AML 4 TEXT AND DATA

Report

Objective:

The goal is to develop an embedding model and pre trained word embedding that can learn from the dataset and effectively distinguish between positive and negative reviews based on their characteristics. The model should be capable of analyzing various aspects of the reviews and making a binary prediction of whether each review is positive or negative

EMBEDDING MODEL:

number	traning	loss	validation	Performance on Test Set (Loss, Accuracy)
Model 2	100	0.4512	10000	0.7894
Model 3	500	0.4013	10000	0.8209
Model 4	2000	0.3772	10000	0.8331
Model 5	5000	0.385	10000	0.8921

PRETRAINED WORD EMBEDDING:

number	training	loss	validation	Performance on Test Set (Loss, Accuracy)
Model 2	100	0.4823	10000	0.7700
Model 3	500	0.3826	10000	0.827
Model 4	2000	0.3709	10000	0.8382
Model 5	5000	0.3654	10000	0.8348

AML 4 TEXT AND DATA

Outcome:

The purpose of this study was to evaluate the performance of an embedding model with a pretrained word embedding model for binary classification of positive and negative reviews on the IMDB reviews datase The models were trained with varying amounts of training samples and then examined on the test set based on their loss and accuracy
Four distinct embedding models were developed using 100, 500 2000 and 5000 training data Model 5 had the best accuracy of 0.8921 with a loss of 0.385 on the test set while Model 2 had the lowest accuracy of 0.7894 with a loss of 0.4512 The model's performance improved as the number of training samples rose demonstrating that a bigger dataset can help the model learn more about the positive and negative features of reviews and enhance its generalization ability
Four other models were also trained with 100 500 2000 and 5000 training samples for the pre trained word embedding model On the test set Model 5 had the best accuracy of 0.8348 with a loss of 0.3654 while Model 2 had the lowest accuracy of 0.77 with a loss of 0.4823 The performance of the pre-trained word embeding model improved with the amount of training samples similar to the embedding model
When the two models were compared in terms of accuracy the pre-trained word embedding model outperformed the embedding model with the highest performing models obtaining an accuracy of 0.8348 and 0.8921 respectively In addition the pre-trained word embedding model exhibited lower loss values than the other models indicating that it was better at reducing the disparity between predicted and real labels
The pretrained word embedding strategy appears to perform better than the simple embedding layer approach. This is demonstrated by the test set performance, where the pretrained word embedding models consistently outperform the embedding models across all four models. Furthermore, pretrained word embedding models exhibit lower training loss and greater validation accuracy than embedding models, implying that they are able to apply to new data better.

CONCLUSION:

In conclusion this study demonstrates that a pretrained word embedding model outperforms an embedding model in binary classification of positive and negative reviews Furthermore the performance of both models improved as the number of training samples increased demonstrating the relevance of having a big and diverse dataset for training machine learning models Future research might look into various pre-trained word embedding models or hybrid models that mix several approaches to increase sentiment analysis task accuracy