Face. Retrieved October 5, 2023, from https://huggingface.co/docs/transformers/main_classes/tokenizer

Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural Language Processing with Transformers*. O'Reilly Media.