ISSN 2278-3091
Volume 10, No.5, September - October 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse201052021.pdf
https://doi.org/10.30534/ijatcse/2021/201052021

A Novel Approach for Detection of Fake News using Long Short Term Memory (LSTM)

K. Kranthi Kumar[1], S. Hanumantha Rao[2], G. Srikar[3], M. Bharat Chandra[4]
[1]Associate Professor, Dept of IT, Sreenidhi Institute of Science and Technology, Hyderabad, India, kranthikumark@sreenidhi.edu.in
[2]B.Tech Student, Dept of IT, Sreenidhi Institute of Science and Technology,Hyderabad,India,hanumanthsomaraju@gmail.com
[3]B.Tech Student, Dept of IT, Sreenidhi Institute of Science and Technology,Hyderabad,India,govardhanasrikee@gmail.com
[4]B.Tech Student, Dept of IT, Sreenidhi Institute of Science and Technology,Hyderabad,India,bharatchandra888@gmail.com

Received Date : August 07, 2021 Accepted Date : September 16, 2021 Published Date : October 06, 2021

ABSTRACT

Online media for news consumption has doubtful advantages. From one perspective, it has minimal expense, simple access, and fast dispersal of data which leads individuals to search out and devour news from online media. On the other hand, it increases the wide spread of "counterfeit news", i.e., inferior quality news with purposefully bogus data. The broad spread of fake news contrarily affects people and society. Hence, fake news detection in social media has become an emerging research topic that is drawing attention from various researchers.

In past, many creators proposed the utilization of text mining procedures and AI strategies to examine textual data and helps to foresee the believability of news. With more computational capacities and to deal with enormous datasets, deep learning models present a better presentation over customary text mining strategies and AI methods. Normally deep learning model, for example, LSTM model can identify complex patterns in the data. Long short term memory is a tree organized recurrent neural network (RNN) used to examine variable length sequential information. In our proposed framework we set up a fake news identification model dependent on LSTM neural network. Openly accessible unstructured news datasets are utilized to evaluate the exhibition of the model. The outcome shows the prevalence and exactness of LSTM model over the customary techniques specifically CNN for fake news recognition.

Key words: LSTM (Long short term memory), RNN, Artificial Intelligent, Fake News and Recognition.

1. INTRODUCTION

Tremendous increase in the affordability and advancement of the hand-held gadgets and access to internet expanded the quantity of computerized media clients. We can access any data in the planet with a single click. On the other hand Counterfeit news is the major problem faced by the computerized world and it reduces the reliability of the news. Fake news can easily spread through online by utilizing social platforms such as twitter, whatsapp, snapchat and facebook to change public discernments. The expectation of taking this theme is to make an execution that diminishes fake news on the web. It is prompting an awful environment on the web and furthermore making issues for individuals. Consequently identifying fake news is viewed as quite difficult for the current content based methods. There is an earnest requirement for examining AI ways to deal with identify fake news. A few fake news recognition models are used for detecting the counterfeit news namely text mining and traditional learning methods [1, 2] and deep learning models [3, 4]. The above machine learning models are used in the following categories they are content, social setting, and propagation [5]. Current neural network models show better performance than the customary ones because of the extraordinary capacity in extracting the important features from the data and identifying meaningful insights from the data, yet, the above strategies can't distinguish fake news, recently emerged news, and time-basic occasions [6].

There are many ways to deal with the issue of falsehood via online media. Measurable strategies are utilized to recognize the relationship between different components of the data by breaking down the originator of the data, and dissecting examples of dispersal. AI calculations are utilized for characterization of questionable substance and dissecting the records that shares the similar content.

Our methodology is to foster a model where it will recognize whether the given news is fake or genuine utilizing LSTM (Long short term memory) and other concepts such as tokenization, NLP and word embedding. In our proposed system the model will predict the outcomes for the given dataset with a performance of 91.73%.

2. RELATED WORK

One of the previous investigations on counterfeit news recognition and programmed reality checking with in excess of 1,000 examples was finished by [7] using LIAR named dataset. The set has 12,800 named small proclamations of human from the website POLITIFACT. The proclamations named in 6 unique classifications, for example, pants fire, bogus, scarcely obvious, half evident, for the most part obvious, and valid. The study utilized a few classifiers like calculated relapse, SVM, a long momentary memory which is Bi-directional (Bi-LSTM), and used CNN method. For Calculated Relapse and Support Vector Machine, the investigation utilized a tool for Lib Short Text and it has given critical execution on small text characterization issues. The work looked at a few strategies utilizing text includes just and accomplished a precision of 0.205 and 0.209 on the approval and in the testing ones. Because of excessive fitting, the Bidirectional LSTMs didn't perform great execution. In any case, the CNN outflanked all models, bringing about a precision of 0.270 on the holdout information parting.

Many creators present utilization of methods which are related to mining of text, AI procedures so as to investigate printed information to anticipate validity of news-articles. By growing computing capacities thereby dealing with enormous data, a better presentation is given by models which are learning than that of customary message methods (for mining), AI procedures. CNN, RNN are generally investigated Deep Neural Network (DNN) designs to tackle different NLP assignments [8][9].The momentum work is identified with number of examination regions like message grouping, talk discovery, spammer location, and opinion examination. Counterfeit news can be distinguished utilizing distinctive AI strategies. Writers [10] proposed a straightforward methodology for counterfeit news recognition utilizing innocent Bayes classifier is tried against an informational index of new posts which are from Face book [11].

The two new datasets that covered seven news areas was built by Perez Rosas [12]. First set of data was gathered with public support that has included informational index pertaining to news. Next set of data from a dataset was gathered straightforwardly from the internet and it included a space. Then, performed an analysis to validate and differentiate false news from the real news. They assembled a phony news finder utilizing direct SVM classifier and five-overlay cross validation dependent on the blend of lexical, syntactic, and semantic data as etymological components. In light of their outcomes, their model accomplished exactness up to 78%.

A model was given by Davis and Delegate [13] in which they used 3 layered Perceptron which has got multiple layers. The work got better outcomes compared to the other ones. The popular FNC-1 was utilized so that it uses some programmed techniques to identify counterfeit news. Aim is to group the data that comprises feature sets which are of text data as irrelevant, concurring, and dissenting. It gained 92% precision. What's more, Mill operator, Oswalt [14] utilized equivalent set of data for recognizing counterfeit newscast utilizing an organization framework with consideration system. Then constructed an organization design utilizing various Bidirectional LSTMs and a consideration component. Excellent outcome was accomplished by the combination of Bi-directional LSTM and Multi Facet Perceptron with 56% precision.

A structure was proposed in 2018 named Occasion Antagonistic Neural Network by Wang et al.[15] could recognize counterfeit articles dependent on modular components(multi type) and understands adaptable element. They checked the efficacy of the model by taking news articles from Twitter and weibo, and the outcome said that it's the best in class then. Then assembly of a profound organization named FACE Finder by Zhang et al. [16] depended upon a bunch of unequivocal, inert elements removed out of printed data. This method has been placed one among the cutting edge models because of it's extraordinary exhibition.

A geometric based model was developed for detecting false news by Monty et al. [17] in the year 2019. For demonstration, they gathered news articles from various fact checking websites and they sifted through all information which not-contained something like one URL reference on Twitter. Their trials showed that Social network features, like design and propagation, accomplished high exactness (91.7 %) on false news detection model.

3. PROPOSED SYSTEM

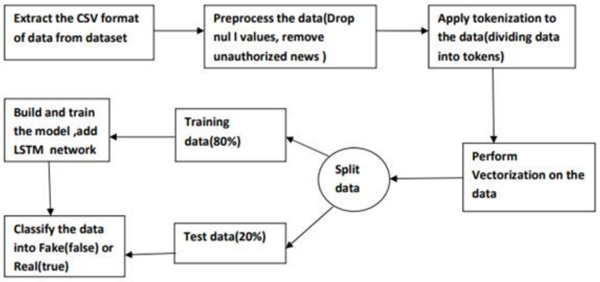

Extract the CSV format of data from dataset
Preprocess the data(Drop nul l values, remove unauthorized news)
Apply tokenization to the data(dividing data into tokens)
Build and train the model ,add LSTM network
Training data(80%)
Split data
Perform Vectorization on the data
Classify the data into Fake(false) or Real(true)
Test data(20%)

Figure 1: System Architecture.

In the above Figure 1 it describes the architecture of the proposed system.

3.1 Algorithm

1. Import the important packages (NLTK, wordtovec) and preprocess the data by removing null values and unauthorized news.

2. Perform exploratory data analysis using word cloud and Identify the frequency of the words in the given text.

3. Split the data for Training and Testing.

4. Perform the process of Tokenization by splitting each sentence into individual tokens.

5. Perform the process of Vectorization by calculating weights to each tokens and these weights are used to train the model.

6. Train the model and calculate loss. Increase epochs to minimize losses.

7. Evaluate the Model.

8. Calculate accuracy score and view classification report.

3.2. Preprocessing

In the below Figure 2 it describes about the steps involved in pre-processing. In this Pre-processing step, the input dataset is taken from kaggle machine learning repository. Firstly we change the format of the dataset to CSV format and we next identify and remove the noise data. The noise data could be of any type such as null values and unauthorized news article and redundancy of the news items. Secondly we identify the above mentioned noise data in dataset and we remove them and we use NLTK tool kit to remove punctuations and stop words. This step results the pre processed data which is suitable for the process of tokenization.

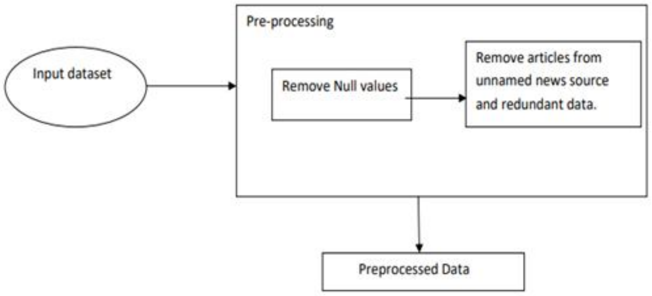


Figure 2: Steps of Preprocessing.

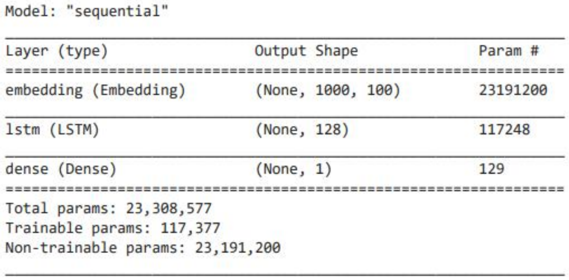
3.3 Tokenization

Tokenization refers to dividing large set of data into individual words or smaller lines. In our proposed system we use word tokenization. This step is performed next to pre processing by importing the tokenizer package from Keras. The large text can be converted to individual tokens by using word tokenize function.

3.4 Word embedding and Vectorization

This step is performed next to the process of tokenization. While training the model the input is given in the numerical form so we need to convert the words into vectors and this process is called as Vectorization. Word embedding is capable of getting the insights of words and identifies the syntactic and semantic similarity between them. In our proposed system we use Word2vec ( ) method for converting words into vectors. We use most similar ( ) method to identify the similarity between words. All the words are converted into sequence of vectors and it is used in training the model.

3.5 Model



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 1000, 100)         23191200
_________________________________________________________________
lstm (LSTM)                  (None, 128)               117248
_________________________________________________________________
dense (Dense)                (None, 1)                 129
=================================================================
Total params: 23,308,577
Trainable params: 117,377
Non-trainable params: 23,191,200
_________________________________________________________________
```

Figure 3: Model of proposed system.

In the above Figure 3 it describes the model of the proposed system. The input to the given model is the output of the word embedding process. The type of the machine learning model used for implementation in our proposed system is a sequential model which contains embedding as the first layer in which it consists of values vocabulary size, number of features and length of sentence. The next is LSTM with 128 neurons for each layer, followed by dense layer with sigmoid activation function in order to get final output. In our proposed system we have used Adam optimizer for adaptive estimation and finally adding drop out layer in between so that over fitting is avoided. Then training of the model is done followed by testing. After testing we will get final output in terms of numerical value ranging from 0 to 1.If the final predicted value is greater than 0.5 then it is classified and true and it can be termed as the real news else it is classified as false and it can be termed as fake news.

4. EXPERIMENTAL SET UP

The investigations were performed on a Core™ Intel processor and the specifications of central processing unit (CPU) are i7-4790U 3.60 GHz with 16 GB random access memory (RAM). The model was implemented by using python 3.9.0 using various libraries namely sklearn, Keras, nltk, tokenizer. Dataset which is used is taken from machine learning repository at kaggle. The dataset consists of 27436 records. Preprocessing is the initial step that is performed on the data which removes null values and news articles from unauthorized publishers and stop words. Tokenization is the important step in the experiment as it divides the sentences into tokens and this process is implemented by the module tokenizer and tokenizes function. These tokens are in linguistic form which is not suitable for the model to learn the insights from the data hence these tokens are converted to numeric form with the process of word embedding and Vectorization, word2vec is the important inbuilt function that converts the words into vectors. We use the sequential model as a machine learning model and later we add the LSTM network and various dense layers. Sigmoid function is used as an activation function and Adam optimizer upgrades the efficiency of the system by minimizing losses.

5. RESULTS

In Figure 4 it shows the graphical visualization of different types of news present in the dataset. The exploratory data analysis is performed by using matplotlib library in python which mainly extracts the insights of the data through various types of statistical graphs. There are approximately 9000 news items, 6500 politics news, 1800 government news, nearly 4000 left-news, less than 1000 US news and Middle East news items.

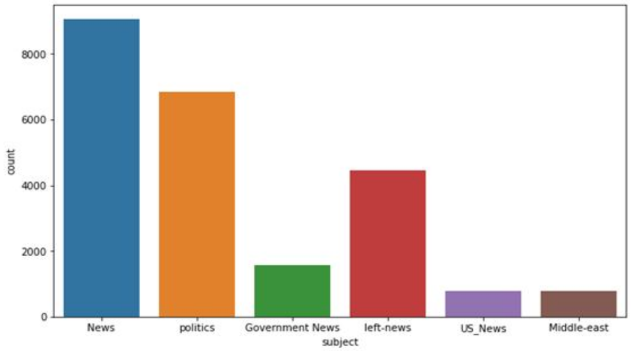


Figure 4: Visualization of types of news items.

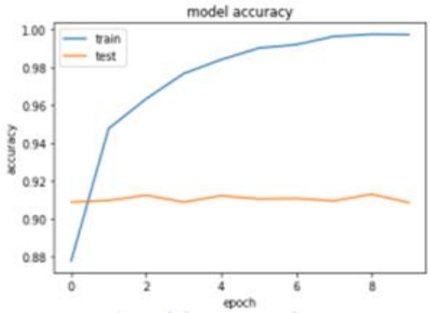


Figure 5: Variation of accuracy vs epochs.

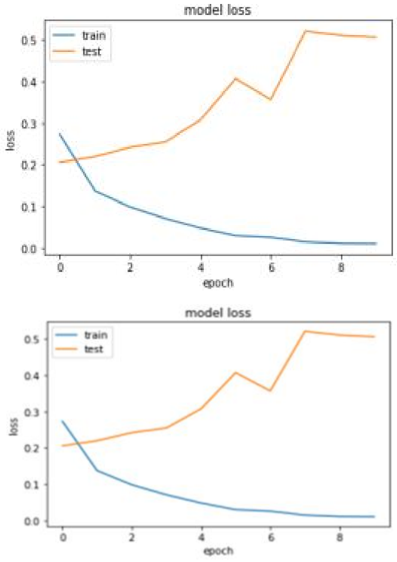


Figure 6: Variation of loss vs. epochs.

In the above Figure 5, it represents variation in accuracy and epochs. The proposed system shows an increase in the accuracy of the model for the training set with increase in the number of epochs. In Figure 6 the graph depicts the variation of losses and number of epochs. The training losses significantly decreased with increase in epochs and these losses are further decreased in the proposed system by using Adam optimizer. ‘

$$Accuracy_{score} = (TP + TN)/(TP + FP + FN + TN)$$

Equation 1: Formula of Accuracy score.

In the above Equation 1, it describes accuracy equation of the built model for the given test data. The proposed system shows an accuracy of 91.73% in classification of fake and real news. News article is provided as an input to model and the model classifies it as fake or real. The final output displays a numerical value ranging from 0 to 1.If the final predicted value is greater than 0.5 then it is classified as true and it can be termed as the real news else it is classified as false and it can be termed as fake news.

6. CONCLUSION

In conclusion, the main objective of the proposed system is to detect the fake news using deep learning models. The proposed system followed a sequence of operations in which word Vectorization is applied to change over any word present in the message of information into a vector any ideal measurement. Word Vectorization manages change of the high dimensionality of data. The exactness of LSTM model when contrasted with CNN, RNN models is high and roughly it is 91.73%. CNN performs better for separating nearby and position-invariant elements while LSTM-RNN is appropriate for a long-range semantic reliance based arrangement. RNN turn out better for undertakings where successive displaying is more significant. The outcomes show LSTM model is altogether more compelling than different models. The tests additionally show that the decision of the versatile learning rate calculation assumes a significant part in the yield to deal with vanishing gradient issue of RNN. The proposed model functions admirably for the reasonable and imbalanced high dimensional data. More exhaustive investigations will be needed in the future to additionally see how profound learning model with consideration can assist with assessing the programmed believability examination of news.

REFERENCES

1. N. J. Conroy, V. L. Rubin, and Y. Chen, **Automatic deception detection: Methods for finding fake news,** Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, pp. 1–4, 2015.
2. Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, **Novel visual and statistical image features for microblogs news verification,** IEEE transactions on multimedia, vol. 19, no. 3, pp. 598–608, 2017.
3. J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F.Wong, and M. Cha, **Detecting rumors from microblogs with recurrent neural networks**. in Ijcai, 2016, pp. 3818–3824,2016.

4. N. Ruchansky, S. Seo, and Y. Liu, **Csi: A hybrid deep model for fake news detection,** in Proceedings of the 2017 ACM on Conference on Information and knowledge Management. ACM, pp. 797–806. 2017.
5. X. Zhou and R. Zafarani, **Fake news: A survey of research, detection methods, and opportunities**, arXiv preprint arXiv:1812.00315, 2018.
6. K. Shu, A. Sliva, S.Wang, J. Tang, and H. Liu, **Fake news detection on social media: A data mining perspective,** ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.
7. W. Y. Wang, **Liar, liar pants on fire: a new benchmark dataset for fake news detection**, in Proceedings of the Annu. Meet. Assoc. Comput. Linguist, pp. 422–426, Vancouver, Canada, July 2017.
8. Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. (2017). **Comparative Study of CNN and RNN for Natural Language Processing**, 2017.
9. Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao. (2015). **Recurrent Convolutional Neural Networks for Text Classification**,Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.
10. Granik, Mykhailo ,Volodymyr Mesyura.(2017). **Fake News Detection using Naive Bayes Classifier**. IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON):900-903, 2017.
11. S. Gilda. (2017). **Evaluating Machine Learning Algorithms for Fake News Detection**, IEEE 15th Student Conference on Research and Development (SCOReD), Putrajaya: 110-115, 2017.
12. V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, **Automatic detection of fake news**, in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3391–3401.
13. R. Davis and C. Proctor, **Fake news, real consequences: Recruiting neural networks for the fight against fake news,** 2017.
14. K. Miller and A. Oswalt, **Fake news headline classification using neural networks with attention**, tech. rep., California State University, year, Tech. Rep., 2017.
15. Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, **Eann: Event adversarial neural networks for multi-modal fake news detection**, in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2018, pp.849–857.
16. J. Zhang, L. Cui, Y. Fu, and F. B. Gouza, **Fake news detection with deep diffusive network model**, arXiv preprint arXiv:1805.08751, 2018.
17. F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, **Fake news detection on social media using geometric deep learning**, arXiv preprint arXiv:1902.06673, 2019.