A STUDY OF DEEP LEARNING MODELS USED FOR TEXT CLASSIFICATION

Y. Sri Lalitha , Shaik Abdul Hameed

Department of Information Technology

Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad

ABSTRACT

Text is a remarkably rich source of information. However, because text is unstructured, it can be challenging and time-consuming to extract information from it. Content tagging, news categorization, spam detection, comment classification, topic recognition, and language detection are a few of the research areas in text classification (for example, knowing if an incoming support ticket is written in English or Spanish). Tickets are immediately routed to the team in Spanish. To categorise the text, either manual annotation or automatic tagging might be employed. With more text data being used in industrial applications, automatic text classification is becoming increasingly crucial. This reveals some clever NLP-based word processing tools.

Machine learning is a self-learning engine that can be enabled by developers and employs a training set and some initial data inputs. Deep learning has become a new subset of machine learning as a result of recent developments in the field. Deep learning is the use of deep artificial neural networks. A neural network has many layers, and the intrinsic layers that are closest to the data it contains are referred to as "deep" layers. There are various deep learning model types available for text classification, including:

Convolutional neural networks (CNN) and RNN (recurrent neural network)

- LSTM (long term memory)

We will investigate the CNN and LSTM models for IMDB movie reviews in this research.

INTRODUCTION

Sentiment analysis is a method that uses computational linguistics, text analysis, and natural language processing to extract and identify subjective information in source materials. Its goal is to ascertain the general contextual polarity of the document as well as the speaker's or writer's feelings toward a subject. In the preceding decade, the emergence of social networks including blogs and social networks sparked interest in sentiment analysis. Online feedback has evolved into a type of virtual currency for businesses wanting to market their goods, attract new customers, and uphold their reputation. This is due to the growth of data from reviews, ratings, ideas, and other expressions online. Many individuals automate noise removal, language understanding, content selection, and pertinent actions by using sentiment analysis. In this research, we develop text categorization models using CNN and LSTM and evaluate their performance in terms of accuracy.

LITERATURE SURVEY

Machine Learning based classification for Sentimental analysis of IMDb reviews

The purpose of this paper is to look at the emotional representations of IMDb assessments utilising machine learning-based assessments at the document level. To boost ranking performance, the report will first remove noisy phrases and then normalise words in IMDb ranks. The ratings will be translated to a set of words that characterise the assessment's qualities in the following step of the report. Finally, a number of strategies are employed to train and test the word matrix in order to determine which strategy is the most effective for categorizing these words.(logistic regression, SVM, Naive Bayes, random forest).

Logistic Regression

This method of categorization is widely used in the field of generalised linear models. Logistic regression is used to model the likelihood that characterises a test result. Maximum Entropy is another name for this technique. Naive

Bayes, Hidden Markov Models, etc. are examples of generative classification models. Naive Bayes, Hidden Markov Models, etc. are examples of discriminative classification models (Logistic Regression, SVM, etc.). Both models then attempt to determine p (class | features) or p (y | x). A discriminative model models p(y|x) right away while a generative model first attempts to model the joint probability distribution p(x, y) before using Baye's theorem to determine the conditional probability p(y|x).

SVM

Problems involving classification and regression can be resolved using supervised machine learning (SVM). Regression predicts a continuous value, whereas classification predicts a label or group. The SVM locates the hyperplane that divides the classes that we display in n-dimensional space and then classifies the data. SVM creates this hyperplane by applying "Kernel" mathematical functions to our data. In addition to linear and sigmoidal kernels, there are also nonlinear, polynomial, RBF, and other varieties. For separable linear problems, the "linear" kernel is employed. As a result, we use "linear SVM" to solve our linear problem (positive and negative values only).

Naive Bayes

Naive Bayes is the most straightforward and quick categorization algorithm for big amounts of data. Spam filtering, text categorization, sentiment analysis, and recommendation systems have all used the Naive Bayes classifier. The Bayes probability theorem can be used to predict an unknown class. The Naive Bayes scoring technique in machine learning offers simple but efficient scoring operations. The naïve Bayesian classification is based on applying Bayes' theorem with a high assumption of feature independence. Naive Bayes classification performs well when used for textual data analysis, such as natural language processing. Other names for Bayesian ship models are independent Bayesian models and simple Bayesian models.

Random Forest

The random forest methodology is an extension of the bagging method that creates a forest of disjointed decision trees utilising resource pooling and randomization. In order to produce a random set of features, feature randomization, sometimes referred to as "feature packing" or "random space approach," ensures that decision trees have the least amount of association. Decision trees and random forests differ significantly from one another. Decision trees take into account all potential subsets, whereas random forests just choose a subset of traits.

The most significant feature of the Random Forest algorithm is its propensity to change data sets with continuous and categorical variables, which is useful in regression and classification. It performs better than its rivals when it comes to classifying assets.

PROPOSED WORK

Deep learning is a form of artificial intelligence that mimics the human brain's data processing and decision-making processes. Deep learning is a subset of machine learning artificial intelligence in which networks are used to interpret unstructured or unsigned material without the assistance of humans. Another name for this is deep convolutional neural learning.

For text classification, the suggested method employs two Deep Learning models. The IMDB movie review rating is used to evaluate these deep learning models.

CNN

CNN is a particular kind of neural network, but what sets it apart is its convolutional layer. CNN analyses the pixel array's angles, vectors, and dimensions. As a result, CNN can handle data in matrix form better and can access all of the matrix's characteristics. Edge detection, corner detection, and texture warping are some of the features of convolutional layers, a different tool for CNN modelling. This layer can recognise every feature of the image by navigating through its matrix. This demonstrates how the network's convolutional layers can recognise increasingly complex features. The size of the convolutional layer must increase as the function does. Sequential data, time series data, and one-dimensional arrays are examples of text data. We'll use a one-dimensional convolution layer for this. The model's principle is nearly identical, but the convolution layers' data type and size have altered. To use Text CNN, you'll need a word embedding layer and a one-dimensional convolutional network.

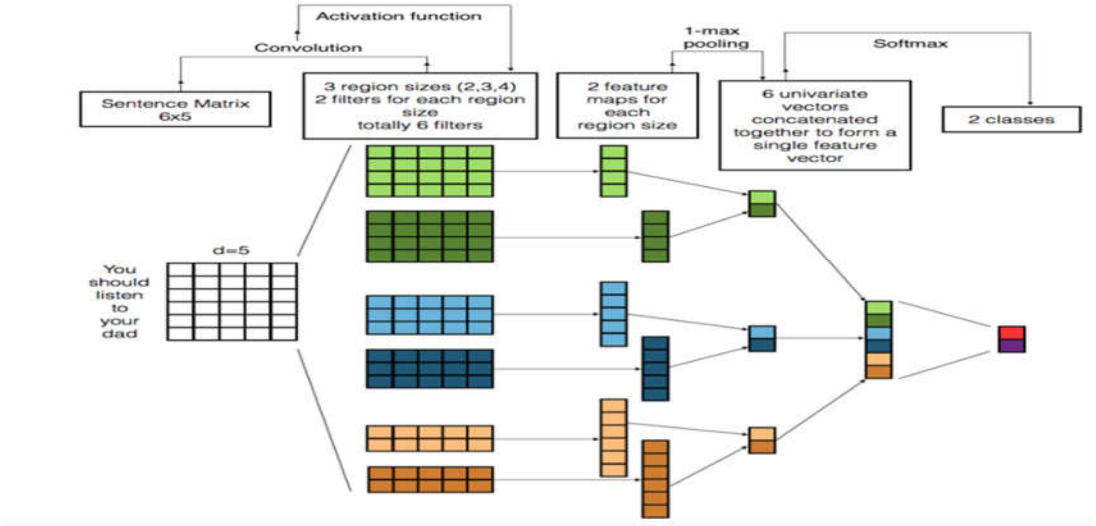
LSTM

An example of a deep learning recurrent artificial neural network is long-term memory. Unlike traditional feedback neural networks, LSTM has feedback loops.
In contrast to a traditional recurrent neural network, which recalls all data, LSTM only retrieves the most recent data. A more advanced recurrent neural network that overcomes the gradient dispersion issue is the LSTM. It has a memory cell at the top that facilitates effective data transmission from one location to another. As a result, it doesn't have the dispersion gradient issue and keeps a lot of data from earlier RNN states. Valve technology allows for the addition of data to or removal of data from memory cells.

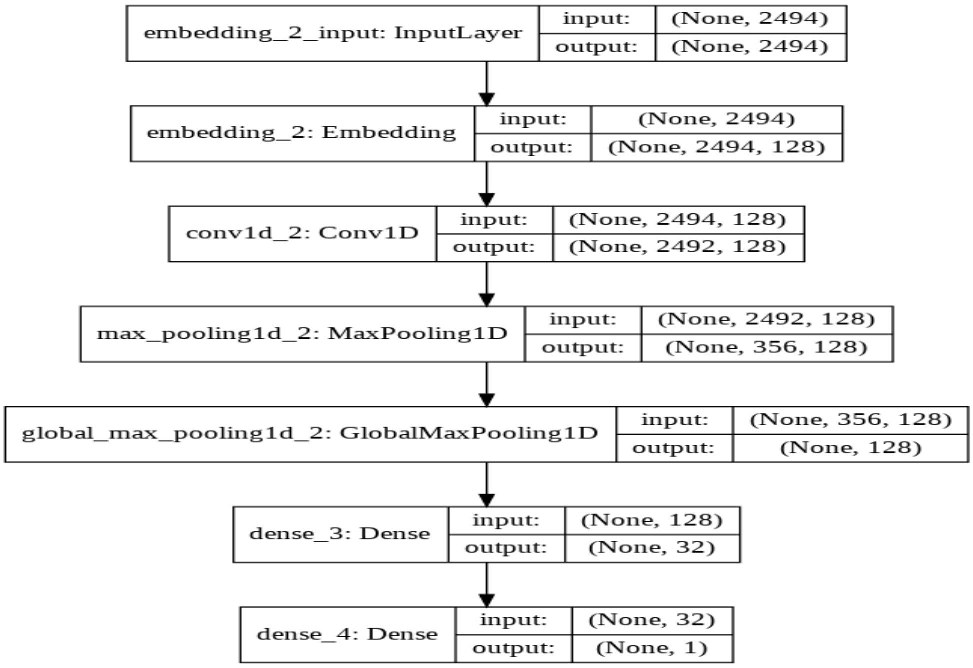
METHADOLOGY

Convolutional neural network

A convolutional neural network (CNN) is a deep learning system that can collect an input image, prioritise and differentiate between different features within it.


Activation function
Convolution
1-max pooling
Softmax
Sentence Matrix 6x5
3 region sizes (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
You should listen to your dad

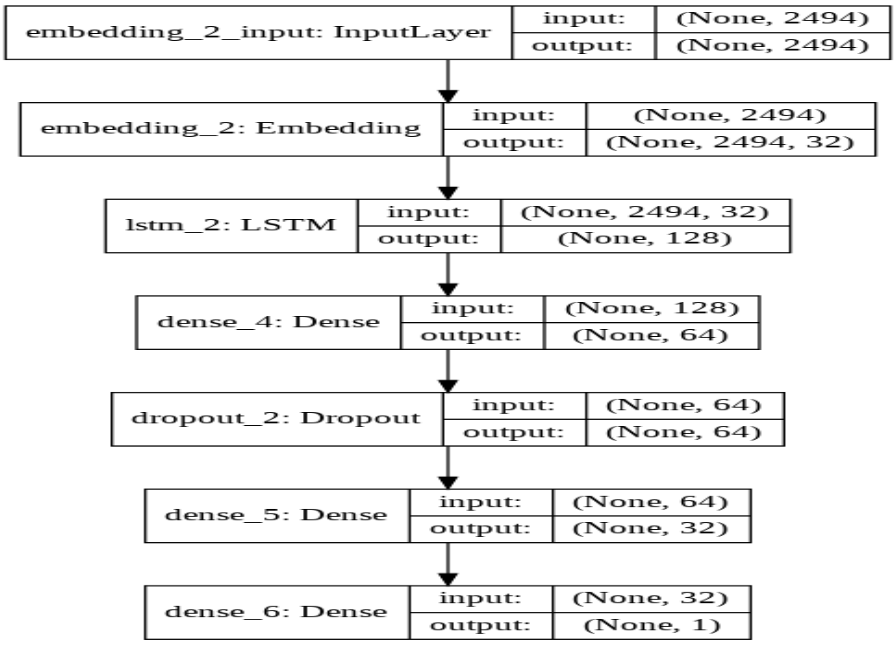
Architecture



LSTM

When it comes to memory, long-term memory (LSTM) is a sort of recurrent neural network that performs better than conventional recurrent neural networks. When they have sufficient control over the storage of some models, LSTMs perform noticeably better. Similar to any other NN, the LSTM can have several hidden layers. As it traverses each layer, unnecessary information is discarded while significant information is saved in each cell.

Architecture

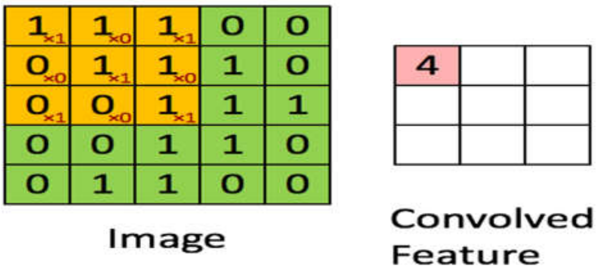


Convolutional neural network

When a convolutional neural network processes an image, different weights are assigned to different regions of the image to identify them.

ConvNet requires much less preprocessing than standard classification techniques. They are created manually using simple methods, but with enough training, ConvNets can learn how to create these filters.

The visual cortex's structure, which is akin to the concept of a neural network in the human brain, inspired ConvNet. Individual neurons only respond to changes in the receptive field, which is a small portion of the visual fields.



1 ×1	1 ×0	1 ×1	0	0
0 ×0	1 ×1	1 ×0	1	0
0 ×1	0 ×0	1 ×1	1	1
0	0	1	1	0
0	1	1	0	0

4		

Padding

Padding around the input frequently prevents the function map from shrinking. If padding is not given, resource maps with a large number of input items will begin to shrink, resulting in relevant information being lost out of boundaries. Low-level elements like as borders, colour, gradient direction, and so on are generally acquired in the first level of convolution. As more layers are added, the design accommodates higher-level functions, resulting in a network that embeds photos into the dataset in the same way we do.

Pooling Layer

The pooling level is in charge of decreasing the spatial dimension of the convolutional element, just like the convolutional level. The processing power needed to process the data is decreased by lowering its dimensionality. Extraction of the principal position and rotation invariant functions, which support model training, is also advantageous.

The two types of clustering are maximum clustering and average clustering. The maximum value of the portion of the image that the Kernel covers is returned by Max Pooling. On the other hand, average pooling gives the image's central region's average value.

Dense Layer

A layer that is densely linked to the layer below it has all of its neurons coupled to all of the neurons in the layer below it. The layer that artificial neural networks employ the most is this one. According to one concept, each neuron in the dense layer receives its output before performing matrix vector multiplication. The row vector of the output from the preceding layers is equivalent to the column vector of the dense layer by multiplying the matrix vector. When multiplying vector matrices, the row vector must have the same number of columns as the column vectors.

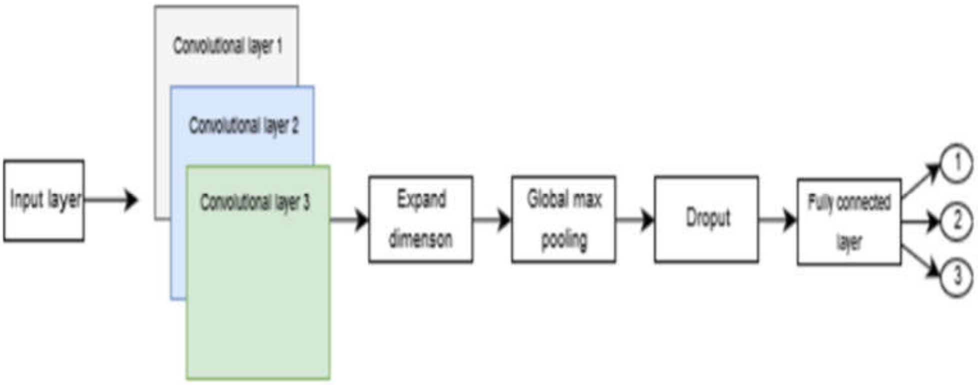
Dropout Layer

Another characteristic of CNNs is their level of carelessness. The Dropout layer is a mask that prevents some neurons from contributing to the next level while allowing others to do so. To remove part of the functions from the input vector or to remove some of the hidden neurons, we can apply a dropout layer or a hidden layer. Dropout rates are important in CNN training because they prevent the training data from being overfitted. In the absence of the first batch of training samples, learning is disproportionately affected. As a result, it would be unable to learn traits that only appear in later examples.

Fully Connected Layer

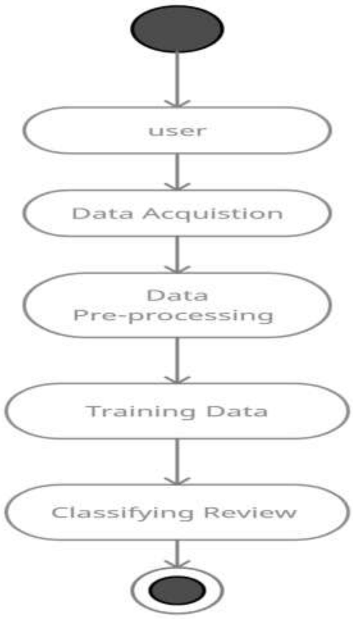
Learning nonlinear combinations of high-level functionality represented by convolutional-level outputs using a fully connected layer is (usually) a useful technique. You are learning a fictitious model of a nonlinear function in that space on a fully connected level.

After changing the shape to fit our multilayer perceptron, let's flatten the image into a column vector. Backpropagation takes place in each training cycle following the delivery of the flat output to a feedback neural network.



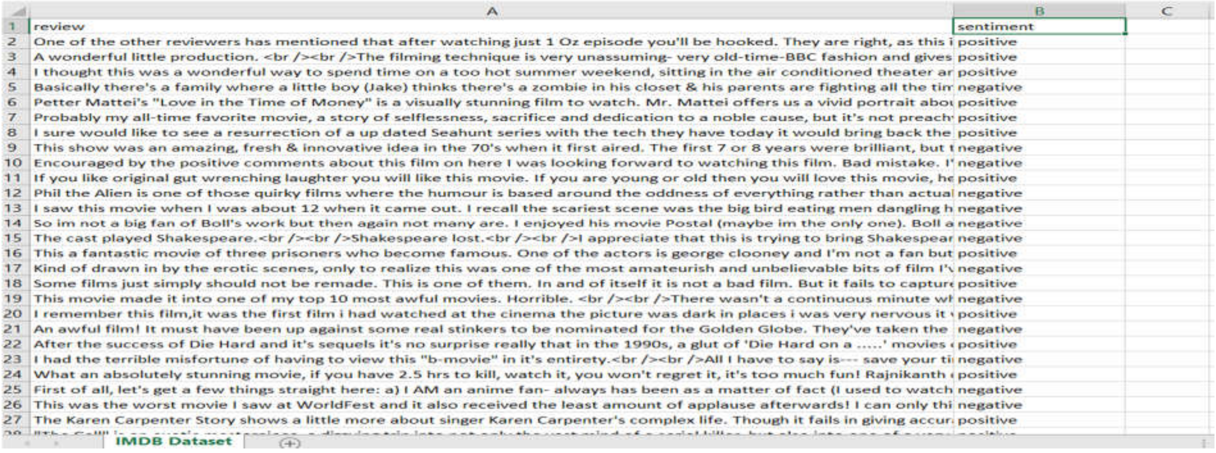
Project Architecture

This project's architecture depicts the system's flow and refers to the layers that were used in this case. Simulate an activity using diagrams of sequential and simultaneous actions. In a nutshell, we use an active schema to visually depict workflows. The schema operation is in the same state as when it was deleted and in the same order as when it was removed. We utilise the activity of a pattern to express or reflect the causes of a specific event.



Dataset

This is a binary sentiment evaluation dataset that contains far more information than prior benchmark datasets. We provide a series of 25,000 highly polarised film reviews for training and testing. There's also data that hasn't been classified. Raw text formats and pre-processed word packs are available in two forms. The pictures that served as a database for this search are depicted in the diagram below.



	A	B	C
1	review	sentiment	
2	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. They are right, as this i	positive	
3	A wonderful little production. <br /><br />The filming technique is very unassuming- very old-time-BBC fashion and gives	positive	
4	I thought this was a wonderful way to spend time on a too hot summer weekend, sitting in the air conditioned theater ar	positive	
5	Basically there's a family where a little boy (Jake) thinks there's a zombie in his closet & his parents are fighting all the tim	negative	
6	Petter Mattei's "Love in the Time of Money" is a visually stunning film to watch. Mr. Mattei offers us a vivid portrait abou	positive	
7	Probably my all-time favorite movie, a story of selflessness, sacrifice and dedication to a noble cause, but it's not preachy	positive	
8	I sure would like to see a resurrection of a up dated Seahunt series with the tech they have today it would bring back the	positive	
9	This show was an amazing, fresh & innovative idea in the 70's when it first aired. The first 7 or 8 years were brilliant, but t	negative	
10	Encouraged by the positive comments about this film on here I was looking forward to watching this film. Bad mistake. I'	negative	
11	If you like original gut wrenching laughter you will like this movie. If you are young or old then you will love this movie, he	positive	
12	Phil the Alien is one of those quirky films where the humour is based around the oddness of everything rather than actual	negative	
13	I saw this movie when I was about 12 when it came out. I recall the scariest scene was the big bird eating men dangling h	negative	
14	So im not a big fan of Boll's work but then again not many are. I enjoyed his movie Postal (maybe im the only one). Boll a	negative	
15	The cast played Shakespeare.<br /><br />Shakespeare lost.<br /><br />I appreciate that this is trying to bring Shakespear	negative	
16	This a fantastic movie of three prisoners who become famous. One of the actors is george clooney and I'm not a fan but	positive	
17	Kind of drawn in by the erotic scenes, only to realize this was one of the most amateurish and unbelievable bits of film I'v	negative	
18	Some films just simply should not be remade. This is one of them. In and of itself it is not a bad film. But it fails to capture	positive	
19	This movie made it into one of my top 10 most awful movies. Horrible. <br /><br />There wasn't a continuous minute wh	negative	
20	I remember this film,it was the first film i had watched at the cinema the picture was dark in places i was very nervous it v	positive	
21	An awful film! It must have been up against some real stinkers to be nominated for the Golden Globe. They've taken the	negative	
22	After the success of Die Hard and it's sequels it's no surprise really that in the 1990s, a glut of 'Die Hard on a .....' movies c	positive	
23	I had the terrible misfortune of having to view this "b-movie" in it's entirety.<br /><br />All I have to say is--- save your ti	negative	
24	What an absolutely stunning movie, if you have 2.5 hrs to kill, watch it, you won't regret it, it's too much fun! Rajnikanth c	positive	
25	First of all, let's get a few things straight here: a) I AM an anime fan- always has been as a matter of fact (I used to watch	negative	
26	This was the worst movie I saw at WorldFest and it also received the least amount of applause afterwards! I can only thi	negative	
27	The Karen Carpenter Story shows a little more about singer Karen Carpenter's complex life. Though it fails in giving accur	positive	

IMDB Dataset

RESULTS AND COMPARISION

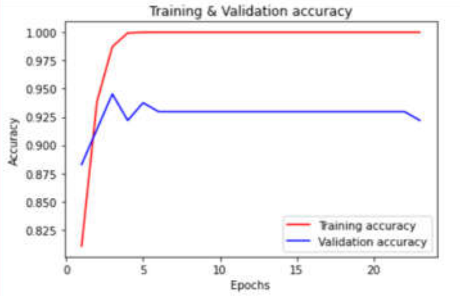
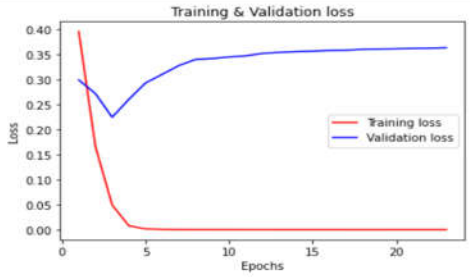
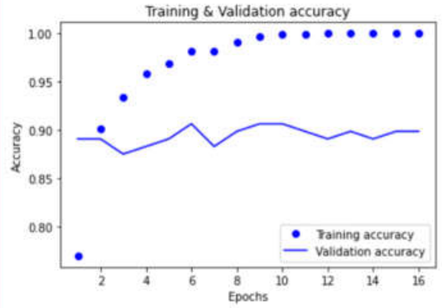
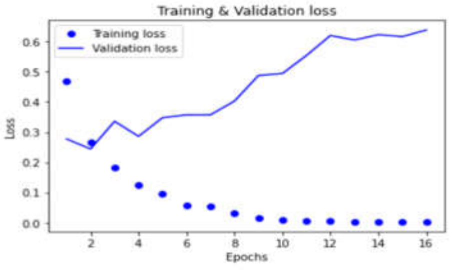
MODELS	TRAINING AND VALIDATION ACCURACY	TRAINING AND VALIDATION LOSS
CNN	 	
LSTM	 	

Table contains Training and validation Accuracy values and Loss values of CNN and LSTM models

Model	Accuracy
CNN	89%
LSTM	85%

Table contains Accuracy values of CNN and LSTM models.

CONCLUSION

Techniques of sentiment analysis are one of the most essential foundations in decision-making. Many individuals rely on sentiment analysis to determine which products or services are most beneficial. In order to produce IMBD movie reviews, we begin with a reasonable model.

The CNN and LSTM models attained test accuracy of 89 percent and 85 percent, respectively, in evaluation tests. This demonstrates that in the sentiment analysis of movie reviews, the CNN model outperformed the LSTM.

Convolutional neural networks can be used to assess any type of data in which surrounding data is likely to be significant. CNNs have a rigid and advanced structure, yet they are fairly limiting.

Because an LSTM is frequently used to process and forecast given data streams, it is designed to perform differently than a CNN. A CNN, on the other hand, is meant to look for "spatial correlation" in data.

If you want to perform things like:

- categorization in strings of varying lengths (from N to 1)

- try to make another string that doesn't have a set length to input length ratio (from N to M).

Neural networks with only one forward direction will fail (due to size inconsistency). In this instance, a recursive neural network, such as an LSTM, is required.

FUTURE ENHANCEMENT

These models are giving an accuracy of 85% - 89% which is good in sentiment analysis. So, even though the accuracy is good, the efficiency of the model needs to be still improved. To improve the accuracy, I would like to implement RNN algorithm sequentially, in future, we want to compare the present model with RNN and LSTM models.

REFERENCES

1. Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method

2. Hybrid Deep Learning Models for Sentiment Analysis

3. https://www.d2l.ai/chapter_natural-language-processing- applications/sentiment-analysis-cnn.html

4. http://www.diva-portal.org/smash/get/diva2:1105494/FULLTEXT01.pdf

5. https://towardsdatascience.com/cnn-sentiment-analysis-9b1771e7cdd6