iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.35687

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

Fake News Detection using RNN-LSTM

Samarth Mengji[1], Saransh Ambarte[2], Sai Viswa Teja Arumilli[3], Shankar Mhamane[4], Rashmi Rane[5]
[1, 2, 3, 4, 5]*School of CET Maharashtra Institute of technlogy World Peace University*

Abstract: ***Fake news distribution is a social phenomenon that can't be avoided on a personal level or through web-based social media like Facebook and Twitter. We're interested in counterfeit news because it's one of many sorts of double dealing in online media, but it's a more severe one because it's designed to deceive people. We're concerned about this now that we've seen what's going on. We are concerned about this issue because we have seen how, through the usage of social correspondence, this marvel has recently caused a shift in the direction of society and people groupings, as well as their opinions. Along these lines, we chose to confront and decrease this wonder, which is as yet the principal factor to pick a large portion of our choices. Our objective in this study is to develop a detector that can predict if a piece of news is false based just on its content, and then attack the problem using RNN method models LSTMs and Bi-LSTMs to tackle the problem from a basic deep learning viewpoint.***
Keywords: RNN (Recurrent Neural Networks), LSTM (Long Short-Term Memory), Fake news detection, Deep learning

I. INTRODUCTION

Fake news is used to fool people with false information. With the proliferation of fake news from online media and other sources, the ability to discern between real and false news is becoming increasingly crucial. Fake news is a major contributor to riots, violence, mass lynching, and other social and financial upheavals. Any item of false news can be made to intentionally mislead, push a one-sided viewpoint, specific cause or goal, or, in any case, for entertainment. It should be called attention to that this is a general issue which influences every one individual all throughout the planet. What's more, it has likewise been there since forever ago. Despite the fact that in prior occasions as there was no admittance to overall data, the discovery of phony news was moderately troublesome and cost wasteful. Be that as it may, these days it is exceptionally simple, attainable and qualified to distinguish whether a news piece is phony or genuine. Which carries us to the genuine issue of examining and distinguishing the phony news from genuine news in Gigabytes of data.

Profound learning strategies like Convolutional Neural Networks (CNN) and RNNs are ordinarily used to perceive muddled models in artistic information. LSTM is a tree-coordinated discontinuous neural association for examining sequential material. You can watch specific sequences both front-to-back and back-to-front using the bi-directional LSTM. The model's presentation is evaluated using Kaggle datasets of unstructured news stories that are freely available.

In this paper, we offer a system for distinguishing between real and fraudulent news reports. Before applying NLP, information is scraped from Kaggle and preprocessed. Stop word evacuation and stemming are both done with NLTK. The model is built using LSTM and a RNN model.

The following is how the paper is organized: In the next section, we'll discuss the methodologies employed by different researchers to detect false news, as well as a brief literature survey we conducted for this project. The following section will provide a thorough knowledge of the RNN and LSTM models that we employed in our project, as well as our recommended approach and architecture. Then we'll go over our experiments and the results we got with the dataset we used, as well as the processes we took to prepare our dataset before developing our LSTM models.

II. RELATED WORK

Fake news frequently incorporates propaganda, hatred, and other heinous motives [1]. A news story is a collection of words [2]. Authenticity and aim are two crucial aspects in a narrow definition of false news [3]. As a result, many researchers in this field have advocated for the employment of text mining and machine learning approaches. Authors today claim that deep learning models perform better than older methodologies [2]. [2] presents a model for identifying false news that is built on a bi-directional LSTM-RNN. The model's performance is assessed using two datasets of unstructured news articles that are freely available. The results show that the Bi-directional LSTM model beats other approaches for detecting false news in terms of accuracy, including CNN, vanilla RNN, and unidirectional LSTM. In [4,] the author uses RNN method models (vanilla, GRU) and LSTMs to create a classifier that can predict whether a piece of news is true or false based solely on its text, approaching the problem from a deep learning perspective. The author of [5] wants to evaluate and analyze a variety of ways to solving this problem, including traditional machine learning algorithms like Naive Bayes and popular deep learning approaches like hybrid CNN and RNN.

This paper sets the groundwork for choosing a machine learning or deep learning approach for problem resolution that strikes the right mix of accuracy and portability. The nave bayes model, decision tree, random forest, k closest neighbor, LSTM, CNN&LSTM, and CNN&LSTM are among the techniques investigated by the author of [6]. [7] The paper has checked and analyzed a number of research publications as well as a number of survey pieces, and has compiled this document to provide readers a brief overview of what fake news is, its many flavors in the news spectrum, its characteristics, and basic identification. Identifying future news elements that have been labelled as phony. The paper [8] successfully tested for classification using RNNs, long short-term memories, and grated recurrent units. The Tensor board, which also acts as a visualization tool for the neural network, is used to execute the recommended design. The LSTM model may reach an accuracy score of up to 94 percent, according to the confusion matrix and classification reports.

III. BI-DIRECTIONAL LSTM-RNN MODEL

The suggested technique detects false news by analyzing the inclination of a constructed news story title and the relationship between the news story title and the article body to see if the content in the article is correct. This section depicts the information processing approach and the model design for the experiments.

A. RNNs

A RNN is a feed-forward neural network. RNNs are a form of Neural Network that considers the yield from past advancements as a contribution to the current advancement. In RNNs, a recurrent hidden layer is utilized to manage a variable whose activation is dependent on the prior time.

RNNs have a "memory" that keeps track of everything they've learnt. To create the output, it utilizes the same limits for each contribution and executes the same tasks on all of the data sources or hidden layers. This, in contrast to other neural systems, lowers boundary complexity.

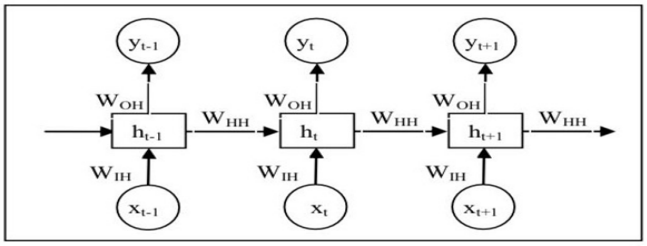


Fig 1. Basic architectural flow of RNN

An essential RNN updates the secret state (h1,h2,...,hT) and yields for each timestamp given an information grouping (x1,x2,...,xT) (y1, y2,...,yT). Figure 1 depicts the full engineering of essential RNN. The vectors xt and yt are the information and create vectors at timestamp t. WIH, WHH, and WOH are three association weight networks that individually handle the weight associated with inclusion, concealment, and vector production.

1) Formula for applying Activation function (Tanh)

$$h_t = \tanh (W_{hh}h_{t-1} + W_{xh}x_t)$$

Where h_t denotes current state
h_{t-1} denotes previous state
x_t denotes input state
w_{hh} denotes weight at recurrent neuron
w_{xh} denotes weight at input neuron

2) Formula for Calculating Output

$$y_t = W_{hy}h_t$$

Y_t denotes output
W_{hy} denotes weight at output layer

B. LSTM(LSTMs)

RNNs that can learn long-term dependence connections are known as long-term memory networks (LSTM). The LSTM algorithm is a powerful tool for dealing with the vanishing gradient problem. The buried layer of a basic RNN is replaced with an LSTM cell in LSTM-RNN. When the output is created by passing the input through a high number of hidden layers, the problem of decreasing gradient descent arises. Multiple weight changes occur because the RNN layer's output is decided by back propagation on all of its starting values. As a result, the "first" memory is "forgotten." To solve this problem, LSTM units with gates are employed.

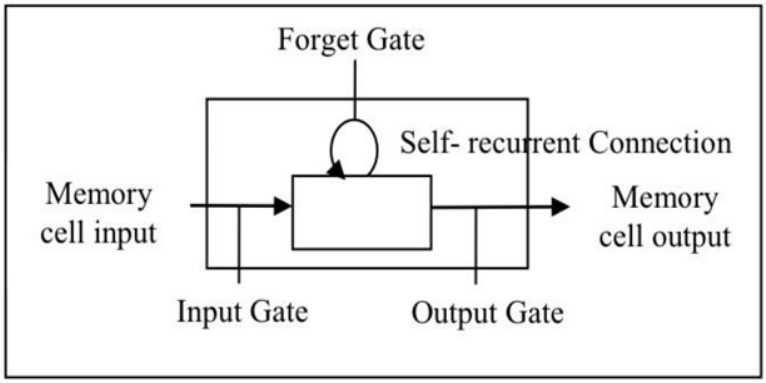


Fig 2. LSTM Cell

C. Bi-directional LSTM - RNN

LSTMs help in the preserving the liabilities that might lay out backwards in time and via bottom layers of a deep network. For large-scale sequential text prediction and categorization of the text, bi-directional processing is an obvious choice. A Bi-Directional LSTM network, as shown in Figure 3, in both directions at the same time, runs through the input sequence.

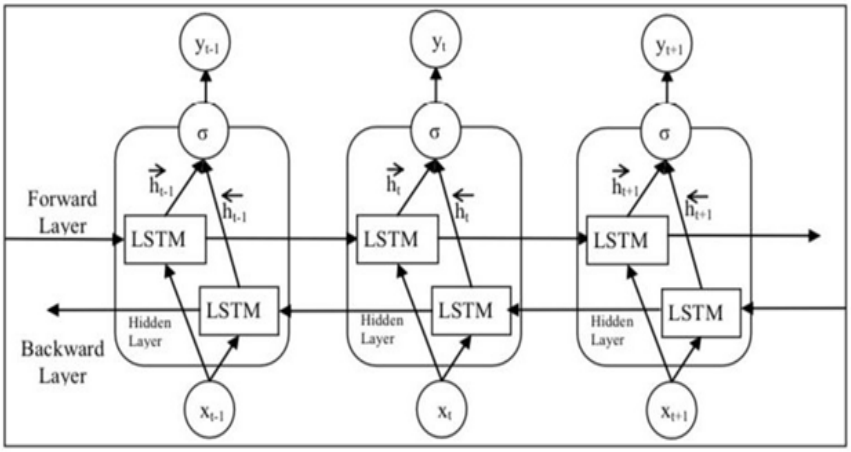


Fig 3. Basic architectural flow of Bi-directional LSTM- RNN

Figure 4 depicts the suggested false news detection model, which is built on a bi-directional LSTM-RNN. First, the news reports are pre-processed. Each piece of news is given a binary number, with 1 indicating fake news and 0 indicating genuine news. The two columns 'title' and 'text' are merged to produce a new column on which preprocessing may be done right away. The NLTK porter stemmer removes punctuation and stop words from the input news articles, while gensim removes words with fewer than two characters. The title and content text of news reports are transformed into padded sequences of words separated by spaces. Tokenization is used to break these sequences down even further into lists of tokens. The converted vector represented data is partitioned into train, validation, and test data with a test size of 0.2. The course is based on a collection of news stories. The validation data set is used to fine-tune the model. Using the trained model, the test data is also utilized to estimate the anticipated label of a news article. Two models are trained – LSTM and Bi-LSTM and comparison of both model is done.
Given below is the architecture of our proposed model.

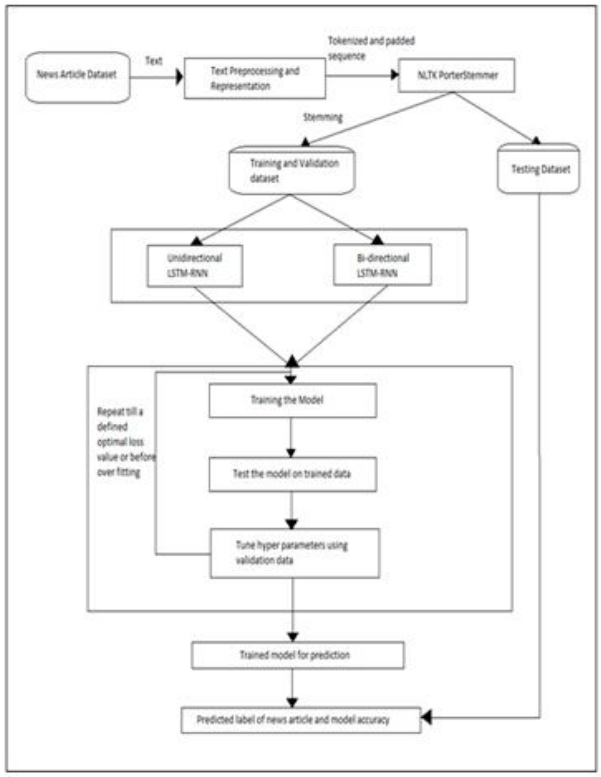


Fig. 4: Proposed architecture flow for detecting fake news

IV. EXPERIMENTS AND RESULTS

TensorFlow r0.12 is used in all of the codes, which are drafted in Python 2.7. All experimentation were made on an Intel(R) CoreTM i5-7200U CPU running at 2.50 GHz and 2.71 GHz with 8 GB RAM.

The two datasets used in this study were obtained from Kaggle's open Machine Learning Repository.

The title, text, Subject, and date of each news storey in this dataset are all fake. This dataset's vocabulary is roughly 59.8 MB in size. Table 1 shows the specifications for a fake dataset.

Attribute	Type
Title	Text
Text	Text
Subject	Text
Date	Numeric

Table 1. Specification of Fake articles dataset

True: This dataset contains the title, content, Subject, and date for each news story. This dataset's vocabulary is roughly 51.1 MB in size. Table 2 shows the True Dataset specification.

Attribute	Type
Title	Text
Text	Text
Subject	Text
Date	Numeric

Table 2. Specification of True articles dataset

The proposed model is trained on the combination of Fake and True news datasets, which contain a total of 44896 news items.

Given below is a snapshot of dataset after combining both Fake and True dataset.

```
In [7]: # Concatenating Real And Fake News
        df = pd.concat([df_true, df_fake]).reset_index(drop = True)
        df
```

Out[7]:

	title	text	subject	date	isfake
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1
...	...	...	...	...	...
44893	McPain: John McCain Furious That Iran Treated ...	21st Century Wire says As 21WIRE reported earl...	Middle-east	January 16, 2016	0
44894	JUSTICE? Yahoo Settles E-mail Privacy Class-ac...	21st Century Wire says It s a familiar theme. ...	Middle-east	January 16, 2016	0
44895	Sunnistan: US and Allied 'Safe Zone' Plan to T...	Patrick Henningsen 21st Century WireRemember ...	Middle-east	January 15, 2016	0
44896	How to Blow $700 Million: Al Jazeera America F...	21st Century Wire says Al Jazeera America will...	Middle-east	January 14, 2016	0
44897	10 U.S. Navy Sailors Held by Iranian Military ...	21st Century Wire says As 21WIRE predicted in ...	Middle-east	January 12, 2016	0

44898 rows x 5 columns

```
In [8]: df.drop(columns = ['date'], inplace = True)
```

Fig 5. Snapshot of dataset after combining

A. Preprocessing

After combining both the datasets into one, we pre-process our dataset. We add “isFake” Boolean column to our dataset and one more column which is combination of “title” and “text” as those are the columns we used for predicting whether the news is fake or not. We perform data cleaning on our combined column named “original”. By creating a method which include stop words from nltk.corpus and genism, we also remove words which have length less than 2.

We make new column named “clean” which is cleaned version of our “original” column after removing stopwords.

After data cleaning, we split our data using train test split method with test size is 0.2. We perform tokenization using word_tokenize from nltk and create a list of tokenized words and add padding to create padded sequences.

```
# Adding Padding
padded_train = pad_sequences(train_sequences,maxlen = 40, padding = 'post', truncating = 'post')
padded_test = pad_sequences(test_sequences,maxlen = 40, truncating = 'post')
```

```
for i,doc in enumerate(padded_train[:5]):
    print("The padded encoding for document",i+1," is : ",doc)
```

```
The padded encoding for document 1  is :  [ 367  556 5268  232   25    9   66 1066  580  367 6106  308  556    2
   83  265  569   38  232  776   10    1  805   60    3 6106 1359  580
    1 1382  188  411  560   46   48   13 8252    2    1   45]
The padded encoding for document 2  is :  [ 1024  8203  4510     1  2681   738   383    87  1024 43185     3     1
  178 12077   738   383   323   538  2146  2532 14544  2916  8619  3816
 30462  6513    18 11691  9874   170  2514 11576 35160   360  2274  1206
   10     1   955  5516]
The padded encoding for document 3  is :  [  749     6  1967  1644   730   121   476    41  6081   698  6892  9949
  3798   289  6731  1184 13731  7406 11229     3     6   953   730   122
  1212    66    11  1577  1471  1528    15     8  2295   277  1465  1065
    29     6   721  7130]
The padded encoding for document 4  is :  [  534    30  6859   190  6654   317    13   385  3593  1071   317     9
  6732  6830  1454  4535   137   611  2319    30  6859  1763  1128  3593
  1071    13   385  1057    72   594  1608 38248  7063   534   944  7063
    19 11693    13  2051]
The padded encoding for document 5  is :  [  317  1510  2730  1215  2097  1736   110     1  8742  7294   109    21
   494   317  6929    56  3593  9530   547  1510 13207 27345   378    42
  9112  2126    35   133   737  1510   110     1  8742  1905  2097   503
   242 14777  6173   209]
```

Fig 6. Tokenized and padded sequence of dataset

This is how our tokenized and padded dataset looks before start training our LSTM model.

B. Dataset Modelling

The tensorflow keras sequential model is the foundation of our model. For a basic stack of layers with precisely one input tensor and one output tensor for each layer, a sequential approach works well, which is appropriate for our model because it will have a single input – a combination of news storey title and news article – and a single output of 0 or 1.The first layer we add to our model is embedding which is offered by keras that can be used for neural networks on text data. We specify two arguments to the embedding – 1)input_dim: total number of unique words in our cleaned dataset

2)output_dim: Words will be embedded in a vector space of this size

The second layer is our LSTM layer. We build two similar model first one being unidirectional LSTM and second one bidirectional LSTM.

Then we add dense layers. The thick layer is a conventional highly connected neural network layer since each neuron receives input from all neurons in the preceding layer.
Two dense layers were employed, each with a different activation function: a linear unit rectified and a sigmoid function.
The activation function of the ReLU (Corrected Linear Unit) is half corrected from the bottom up. When x is less than zero, F(x) is zero, and when x is more than or equal to zero, F(x) is x.
The sigmoid function occurs between 0 and 1. As a consequence, it's suitable for models in which we need to forecast probability as a result. Because a news storey may be either false or true, with a value of 0 or 1, sigmoid is the best option.

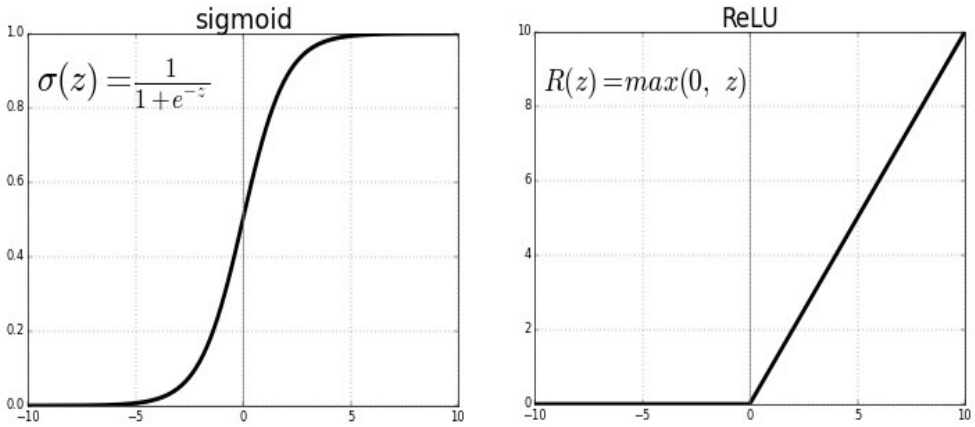


Fig 7. Activation function ReLU and Sigmoid

We then train both the models on our dataset and the results are displayed in the next section.

C. Results

After training the model for 10 epochs we got the following result

Model	Training Accuracy	Validation Accuracy	Test Accuracy
Unidirectional LSTM-RNN	0.997	0.89	0.92
Bi-directional LSTM-RNN	1	0.97	0.99

Table 3. Accuracy of two models

Given below are the graphs of training and testing accuracy and loss of both the models over 10 epochs.

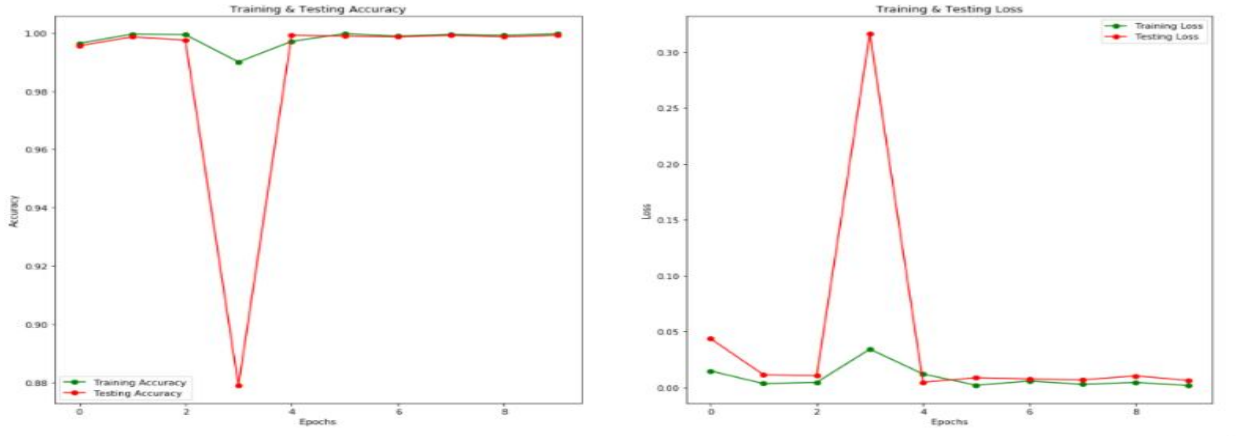


Fig 8. Unidirectional LSTM model accuracy and loss

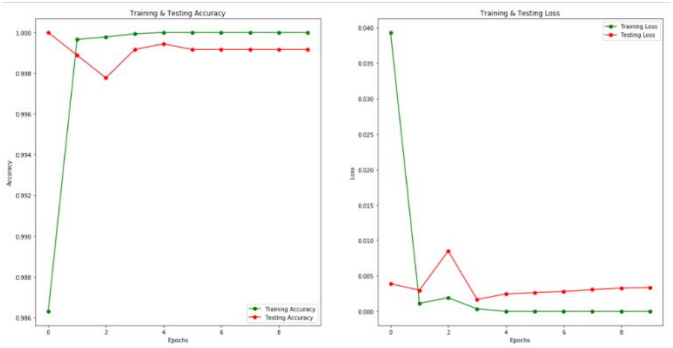


Fig 9. Bi-directional LSTM model accuracy and loss

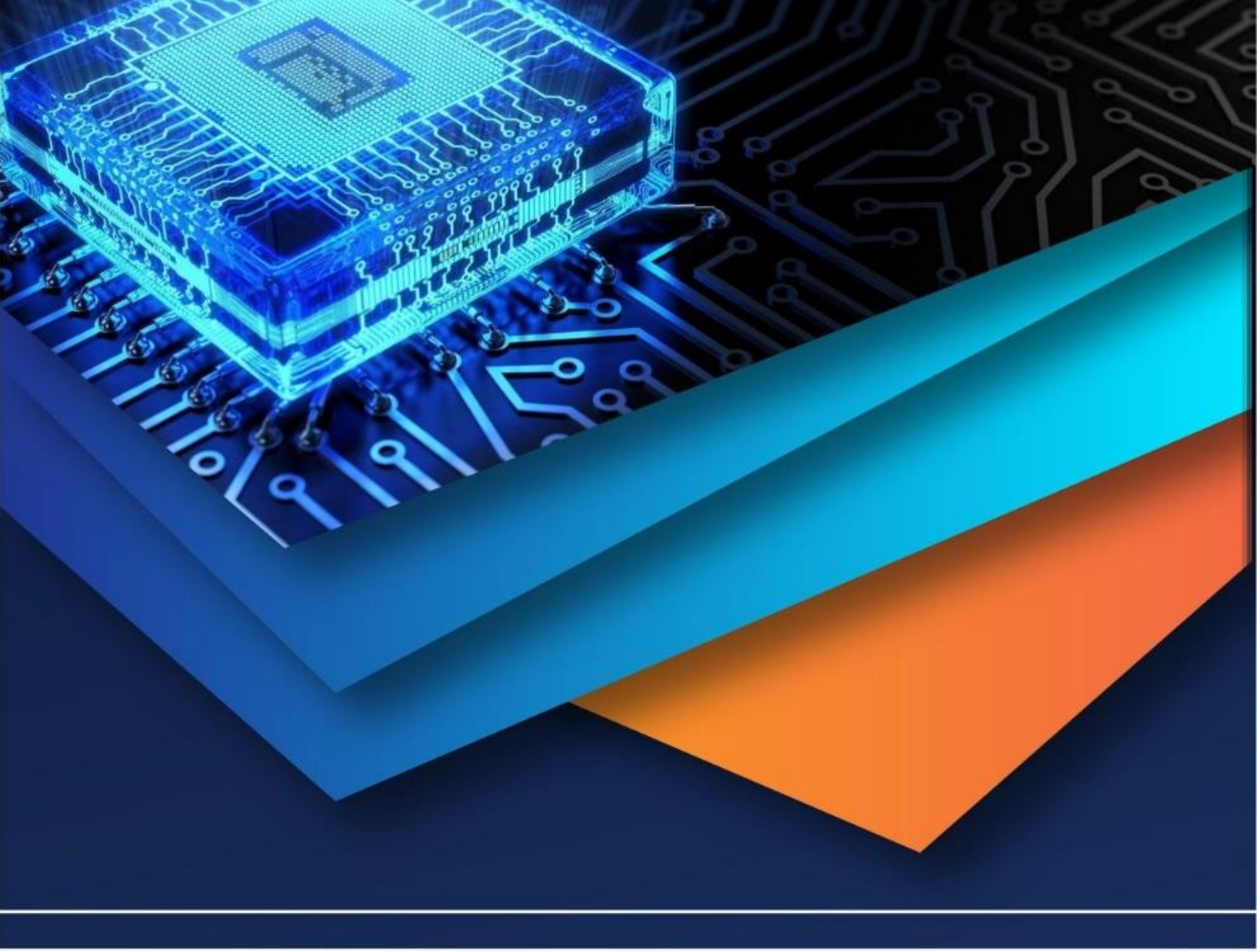
During model training, the RNN model suffers from a vanishing gradient issue due to the deep network structure. LSTM-RNN is used to address the vanishing gradient problem. From the graphs we can say that the second model has less loss and more accuracy compared to the first but the difference is minimal.

V. CONCLUSION

The methodology for the purpose of creating a fake news detector that can predict the fake news articles has been outline of this paper. The methodology implemented to utilize the input attributes of the dataset such as title along with the article content to achieve the effective prediction of the fake news. The prediction is effectively implemented through the use of RNN approaches that can provide highly accurate predictions of the fake news using the input variables. The presented technique utilizes NLTK porter stemming NLP, word tokenizer, word embedding, unidirectional and bidirectional LSTM using RNN.

REFERENCES

[1] Ghinadya and Suyanto,"Synonyms-Based Augmentation to Improve Fake News Detection using Bidirectional LSTM" in 2020 8th International Conference on Information and Communication Technology (ICoICT)
[2] Pritika Bahada,Preeti Saxenaa,Raj Kamalb,"Fake News Detection using Bi-directional LSTM-Recurrent NeuralNetwork" in International Conference on recent trends in advanced computing 2019, ICRTAC 2019
[3] Manoj Kumar Balwant,"Bidirectional LSTM Based on POS tags and CNN Architecture for Fake News Detection" in 10th ICCCNT 2019 July 6-8, 2019, IIT - Kanpur Kanpur, India
[4] Eslam Amer,Mahmoud Gadallah,Sherry Girgis,"Deep Learning Algorithms For Detecting Fake News in Online Text" in 13th IEEE International Conference on Computer Engineering and Systems
[5] Wenlin Han and Varshil Mehta,"Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation" in 2019 IEEE International Conference on Industrial Internet (ICII).
[6] Rohit Kumar Kaliyar,"Fake News Detection Using A Deep Neural Network" in 2018 4th International Conference on Computing Communication and Automation (ICCCA).
[7] Sunidhi Sharma and Dilip Kumar Sharma,"Fake News Detection: A long way to go" in 2019 4th International Conference on Information Systems and Computer Networks (ISCON) GLA University, Mathura, UP, India. Nov 21-22, 2019.
[8] Abhishek Verma, Vanshika Mittal, Suma Dawn,"FIND: Fake Information and News Detections using Deep Learning"in 2019 Twelfth International Conference on Contemporary Computing (IC3).
[9] Akshay Jain and Amey Kasbe, “Fake News Detection”, in 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Sciences
[10] Rohit Kumar Kaliyar,” Fake News Detection Using A Deep Neural Network”, in 2018 4th International Conference on Computing Communication and Automation (ICCCA)
[11] Shivam B. Parikh and Pradeep K. Atrey, “Media-Rich Fake News Detection: A Survey”, in 2018 IEEE Conference on Multimedia Information Processing and Retrieval
[12] Syed Ishfaq Manzoor, Dr Jimmy Singla and Nikita,” Fake News Detection Using Machine Learning approaches: A systematic Review”, in Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)
[13] Kartik Rajesh,Aditya Kumar and Dr. Rajesh Kadu,” Fraudulent News Detection using Machine Learning Approaches”, in 2019 Global Conference for Advancement in Technology (GCAT) Bangalore, India. Oct 18-20, 2019
[14] Philogene Kyle Dimpas, Royce Vincent Po and Mary Jane Sabellano,”Filipino and English clickbait detection using a long short term memory RNN”, in 2017 International Conference on Asian Language Processing (IALP)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY