FAKE NEWS DETECTION IN SHORT MESSAGE SERVICE WITH LSTM-RNN OVER RANDOM FOREST ALGORITHM

Kasturi pogiri[1], S.John Justin Thangaraj[2*]

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract

Aim: This proposed research is to develop fake news detection in SMS (Short Message Service)service using the LSTM-RNN model and improve the accuracy with neural networks in contrast to random forest model. **Materials and Methods:** The LSTM-RNN model is applied on data, which is a text file containing sequences.a collection of words A random forest for predicting the accuracy of fake news that compares two sources. Model of randomforest It has been suggested and developed to have LSTM. The size of the sample. The G Power value of 0.8 was used to calculate the number of people in each category. The precision was excellent**.** LSTM-RNN (56 percent) was the most effective in spotting bogus news. When compared to randomforest, the least mean error is (43%). **Results:** The accuracy was maximum in detecting the fake news in social media using LSTM 51% with long short term memory model 40% for the same dataset. **Conclusion:** The study proves that LSTM exhibits better accuracy than random forest in detecting the fake news on SMS(Short Message Service) service.

Keywords: Machine Learning, Innovative detection, Randomforest , LSTM, Neural Networks.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

[2*]Project Guide, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

1. Introduction

The use of computer-mediated technology to create and share information is made easier with social media (Sarkar 2019). The way individuals interact and communicate has changed as a result of the media. It provides them with low-cost, easy-to-access information that is distributed quickly. The majority of people use social media to find and consume news rather than mainstream marketing (Rana et al. 2019). These days, traditional news institutions are a thing of the past. On the one hand, social media has evolved into a tremendous source of information.On the one hand, knowledge and bringing people together has a negative impact on society (Schwab 2017). Take a look at these examples.here are a few instances (Rana et al. 2019).WhatsApp, a prominent messaging app owned by Facebook Inc, was used as a political battleground in Brazil's election. For campaigning, false rumors, edited photographs, de-contextualized videos, and audio jokes were used (D'Ulizia et al. 2021) we use innovative detention methods with the help of machine learning algorithms. These types of things went viral on the internet without anyone keeping track of where they came from or how far they spread (Harvard Business Review et al. 2019). After repeated terrorist attacks in 2019, Sri Lanka implemented a statewide block on major social media and messaging services such as Facebook and Instagram(Prabhu et al. 2019) . The administration said that there were "false news reports" circulating on the internet. This demonstrates the difficulties that the world's most powerful internet corporations have in combating misinformation (United Nations Publications 2019). Detecting fake news on social media is difficult for a number of reasons. To begin with, gathering data on fake news is challenging.hence innovative detection methods are introduced with the help of machine learning algorithms.

Furthermore, manually labeling bogus news is tough. Because they are written with the goal of deceiving readers, it is difficult to detect them solely based on news content . Facebook, Whatsapp, and Twitter, on the other hand, are closed messaging apps. As a result, it's difficult to label misinformation spread by respectable news organizations or friends and family as phony. Because they are insufficient to train the application dataset, it is difficult to validate the trustworthiness of freshly appearing and time-bound news(United Nations Publications 2019; Rana et al. 2019). Significant methods for differentiating credible users, extract useful news features and develop authentic information dissemination systems are some useful domains of research and need further investigations.There are numerous techniques to dealing with the problem of social media disinformation. Statistical approaches are used to determine the relationship between various aspects of the data while also examining the source of the data. studying patterns of transmission of information For classification, machine learning techniques are utilised.untrustworthy content, as well as the accounts that share it.Various approaches concentrate on the development of children.approaches for information authentication, as well as specific case studies. To eradicate fake news some innovative detention methods are introduced from artificial intelligence and machine learning algorithms like, Accuracy, Naive Bayes Classifiers, Deep Neural Networks.Our team has extensive knowledge and research experience that has translated into high quality publications(Pandiyan et al. 2022; Yaashikaa, Devi, and Kumar 2022; Venu et al. 2022; Kumar et al. 2022; Nagaraju et al. 2022; Karpagam et al. 2022; Baraneedharan et al. 2022; Whangchai et al. 2022; Nagarajan et al. 2022; Deena et al. 2022)

Naïve Bayes Classifier is an innovative detection algorithm as it helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. The Naive Bayes Classifier is a part of a deep neural network. A deep neural network (DNN) is an artificial neural network (ANN) with multiple layers between the input and output layers. There are different types of neural networks but they always consist of the same components: neurons, synapses, weights,

2. Materials and Methods

This study was implemented using jupyter notebook software ,and hardware configurations are intel i3 core processor, 64GB HDD, 4GB RAM, and the software configurations are windows OS, python jupyter notebook. The work was carried out on 6328 records from the text file from online dataset kaggle website. The accuracy in predicting the next word was performed by evaluating two groups. A total 150 epochs were performed on each group to achieve better accuracy. The study uses a dataset downloaded from kaggle website .

LSTM

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.Import the python libraries required for the fake news detection.The pseudocode for LSTM are:

Step 1: import libraries
Step2: import dataset

Step3: creating the target column
Step4: concatenating the title text of the news
Step5: converting data column to data time format
Step6: appending two data sets

Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Random forest is straightforward. Import the python libraries required for the fake news detection.The pseudocode for Random forest are

Step 1: import libraries
Step2: import dataset
Step3: creating the target column
Step4: concatenating the title text of the news
Step5: converting data column to data time format
Step6: appending two data sets

Statistical analysis

The SPSS statistical software was used in the research for statistical analysis. Variables like test data are independent whereas predicted data is dependent on test data. Group statistics and independent sample tests were performed on the experimental results and the graph was built for two graphs with two parameters under the study. The analysis of the experiment is represented in bar graph (comparison between LSTM and random forest algorithm). A table for comparison of loss of accuracy is drawn from the spss tool. The above analysis paves a path to conclude the effectiveness of the algorithm and final conclusion is drawn.

3. Results

The proposed LSTM technique and the existing random forest algorithm were run in a jupyter notebook one at a time. The accuracy and loss values of LSTM and random forest increase as the sample sets are run for a number of iterations. Table 1 shows the significant levels for LSTM and P=0.01 was used to evaluate random forest models. With a 43.50% chance of being correct, both LSTM and random forest have a less significant level less than 0.05.

4. Discussion

The proposed innovative detection model is illustrated in using a Bi-directional LSTM-recurrent neural network. First, the news articles are pre-processed(Hassanien, Elghamrawy, and Zelinka 2021). Each news piece is given a binary designation, with 1 indicating fake news and 0 indicating true news(United Nations 2019). Punctuation and stop words are removed from the input news items before they are converted to UTF-8 format(Hassanien, Elghamrawy, and Zelinka 2021).The title and content text of news stories are converted into space-separated padded sequences of words(Cristianini, Shawe-Taylor, and Department of Computer Science Royal Holloway John Shawe-Taylor 2000).

These Sequences are further subdivided into token lists.Stanford NLP team provides Global Vectors for Word Representation (GloVe) embeddings (Rana et al. 2019). It is a method for producing vector representations for words that is based on unsupervised learning(Ireton and Posetti 2018). To deal with the high dimensional news items, pre-trained GloVe word embeddings are used. Instead of loading random weights, the embedding layer will load weights from GloVe. GloVe uses global aggregated co-occurrence statistics throughout the entire corpus of news articles (Sarkar 2019). Significant linear substructures of the word vector space are formalised as a result of the representations. The transformed vector-represented data is divided into three categories: train, validation, and test(Shu and Liu 2019).The training is based on a corpus of news articles(Gunjan et al. 2020).

The model is fine-tuned using the validation data set. The test data is also used to determine the expected label of a news article using the trained model. To detect bogus news, researchers used CNN, RNN variations such as Vanilla RNN, LSTM-RNN, and Bi-directional LSTM-RNN, and Bi-directional LSTM-RNN (Office of the Director of National Intelligence Council 2017). Each embedding layer corresponding to training data is supplied into CNN if CNN is chosen as the model. For evaluating CNN performance, several filter sizes are utilised (D'Ulizia et al. 2021).

Although the proposed methodology attained satisfactory results,the limitation in the proposed approach is that there needs to be improved accurate news detection. In future this can be combined with more data text files which can produce better results.

5. Conclusion

The results show that the proposed LSTM outperforms random forest in terms of accuracy and loss for innovative detection of fake news. The proposed LSTM proves with better accuracy (56%) when compared with random forest for detecting fake news in sms services.

DECLARATIONS

Conflicts of Interests

No conflict of interests in this manuscript.

Authors contributions

Author PK was involved in data collection, data analysis implementation,algorithm forming and manuscript writing. Author JJT was involved in

designing the workflow, guidance and review of manuscript.

Acknowledgements
We would like to acknowledge saveetha school of engineering ,saveetha institute of medical and technical sciences.(formerly known as saveetha university)for providing facilities and constant services to this study.

Funding
thank the following organizations for providing financial support that enabled us to complete the study.
1.Vee Eee Technologies and Solutions Pvt.Ltd
2.Saveetha university
3.Saveetha institute of medical and technical sciences
4.Saveetha school of engineering

6. References

Baraneedharan, P., Sethumathavan Vadivel, C. A. Anil, S. Beer Mohamed, and Saravanan Rajendran. 2022. "Advances in Preparation, Mechanism and Applications of Various Carbon Materials in Environmental Applications: A Review." *Chemosphere*. https://doi.org/10.1016/j.chemosphere.2022.134596.

Cristianini, Nello, John Shawe-Taylor, and Department of Computer Science Royal Holloway John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press.

Deena, Santhana Raj, A. S. Vickram, S. Manikandan, R. Subbaiya, N. Karmegam, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2022. "Enhanced Biogas Production from Food Waste and Activated Sludge Using Advanced Techniques – A Review." *Bioresource Technology*. https://doi.org/10.1016/j.biortech.2022.127234.

D'Ulizia, Arianna, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. "Fake News Detection: A Survey of Evaluation Datasets." *PeerJ. Computer Science* 7 (June): e518.

Gunjan, Vinit Kumar, Sabrina Senatore, Amit Kumar, Xiao-Zhi Gao, and Suresh Merugu. 2020. *Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies*. Springer Nature.

Harvard Business Review, Don Tapscott, Marco Iansiti, and Karim R. Lakhani. 2019. *Blockchain: The Insights You Need from Harvard Business Review*. Harvard Business Press.

Hassanien, Aboul-Ella, Sally M. Elghamrawy, and Ivan Zelinka. 2021. *Advances in Data Science and Intelligent Data Communication Technologies for COVID-19: Innovative Solutions Against COVID-19*. Springer Nature.

Ireton, Cherilyn, and Julie Posetti. 2018. *Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training*. UNESCO Publishing.

Karpagam, M., R. Beaulah Jeyavathana, Sathiya Kumar Chinnappan, K. V. Kanimozhi, and M. Sambath. 2022. "A Novel Face Recognition Model for Fighting against Human Trafficking in Surveillance Videos and Rescuing Victims." *Soft Computing*. https://doi.org/10.1007/s00500-022-06931-1.

Kumar, P. Ganesh, P. Ganesh Kumar, Rajendran Prabakaran, D. Sakthivadivel, P. Somasundaram, V. S. Vigneswaran, and Sung Chul Kim. 2022. "Ultrasonication Time Optimization for Multi-Walled Carbon Nanotube Based Therminol-55 Nanofluid: An Experimental Investigation." *Journal of Thermal Analysis and Calorimetry*. https://doi.org/10.1007/s10973-022-11298-4.

Nagarajan, Karthik, Arul Rajagopalan, S. Angalaeswari, L. Natrayan, and Wubishet Degife Mammo. 2022. "Combined Economic Emission Dispatch of Microgrid with the Incorporation of Renewable Energy Sources Using Improved Mayfly Optimization Algorithm." *Computational Intelligence and Neuroscience* 2022 (April): 6461690.

Nagaraju, V., B. R. Tapas Bapu, P. Bhuvaneswari, R. Anita, P. G. Kuppusamy, and S. Usha. 2022. "Role of Silicon Carbide Nanoparticle on Electromagnetic Interference Shielding Behavior of Carbon Fibre Epoxy Nanocomposites in 3-18GHz Frequency Bands." *Silicon*. https://doi.org/10.1007/s12633-022-01825-1.

Office of the Director of National Intelligence Council. 2017. *Global Trends 2030: Alternative Worlds*. Createspace Independent Publishing Platform.

Pandiyan, P., R. Sitharthan, S. Saravanan, Natarajan Prabaharan, M. Ramji Tiwari, T. Chinnadurai, T. Yuvaraj, and K. R. Devabalaji. 2022. "A Comprehensive Review of the Prospects for Rural Electrification Using Stand-Alone and Hybrid Energy Technologies." *Sustainable Energy Technologies and Assessments*. https://doi.org/10.1016/j.seta.2022.102155.

Prabhu, C. S. R., Aneesh Sreevallabh Chivukula, Aditya Mogadala, Rohit Ghosh, and L. M.

Jenila Livingston. 2019. *Big Data Analytics: Systems, Algorithms, Applications*. Springer Nature.

Rana, Nripendra P., Emma L. Slade, Ganesh P. Sahu, Hatice Kizgin, Nitish Singh, Bidit Dey, Anabel Gutierrez, and Yogesh K. Dwivedi. 2019. *Digital and Social Media Marketing: Emerging Applications and Theoretical Development*. Springer Nature.

Sarkar, Dipanjan. 2019. *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*. Apress.

Schwab, Klaus. 2017. *The Fourth Industrial Revolution*. Penguin UK.

Shu, Kai, and Huan Liu. 2019. *Detecting Fake News on Social Media*. Morgan & Claypool Publishers.

United Nations. 2019. *World Migration Report 2020*. United Nations.

United Nations Publications. 2019. *Digital Economy Report 2019: Value Creation and Capture - Implications for Developing Countries*.

Venu, Harish, Ibham Veza, Lokesh Selvam, Prabhu Appavu, V. Dhana Raju, Lingesan Subramani, and Jayashri N. Nair. 2022. "Analysis of Particle Size Diameter (PSD), Mass Fraction Burnt (MFB) and Particulate Number (PN) Emissions in a Diesel Engine Powered by Diesel/biodiesel/n-Amyl Alcohol Blends." *Energy*. https://doi.org/10.1016/j.energy.2022.123806.

Whangchai, Niwooti, Daovieng Yaibouathong, Pattranan Junluthin, Deepanraj Balakrishnan, Yuwalee Unpaprom, Rameshprabu Ramaraj, and Tipsukhon Pimpimol. 2022. "Effect of Biogas Sludge Meal Supplement in Feed on Growth Performance Molting Period and Production Cost of Giant Freshwater Prawn Culture." *Chemosphere* 301 (August): 134638.

Yaashikaa, P. R., M. Keerthana Devi, and P. Senthil Kumar. 2022. "Advances in the Application of Immobilized Enzyme for the Remediation of Hazardous Pollutant: A Review." *Chemosphere* 299 (July): 134390.

TABLES AND FIGURES

Table 1: Comparison of accuracy and loss obtained between Randomforest and LSTM

	Algorithm	N	mean	Std.deviation	std.Error mean
Accuracy	LSTM	3	39.6667	14.66648	9.62239
	RandomForest	3	31.333	10.0664	5.81187
Loss	LSTM	3	33.8667	12.00181	6.92925
	RandomForest	3	51.8733	3.39837	1.96205

Table 2:Independent samples test analysis

Accuracy	**Independent Samples Test**								
	Levene's Test for Equality of Variances		**T-test for Equality of Means**						
	F	**Sig**	**t**	**df**	**Sig(2-tailed)**	**Mean Difference**	**Std.Error Difference**	**95% Confidence Interval of the Difference**	
								Lower	**Upper**
Equal variances assumed	0.267	0.611	4.901	18	0.438	0.7790	0.15894	0.44509	1.11291
Equal variances not			4.901	17.201	0.439	0.7790	0.15894	0.44397	1.11403

assumed									

Figure 1: Comparison of mean accuracy

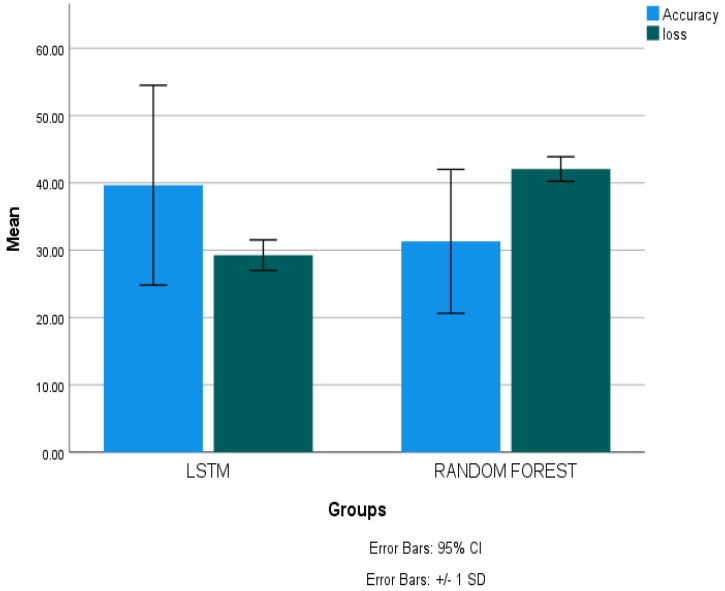

Fig. 1. Comparison of mean accuracy and loss of both LSTM and random forest. The standard error appears to be less in LSTM compared to random forest also the standard error appears +/-2 SD. X-axis: LSTM vs random forest algorithm. Y-axis : mean accuracy.