British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 7, Issue 1, 2024 (pp. 115-121)
www.abjournals.org

FAKE NEWS DETECTION SYSTEM USING LOGISTIC REGRESSION, DECISION TREE AND RANDOM FOREST

Oni Oluwabunmi Ayankemi[1*], Idowu Oluwaferanmi Ruth[2], and Bassir Abdullai Abiye[3]

[1-3]Department of Computer Studies, Faculty of Science, The Polytechnic Ibadan, Ibadan.

[*]Corresponding Author's Email: onioluwabunmia@gmail.com

Cite this article:

Oluwabunmi O. A., Oluwaferanmi I. R., Abdullai B. A. (2024), Fake News Detection System Using Logistic Regression, Decision Tree and Random Forest. British Journal of Computer, Networking and Information Technology 7(1), 115-121. DOI: 10.52589/BJCNIT-IOYRPY7G

Manuscript History

Received: 18 Jan 2024

Accepted: 23 Apr 2024

Published: 17 May 2024

ABSTRACT: *The purpose of this study is to design a fake news detection system with these three machine learning models, namely: Decision Tree, Random Forest, and Logistic Regression. These three different models were analysed to determine the most efficient model for accurately detecting fake news. The result obtained showcased Logistic Regression with an accuracy of 98.80%, Decision Tree with an accuracy of 99.64% and Random Forest with an accuracy of 99.23%. It is evident as deduction from the comparative analysis that our best model came out to be Decision Tree with an accuracy of 99.64%.*

KEYWORDS: Internet, Social media, Fake news, Classification, Machine learning, Data set, Computational technique.

Article DOI: 10.52589/BJCNIT-IOYRPY7G
DOI URL: https://doi.org/10.52589/BJCNIT-IOYRPY7G

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 7, Issue 1, 2024 (pp. 115-121)

www.abjournals.org

INTRODUCTION

Fake news is a main consideration in triggering riots, mob lynchings, and other social-monetary aggravations. The severity of the problem has increased substantially since 2019 was declared the year of fake news. News and media inclusion get enormously misshapen because of the introduction and spread of phoney news. Where news can be a shelter, counterfeit news is a plague to the general public. In any case, the distinction between verifiable news and fake news is troublesome. Its examination consequently assumes a vital role in the current situation.

Fortunately, there are several computational methods that could be applied to recognize some posts as fraudulent based solely on the text in them. Most of these techniques rely on websites for fact-checking. Researchers keep a range of archives that contain lists of previously visited websites that have been labelled as dubious and false

Due to the spread of misleading information on the internet, particularly in news blogs, feeds, and social media channels, the detection of fake news has recently sparked a growing amount of public curiosity and interest from scholars. Studies have mostly concentrated on identifying and classifying bogus news on social platforms like Twitter and Facebook. Fake news has been conceptually divided into various types, and this knowledge has been disseminated to enable the generalisation of machine learning (ML) models across various domains. Some of these models include Support Vector Machine (SVM) and Logistic Regression (LR), whose accuracy was the highest (92%) among K-nearest neighbour (KNN), Random Forest, Linear Support Vector Machine (LSVM), Decision Tree (DT), and Stochastic Gradient Descent (SGD). There searchers found that the total accuracy decreased as the number of n-grams calculated for a particular article rose. The main aim of this project is to design a fake news detection system with machine learning models. We will be using three of these methods (Decision Tree, Random Forest, and Logistic Regression) for this project. The programming language to be used is Python. In this project, we are proposing three machine learning methods that can determine whether an article is credible using datasets obtained from the Kaggle dataset. The system will help users detect if it is real or fake news. The three different models will be analysed to determine the most efficient model for accurately detecting fake news.

RELATED WORKS

A survey on spam detection methods was undertaken by Sharma et al. in 2014. In this study, the authors discovered that in order to classify emails as spam or not starting from a certain dataset, an artificial neural network (ANN) must first be trained. After these features have been gathered, they can be categorised using classifiers such as Naive Bayes, Support Vector Machines, TF-IDF, or K-nearest neighbours. An earlier study on fake news identification that is more directly related to utilising a text-only approach to build a classification is presented in Genes' *Detecting False News with NLP*, published in 2017. The writers did not only produce a new benchmark dataset of utterances with 12,800 brief statements that have been manually tagged on various subjects. Baseline algorithms such as logistic regression, support vector machines (SVM, LSTM, CNN), and an enhanced CNN that included metadata were runned by the database's developers.

Perez-Rosas et al. (2018) paid most attention to the automatic detection of false contents in online news in their paper. They presented two distinct datasets for this purpose: one gathered

Article DOI: 10.52589/BJCNIT-IOYRPY7G
DOI URL: https://doi.org/10.52589/BJCNIT-IOYRPY7G

through information from crowds and spanning six news domains (sports, business, entertainment, politics, technology, and education), and the other gathered from the web and focusing on celebrities. They created some classification models that depend on a combination of lexical, syntactic, and semantic data as well as characteristics representing text readability properties, which are similar to the human capacity to detect forgeries, using a linear sum classifier and fivefold cross verification with accuracy, precision, recall, and FI measures averaged over the five iterations.

In their study, Vedova et al. (2018) first proposed a novel ML fake news detection technique that performs better than the existing methods in the literature by combining news content and social context features, which increased its accuracy up to 78.8 percent. Secondly, they applied said technique within a Facebook Messenger chatbot and verified it with a practical process, yielding an 81.7 percent accuracy in detecting fake news. They first discussed the datasets they utilised for their test, then showed the content-based approach they employed and the way they suggested combining it with a social-based strategy that had previously been proposed in the literature. Their objective was to categorise a news item as reliable or phoney. The last dataset consists of 15,500 posts from 32 pages (14 conspiracy pages, 18 scientific pages), with more than 2,300,000 likes from more than 900,000 individuals. 6,577 posts (42.4% of all posts) and 8,923 (57.6%) are not hoaxes.

Mykhailo Granik et al. (2017) demonstrate a straightforward method for detecting bogus news using a naive Bayes classifier in their study. This strategy was put into practice as a software system and evaluated using a set of Facebook news postings as the test set. They were gathered from three sizable left-leaning and right-leaning Facebook pages, as well as three sizable mainstream news sources for politics (Politico, CNN and ABC News). They succeeded in classifying objects with an accuracy of about 74%. Fake news classification accuracy is marginally worse. The dataset's skewness: only 4.9% of it contains bogus news—could be to blame for this.

Gupta et al. (2018) provided a framework based on several machine learning approaches that addresses a number of problems, such as the lack of accuracy, time lag (Bot Maker), and the lengthy processing time required to handle a thousand tweets in a single second. They started by gathering 400,000 tweets from the HSpam14 dataset. Then they describe in more detail the 250,000 non-spam tweets and the 150,000 spam tweets. Along with the Top-30 terms from the Bag-of-Words model that offer the largest information gain, they also deduced several lightweight features. They outperformed the old solution by almost 18% and were able to reach an accuracy of 91.65 percent.

METHODOLOGY

This project presents the methodology used for the classification. In this method, supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection. The dataset used for classification was drawn from a public domain. Fake news articles were collected from an open source Kaggle dataset that was published during the 2016 election cycle. The news articles were collected from news organisations: Guardians and Bloomberg during the election period. The dataset is already sorted out qualitatively into fake and real where we have 23,481 fake articles and 21,417 real articles. Followed by data

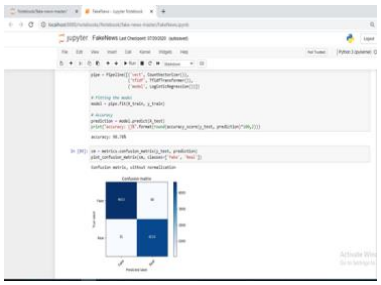
Article DOI: 10.52589/BJCNIT-IOYRPY7G
DOI URL: https://doi.org/10.52589/BJCNIT-IOYRPY7G

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 7, Issue 1, 2024 (pp. 115-121)
www.abjournals.org

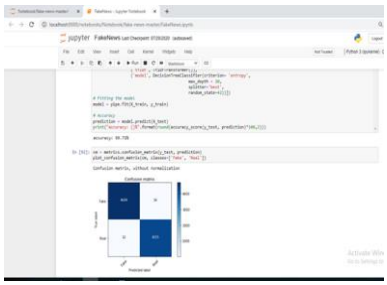
preprocessing, implementing features selection, then performing the training and testing of the dataset and finally running the classifiers. The methodology is based on conducting various experiments on dataset using the three machine algorithms, namely: Logistic Regression, Random Forest and Decision Tree to respond by informing the user of the authenticity of the news.

RESULTS

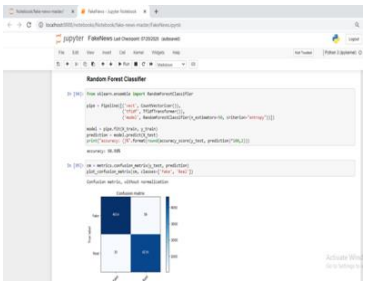
In order to ascertain, the comparison of machine learning models is done to know which models are best to deploy a system, based on different metrics such as accuracy score and confusion matrix to determine the best model for our system depending on what we seek to optimise. However, for the purpose of this research, a comparative analysis on Logistic Regression, Decision Tree and Random Forest for Fake News Detection system was carried out with the focus on seeking the classifier with the highest performance rate. Training, testing and validation of the three different models on the same dataset from the Kaggle dataset was also carried out. After training and testing the models, the scores are saved in the variable model scores, the percentage of each model`s accuracy is then shown in a confusion matrix.



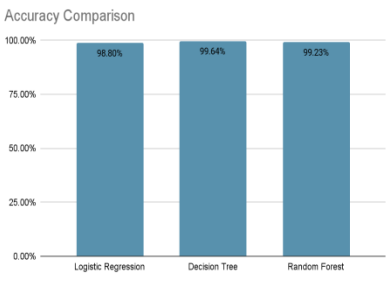
Confusion matrix (Logistic Regression)

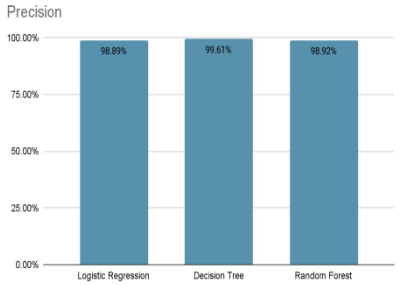


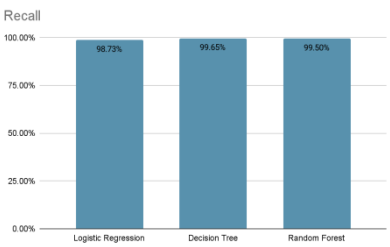
Confusion matrix (Decision tree)

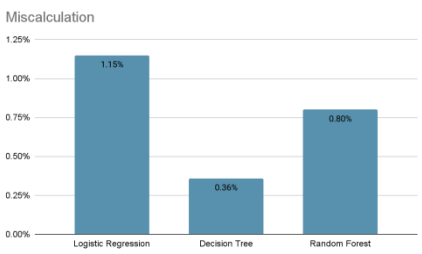


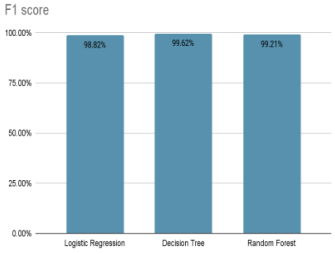
Confusion matrix (Random Forest)











Article DOI: 10.52589/BJCNIT-IOYRPY7G
DOI URL: https://doi.org/10.52589/BJCNIT-IOYRPY7G

Classifier	Accuracy	Miscalculation	Precision	Recall	F1- Score
Logistic Regression	98.80%	1.15%	98.89%	98.73%	98.82%
Decision Tree	99.64%	0.36%	99.61%	99.65%	99.62%
Random Forest	99.23%	0.80%	98.92%	99.50%	99.21%

As evident above, our best model came out to be Decision Tree with an accuracy of 99.64%. Hence we can say that if a user feeds a particular news article or its headline in this model, there are at least 90% chances that it will be classified to its true nature.

SUMMARY

It is significant to find the accuracy of news which is available on the internet. In this project, we have used three of Machine Learning prediction models: Logistic Regression with an accuracy of 98.80%, Decision Tree with an accuracy of 99.64% and Random Forest with an accuracy of 99.23% which the best model came out to be Decision Tree so as to carry out a comparative analysis on effective and efficient detection of fake news. The technique used focuses on using machine learning to statistically analyse given news articles and does not rely on a “blacklist” of news stories from various unverified sources. As with all blacklists, a fake news site that has not been seen previously will not be correctly identified as True News by the Detector. Hence, this project will create more awareness to people. It will contribute to start a much-justified war against one of the most prevalent hazards i.e. fake news in society. It will serve as root and branch eradication of fake news in the society as a whole.

Article DOI: 10.52589/BJCNIT-IOYRPY7G
DOI URL: https://doi.org/10.52589/BJCNIT-IOYRPY7G

REFERENCES

1. Aayush Ranjan, " Fake News Detection Using Machine Learning", Department Of Computer Science & Engineering Delhi Technological University, July 2018.
2. Agrawal, Srishti. "FAKE NEWS DETECTION USING ML." International Research Journal of Engineering and Technology (IRJET), vol. 07, no. 05, 2020, p. 6, https://www.irjet.net/archives/V7/i5/IRJET-V7I51102.pdf.
3. Ahmed H., Traore I., Saad S. (2017) Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments. ISDDC 2017. Lecture Notes in Computer Science, vol 10618. Springer, Cham.
4. Alim A. A. A., Ayman A., Praveen K., D., & Myung S.C., (2021) Detecting Fake News Using Machine Learning: A Systematic Literature Review.
5. Allcott S.H and M. Gentzkow. Social media and fake news in the 2016 election. Journal of Economic Perspectives, 2017
6. Anagha S. A., Aneeta S., Berin J. ,Vineeta S., & Asst Prof Teenu J.(2021) Fake News Detection System Using Machine Learning. International Journal of Advances in Computer Science and Technology (10) 6, 12-15.
7. Anjali J., Harsh K & Avinash S.,(2019) A Smart System For Fake News Detection Using Machine Learning. International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) (1)3, 1-6.
8. Buntain C. and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," 2017 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2017, pp. 208-215.
9. Conroy, N., Rubin, V. and Chen, Y. (2015). "Automatic deception detection: Methods for finding fake news" at Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.
10. Dataset- Fake News detection William Yang Wang. " liar, liar pants on _re": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648, 2017.
11. Della M. L Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2018, pp. 272- 279.
12. Fake news websites. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017
13. Granik M. and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.
14. Great moon hoax. https://en.wikipedia.org/wiki/Great Moon Hoax. [Online; accessed 25-September-2017].
15. Gupta H., Jamal S.H, Madisetty S. and Desarkar M. S. , "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383
16. Iftikhar A., Muhammad Y., Suhail Y.,& Muhammad O.A., (2020). Fake News Detection
Using Machine Learning Ensemble Methods. Complexity (2020) Article ID 8885861, 10-14.
17. Julie Posetti and Alice Matthews (2018). A short guide to the history of 'fake news' and

Article DOI: 10.52589/BJCNIT-IOYRPY7G
DOI URL: https://doi.org/10.52589/BJCNIT-IOYRPY7G

Disinformation. ICFJ International Center For Journalists (4) 1-19.
18. Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, "Fake News Detection using Machine Learning and Natural Language Processing," International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6, March 2019
19. Parab L. S., Sawant G. S., &Gorivale S. M.,(2021) Fake News and Message Detection Project. Department of Computer Engineering PadmabhushanVasantdada Patil Pratishthan's College of Engineering, University of Mumbai.
20. Parikh S. B and Atrey P. K, "Media-Rich Fake News Detection: A Survey," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 436-441
21. Patil S.M., Malik A.K. (2019) Correlation Based Real-Time Data Analysis of Graduate Students Behaviour. In: Santosh K., Hegadi R. (eds) Recent Trends in Image Processing and Pattern Recognition. RTIP2R 2018. Communications in Computer and Information Science, vol 1037. Springer, Singapore.

22. Shailesh-Dhama,"Detecting-Fake-News-with-Python", Github, 2019
23. Shankar M. Patil, Dr. Praveen Kumar, "Data mining model for effective data analysis of higher education students using MapReduce" IJERMT, April 2017 (Volume-6, Issue-4).
24. Srivastava, Aman. "Real Time Fake News Detection Using Machine Learning and NLP." International Research Journal of Engineering and Technology (IRJET),vol. 07, no. 06, 2020, p. 5, https://www.irjet.net/archives/V7/i6/IRJET-V7I6688.pdf
25. Syed I., M., Dr Jimmy S., Nikita, (2019). Fake News Detection Using Machine Learning approaches: A systematic Review. Proceedings of the Third International Conference on Trends in Electronics and Information (ICOEI 2019).230-234.
Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective" arXiv:1708.01967v3 [cs.SI], 3 Sep 2017
26. Tijare, Poonam. (2019). A Study on Fake News Detection Using Naïve Bayes, SVM, Neural Networks and LSTM. Journal of Advanced Research in Dynamical and Control Systems. 11, 06-Special Issue, 2019
27. Ultimate guide to deal with Text Data (using Python) – for Data Scientists and Engineers by Shubham Jain, February 27, 2018
28. Uma Sharma, Sidarth Saran, Shankar M. Patil (2020) Fake news detection using machine language Algorithms. International Journal of Engineering Research and Technology (IJERT) (8)3, 509-518
29. Understanding the random forest by Anirudh Palaparthi, Jan 28, at analytics vidya.
30. What is a Confusion Matrix in Machine Learning by Jason Brownlee on November 18, 2016 in Code Algorithms From Scratch

Article DOI: 10.52589/BJCNIT-IOYRPY7G
DOI URL: https://doi.org/10.52589/BJCNIT-IOYRPY7G