MOVIE REVIEW

A Project Report in partial fulfillment of the degree

Bachelor of Technology

in

Electronics & Communication Engineering/Computer Science & Engineering

By

19K41A04G9	NADA TAHANI
19K41A05G6	R. SUDHAN JEE
19K41A05G7	R. AKSHITHA

Under the Guidance of

D. RAMESH

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
S R ENGINEERING COLLEGE, ANANTHASAGAR, WARANGAL
(Affiliated to JNTUH, Accredited by NBA)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled **"Movie Review"** is a record of bonafide work carried out by the student(s) Nada Tahani, R. Sudhan Jee, R. Akshitha bearing Roll No(s) 19K41A04G9, 19K41A05G6, 19K41A05G7 during the academic year 2022-23 in partial fulfillment of the award of the degree of ***Bachelor of Technology*** in **Electronics & Communication /Computer Science Engineering** by the Jawaharlal Nehru Technological University, Hyderabad.

.

Supervisor **Head of the Department**

External Examiner

ABSTRACT

Movie reviews play a major role not just because they are important or interesting it's because they influence people's mindsets. People chose movies based on these ratings or reviews. So, these ratings or reviews should be accurate or well-aimed. So, we will develop an application that collects tweets related to a movie from Twitter and prepares a dataset. We will send this dataset to a trained LSTM model. This model gives us a polarity of these tweets. Based on the polarity we calculate the average of the polarity of these statements. This average ranges from 0-5. This rating will be available to the user in the form of a web page.

Table of Contents

1. INTRODUCTION

Humans are subjective creatures and their opinions are important because they reflect their satisfaction with products, services, and available technologies. Being able to interact with people on that level has many advantages for information systems; such as enhancing product quality, adjusting marketing and business strategies, improving customer services, managing crises, and monitoring performances.

A movie review is an article reflecting its writers' opinion about a certain movie and criticizing it positively or negatively, which enables everyone to understand the overall idea of that movie and make the decision whether to watch it or not. A movie review can affect the whole crew who worked on that movie. A study illustrates that in some cases, the success or failure of a movie depends on its reviews. Therefore, a vital challenge is to be able to classify movie reviews to capture, retrieve, quantify and analyze watchers more effectively.

Nowadays, if you want to a successful business, it is very important to act according to your viewers' comments. When we look at today's most prominent and successful companies like Amazon or Netflix, we can see that they are the companies that use data best and know their customers best. These days, people have become smart enough to read or watch a movie review investing their money in a ticket. Movies shape the minds of many. Film reviewing is a creative job but is also a responsible one. A film critic cannot give biased opinions. film reviewers can manipulate the audience. Movie review classification into positive or negative reviews is connected with words occurrences from the text of the review, and whether those words have been used before in a positive or a negative context. These factors help enhance the review understanding process.

Before seeing a movie, we read public reviews. Priority is given to this movie rating. We will thus develop a service that gives a movie review. Based on the public tweets, this project rates the movies. So, we will develop an application that collects tweets related to a movie from Twitter and prepares a dataset. We will send this dataset to a trained LSTM model. We gather tweets on a certain movie and determine the polarity of each one. This model gives us a polarity of these tweets. The rating for the film is between 0 and 5, depending on the polarity. The movie's rating will be made publicly available through a web page.

2. LITERATURE REVIEW

[1] This paper includes the details of two proposed deep learning architectures CNN-LSTM and LSTM-CNN method. The proposed system uses IMDB movie review data set which contains 1000 positive reviews and 1000 negative reviews. For training and validation, the full training examples have been arbitrarily split. Input dataset divided into two, training dataset and the validation dataset. Keras provide two methods to evaluate deep learning model. The first one is automatic verification of dataset and second is manual verification of dataset. Keras separates one portion of training data into validation data and then assesses the performance of model on that validation dataset on every epoch.

[2] Micro-blog has become an important place for people to talk , Sentiment analysis in the application of mass data will help to improve the Internet public opinion monitoring system. Therefore, the research scheme proposed in this paper is the use of deep learning CNN to avoid the explicit feature extraction, and implicitly learned from the training data. The practice proves that the deep learning method is feasible to improve the accuracy of emotion analysis.

[3] Sentiment Analysis (SA) is the task of inferring polarity of an opinion in a text. Though most of the work in SA is for English, there has been work in other languages as well such as Chinese, Japanese, German and Spanish). To perform SA on these languages, cross-lingual approaches are often used due to the lack of annotated content in these languages. In Cross-Lingual Sentiment Analysis (CLSA), the training corpus in one language (call it Ltrain) is used to predict the sentiment of documents in another language (call it Ltest). Machine Translation is often employed for CLSA. A document in Ltest is translated into Ltrain and is checked for polarity using the classifier trained on the polarity marked documents of Ltrain. This paper presented an approach to cross-lingual SA that uses WordNet synset identifiers as features of a supervised classifier. The sense-based approach provides a cross-lingual classification accuracy of 72% and 84% for Hindi and Marathi respectively, which is an improvement of 14% - 15% over the baseline based on a cross-lingual approach using a naïve translation of the training and test corpus.

[4] Presents recent research on Automation Control Theory Perspectives in Intelligent Systems Proceedings of the 5th Computer Science On-line Conference 2016 (CSOC2016), Vol2 Automation Control Theory Perspectives in Intelligent Systems. The proceedings are divided in three volumes Vol1: Artificial Intelligence Perspectives in Intelligent Systems, Volume 2: Automation Control Theory Perspectives in Intelligent Systems, and Volume 3: Software Engineering Perspectives and Application in Intelligent Systems. It contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing.

[5] Of late, most of the research works on SA in natural language processing (NLP) are focused on English language. However, it is noted that Bangla does not have a proper dataset that is both large and standard. In this work, a substantial textual dataset of both Bangla and Romanized Bangla texts have been provided which is first of this kind and post-processed, multiple validated, and ready for Sa implementation and experiments. Further, this dataset has been tested in Deep Recurrent model, specifically, Long Short-Term Memory (LSTM), using two types of loss functions — binary cross-entropy and categorical cross-entropy.

3. DESIGN:

3.1 Requirement Specifications (S/W & H/W)

Hardware Requirements

- ✓ **System** : Processor Intel(R) Core (TM) i5-8265U CPU @ 1.60GHz, 1800 MHz, 4 Cores, 8 Logical Processors
- ✓ **RAM** : 4 GB
- ✓ **Hard Disk** : 557 GB
- ✓ **Input** : Keyboard and Mouse
- ✓ **Output** : PC

Software Requirements

- ✓ **OS** : Windows 10
- ✓ **Deployment software** : Google Colaboratory (Online Compiler)
- ✓ **Program Language** : Python

3.2 Flowchart

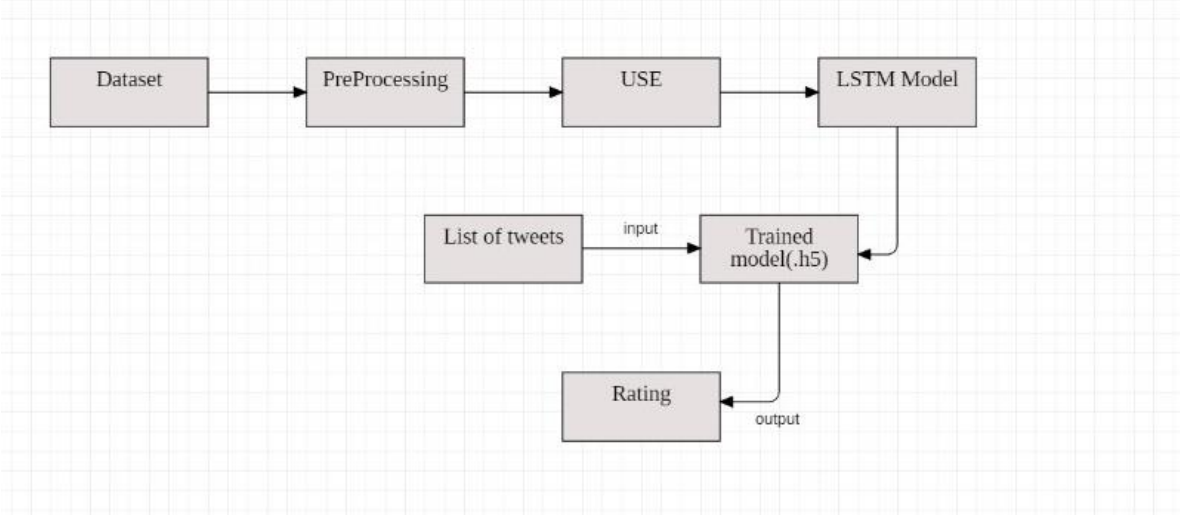


Fig 1: Flow chart of our capstone project. In the above flow chart, we described model workflow in this project

4. DATASET:

We collected the dataset from Kaggle. We use the same dataset for training and testing. The dataset consists of a total of 20,000 texts or statements.

Input features:

Training dataset:

The training dataset consists of 15,000 statements or reviews related to web series and movie. This dataset consists of two columns. The first column consists of text and the second column consists of the polarity of the text.

First column: **Text**

This column includes the statements or comments.

Second column: **Polarity**

This column includes the polarity of the text.

Testing dataset:

The training dataset consists of 5,000 statements or reviews related to a web series and movie. This dataset consists of two columns. Same as the training dataset.

First column: **Text**

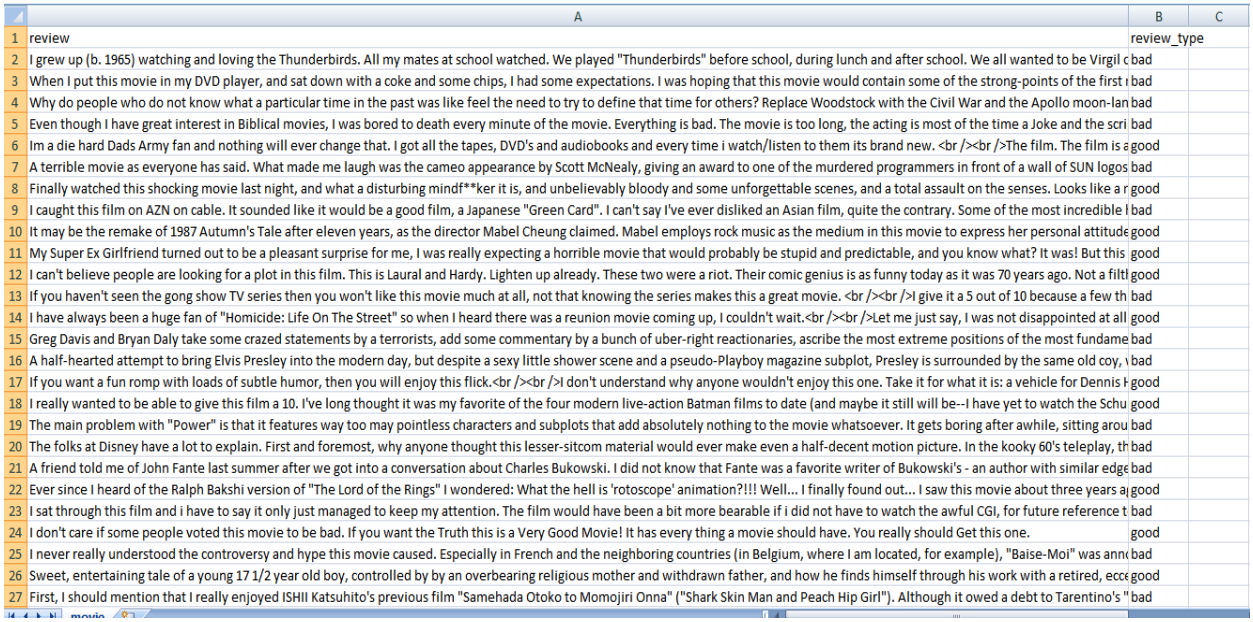
This column includes the statements or comments.

Second column: **Polarity**

This column includes the polarity of the text.

Output feature:

- Rating
- Value: Ranges from 0 to 10.



	A	B	C
1	review	review_type	
2	I grew up (b. 1965) watching and loving the Thunderbirds. All my mates at school watched. We played "Thunderbirds" before school, during lunch and after school. We all wanted to be Virgil c	bad	
3	When I put this movie in my DVD player, and sat down with a coke and some chips, I had some expectations. I was hoping that this movie would contain some of the strong-points of the first i	bad	
4	Why do people who do not know what a particular time in the past was like feel the need to try to define that time for others? Replace Woodstock with the Civil War and the Apollo moon-lan	bad	
5	Even though I have great interest in Biblical movies, I was bored to death every minute of the movie. Everything is bad. The movie is too long, the acting is most of the time a Joke and the scri	bad	
6	Im a die hard Dads Army fan and nothing will ever change that. I got all the tapes, DVD's and audiobooks and every time i watch/listen to them its brand new. The film. The film is a	good	
7	A terrible movie as everyone has said. What made me laugh was the cameo appearance by Scott McNealy, giving an award to one of the murdered programmers in front of a wall of SUN logos	bad	
8	Finally watched this shocking movie last night, and what a disturbing mindf**ker it is, and unbelievably bloody and some unforgettable scenes, and a total assault on the senses. Looks like a r	good	
9	I caught this film on AZN on cable. It sounded like it would be a good film, a Japanese "Green Card". I can't say I've ever disliked an Asian film, quite the contrary. Some of the most incredible l	bad	
10	It may be the remake of 1987 Autumn's Tale after eleven years, as the director Mabel Cheung claimed. Mabel employs rock music as the medium in this movie to express her personal attitude	good	
11	My Super Ex Girlfriend turned out to be a pleasant surprise for me, I was really expecting a horrible movie that would probably be stupid and predictable, and you know what? It was! But this	good	
12	I can't believe people are looking for a plot in this film. This is Laural and Hardy. Lighten up already. These two were a riot. Their comic genius is as funny today as it was 70 years ago. Not a filt	good	
13	If you haven't seen the gong show TV series then you won't like this movie much at all, not that knowing the series makes this a great movie. I give it a 5 out of 10 because a few th	bad	
14	I have always been a huge fan of "Homicide: Life On The Street" so when I heard there was a reunion movie coming up, I couldn't wait. Let me just say, I was not disappointed at all	good	
15	Greg Davis and Bryan Daly take some crazed statements by a terrorists, add some commentary by a bunch of uber-right reactionaries, ascribe the most extreme positions of the most fundame	bad	
16	A half-hearted attempt to bring Elvis Presley into the modern day, but despite a sexy little shower scene and a pseudo-Playboy magazine subplot, Presley is surrounded by the same old coy,	bad	
17	If you want a fun romp with loads of subtle humor, then you will enjoy this flick. I don't understand why anyone wouldn't enjoy this one. Take it for what it is: a vehicle for Dennis H	good	
18	I really wanted to be able to give this film a 10. I've long thought it was my favorite of the four modern live-action Batman films to date (and maybe it still will be--I have yet to watch the Schu	good	
19	The main problem with "Power" is that it features way too may pointless characters and subplots that add absolutely nothing to the movie whatsoever. It gets boring after awhile, sitting arou	bad	
20	The folks at Disney have a lot to explain. First and foremost, why anyone thought this lesser-sitcom material would ever make even a half-decent motion picture. In the kooky 60's teleplay, th	bad	
21	A friend told me of John Fante last summer after we got into a conversation about Charles Bukowski. I did not know that Fante was a favorite writer of Bukowski's - an author with similar edge	bad	
22	Ever since I heard of the Ralph Bakshi version of "The Lord of the Rings" I wondered: What the hell is 'rotoscope' animation?!!! Well... I finally found out... I saw this movie about three years a	good	
23	I sat through this film and i have to say it only just managed to keep my attention. The film would have been a bit more bearable if i did not have to watch the awful CGI, for future reference t	bad	
24	I don't care if some people voted this movie to be bad. If you want the Truth this is a Very Good Movie! It has every thing a movie should have. You really should Get this one.	good	
25	I never really understood the controversy and hype this movie caused. Especially in French and the neighboring countries (in Belgium, where I am located, for example), "Baise-Moi" was ann	bad	
26	Sweet, entertaining tale of a young 17 1/2 year old boy, controlled by by an overbearing religious mother and withdrawn father, and how he finds himself through his work with a retired, ecce	good	
27	First, I should mention that I really enjoyed ISHII Katsuhito's previous film "Samehada Otoko to Momojiri Onna" ("Shark Skin Man and Peach Hip Girl"). Although it owed a debt to Tarentino's "	bad	

Fig 2: Dataset

5. DATA PREPROSSESSING:

We have used python inbuilt model i.e re and universal sentence encoder to complete this process, where the main goal is to convert the data into a vector, perform embedding on it. These are the steps taken for data Pre-processing.

➢ **Text cleaning**

In any machine learning task, cleaning or pre-processing the data is as important as a model building if not more and when it comes to unstructured data like text, this process is even more important. The objective of this kernel is to understand the various text pre-processing steps with code examples. Some of the common text pre-processing / cleaning steps are:

- Lower casing
- Removal of Punctuations
- Removal of Stop words
- Removal of Frequent words
- Removal of Rare words
- Stemming
- Lemmatization
- Removal of emojis
- Removal of emoticons
- Conversion of emoticons to words
- Conversion of emojis to words
- Removal of URLs
- Removal of HTML tags
- Chat words conversion
- Spelling correction

	review	review_type
0	I grew up (b. 1965) watching and loving the Th...	bad
1	When I put this movie in my DVD player, and sa...	bad
2	Why do people who do not know what a particula...	bad
3	Even though I have great interest in Biblical ...	bad
4	Im a die hard Dads Army fan and nothing will e...	good
...	...	...
19995	I was required to watch the movie for my work,...	bad
19996	"White Noise" had potential to be one of the m...	bad
19997	The Five Deadly Venoms is a great kung-fu acti...	good
19998	Ali G Indahouse has got to be one of the funni...	good
19999	I found myself at sixes and sevens while watch...	good

20000 rows × 2 columns

Fig 3: Before Pre-processing

	review_type	review
0	bad	i grew up b 1965 watching and loving the thund...
1	bad	when i put this movie in my dvd player and sat...
2	bad	why do people who do not know what a particula...
3	bad	even though i have great interest in biblical ...
4	good	im a die hard dads army fan and nothing will e...
...	...	...
19995	bad	i was required to watch the movie for my work ...
19996	bad	white noise had potential to be one of the mo...
19997	good	the five deadly venoms is a great kung - fu ac...
19998	good	ali g indahouse has got to be one of the funni...
19999	good	i found myself at sixes and sevens while watch...

20000 rows × 2 columns

Fig 4: After Pre-processing

6. METHODOLOGY:

This section talks about the Universal sentence encoder and LSTM models used for the project.

❖ UNIVERSAL SENTENCE ENCODER

The universal sentence encoder makes looking up embeddings at the sentence level as simple as it has previously been to look up embeddings at the word level. Then, using less supervised training data, the sentence embeddings can be easily employed to compute sentence-level meaning similarity and improve performance on subsequent classification tasks. The universal sentence encoder model converts textual information into numerically represented, high-dimensional vectors called embeddings. It aims to transfer learning, especially to other NLP tasks like text categorization, semantic similarity, and clustering. The freely accessible universal sentence encoder is listed in the Tensor flow hub. To learn for a wide range of jobs, it is trained on a number of data sources.

On a high level, the idea is to design an encoder that summarizes any given sentence to a 512-dimensional sentence embedding. We use this same embedding to solve multiple tasks and based on the mistakes it makes on those, we update the sentence embedding. Since the same embedding has to work on multiple generic tasks, it will capture only the most informative features and discard noise. The intuition is that this will result in a generic embedding that transfers universally to a wide variety of NLP tasks such as relatedness, clustering, paraphrase detection, and text classification.

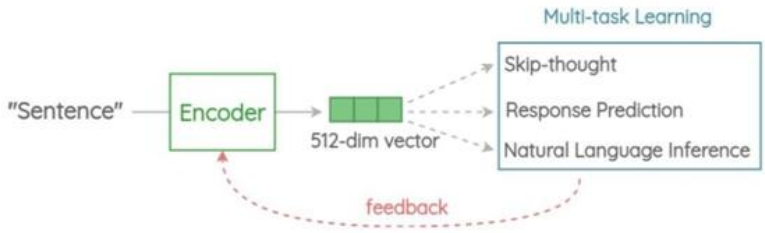


Fig 5: Architecture

Encoder:

This is the component that encodes a sentence into fixed-length 512-dimension embedding. In the paper, there are two architectures proposed based on trade-offs in accuracy vs inference speed.

Variant 1: Transformer Encoder

In this variant, we use the encoder part of the original transformer architecture. The architecture consists of 6 stacked transformer layers. Each layer has a self-attention module followed by a feed-forward network. The self-attention process takes word order and surrounding context into account when generating each word representation. The output context-aware word embeddings are added element-wise and divided by the square root of the length of the sentence to account for the sentence-length difference. We get a 512-dimensional vector as output sentence embedding. This encoder has better accuracy on downstream tasks but higher memory and computes resource usage due to complex architecture. Also, the compute time scales dramatically with the length of the sentence as self-attention has O(n2) time complexity with the length of the sentence. But for short sentences, it is only moderately slower.

Variant 2: Deep Averaging Network (DAN)

In this simpler variant, the encoder is based on the architecture. First, the embeddings for words and bi-grams present in a sentence are averaged together. Then, they are passed through a 4-layer feed-forward deep DNN to get 512-dimensional sentence embedding as output. The embeddings for word and bi-grams are learned during training. It has slightly reduced accuracy compared to the transformer variant, but the inference time is very efficient. Since we are only doing feedforward operations, the compute time is of linear complexity in terms of the length of the input sequence.

❖ LSTM

Long short-term memory (LSTM) is an artificial neural network used in the fields of artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Such a recurrent neural network (RNN) can process not only single data points (such as images) but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition, machine translation, robot control, video games, and healthcare. LSTM has become the most cited neural network of the 20th century.

The name LSTM refers to the analogy that a standard RNN has both "long-term memory" and

"short-term memory". The connection weights and biases in the network change once per episode of training, analogous to how physiological changes in synaptic strengths store long-term memories; the activation patterns in the network change once per time-step, analogous to how the moment-to-moment change in electric firing patterns in the brain store short-term memories. The LSTM architecture aims to provide a short-term memory for RNN that can last thousands of timesteps, thus "long short-term memory".

A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

LSTM networks are well-suited to classifying, processing, and making predictions based on time series data since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models, and other sequence learning methods in numerous applications.

At a high-level LSTM works very much like an RNN cell. Here is the internal functioning of the LSTM network. The LSTM consists of three parts, as shown in the image below and each part performs an individual function. The first part chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. In the second part, the cell tries to learn new information from the input to this cell. At last, in the third part, the cell passes the updated information from the current timestamp to the next timestamp. 11 These three parts of an LSTM cell are known as gates. The first part is called Forget gate, the second part is known as the Input gate and the last one is the Output gate. Just like a simple RNN, an LSTM also has a hidden state where H(t-1) represents the hidden state of the previous timestamp and Ht is the hidden state of the current timestamp. In addition to that LSTM also have a cell state represented by C(t-1) and C(t) for the previous and current timestamp respectively.

Model Architecture:

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 1, 256)            787456
_________________________________________________________________
lstm_1 (LSTM)                (None, 1, 128)            197120
_________________________________________________________________
lstm_2 (LSTM)                (None, 1, 64)             49408
_________________________________________________________________
dense (Dense)                (None, 1, 1)              65
=================================================================
```

Fig 6: Model parameters

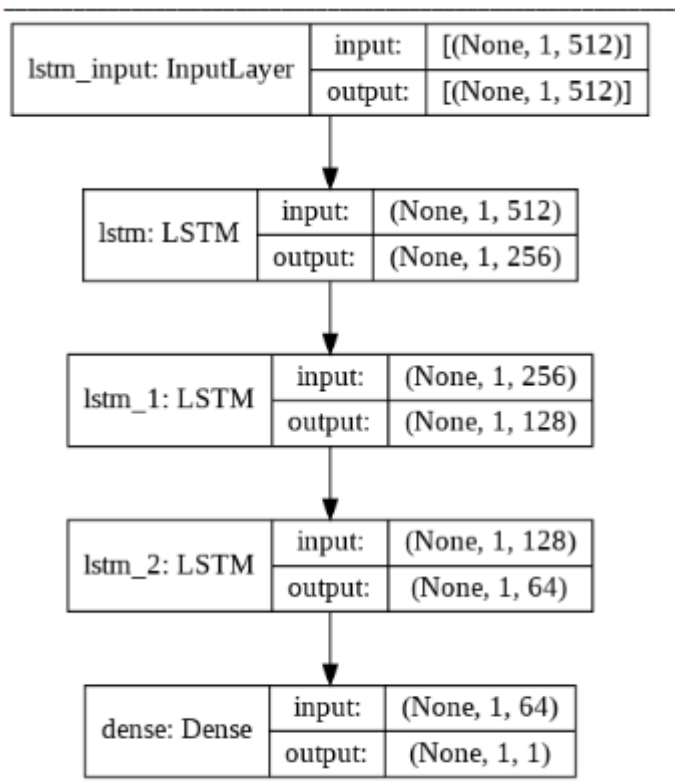


Fig 7: Model layers

7. RESULTS

We used 3 LSTM layers to improve the accuracy of the model. The accuracy of LSTM model was 84%. As the model has good accuracy, we use this model to give rating to movies or series.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm (LSTM)                  (None, 1, 256)            787456
_________________________________________________________________
lstm_1 (LSTM)                (None, 1, 128)            197120
_________________________________________________________________
lstm_2 (LSTM)                (None, 1, 64)             49408
_________________________________________________________________
dense (Dense)                (None, 1, 1)              65
=================================================================
Total params: 1,034,049
Trainable params: 1,034,049
Non-trainable params: 0
_________________________________________________________________
157/157 [==============================] - 1s 5ms/step - loss: 0.3971 - accuracy: 0.8452
```

Fig 8: Accuracy

Fig 9: Model results

8. CONCLUSION:

Movies are widely appreciated and criticized art forms. They are a significant source of entertainment and lead to web forums like IMDB and amazon reviews for users to give their feedback about the movies and web series. These reviews and feedbackdraw incredible consideration.

Although this information is unstructured, it is very crucial. We were inspired to work on this project to resolve this problem of unstructured movie reviews and that people need not spend a lot of time reading the whole review to understand whether the reviewer thinks about the movie in a positive or negative view.

9. FUTURE SCOPE

We can develop an android application in which the customer will be able to see the movie ratings and its description.

10.REFERENCES:

[1] https://sci-hub.se/10.1109/ICCES45898.2019.9002043

[2] https://ieeexplore.ieee.org/abstract/document/7468968/

[3] https://dl.acm.org/doi/abs/10.1145/2911451.2911490

[4] https://link.springer.com/book/10.1007/978-3-319-33622-0

[5] https://ieeexplore.ieee.org/abstract/document/7860338

[6] https://www.sciencedirect.com/science/article/pii/S1877042811024323

[7] https://ieeexplore.ieee.org/abstract/document/7748849

[8] https://www.researchgate.net/profile/Amna-Noureen/publication/318096420_Sentiment_Analysis_Using_Deep_Learning_Techniques_A_Review/links/596509014585157fcc5e3595/Sentiment-Analysis-Using-Deep-Learning-Techniques-A-Review.pdf

[9] https://www.sciencedirect.com/science/article/abs/pii/S0885230810000616