Original Article

Automatic Identification of Fake News Circulation in Social Media using Logistic Regression over Naïve Bayes and Xg Boost Algorithm to Improve Accuracy

C. Balaji[1], A.Prabhu Chakkaravarthy[2]

[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, Corresponding Author, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India. Pincode: 602105.

Abstract

Aim: To Detect the Fake News in Social Media using Logistic Regression and XGBoost Algorithms. To achieve accuracy a novel logistic regression is used. **Materials and Methods**: The datasets extracted from the kaggle data world and those datasets named as 'TRUE' and 'FAKE'. Accuracy and loss are performed with datasets from kaggle library. The total sample size is 20. The two groups considered were Logistic regression (N=10) and xg boost (N=10). **Results:** Novel logistic regression pops up with the mean accuracy of when contrasted with the Xgboost algorithm. Mean Accuracy value for Logistic Regression and XgBoost Algorithm is 93.68 and 92.93 respectively. Mean Loss value for Logistic Regression and Xg boost is 6.31 and 7.06 respectively. Ultimately Novel logistic regression(NLR) pops up with a better significant value than the Xgboost algorithm. The two algorithms NLR and Xgboost are statistically satisfied with the independent sample T-Test value ($p<0.006$) with confidence level of 95%. **Conclusion**: Detecting fake news significantly seems to be better in Logistic Regression than xg boost.

Keywords: Fake news detection, Machine learning, Novel logistic regression, Xgboost, Logistic regression, Supervised algorithm.

DOI: 10.47750/pnr.2022.13.S04.073

INTRODUCTION

Fake News is a term that was less heard in the previous decade but now due to massive growth of social media networks fake news has become the most surfed term on the internet (Kaliyar, Goswami, and Narang 2021). Considering the huge number of users for the online media i.e. the online detection of fake news on social media networks is probably the only way to take necessary measures and currently getting huge attention from researchers from across the world (Singh and Sharma 2021). Most of the researchers are focused on detecting the fake news of a particular category (ex. Political, organizational, personal). Fake news led to more remarkable problems and threats such as Slandering, confusion etc (Murayama et al. 2021); (Breiman 2017).

In the field of Machine learning, research papers on fake news detection covers journals 75 from IEEE Xplore digital library and 10 articles were published in research gate. In the ongoing research work, fake news detection is exhibited on the basis of repeated words, writing style of content,text classification, sentiment analysis. Knowledge based classification helps in improving the quality of detecting fake news. Binary logistic regression probabilistic model to predict the fake news and Logistic Regression Parameter estimation model to select the Independent variables (Waikhom and Goswami, n.d.). True and Fake news is collected from the previous records and logistic regression is applied to this data to examine the contribution of different features causing Fake news (Alameri and Mohd 2021). Based on the performance and sensitivity measure analysis, Logit model also known as Logistic Regression is chosen as the superior approach in predicting the severity of fake news(Abdulrahman and Baykara 2020). Developed a Binary classification modeling approach to predict the probability that the news is fake (Nikam and Nikam 2021). Xg Boost and Logistic Regression approaches are used for testing the predictive approach. Created classification models using Logistic regression, Naive Bayes, Extreme Gradient Boosting Tree

and Artificial neural networks classification methods (Shaikh and Patil 2020). For this research work five different algorithms are used and found out that Logistic Regression performs well among other algorithms (Baarir and Djeffal 2021).
Our team has extensive knowledge and research experience that has translate into high quality publications(Bhansali et al. 2021; Jayanth et al. 2021; Sudhakar, Ravel, and Perumal 2021; Sathiyamoorthi et al. 2021; Deepanraj et al. 2021; Raju et al. 2021; Arun Prakash et al. 2020; Kamath et al. 2020; Shanmugam et al. 2021; Rajasekaran et al. 2020; Adhinarayanan et al. 2020; Rajesh et al. 2020; Aurtherson et al. 2021). The research gap identified is that the Xg boost algorithm has less accuracy. Detection of fake news using natural language is shown at a very low percentage while analyzing and manual input is not possible to add to the dataset. The aim of research work is to improve accuracy of detecting fake news, and to reduce loss of data while training and testing dataset. The novel Logistic Regression function used to achieve accuracy.

MATERIALS AND METHODS

The ongoing research work is accomplished in the Data Analytics Lab, Saveetha School of Engineering. Experimented laboratory makes use of for assessing the fake news detection from an open dataset available in kaggle. Group 1 is the Novel Logistic Regression algorithm and group 2 is the Xg boost algorithm. Sample size is taken as 10 for each group. Computation is performed using G-power and the obtained G-power is 80% with a confidence interval at 95% and alpha value is 0.05 and beta value is 0.2 (Wang et al. 2021).
Two datasets named TRUE and FAKE containing the data of Fake news in the previous years are retrieved from the Kaggle Data World. After successful preprocessing the dataset consists of 500 rows and columns describing the Text of the news and the subject of the news. The dataset is partitioned into preparing and testing parts. The dataset is read into the program and the data in this dataset is given as input using pandas library. Logistic regression Classifiers present in sklearn library and Xg boost algorithm classifiers present in sequential libraries are imported and from predict() function, accident severity is predicted. The output is displayed either as True or Fake.

Logistic Regression Algorithm
Novel Logistic Regression falls under the Supervised algorithm. It is applied to predicting dependent variables by utilizing a given set of independent variables. Logistic regression algorithm specifies the yield of a dependent variable in the form of either ‘yes’ or ‘no’ and ‘true’ or ‘false’, which means there will be only two classes. Logit model is used to predict the output randomly based on the given independent features. The results produced using this algorithm are in 0’s and 1’s. Logistic Regression is also similar to linear regression except that in case of Logistic Regression, a curve is obtained as a result and in linear regression a straight line is obtained.Table 1 represents the pseudocode for logistic regression. Table 3 represents accuracy of Fake News Detection classification using Logistic Regression (Wang et al. 2021; Masciari et al. 2020).

Xgboost algorithm
Pseudocode for the XGBoost algorithm is given in Table 2. Gradient framework is used by the xg boost ensemble learning method. Xg Boost is designed for speed and performance and is an implementation of gradient boosted decision trees. Xg Boost differentiates in a wide range of applications, Probability, Languages and Cloud. Table 2 represents the pseudocode for the xg boost algorithm. Table 4 represents accuracy of Fake News Detection classification using xg boost algorithm. Equation 1 shows the calculation of accuracy.

$$\text{Accuracy} = TP+TN/\ TP+TN+FP+FN \qquad (1)$$

Where,

TP represents no. of true positives classified by model.
FP represents no.of false positives classified by model.
TN represents no.of true negative classified by model.
FN represents no.of false negatives classified by model.

For executing this work a jupyter notebook is essential. Algorithms are compatible with 64-bit, the System should carry a minimum of 8GB RAM and 256GB SSD + 1TB HDD ROM for processing the data, and Intel i5 Processor. The language used for executing each cell in the proposed system is python.

Statistical Analysis
Statistical tool used for our study is IBM SPSS. Independent sample t test is performed and significance values are obtained. The independent variables are Title, Text, Subject, Date and the dependent variable is Target which describes the detection of fake news in True or Fake format.

RESULTS

The experimental results are carried out on the Logistic Regression Algorithm and Naive algorithm where the performance is measured based on accuracy. Table 5 represents the comparison of the Logistic Regression Algorithm and Xg boost algorithm. The accuracy of the Logistic Regression Algorithm is 93.68% and the Xg boost algorithm is 90.05%.
Table 6 and Table 7 represents the independent sample test that has been performed on novel Logistic model and Xg boost algorithms for calculating the equal variance assumed and equal variance not assumed and it also shows mean difference, standard error differences with a confidence level of 95%. Mean Accuracy value for Logistic Regression and XgBoost Algorithm is 97.5 and 96.12 respectively. Mean Loss value for Logistic Regression and Decision Tree is 2.4843 and 3.7333 respectively.
Fig. 1 shows bar chart comparison of two algorithms. Mean Accuracy value for Logistic Regression and XgBoost Algorithm is 93.68 and 92.93 respectively. Mean Loss value for Logistic Regression and Xg boost is 6.31 and 7.06 respectively. The error bars are shown in the graph and error rate is less for Logistic regression algorithm compared to Xg boost algorithm.

DISCUSSION

In this study, accuracy of the Novel Logistic Regression algorithm is significantly higher than Xg boost Algorithm. Outcome accuracy value of novel logistic regression is 98.76% with higher better value than Xg boost algorithm value of 94.72%.
To support this research work, it is shown that fake news detection is classified using Decision Tree, Random Forest, Support vector machines and Logistic regression algorithms and the comparison of accuracies among these three algorithms provided that Logistic Regression algorithm achieved higher accuracy of 90% (Han and Mehta 2019). The compared prediction performances of Neural Networks, Multinomial Logistic regression and Bayesian Logistic regression (Ngada and Haskins 2020). These comparison results showed that the Multinomial Logistic regression model produced higher accuracy of 82.74%(Shu and Liu 2019). Predicted fake news using the comparison between XG Boost and Novel Logistic Regression and Logistic Regression resulted in the less mean error rate (9.25% to 11.78%). To oppose this research work, when compared to XGBoost performed the fake news detection using Logistic Regression algorithm and achieved an accuracy of 92.7% (Patel and Meehan 2021). The XGBoost algorithm is used to determine the fake news severity(Konagala and Bano 2021). The TFIDFtransformer, count vectorizer and trained the data on four algorithms and found out that Logistic Regression achieved the best possible results (Akulwar and Khobragade 2021).

CONCLUSION

Accurate and effective model has been developed for fake news detection on social media using the Novel logistic regression model. The proposed model type reveals significantly higher accuracy than the Xgboost algorithm.

DECLARATIONS

Conflict of Interests
No conflict of interest in this manuscript.

Authors Contribution
Author CB was involved in data collection, data analysis, Manuscript writing. Author APC assisted in conceptualization, data validation and critical review of manuscript.

Acknowledgement
Author would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, Saveetha University for providing the opportunities and facilities to carry out research study.

Funding
We thank the following organization for providing financial support that enabled us to complete study.

1. Insysness Technology Pvt. Ltd. Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Abdulrahman, Awf, and Muhammet Baykara. 2020. "Fake News Detection Using Machine Learning and Deep Learning Algorithms." 2020 International Conference on Advanced Science and Engineering (ICOASE). https://doi.org/10.1109/icoase51841.2020.9436605.
2. Adhinarayanan, Rajesh, Aravindh Ramakrishnan, Gopal Kaliyaperumal, Melvinvíctor De Poures, Rajesh Kumar Babu, and Damodharan Dillikannan. 2020. "Comparative Analysis on the Effect of 1-Decanol and Di-N-Butyl Ether as Additive with diesel/LDPE Blends in Compression Ignition Engine." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, June, 1–18.
3. Akulwar, Akanksha, and Anish Khobragade. 2021. "FAKE NEWS DETECTION USING VARIOUS MACHINE LEARNING TECHNIQUES." EPRA International Journal of Research & Development (IJRD). https://doi.org/10.36713/epra6715.
4. Alameri, Saeed Amer, and Masnizah Mohd. 2021. "Comparison of Fake News Detection Using Machine Learning and Deep Learning Techniques." 2021 3rd International Cyber Resilience Conference (CRC). https://doi.org/10.1109/crc50527.2021.9392458.
5. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. "Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite." Biomass Conversion and Biorefinery, August. https://doi.org/10.1007/s13399-020-00938-0.
6. Aurtherson, P. Babu, Bhanu Teja Nalla, Karthikeyan Srinivasan, Kulmani Mehar, and Yuvarajan Devarajan. 2021. "Biofuel Production from Novel Prunus Domestica Kernel Oil: Process Optimization Technique." Biomass Conversion and Biorefinery, May. https://doi.org/10.1007/s13399-021-01551-5.
7. Baarir, Nihel Fatima, and Abdelhamid Djeffal. 2021. "Fake News Detection Using Machine Learning." 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-Being (IHSH). https://doi.org/10.1109/ihsh51661.2021.9378748.
8. Bhansali, Karan J., Kamlesh R. Balinge, Subodh U. Raut, Shubham A. Deshmukh, M. Senthil Kumar, C. Ramesh Kumar, and Pundlik R. Bhagat. 2021. "Visible Light Assisted Sulfonic Acid-Functionalized Porphyrin Comprising Benzimidazolium Moiety for Photocatalytic Transesterification of Castor Oil." Fuel 304 (November): 121490.
9. Breiman, Leo. 2017. Classification and Regression Trees. Routledge.
10. Deepanraj, B., N. Senthilkumar, D. Mala, and A. Sathiamourthy. 2021. "Cashew Nut Shell Liquid as Alternate Fuel for CI Engine—optimization Approach for Performance Improvement." Biomass Conversion and Biorefinery, February. https://doi.org/10.1007/s13399-021-01312-4.
11. Han, Wenlin, and Varshil Mehta. 2019. "Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation." 2019 IEEE International Conference on Industrial Internet (ICII). https://doi.org/10.1109/icii.2019.00070.
12. Jayanth, Bellappu Venkat, Melvin Victor Depoures, Gopal Kaliyaperumal, Damodharan Dillikannan, Dilipsingh Jawahar, Kumaran Palani, and Ganesha Prasad Meravanigee Shivappa. 2021. "A Comprehensive Study on the Effects of Multiple Injection Strategies and Exhaust Gas Recirculation on Diesel Engine Characteristics That Utilize Waste High Density Polyethylene Oil." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, June, 1–18.
13. Kaliyar, Rohit Kumar, Anurag Goswami, and Pratik Narang. 2021. "FakeBERT: Fake News Detection in Social Media with a BERT-Based Deep Learning Approach." Multimedia Tools and Applications, January, 1–24.
14. Kamath, Manjunath, Subha Krishna Rao, Jaison, Sridhar, Kasthuri, Gopinath, Sivaperumal, and Shantanu Patil. 2020. "Melatonin Delivery from PCL Scaffold Enhances Glycosaminoglycans Deposition in Human Chondrocytes – Bioactive Scaffold Model for Cartilage Regeneration." Process Biochemistry 99 (December): 36–47.
15. Konagala, Varalakshmi, and Shahana Bano. 2021. "Fake News Detection Using Deep Learning." Research Anthology on Fake News, Political Warfare, and Combatting the Spread of Misinformation. https://doi.org/10.4018/978-1-7998-7291-7.ch025.
16. Masciari, Elio, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlì. 2020. "Leveraging Machine Learning for Fake News Detection." Proceedings of the 9th International Conference on Data Science, Technology and Applications. https://doi.org/10.5220/0009767401510157.
17. Murayama, Taichi, Shoko Wakamiya, Eiji Aramaki, and Ryota Kobayashi. 2021. "Modeling the Spread of Fake News on Twitter." PloS One 16 (4): e0250419.
18. Ngada, Okuhle, and Bertram Haskins. 2020. "Fake News Detection Using Content-Based Features and Machine Learning." 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). https://doi.org/10.1109/csde50874.2020.9411638.
19. Nikam, Prof Rohit, and Rohit Nikam. 2021. "Location Based Fake News Detection Using Machine Learning." International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/ijraset.2021.34939.
20. Patel, Ankitkumar, and Kevin Meehan. 2021. "Fake News Detection on Reddit Utilising CountVectorizer and Term Frequency-Inverse Document Frequency with Logistic Regression, MultinominalNB and Support Vector Machine." 2021 32nd Irish Signals and Systems Conference (ISSC). https://doi.org/10.1109/issc52156.2021.9467842.
21. Rajasekaran, S., D. Damodharan, K. Gopal, B. Rajesh Kumar, and Melvin Victor De Poures. 2020. "Collective Influence of 1-Decanol Addition, Injection Pressure and EGR on Diesel Engine Characteristics Fueled with diesel/LDPE Oil Blends." Fuel 277 (October): 118166.
22. Rajesh, A., K. Gopal, De Poures Melvin Victor, B. Rajesh Kumar, A. P. Sathiyagnanam, and D. Damodharan. 2020. "Effect of Anisole Addition to Waste Cooking Oil Methyl Ester on Combustion, Emission and Performance Characteristics of a DI Diesel Engine without Any Modifications." Fuel 278 (October): 118315.
23. Raju, P., K. Raja, K. Lingadurai, T. Maridurai, and S. C. Prasanna. 2021. "Glass/Caryota Urens Hybridized Fibre-Reinforced nanoclay/SiC Toughened Epoxy Hybrid Composite: Mechanical, Drop Load Impact, Hydrophobicity and Fatigue Behaviour." Biomass Conversion and Biorefinery, March. https://doi.org/10.1007/s13399-021-01427-8.
24. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." Environmental Progress & Sustainable Energy 40 (6). https://doi.org/10.1002/ep.13696.
25. Shaikh, Jasmine, and Rupali Patil. 2020. "Fake News Detection Using Machine Learning." 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC). https://doi.org/10.1109/isssc50941.2020.9358890.
26. Shanmugam, Rajasekaran, Damodharan Dillikannan, Gopal Kaliyaperumal, Melvin Victor De Poures, and Rajesh Kumar Babu. 2021. "A Comprehensive Study on the Effects of 1-Decanol, Compression Ratio and Exhaust Gas Recirculation on Diesel Engine Characteristics Powered with Low Density Polyethylene Oil." Energy Sources, Part A: Recovery, Utilization, and Environmental Effects 43 (23): 3064–81.
27. Shu, Kai, and Huan Liu. 2019. Detecting Fake News on Social Media. Morgan & Claypool Publishers.
28. Singh, Bhuvanesh, and Dilip Kumar Sharma. 2021. "Predicting Image Credibility in Fake News over Social Media Using Multi-Modal Approach." Neural Computing & Applications, May, 1–15.
29. Sudhakar, M. P., Merlyn Ravel, and K. Perumal. 2021. "Pretreatment and Process Optimization of Bioethanol Production from Spent Biomass of Ganoderma Lucidum Using Saccharomyces Cerevisiae." Fuel 306 (December): 121680.

30. Waikhom, Lilapati, and Rajat Subhra Goswami. n.d. "Fake News Detection Using Machine Learning." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3462938.
31. Wang, Yuhang, Li Wang, Yanjie Yang, and Tao Lian. 2021. "SemSeq4FD: Integrating Global Semantic Relationship and Local Sequential Order to Enhance Text Representation for Fake News Detection." Expert Systems with Applications 166 (March): 114090.

TABLES AND FIGURES

Table 1. Pseudo code for Logistic Regression

S.No	Steps
1	Assign variable name to dataset
2	Clean data i.e removing stopwords, punctuation etc.
3	Importing word cloud to count occurrences of a word
4	Prepare model and train data
5	By using function pipeline() , activate model
6	Predicted Target is returned as output
7	Output: Fake news detection

Table2: Pseudocode for XG Boost algorithm

S.No	Steps
1	Import all packages required (i.e. Pandas, Numpy, Scikit learn.
2	Variable names are set for imported datasets
3	Using pandas package input commands to clean data
4	Explore data using word cloud package
5	Note, most repeated words from fake news data
6	Import XgBoost library from Scikit Learn package
7	Using functions countvectorizer, tfidftransformer and xgboost classifier find output

Table 3. Accuracy of Fake News Detection classification using Logistic Regression (mean accuracy=93.6890%, mean loss=6.3%)

Test	**Accuracy**	**Loss**
Test 1	98.76	1.24
Test 2	97.55	2.45
Test 3	96.24	3.76
Test 4	95.33	4.67

Test 5	94.22	5.78
Test 6	93.89	6.11
Test 7	92.45	7.55
Test 8	90.15	9.85
Test 9	89.75	10.25
Test 10	88.54	11.46

Table 4. Accuracy of Fake News Detection classification using XGBoost (mean accuracy= 92.93%, mean loss=7.06%)

Test	Accuracy	Loss
Test 1	97.82	2.18
Test 2	96.20	3.80
Test 3	95.30	4.70
Test 4	94.50	5.50
Test 5	93.40	6.60
Test 6	92.60	7.40
Test 7	91.80	8.20
Test 8	90.70	9.30
Test 9	89.90	10.10
Test 10	87.10	12.90

Table 5. Group Statistics analysis for both algorithms based on Accuracy and Loss.

Algorithm		N	Mean	Std.Deviation	Std.Error Mean
Accuracy	LR	10	93.6890	3.43564	1.08645
	XGB	10	92.9320	3.20479	1.01344
Loss	LR	10	6.3120	3.43564	1.08645
	XGB	10	7.0680	3.20479	1.01344

Table 6. An Independent sample T-Test of accuracy for detecting Fake news using Logistic Regression and Xg boost algorithm. Logistic Regression algorithm appears to perform significantly better than the Xg boost algorithm based on accuracy and loss. (p=0.7)

		F	Sig	t	df	sig(2-tailed)	Mean Difference	Std Error difference	Lower	Upper
Accuracy	Equal Variance assumed	0.104	0.751	0.509	18	0.617	0.75600	1.48574	-2.36543	3.87743
	Equal Variance not assumed			0.509	17.914	0.617	0.75600	1.48574	-2.36543	3.87851

Table 7. An Independent sample T-Test of accuracy for detecting Fake news using Logistic Regression and Xg boost algorithm. Logistic Regression algorithm appears to perform significantly better than the Xg boost algorithm based on accuracy and loss. (p=0.7)

		F	Sig	t	df	sig(2-tailed)	Mean Difference	Std Error difference	Lower	Upper
Loss	Equal Variance assumed	0.104	0.751	-0.509	18	0.617	-0.75600	1.48574	-3.87743	2.36543
	Equal Variance not assumed			-0.509	17.914	0.617	-0.75600	1.48574	-3.87743	2.36651

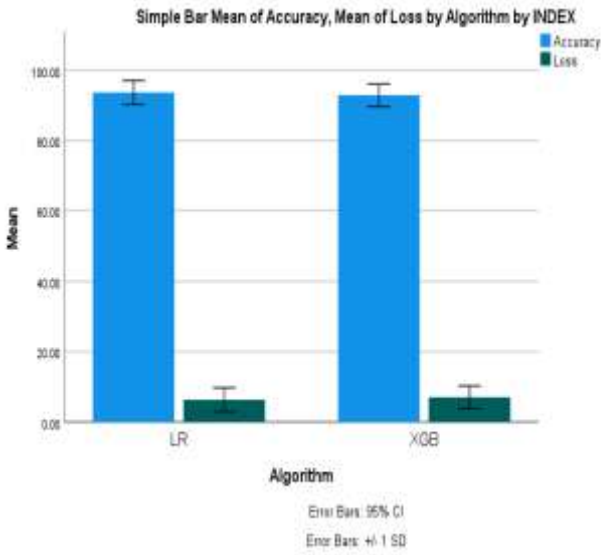

Fig. 1. Comparison of Logistic regression algorithm and Xg boost Algorithm in terms of mean accuracy. Mean accuracy of Logistic Regression is better than Xg boost algorithm and standard deviation of Logistic Regression is slightly better than Xg boost. X Axis: Logistic Regression vs Xg boost. Y Axis : Mean Accuracy of detection = +/- 1 SD.