TF-IDF to BERT: A Primer for Supervised Learning Techniques Applied To Sentiment Classification.

1. Introduction

- ABOUT

Sentiment analysis has been one of the many foundational problems that the field of Natural Language Processing has tried to solve since its initial years. In these exercises, we consider a similar sentiment classification task related to movie reviews.

Sentiment analysis is very popular and a useful business use case for companies as it tries to automate several difficult tasks. Through sentiment classification, we can understand opinions on a large scale. This can have many applications for businesses, such as finding insights in customer feedback, keeping an eye on brand reputation, and spotting marketing trends and opportunities.

- DATASET

We consider the IMDB movie review dataset for the purpose of all experiments. The labelled data set consists of 50,000 movie reviews, specially selected for sentiment analysis. The target variable is binary. A rating 5 results in a sentiment score of 0, while a rating >=7 implies a sentiment score of 1. No individual movie has more than 30 reviews. The 25,000 reviews labelled training set does not include any of the same movies as the 25,000 review test set. We build a classifier that can correctly identify the sentiment (positive/negative) of any review given to a movie.

- APPROACH

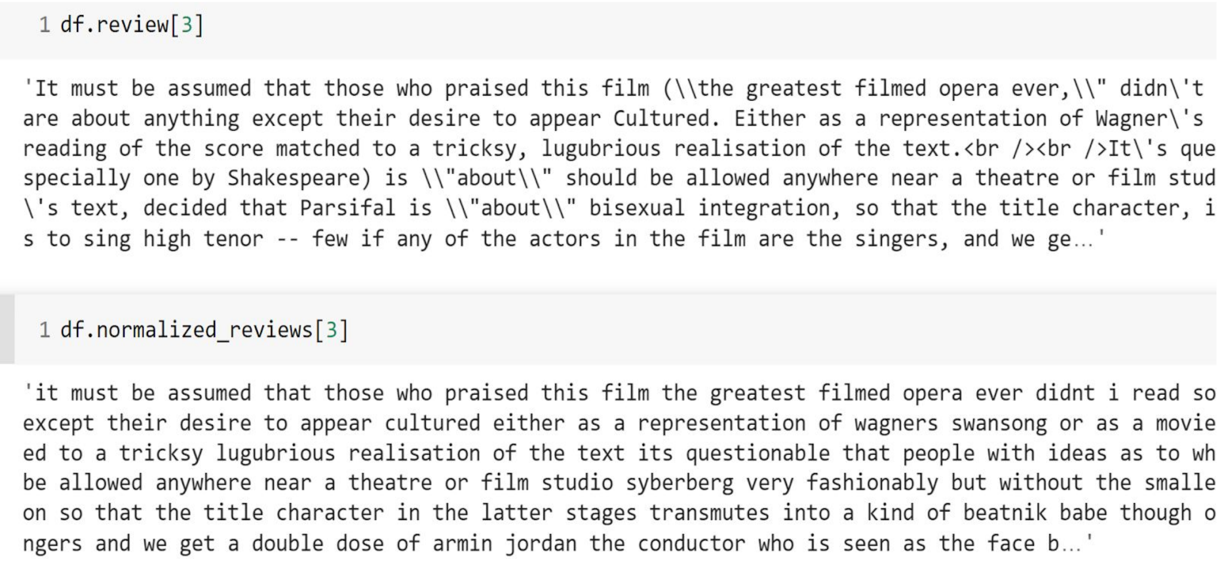
To solve this problem, We first start off with traditional methods to solve Text classification problems, by using Statistical learning algorithms like Naive Bayes with TF-TDF used for feature engineering. This would serve as our baseline. Then we make use of the deep learning based approaches namely LSTMs, BI-LSTMs, LSTM-Attention, CNNs and finally BERT.

2. Data Cleaning

We observed that the reviews consisted of all sorts of special characters, punctuations and word contractions. The reviews should be free from all such characters to get processed as input.

Steps involved:

- Removal of HTML tags
- Converting reviews to lower case
- Removal of all special characters



```
df.review[3]
```

```
'It must be assumed that those who praised this film (\\the greatest filmed opera ever,\\" didn\'t
are about anything except their desire to appear Cultured. Either as a representation of Wagner\'s
reading of the score matched to a tricksy, lugubrious realisation of the text.<br /><br />It\'s que
specially one by Shakespeare) is \\"about\\" should be allowed anywhere near a theatre or film stud
\'s text, decided that Parsifal is \\"about\\" bisexual integration, so that the title character, i
s to sing high tenor -- few if any of the actors in the film are the singers, and we ge...'
```

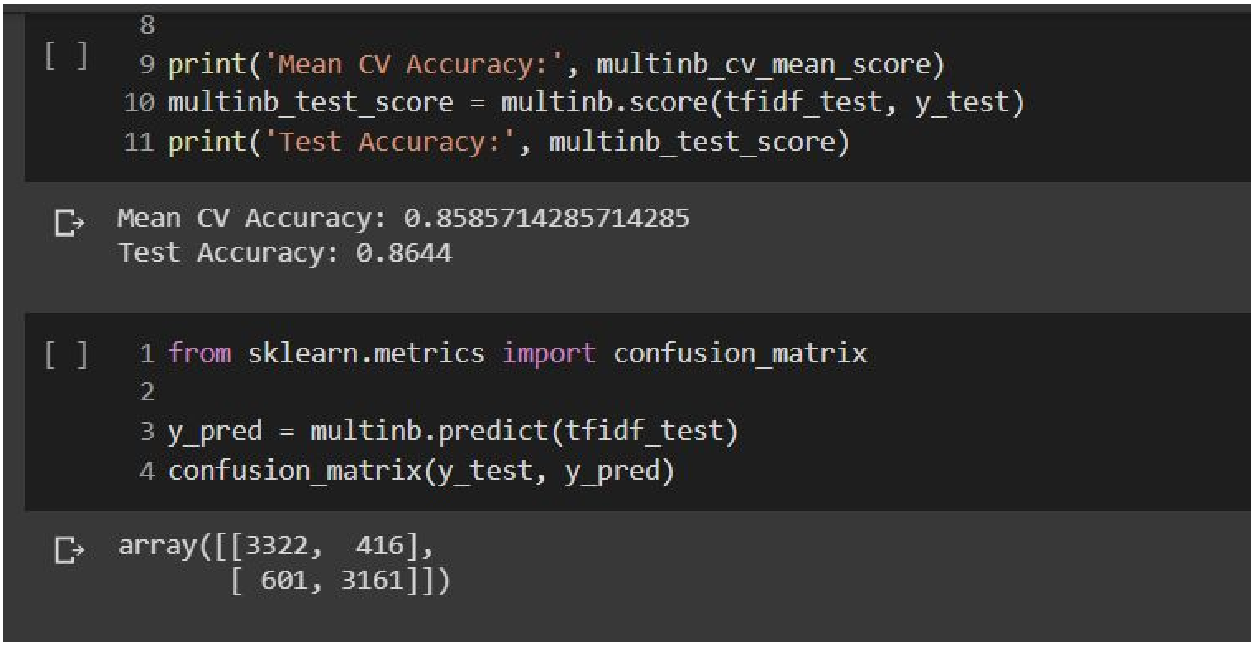
```
df.normalized_reviews[3]
```

```
'it must be assumed that those who praised this film the greatest filmed opera ever didnt i read so
except their desire to appear cultured either as a representation of wagners swansong or as a movie
ed to a tricksy lugubrious realisation of the text its questionable that people with ideas as to wh
be allowed anywhere near a theatre or film studio syberberg very fashionably but without the smalle
on so that the title character in the latter stages transmutes into a kind of beatnik babe though o
ngers and we get a double dose of armin jordan the conductor who is seen as the face b...'
```

3. Classical NLP

The first attempt to model the problem begins with the use of classic Machine Learning algorithms like Naive bayes, which is a popular initial approach. For feature engineering, we made use of TF-IDF. The purpose of TF-IDF is to make a statistical measure as to how relevant a word is to a document in a collection of documents.In our application, It enables us to associate each word in a review with a number that represents how relevant each word is in that review.

The training data is split into train and test sets and the model is trained using 5 fold cross validation to give the following result.



```

print('Mean CV Accuracy:', multinb_cv_mean_score)
multinb_test_score = multinb.score(tfidf_test, y_test)
print('Test Accuracy:', multinb_test_score)
```

```
Mean CV Accuracy: 0.8585714285714285
Test Accuracy: 0.8644
```

```
from sklearn.metrics import confusion_matrix

y_pred = multinb.predict(tfidf_test)
confusion_matrix(y_test, y_pred)
```

```
array([[3322,  416],
       [ 601, 3161]])
```

Now this serves as the baseline accuracy that we would like to beat by making use of the newer and latest advancements in deep learning which have become the default method to tackle unstructured data like texts and images.

4. Deep Learning in NLP

The important architectures relevant for the above task are listed below.

- Convolutional Neural Networks
- Recurrent Neural Networks
- Long Short Term Memory
- Bidirectional LSTMs
- LSTM with Attention
- BERT

A similar flow has been followed and the above architectures are employed, along with some variants, to model the task at hand.

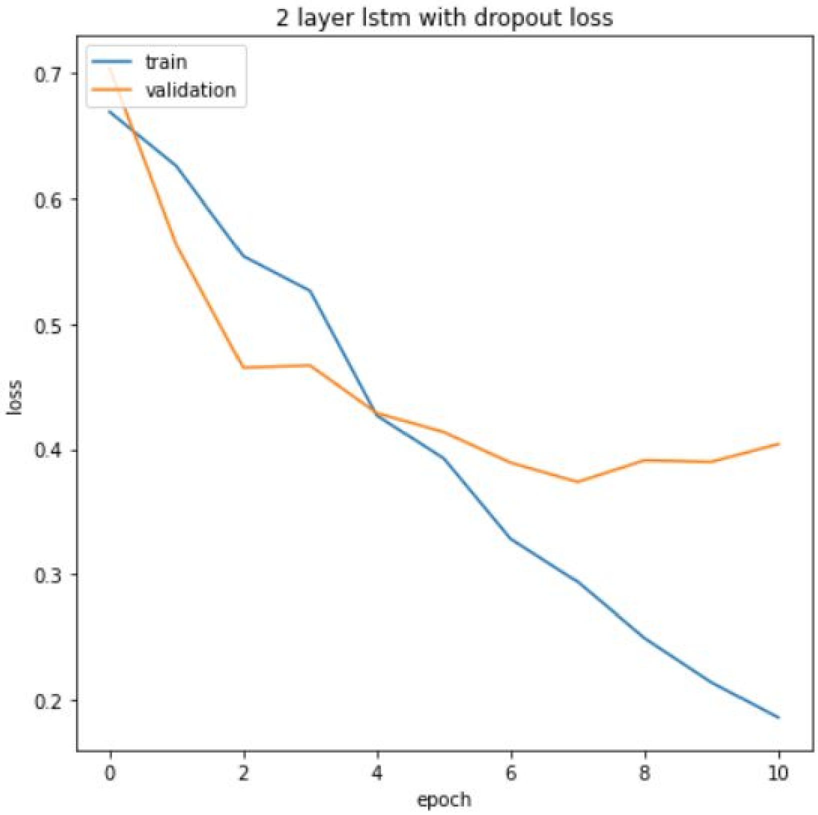
Before the reviews can be fed into the above architecture we need to convert the text into a representation. We convert our entire corpus of reviews into a dictionary, where each word is a 200-dimensional vector of real-valued integers. The distributed representation is learned based on the words in the corpus of reviews and the Word2Vec algorithm is trained from scratch instead of using pre-trained embeddings.

As the corpus of this task is sufficiently big, meaningful word vectors can be generated by solely training Word2Vec on the training dataset.

Since LSTMs were invented to process data which is sequential, it would be logical to start the series of experiments on the reviews with it.

LSTM

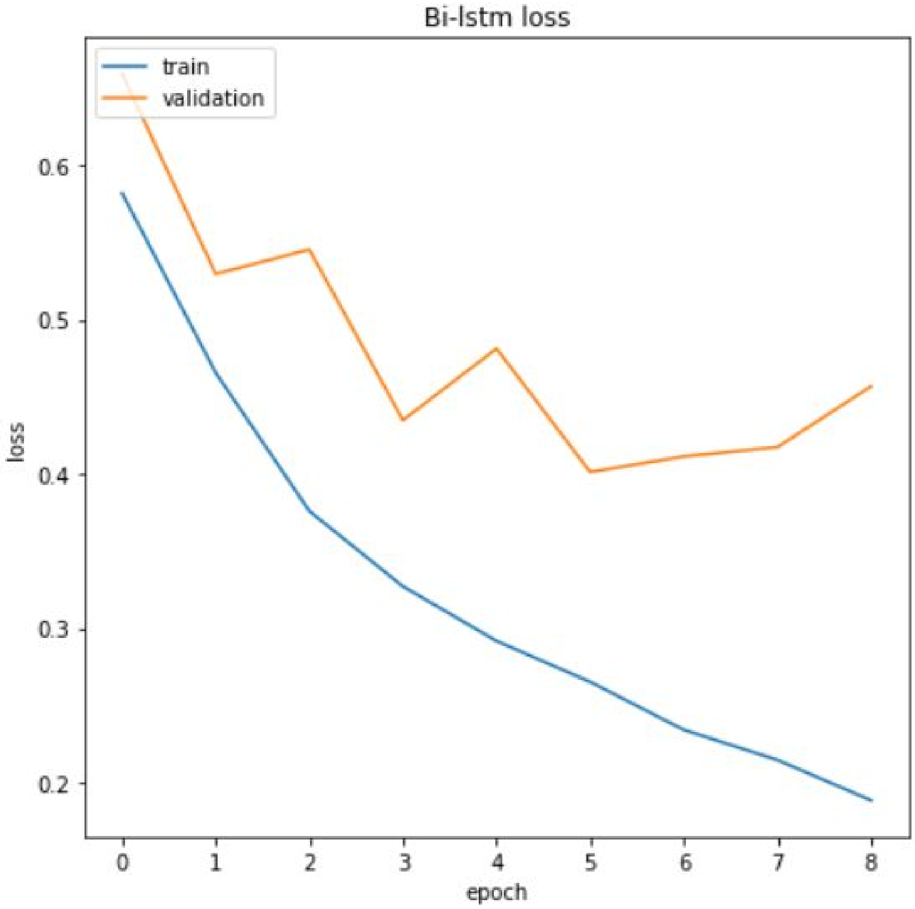
Different variations of this sequential architecture were tried namely simple LSTM, LSTM with dropout, Stacked LSTM(with dropout). The one that turned out to perform the best was stacked LSTM. In this model, two simple LSTM layers are put one on top of another, with penalties so that the model learns effectively. The result of the training came out as.



The validation accuracy of the model was 85.14%. The above loss curve suggests that the model is overfitting the data and has not generalized well to the unseen data. It could be possible due to the uneven length of the movie reviews present in the training set to the ones in the validation set. Bi-LSTMs should solve this problem.

Bi-LSTM

Now the training data is passed into a Bi-LSTM architecture to solve the overfitting problem faced by LSTMs. Bi-LSTM is similar to normal LSTM with the key difference that they learn the sequence in both the directions. The result of the experiment is as follows.



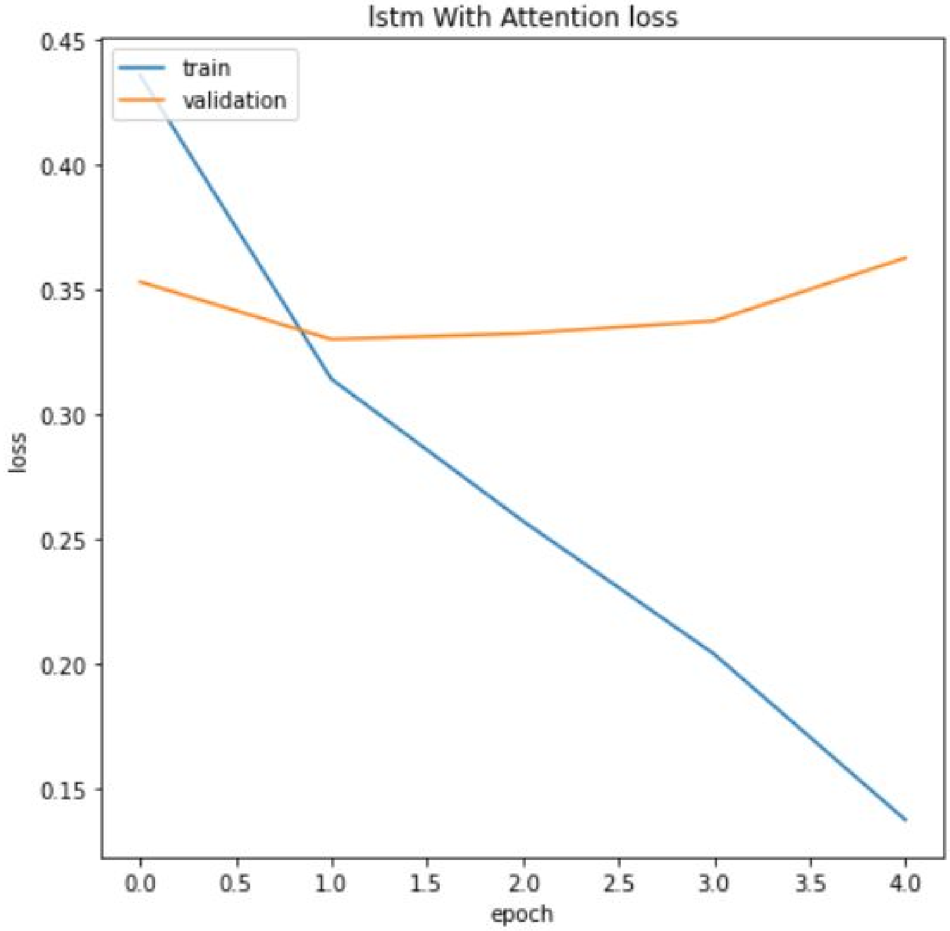
The validation accuracy of this model came out to be 85.71% but the loss curve suggests a similar problem - overfitting. It seems like the length of the reviews could be a problem. The median length of the reviews was calculated to be around 170 tokens, which is quite lengthy for LSTMs. Maybe LSTM with Attention could resolve the issue.

LSTM with Attention

Attention in deep learning can be broadly interpreted as a vector of importance weights: to predict or infer one element, such as a pixel in an image or a word in a sentence, we estimate using the attention vector how strongly it is correlated with other elements and take the sum of their values weighted by the attention vector as the approximation of the target

[https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html].

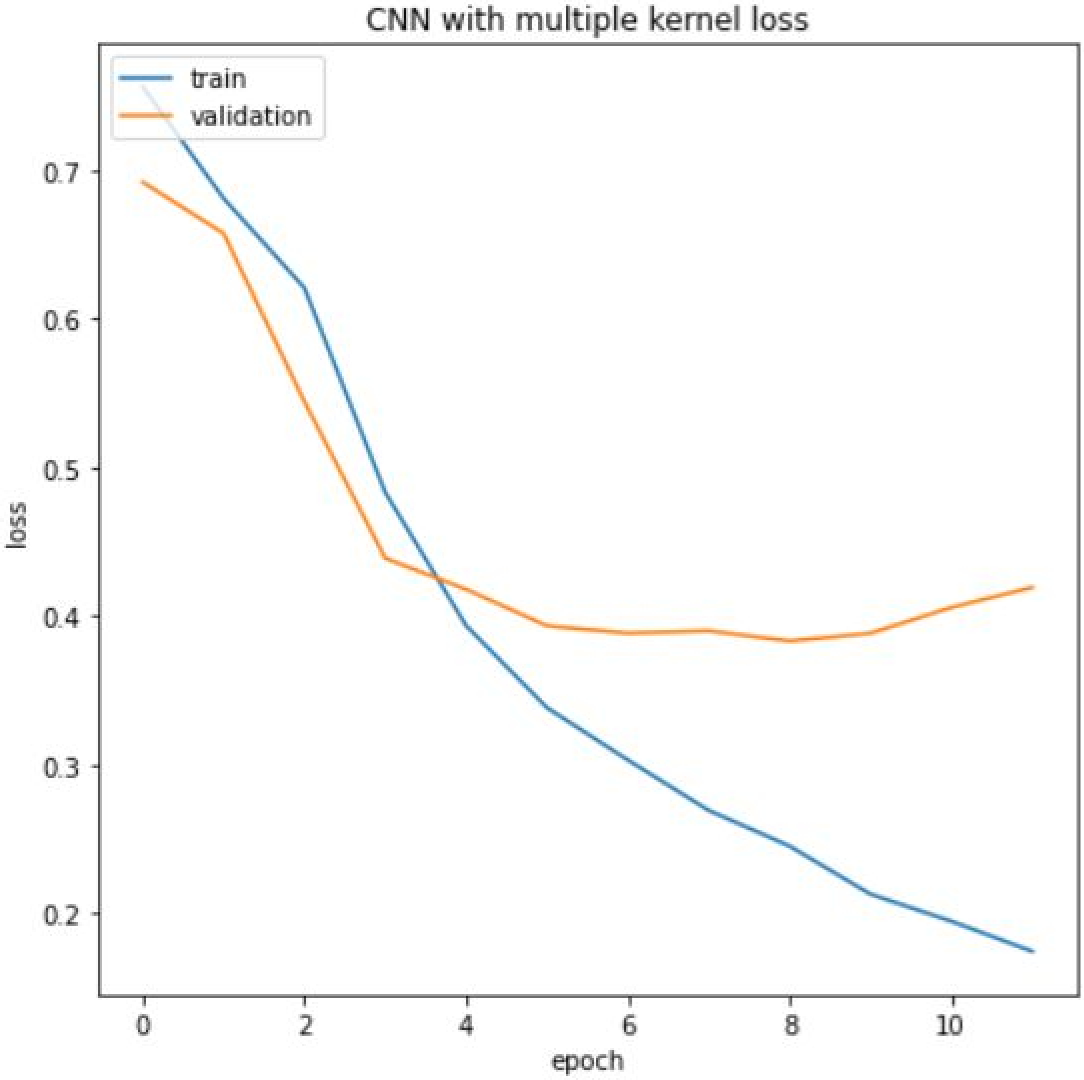
This means it would pay attention to only certain words in a sentence to make its classification.



Attention overfits the training data quite quickly, within 1 epoch. Even though the validation loss is quite low at the beginning, the model overfits fast and only seems to get worse post the 1st epoch.

CNN

CNN’s have also been proved to be useful for the task of sentiment analysis. The way the CNNs work is that the words are broken up into features and are fed into a convolutional layer. The results of the convolution are “pooled” or aggregated to a representative number. This number is fed to a fully connected neural structure, which makes a classification decision based on the weights assigned to each feature within the text.

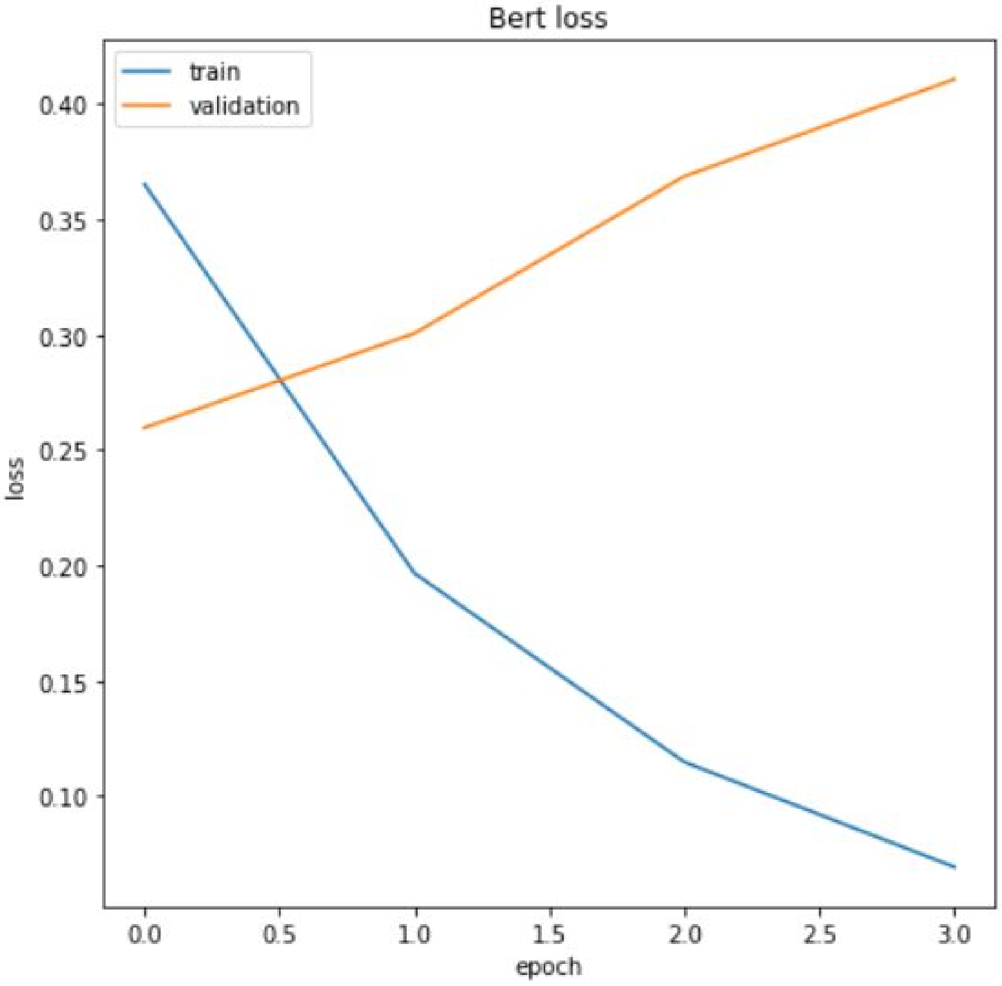


Even though CNNs are much faster to train and lighter and don't take in the inputs sequentially as linear models do, a still similar outcome is seen with CNNs with model overfitting and giving a validation accuracy of 85.18%.

BERT

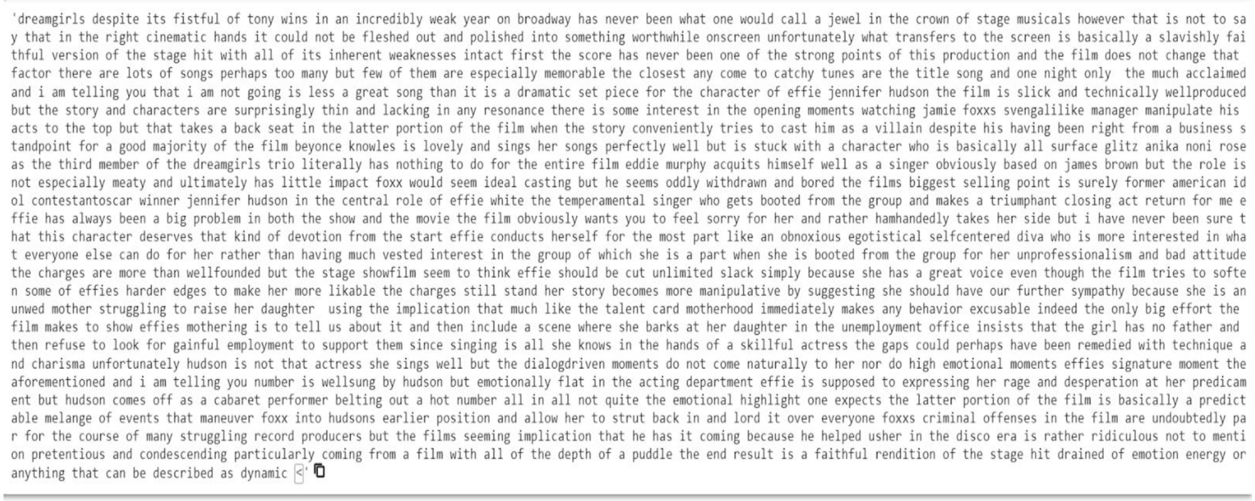
Bidirectional Encoder Representations from the Transformer is the state-of-the-art of the model that is the recent milestone development in the field of NLP, which came in 2018. This architecture builds on the architecture of transformers, which builds on the concept of attention,but it finds its main application in Machine Translation.The encoder layer of the transformer architecture was taken, which was further developed to produce BERT. The main advantage of BERT is that it doesn't read inputs sequentially, which means that all the tokens are available to be processed at once unlike in the case of linear models like LSTMs. This proved to be a huge improvement in training speed. Another big advantage is that it is first pre-trained on a large corpus of text and then uses the knowledge it has gained to solve the task at hand by the process of fine-tuning.

Now we make use of BERT for the sentiment classification task.



BERT is the most powerful out of all the models that have been implemented so far and it overfits the data almost immediately. It can be concluded that as the complexity of the model increases the model seems to overfit on the train data with more ease.

5. Results and Analysis



```
'dreamgirls despite its fistful of tony wins in an incredibly weak year on broadway has never been what one would call a jewel in the crown of stage musicals however that is not to sa
y that in the right cinematic hands it could not be fleshed out and polished into something worthwhile onscreen unfortunately what transfers to the screen is basically a slavishly fai
thful version of the stage hit with all of its inherent weaknesses intact first the score has never been one of the strong points of this production and the film does not change that
factor there are lots of songs perhaps too many but few of them are especially memorable the closest any come to catchy tunes are the title song and one night only  the much acclaimed
and i am telling you that i am not going is less a great song than it is a dramatic set piece for the character of effie jennifer hudson the film is slick and technically wellproduced
but the story and characters are surprisingly thin and lacking in any resonance there is some interest in the opening moments watching jamie foxxs svengalilike manager manipulate his
acts to the top but that takes a back seat in the latter portion of the film when the story conveniently tries to cast him as a villain despite his having been right from a business s
tandpoint for a good majority of the film beyonce knowles is lovely and sings her songs perfectly well but is stuck with a character who is basically all surface glitz anika noni rose
as the third member of the dreamgirls trio literally has nothing to do for the entire film eddie murphy acquits himself well as a singer obviously based on james brown but the role is
not especially meaty and ultimately has little impact foxx would seem ideal casting but he seems oddly withdrawn and bored the films biggest selling point is surely former american id
ol contestantoscar winner jennifer hudson in the central role of effie white the temperamental singer who gets booted from the group and makes a triumphant closing act return for me e
ffie has always been a big problem in both the show and the movie the film obviously wants you to feel sorry for her and rather hamhandedly takes her side but i have never been sure t
hat this character deserves that kind of devotion from the start effie conducts herself for the most part like an obnoxious egotistical selfcentered diva who is more interested in wha
t everyone else can do for her rather than having much vested interest in the group of which she is a part when she is booted from the group for her unprofessionalism and bad attitude
the charges are more than wellfounded but the stage showfilm seem to think effie should be cut unlimited slack simply because she has a great voice even though the film tries to softe
n some of effies harder edges to make her more likable the charges still stand her story becomes more manipulative by suggesting she should have our further sympathy because she is an
unwed mother struggling to raise her daughter  using the implication that much like the talent card motherhood immediately makes any behavior excusable indeed the only big effort the
film makes to show effies mothering is to tell us about it and then include a scene where she barks at her daughter in the unemployment office insists that the girl has no father and
then refuse to look for gainful employment to support them since singing is all she knows in the hands of a skillful actress the gaps could perhaps have been remedied with technique a
nd charisma unfortunately hudson is not that actress she sings well but the dialogdriven moments do not come naturally to her nor do high emotional moments effies signature moment the
aforementioned and i am telling you number is wellsung by hudson but emotionally flat in the acting department effie is supposed to expressing her rage and desperation at her predicam
ent but hudson comes off as a cabaret performer belting out a hot number all in all not quite the emotional highlight one expects the latter portion of the film is basically a predict
able melange of events that maneuver foxx into hudsons earlier position and allow her to strut back in and lord it over everyone foxxs criminal offenses in the film are undoubtedly pa
r for the course of many struggling record producers but the films seeming implication that he has it coming because he helped usher in the disco era is rather ridiculous not to menti
on pretentious and condescending particularly coming from a film with all of the depth of a puddle the end result is a faithful rendition of the stage hit drained of emotion energy or
anything that can be described as dynamic <'
```

On performing a number of experiments, ranging from LSTMs to the State-of-the-Art BERT, it is observed that all of the models have overfit the training data. As it can clearly seen in the above image, the reviews are quite complex in structure.

It is our assumption that the Deep learning approaches would perform fairly well on a dataset like this. If modelling mistakes are to be ruled out, we can have a look at the training data itself by pulling out a few samples and review them to check if the labels that they have been given are consistent with human estimation.

We extract two sets of samples. The first set of samples would be the case where the ground truth label is consistent with all the labels given by all the models. The second set comprises samples where the ground truth is the opposite of what all the models have predicted. For example, if the ground truth is 1 and all the models say it's a 0 and vice-versa.

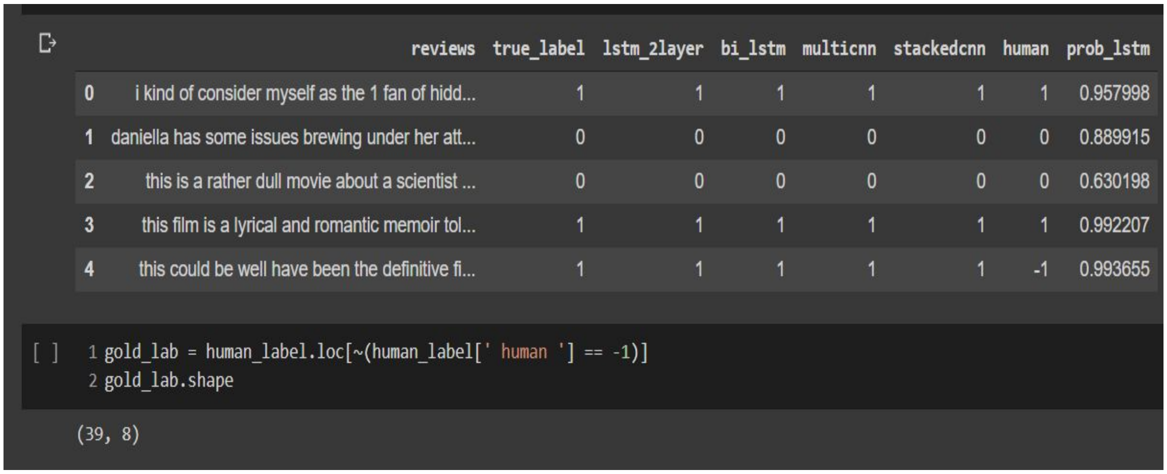
Each bin would consist of 30 examples and all of the 60 reviews would be manually re-annotated by a human. If the review is truly positive, it would be given a value of 1 and 0 in case of a negative example. In case the review is not clear in its sentiment, we flag the review with a -1.

In the first set of examples, where the model's prediction and the ground truth labels are consistent. Manual labels also seemed to be consistent with the ground truths, except for 4 instances where the review was mixed. That results in 26/30 samples as gold.

In the second set of confusing examples. It proved to be difficult for even a human to annotate the reviews, with 17/30 reviews turning out to be mixed in nature.

So out of the 60 samples pulled out for human labelling, 21 proved to be mixed. That's a lot of mixed reviews. Hence it explains the overfitting of all the different models. During training, the model is confused and hence cannot ascertain when it sees the test examples.

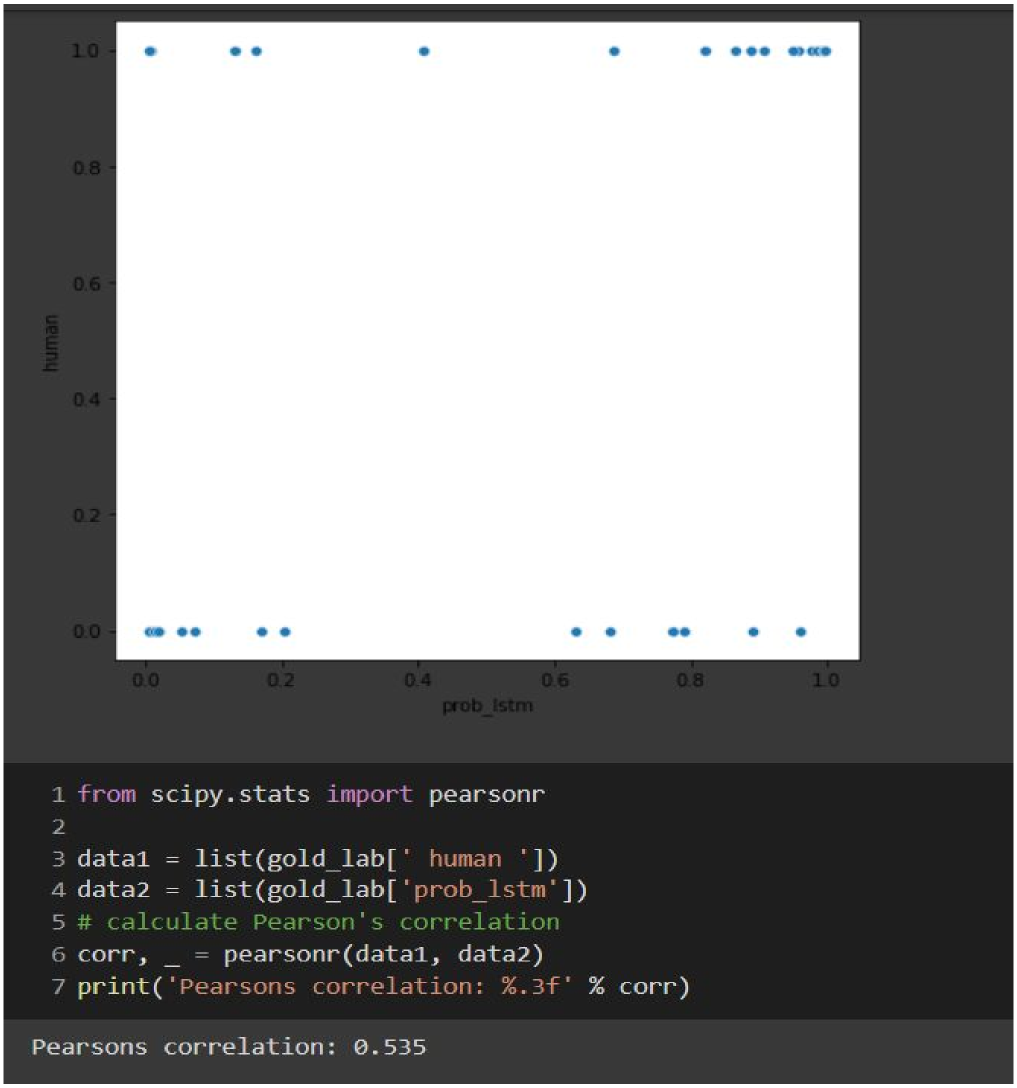
Now we would try to find out how the model is performing in cases where the ground truth labels are legitimate. To assess that, we choose any one of the models and then use it to predict the probability of a review is positive or negative. A higher probability score means the review is positive and vice-versa.
Then we would plot probability score vs human label to see how correlated the two variables are. A higher correlation shows the model is performing well in case of gold training examples.



	reviews	true_label	lstm_2layer	bi_lstm	multicnn	stackedcnn	human	prob_lstm
0	i kind of consider myself as the 1 fan of hidd...	1	1	1	1	1	1	0.957998
1	daniella has some issues brewing under her att...	0	0	0	0	0	0	0.889915
2	this is a rather dull movie about a scientist ...	0	0	0	0	0	0	0.630198
3	this film is a lyrical and romantic memoir tol...	1	1	1	1	1	1	0.992207
4	this could be well have been the definitive fi...	1	1	1	1	1	-1	0.993655

```
[ ]  gold_lab = human_label.loc[~(human_label[' human '] == -1)]
     gold_lab.shape

(39, 8)
```



```python
from scipy.stats import pearsonr

data1 = list(gold_lab[' human '])
data2 = list(gold_lab['prob_lstm'])
# calculate Pearson's correlation
corr, _ = pearsonr(data1, data2)
print('Pearsons correlation: %.3f' % corr)
```

```
Pearsons correlation: 0.535
```

The correlation between the model probabilities and human labels turned out to be about 0.53 which is a high value, as values for strong positive correlation fall between 0.5 and 1. This correlation would have been still higher if the number of confusing examples was minimal.

6. Conclusion

The presence of confusing examples hinders the deep learning models to learn the nuances of the reviews, hence creating a bottleneck in the performance of the deep learning models.

	Training accuracy	Validation accuracy
Stacked LSTM	93.74	85.14
Bi-LSTM	97.31	85.71
LSTM with attention	95.01	86.82
multi-CNN	94.38	85.18

The performance of the above models couldn’t greatly outperform that of the statistical NLP approach. Thus it is recommended to go ahead with the initial Machine Learning approach. Plus the traditional ML model is lightweight so it comes out to be a more feasible solution for the problem.

For future work, we can try performing the same set of experiments with a cleaner dataset to observe how the deep learning approaches fare against the traditional approach .