Comparative analysis using Machine Learning to detect Fake news

Nishanth SM Gowda
Department of Data Science and Computer Applications
Manipal Institute of Technology
Manipal Academy of Higher Education
Manipal,Karnataka,India-576404
nishanth1.mitmpl2023@learner.manipal.edu

***Abstract*—With the development of social media and communication tools, the problem of fake news is expanding and happening more quickly. the identification of fake information is a new field of study that is attracting a lot of attention. Falsified news or propaganda that is disseminated via traditional and online media is referred to as fake news. Broadcasting channels, print, television, and non-traditional broadcast media outlets are all included. Online platforms include social media behemoths like Facebook, Twitter, YouTube, websites and others due to the inability of platforms to recognize and stop the spread of false information. The process of detecting fake news entails locating reports that are purposefully fabricated or hoaxes that are spread via online social media platforms or traditional news outlets (print and television). phony websites and Automatically identifying bogus news entails figuring out the news reports Support Vector Machine (SVM), Logistic Regression , Naive Bayes veracity. Our goal is to create and hone a machine learning model for the purpose of automatically identifying false news. This study describes various machine learning-based classifiers for the detection of fake news. This could aid readers in recognizing the true news for improved understanding.**

***Index Terms*—Machine Learning, Fake news detection, Support Vector Machine, Social media, Phishing website , Naive Bayes , Logistic Regression .**

I. Introduction

The term ”fake news” describes material that is untrue or deceptive but is presented as fact; it is frequently disseminated through traditional news sources,websites, and other media. It is intended to mislead and control viewers or readers, frequently for monetary , ideological, or political motives. While there are many advantages to reading news on social media, the lack of oversight and easy access has resulted in at least 20 deaths in India in 2018 due to false information being shared on these platforms.spread of false information or fake news in [1] OSM networks . False narratives, misrepresented facts, edited photos or videos, and deceptive headlines are just a few examples of the various ways that fake news can appear. Concerns over its influence on public discourse have been raised by its widespread use in the digital age[2], faith in the media, and democratic procedures. Critical thinking, fact-checking, and media literacy abilities are necessary to separate reliable information from inaccurate or misleading content in order to recognize and counteract fake news.

Technology advancements make it easier for people to obtain news on internet platforms, which increases their usefulness as information sources.Preventing the dissemination of false information requires the early detection of fake news[3]. There are still those that fabricate connections in order to deceive other people. Even those who use social media platforms responsibly are nonetheless subjected to inaccurate or misleading news. Eventually, it seems like a waste of time to browse social media. Consumers continue to deal with websites whose existence negatively influences readers' decision to interact with misinformation.

The study of statistical models and methods used by computers to do jobs without explicit instructions is known as machine learning (ML). By utilizing training data, machine learning systems can form opinions or predictions without the need for explicit programming. Machine learning and computational statistics are closely related fields. Artificial neural networks are constructed with web-like connections between neuron nodes, just like the human brain. While standard programs generate analyses using data linearly, deep learning systems hierarchical function enables machines to analyze data non-linearly.

The length of the credit assignment path (CAP) depth, or the number of layers that input is processed through, is a characteristic of deep learning systems. The input-to-output transformation chain is called the CAP. CAPs are a way of expressing likely causal relationships. With results on par with, or sometimes even better than, conventional methods, these deep learning architectures have been applied to a variety of fields, including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bio informatics, drug design, medical image analysis, material inspection, and board game programming.

Because of this research, algorithms including Support Vector Machine, SVM, Linear Regression, and Naive Bayes were used to compare the accuracy of identifying fake news[7]. There are various places to find false information. They are making a deliberate effort to spread inventions, claims, and false information that passes for fact. Their main responsibility is to protect access to information that can be used to

prove whether or not they believe it. Consequently, people's personalities are impacted by fake news. The methods now in use to detect fake news are inaccurate and have a low accuracy rate. In the current classification, the random forest strategy fared better than other detection algorithms. The recommended approach makes use of social media networks AI algorithm.

This research examines online news sources and examines the factors that contribute to the dissemination of false information. Keeping the larger picture in mind, fake news has come to be understood as more than just misinformation spread by propagandists who use it as a tool to cast doubt on the veracity of claims made by people who, for the most part, disagree with them politically and ideologically. Some scholars argue that the approach is faulty since the definition of truth is based on subjective perceptions of reality, suggesting that news is not produced by the news production office but rather is the outcome of meaning-making within each epistemic institution. Today's professional journalism is not ideal, but it has shown to be a remarkably trustworthy source this century. However, anyone may undertake online journalism by using social media, and this has drawbacks such as the propagation of false information. It frequently attempts to look like genuine news, but it is unable to do so since it does not adhere to the journalistic standards and guidelines used by news organizations to confirm the accuracy and legitimacy of their reporting.

The structure of the paper is as follows: A thorough overview of feature selection detection methodologies, false news identification, and performance comparison is provided in Section 3, along with a summary of related research and the history of the models used. The implementation and experimental setup are covered in depth in Section 5. The experimental data are reported in Section 6, together with a description of the comparative analysis.

II. LITERATURE REVIEW

In the digital age, when there is an abundance of information available online, being able to discern fact from fiction is like having superpower-like ability to detect fake news. Scientists and technology specialists are hard at work developing instruments that can spot false narratives, helping us decide which sources to believe. We should investigate these methods as trustworthy friends in the fight against false information.

With the use of machine learning techniques, this literature review attempts to present a thorough overview of the rapidly changing field of fake news detection methodologies.

Nihel Fatima Baarir et al., [11] have worked on the idea that there are various reasons why fake news can be disseminated. Certain ones are created solely to boost the quantity of clicks and traffic to a website. Our suggested method builds a decision-making model using the support vector machine technique based on a set of news data. This algorithm is then applied to classify newly published news stories as authentic or fraudulent. Based on the investigation conducted for this project, the following findings were attained: Text, author, source, date, and sentiment are the key features to use when spotting fake news. The approach that was followed produced a 100 percent detection rate.

Manal Iftikhar et al.'s authors [12] Fake news is most commonly defined as false information that is fabricated to intentionally mislead readers. The primary goal of this work is to compare the outcomes and determine whether model is a superior classifier for this problem and data set using machine learning and a couple deep learning models. We categorize news as factual or deceptive. To ensure that news is real, it is essential to identify fake news. A number of machine learning models were looked at, including Random Forest, Naive Bayes, Decision Trees, and Regression. With training on "Fake News," our model demonstrated an astounding 99 percent accuracy rate.

Swornamalya.M.N et al.'s authors [13] False news detection by hand typically requires the use of a wide range of approaches and procedures. It's probable that websites that verify information will be necessary. As mentioned in the section above, the method entails running a number of trials on datasets utilizing majority voting, Random Forest, and Support Vector Machine (SVM) classifiers. SVM and Random Forest are also used to create ensemble forecasts, which are then combined by a majority vote. As a result, our algorithm is now proficient and well-trained at reliably spotting phony news.

Sakshi Neeraj et al.'s authors [14] The growth of the internet has made it easier for misleading information to spread across all online channels. This work's primary topic is K-Nearest Neighbors Decision Trees in Logistic Regression Earlier studies have illuminated a variety of automated techniques for detecting phony uploads and fake publications. These techniques cover a broad spectrum of dishonest strategies for identifying false news, such as using chatbots to distribute misleading material and clickbait to encourage the propagation of false information. These were only able to reach an accuracy of 63 to 70 percent.

Neetu Mittal et al.'s authors [15] After doing a preliminary analysis, instances of a substantial amount of data on false news are presented for the feasibility study. Review and analyze all of the data. The method and system are intended to identify false information. The definition of fake news is news, and we're figuring out which dataset to use and how to program for it. The dataset was obtained from Kaggle, and every word of the news articles is taken into consideration during the data cleaning procedure. This paper gathers various rational and useful definitions of fake news and provides an initial in-depth investigation of the issue.

K.Tiemtud et al.'s authors [16] We get information on a daily basis from a variety of social media platforms. There is knowledge that is both credible and nonsensical. It is up to the recipients to read it and make sure it is accurate. This work is mostly focused on using neural networks and machine learning.

One supervised learning technique used for tasks like regression, classification, and outlier identification is the Support Vector Machine (SVM) algorithm. It works by estimating which function would be best for classifying the provided data.

Comparing the Word List feature to datasets without the Is Word List feature also yields better accuracy performance.

Bilal Hawashin et al.'s [17] In this study, we compare the optimized performance with the original one and optimize the feature selection process. We employ a number of categorization techniques in this work, including K Nearest Neighbor (KNN), Random Forest, Support Vector Machines (SVM), Logistic Regression, and several machine learning strategies that are compared to social network analysis methodologies. They also conducted a thorough analysis of the various strategies for identifying and countering fake news. In the context of Arabic false news classification, the results obtained without feature selection are on par with deep learning techniques, albeit with notable improvements in model complexity and training time.

The writers of Kripakrishna R. K. et al. (2016) Fake news producers use a variety of dramatic techniques to boost the popularity of their misinformation, one of which is taking advantage of the chance to play with the emotions of the audience. Utilizing methods including machine learning, natural language processing, and artificial intelligence, researchers have classified internet news stories into a variety of binary categories. With the help of these techniques, people can classify news as real or fraudulent and evaluate the reliability of the websites that are distributing it. The model is also evaluated using Random Forest Machine Learning Classifiers and Decision Tree Classifiers. The Random Forest classifier produces an accuracy rate of 99.02 percent, the Decision Tree classifier produces 99.62 percent, and the logistic regression produces 98.98 percent accuracy. Of these three algorithms, With 99 percent accuracy, the Random Forest classifier offers the best results. Users can so confidently assess if the news is real or false.

A.Singh and S.Patida et al.'s authors [19] One of the major risks to the internet community and people's daily lives is phishing. The goal of this attack is to steal confidential data by impersonating a reliable source. The method of feature extraction received little or no attention in a large number of the reviewed literature. In order to identify phrases that are highly connected with fake news, feature extraction techniques can be used to detect fake news. In this study, we investigated different feature extraction methods using a large dataset that combines several false news datasets. Three distinct feature extraction strategies were used in this methodology to enhance the performance of our models, and the result was the development of an SVM model with an accuracy, recall, precision, and 98 percent on the F1 scale.

N.L.S.R. Krishna and M. Adimoolam et al.,[20] This paper's main concept is to compile authentic and fraudulent news stories over time, feed them into machine learning algorithms to build classification models, and then utilize these models to identify authentic or fraudulent news stories.To determine whether the news is true or false, different machine learning models, such as logistic regression, decision trees, and k-nearest neighbor, are created using the pre-processed information. Three distinct algorithms (Naive Bayes, Logistic Regression, and Support Vector Classifier) were combined to create a customized ensemble machine learning technique that showed a 91.5 percent accuracy rate. It showed an 8.49 percent variation and a low bias of 0.1 percent. Additionally, it received a 92 percent score throughout the 10-fold cross-validation process.

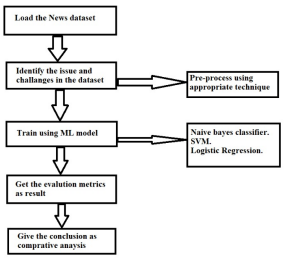

Fig. 1. Block Diagram of fake news detection

III. Methodology

The method this research suggests looks for anomalous activity, or outliers, using machine learning techniques. The following figure serves as a representation of the fundamental rough architecture diagram: The method this research suggests looks for anomalousFirst, we got our dataset from Kaggle, a website that offers datasets for data analysis. For our research, we used two datasets, one labeled "True news" and the other "Fake news Each dataset has four columns: Title, Text, Subject, and Date.We first clean and preprocess the dataset before moving on to the model-building process. It is represented as the outlier in Fig. 1. A collection of modules algorithms process the data.The way these algorithms cooperate is seen in the block diagram above. The different supervised methods are applied to this data after it has been fitted into a model. supervised algorithms :

- Logistic Regression.
- Naive Bayes.
- Support Vector Machine (SVM).

The method this research suggests looks for anomalousThe sklearn package's ensemble module includes a number of ensemble-based techniques and functions designed for regression problems in addition to classification and outlier identification. NumPy, SciPy, and matplotlib all work well with the free and open-source Python module Sklearn, which provides a wealth of straightforward but effective tools for machine learning and data analysis. Designed to effortlessly interface with scientific and numerical libraries, sklearn offers a wide range of methods for classification, clustering, and regression, enabling effective data processing and modeling.

The method this research suggests looks for anomalousSklearn offers an extensive collection of algorithms in the regression area that are designed to predict continuous target variables. Among these algorithms are ensemble techniques like Naive Bayes, Support Vector Machine, and Logistic Regression. To make precise predictions on data that hasn't been seen yet, each of these algorithms models the link between

input features and continuous target variables using a different approach.

We use the Jupyter Notebook platform to illustrate the effectiveness of these regression techniques, taking advantage of its interactive features and user-friendly interface to present the suggested method. Furthermore, this software can be easily moved to the cloud with Google Colab or other platforms that allow Python notebook files, making it simple to collaborate and carry out machine learning workflows.

Below, we provide a detailed explanation of each regression algorithm utilized:

A. Support Vector Machine

The method this research suggests looks for anomalousAn effective supervised learning approach for regression and classification problems is called Support Vector Machine (SVM). The way it operates is by identifying the hyperplane that divides data points into several classes the best. With the introduction of kernel functions, SVM can handle non-linearly separable data and perform well in high-dimensional domains. It is resilient and extensively useful in many disciplines, such as image classification, text analysis, and bioinformatics, thanks to its capacity to maximize the margin between classes.

B. Logistic Regression

The method this research suggests looks for anomalous For applications involving binary classification, a basic machine learning approach is called logistic regression. In spite of its name, classification is used more often than regression. It simulates a binary outcome's likelihood depending on one or more predictor variables. Logistic regression uses a logistic (sigmoid) function to mix the input features linearly in order to estimate the likelihood that an instance will belong to a particular class. It is a well-liked option for baseline classification models and as a building block in more complicated models since it is a straightforward yet powerful algorithm that is well-known for its interpretability and simplicity of implementation.

C. Naive Bayes

The method this research suggests looks for anomalous popular probabilistic machine learning approach for classification problems is called Naive Bayes. It is predicated on the "naive" assumption of feature independence from Bayes' theorem. In fact, Naive Bayes typically performs surprisingly well despite its simplicity, especially for text classification applications like sentiment analysis and spam detection. It uses little training data and is computationally efficient in estimating the parameters needed for classification. Naive Bayes is a well-liked option for real-world applications with constrained computational resources since it is especially helpful when handling high-dimensional data and extensive feature spaces.

RESULTS

For the objective of detecting fake news, we tested the effectiveness of three machine learning models: Naive Bayes, Logistic Regression, and Support Vector Machine (SVM). The results are summarized in the table below:

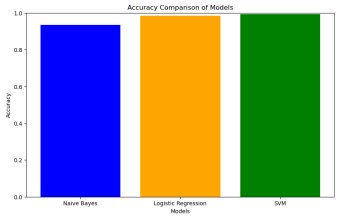

Fig. 2. Accuracy Comparision graph

TABLE I
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Metric	Naive Bayes	Logistic Regression	SVM
Accuracy	0.9350	0.9848	0.9925
F1 Score	0.9320	0.9842	0.9923
Mean Squared Error	0.0650	0.0152	0.0075
Root Mean Squared Error	0.2550	0.1235	0.0864

We also created a graph to compare the accuracy of the models, which you can see below. The graph provides a visual representation of the differences in model performance.

These findings provide information about each model's efficacy and lay the groundwork for assessing approaches for identifying false news.

CONCLUSION

This study demonstrated the efficaciousness of three different machine learning models in identifying false news: SVM, Logistic Regression, and Naive Bayes. Many metrics, such as accuracy, F1 score, mean squared error, and root mean square error, were compared in order to facilitate understanding of each model's performance. These measurements were totaled and presented. The Logistic Regression and SVM models in particular showed promising results in terms of accuracy and F1 score, suggesting that they may successfully distinguish between real and fake news in our dataset.This effort opens up new avenues for research into how to increase the precision of false news detection through the use of ensemble methodologies and deeper machine learning techniques. Furthermore, enhancing the dataset with multilingual and multimedia content and applying these algorithms to social media real-time detection tools on platforms are critical first steps in creating more adaptable and durable strategies to combat misinformation globally.

REFERENCES

[1] X. Jose, S. D. M. Kumar and P. Chandran, "Characterization, Classification and Detection of Fake News in Online Social Media Networks," 2021 IEEE Mysore Sub Section International Conference (MysuruCon), Hassan, India, 2021, pp. 759-765, doi: 10.1109/MysuruCon52639.2021.9641517.

[2] K. R. K and C. K. A, "An Awareness About Phishing Attack And Fake News Using Machine Learning Technique," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 2022, pp. 1-5, doi: 10.1109/ICDCECE53908.2022.9793225.

[3] V. Gupta, R. S. Mathur, T. Bansal and A. Goyal, "Fake News Detection using Machine Learning," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022, pp. 84-89, doi: 10.1109/COM-IT-CON54601.2022.9850560.

[4] K. R. Asish, A. Gupta, A. Kumar, A. Mason, M. K. Enduri and S. Anamalamudi, "A Tool for Fake News Detection using Machine Learning Techniques," 2022 2nd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2022, pp. 1-6, doi: 10.1109/CONIT55038.2022.9848064.

[5] A. Matheven and B. V. D. Kumar, "Fake News Detection Using Deep Learning and Natural Language Processing," 2022 9th International Conference on Soft Computing and Machine Intelligence (ISCMI), Toronto, ON, Canada, 2022, pp. 11-14, doi: 10.1109/ISCMI56532.2022.10068440.

[6] H. E. Wynne and K. T. Swe, "Fake News Detection in Social Media using Two-Layers Ensemble Model," 2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Phuket, Thailand, 2022, pp. 411-414, doi: 10.1109/ITC-CSCC55581.2022.9894967.

[7] N. L. S. R. Krishna and M. Adimoolam, "Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm," 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2022,

[8] N. Kumar and N. Kar, "Approaches towards Fake News Detection using Machine Learning and Deep Learning," 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2023, pp. 280-285, doi: 10.1109/SPIN57001.2023.10117154.

[9] V. Gupta, R. S. Mathur, T. Bansal and A. Goyal, "Fake News Detection using Machine Learning," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022, pp. 84-89, doi: 10.1109/COM-IT-CON54601.2022.9850560.

[10] N. L. S. R. Krishna and M. Adimoolam, "Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm," 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ICBATS54253.2022.9758999. vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[11] N. F. Baarir and A. Djeffal, "Fake News detection Using Machine Learning," 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), Boumerdes, Algeria, 2021, pp. 125-130, doi: 10.1109/IHSH51661.2021.9378748.

[12] M. Iftikhar and A. Ali, "Fake News Detection using Machine Learning," 2023 3rd International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2023, pp. 103-108, doi: 10.1109/ICAI58407.2023.10136676.

[13] S. M. N, K. M. V, S. Verma and S. Rajagopal, "NLP Based Fake News Detection Using Hybrid Machine Learning Techniques," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2022, pp. 818-822, doi: 10.1109/ICESC54411.2022.9885679.

[14] N. Bhargava, S. Sharma, R. Purohit and P. S. Rathore, "Prediction of recurrence cancer using J48 algorithm," 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2017, pp. 386-390, doi: 10.1109/CESYS.2017.8321306.

[15] B. Dhruv, N. Mittal and M. Modi, "Analysis of different filters for noise reduction in images," 2017 Recent Developments in Control, Automation and Power Engineering (RDCAPE), Noida, India, 2017, pp. 410-415, doi: 10.1109/RDCAPE.2017.8358306.

[16] K. Tiemtud and P. Leesutthipornchai, "Thai Fake News Detection Using Machine Learning Model," 2023 International Electrical Engineering Congress (iEECON), Krabi, Thailand, 2023, pp. 346-350, doi: 10.1109/iEECON56657.2023.10126781.

[17] B. Hawashin, A. Althunibat, T. Kanan, S. AlZu'bi and Y. Sharrab, "Improving Arabic Fake News Detection Using Optimized Feature Selection," 2023 International Conference on Information Technology (ICIT), Amman, Jordan, 2023, pp. 690-694, doi: 10.1109/ICIT58056.2023.10225974.

[18] V. Gupta, R. S. Mathur, T. Bansal and A. Goyal, "Fake News Detection using Machine Learning," 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), Faridabad, India, 2022, pp. 84-89, doi: 10.1109/COM-IT-CON54601.2022.9850560.

[19] A. Singh and S. Patidar, "A Survey on Fake News Detection Using Machine Learning," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 327-331, doi: 10.1109/ICAC3N56670.2022.10074450.

[20] N. L. S. R. Krishna and M. Adimoolam, "Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm," 2022 International Conference on Business Analytics for Technology and Security (ICBATS), Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ICBATS54253.2022.9758999.