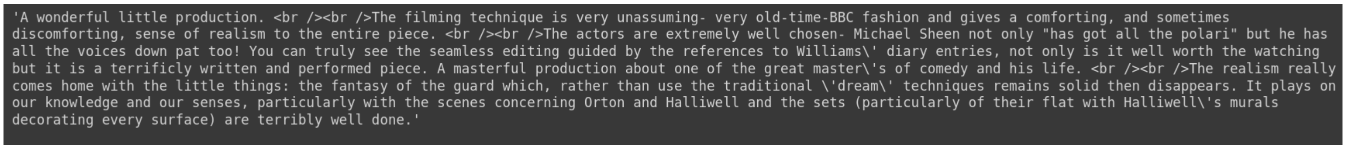
CS60075 : Assignment-2 Report

Debanjan Saha [19CS30014]

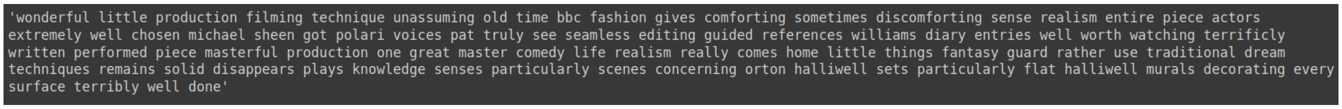
>> Exploratory Data Analysis:

- Initially, I looked at some reviews manually from the given dataset and found out that it contained many unnecessary characters or html tags that needed further processing and cleaning.



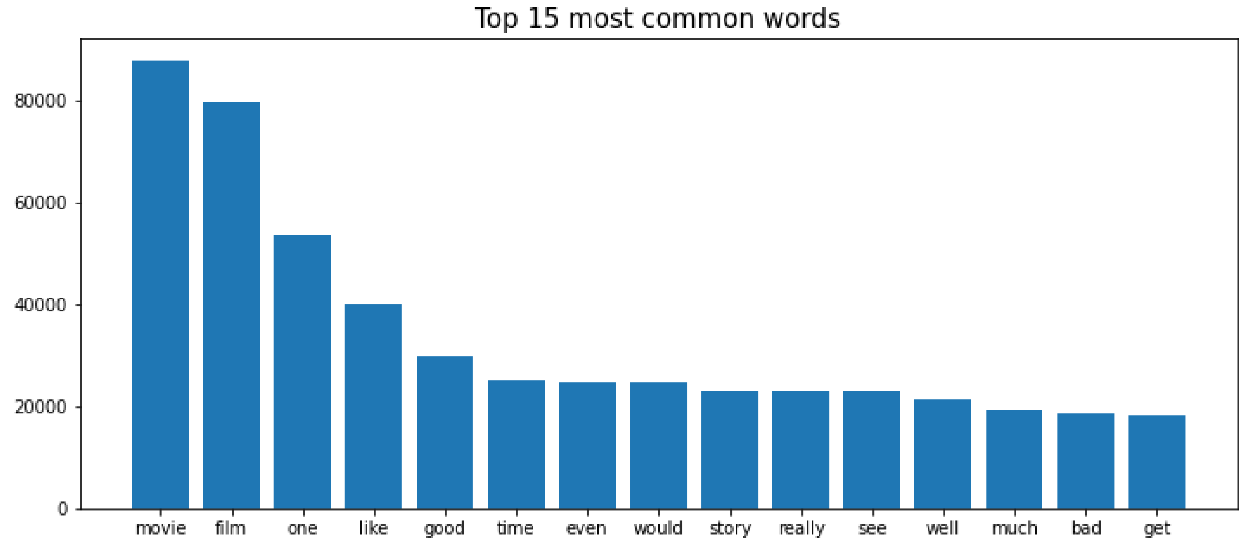
```
'A wonderful little production. <br /><br />The filming technique is very unassuming- very old-time-BBC fashion and gives a comforting, and sometimes
discomforting, sense of realism to the entire piece. <br /><br />The actors are extremely well chosen- Michael Sheen not only "has got all the polari" but he has
all the voices down pat too! You can truly see the seamless editing guided by the references to Williams\' diary entries, not only is it well worth the watching
but it is a terrificly written and performed piece. A masterful production about one of the great master\'s of comedy and his life. <br /><br />The realism really
comes home with the little things: the fantasy of the guard which, rather than use the traditional \'dream\' techniques remains solid then disappears. It plays on
our knowledge and our senses, particularly with the scenes concerning Orton and Halliwell and the sets (particularly of their flat with Halliwell\'s murals
decorating every surface) are terribly well done.'
```

- I wrote a `clean_text` function using which the above text was converted to the text as shown below using preprocessing functions like regex, contractions, stopwords and punctuation removal, and lemmatizers.

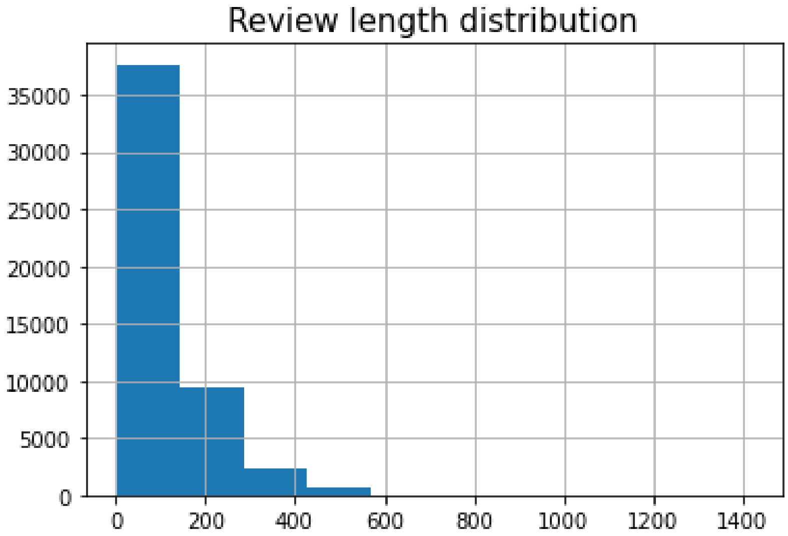


```
'wonderful little production filming technique unassuming old time bbc fashion gives comforting sometimes discomforting sense realism entire piece actors
extremely well chosen michael sheen got polari voices pat truly see seamless editing guided references williams diary entries well worth watching terrificly
written performed piece masterful production one great master comedy life realism really comes home little things fantasy guard rather use traditional dream
techniques remains solid disappears plays knowledge senses particularly scenes concerning orton halliwell sets particularly flat halliwell murals decorating every
surface terribly well done'
```

- After that, I checked what kind of words are present most frequently in the dataset. Here is a plot for that.



- Finally, I tried to check the statistics of the length of the reviews to determine a suitable sequence length for the model. We can see that majority of the reviews have a length between 0-100. The average length of the reviews came out to be = 118.22032. Based on this, I chose the sequence length = 128 for the whole dataset using which I generated features for each review from pretrained word2vec embeddings.



>> Model Configuration:

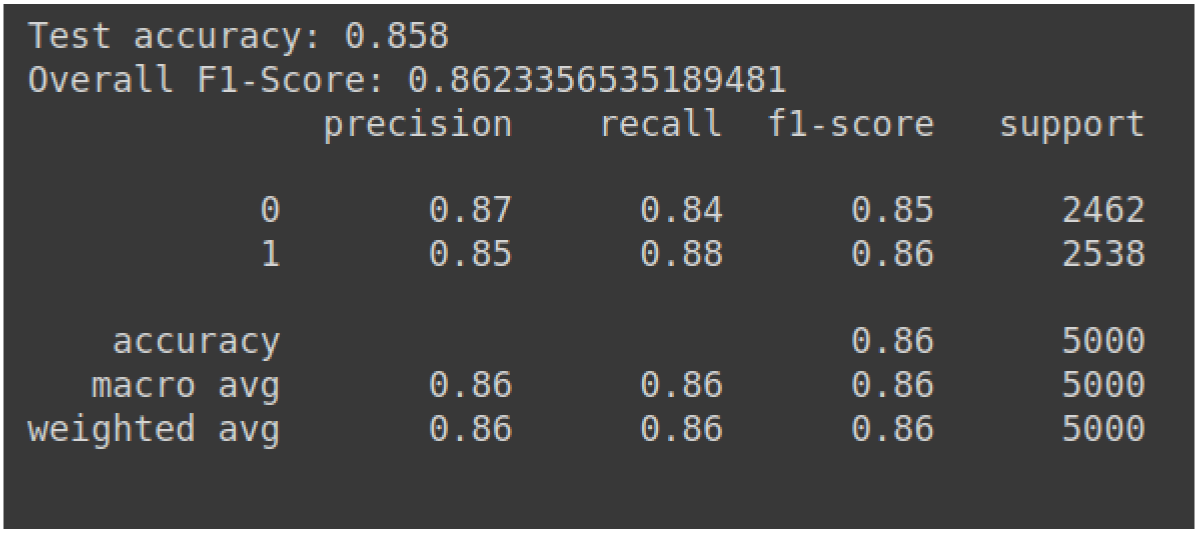
- The model consists of a Bi-LSTM layer.
- There are only 2 unique classes in this task, hence we have a binary classification task. I used Binary Cross Entropy Loss (BCELoss) as the loss function. And I used Adam optimizer as it is shown in many references that it could find convergence-point quickly.
- The model was loaded with pretrained word2vec embeddings of vector dimension 64 which was found to be best after parameter tuning
- Best results given below were obtained at a learning rate=0.001, epochs=5, batch_size=50, hidden_dimension = 256
- The dataset was split into 3 parts (80% train, 10% dev and 10% test). Based on the development set, the best trained model with smallest development loss was saved and further used for testing.

>> Model Link:

- https://drive.google.com/drive/folders/1xtLq4xsT9zQFEBBdEALkAlHCEKPjW29w?usp=sharing

>> Results:

- As shown below, here is a classification report on the test set of the given dataset. Class 0 represents negative sentiment and class 1 represents positive sentiment.
- It is generated using sklearn package’s classification report method



```
Test accuracy: 0.858
Overall F1-Score: 0.8623356535189481
              precision    recall  f1-score   support

           0       0.87      0.84      0.85      2462
           1       0.85      0.88      0.86      2538

    accuracy                           0.86      5000
   macro avg       0.86      0.86      0.86      5000
weighted avg       0.86      0.86      0.86      5000
```

>> Inference:

- As we can see, the prediction accuracy is close to 86% (85.8%) and the overall F1-score achieved is 0.862

- Considering the F1 score of both positive and negative sentiment classes on the test dataset, we can infer that the model has learned well to predict the sentiment given a text of a movie review.