International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 4 April 2023, pp: 1296-1299 www.ijaem.net ISSN: 2395-5252

Fake News Detection Using Machine Learning Algorithms

Shivangi Singh [1], Er. Shilpi Khanna [2], Priyanshu Maurya [3]

[1,3] *Student of Department of Information Technology, Shri Ramswaroop Memorial College of Engineering & Management, Lucknow, Uttar Pradesh, India*

[2] *Assistant Professor, Department of Information Technology, Shri Ramswaroop Memorial College of Engineering & Management, Lucknow, Uttar Pradesh, India*

Date of Submission: 15-04-2023 Date of Acceptance: 25-04-2023

ABSTRACT- On social networking sites, any information may spread extremely fast. However, similar websites also disseminate low-quality news that is riddled with errors (Fake news). The widespread spreading of fake news spreads negativity in society and has a negative influence on people's mental health. As a result, there is an urgent need to focus emphasis on fake news identification. The purpose of this research paper is to examine and compare four well-known machine learning algorithms, namely logistic regression, decision tree classification, gradient boosting classifier, and Random forest classifier, in order to validate the effectiveness of classification performance in detecting fake news. The dataset that we utilised in our research is LIAR, and as a result, we discovered that Gradient boosting classifier performed significantly better than other algorithms.

Keywords- Internet, Social Media, Fake News, Classification, Websites, machine learning, performance comparison.

I. INTRODUCTION

Fake news, which is purposefully false material portrayed as news, has become a big problem throughout the world. The extensive transmission of fake news via social media and other internet platforms has the potential to do harm by influencing public opinion and diminishing faith in the media. People manipulate data and information for a variety of motives, including social, political, and economic ones. As a result, the news is neither entirely genuine nor entirely untrue. The identification of fake news has become a crucial topic of research and development because it has substantial ramifications for democracy, public health, and social cohesion. Effective fake news identification is critical for sustaining the integrity of the news ecosystem and supporting informed decision making among the public. A fake news detection technique seeks to identify intentionally misleading material by studying previously assessed fake and authentic news[1].The current study paper's contributions include an investigation of four machine learning algorithms, and the performance of each machine learning-based model is tested to accurately categorise numerous news items, revealing each model's capacity to enhance its accuracy of identifying false news[2].

II. LITERATURE REVIEW

In their study [3,] Mykhailo Granik et al. provide a simple strategy for detecting fake news using a naïve Bayes classifier. This straightforward technique was turned into a software solution and tested against a data collection of Facebook news postings. The information was gathered from three major Facebook pages as well as three major mainstream political news websites (Politico, CNN, and ABC News). They attained a classification accuracy of about 74%. The dataset's skewness led in lower classification accuracy, with only 4.9% of it being bogus news. Himank Gupta et al. [10] presented a framework based on several machine learning approaches that addresses a variety of issues such as accuracy deficiency, time lag (BotMaker), and high processing time to handle thousands of tweets in one second. They extract 400,000 tweets from the HSpam14 dataset and characterise 150,000 spam tweets and 250,000 non-spam tweets. They also derived several lightweight characteristics, as well as the Top-30 terms generated from the Bag-of-terms model. 4. They achieved an accuracy of 91.65% and outperformed the old solution by almost 18%.

III. METHODOLOGY

A. Logistic Regression

Logistic regression is a statistical approach used to examine the connection between one or more independent variables and a dependent variable. It is a prominent strategy in machine learning and data science for binary classification issues when the outcome of binary interest is binary or categorical such as 'yes' or 'no', 'true' or 'fake', 'success' or 'failure'[5]. It is a predictive analytic method based on the probability notion.

B. Decision Tree Classifier

A decision tree classifier is a form of supervised machine learning algorithm used in machine learning for classification tasks.

A Decision tree is a tree structure that looks like a flowchart, with each internal node representing a test on an attribute, each branch representing a test result, and each leaf node (terminal node) holding a class label. It operates by recursively dividing the data into subsets depending on the values of the independent variables, with the purpose of minimising a measure of impurity or uncertainty in the data[6].

C. Gradient boosting classifier

Gradient boosting classifier is a machine learning approach that belongs to the ensemble learning class, which refers to integrating many models to increase the model's prediction power. This approach works by adding decision trees to the model repeatedly, with each tree correcting the errors created by the preceding trees. This approach computes the gradient of the loss function in relation to the model's predictions, and then uses the gradient of the loss function to update the parameters of the next tree to be added to the model, allowing the new tree to better match the data[12].

D. Random forest classifier

A decision tree is made up of parents and branches with distinct criteria, with each node representing a class for categorization. The random forest classifier is a classification ensemble approach that builds a large number of decision trees. Because the sample generation is random, each decision tree is slightly different from the others. This variety in decision trees aids in reducing overfitting and improving model generalization[9].

IV. IMPLEMENTATION

The steps involved for the conduction of the current research paper includes:

- Data collection
- Data preparation
- Training and testing
- Comparison of algorithms

A. Data collection

News may be found online via a variety of sources, including the internet, social media, and other websites like LIAR, PolitiFact, Fakeddit, and other public datasets for false news categorization are accessible. For the purposes of our research, we used the LIAR dataset.

LIAR: This is a publicly available dataset for detecting false news that comprises hand labelled brief remarks from POLITIFACT.COM.

We utilised csv datasets for this research, namely false.scv and true.csv, and we added a column named "class" to the dataset to differentiate between false and actual news.

B. Data Preparation

The data is prepared by following the given steps:

- Creating a function to convert the text in lowercase, remove the extra space, special characters, URL and links.
- Defining dependent and independent variables x and y.
- Splitting the dataset into training set and testing set
- Convert text to vectors.

C. Training and Testing

We utilised Python 3.6.5 as our programming language for the implementation, and we did random shuffling of the dataframe and employed four distinct machine learning methods. To investigate how well our data fit into the models, we applied the random forest classifier, logistic regression, Gradient boosting classifier, and decision tree algorithms.

Various variations were found during the training of machine learning algorithms. After training, we utilised the learned models of each approach and tested them on the testing set.

D. Comparison of Algorithms

Standard measures were evaluated in order to compare the overall performance of each method. They are compared using the following metrics:

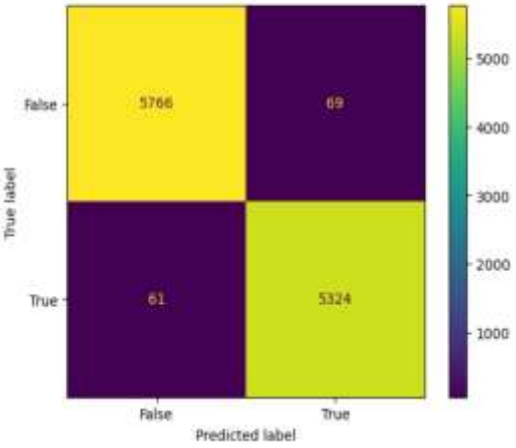
- ➢ Precision
- ➢ Recall
- ➢ F1-score
- ➢ Support
- ➢ Accuracy
- ➢ Confusion metrics

These measures are widely used in the machine learning field and allow us to evaluate a classifier's performance from many angles. Accuracy specifically assesses the resemblance between anticipated fake news and actual false news.

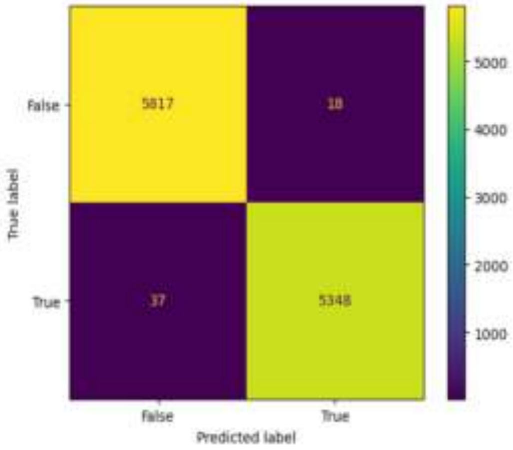
V. RESULTS

Standard measures were evaluated in order to compare the overall performance of each method. The confusion matrix displays the number of predictions (whether accurate or erroneous) with each class. Gradient boosting classifier's confusion matrix demonstrates that this approach identifies all classes more accurately.

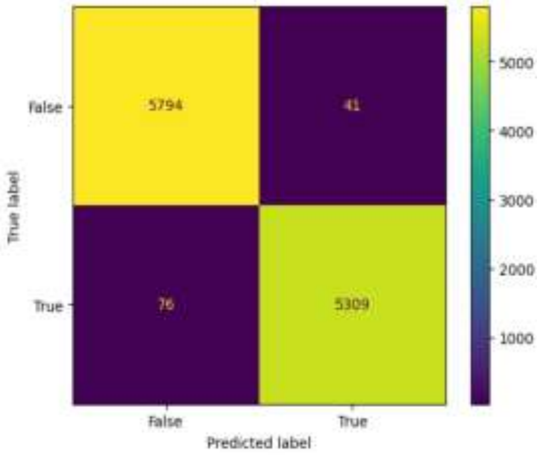
The confusion matrix of several machine learning algorithms are shown below:



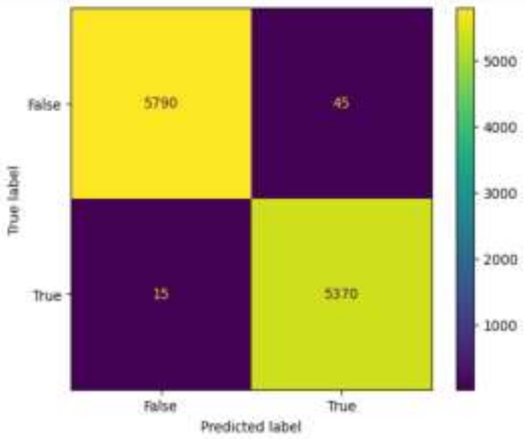
a) Logistic Regression



b) Decision Tree classifier



c) Random Forest Classifier



d) Gradient Boosting Classifier

Performance Report of the algorithms is mentioned in the table below:

Evaluation metrics	Logistic Regression	Decision Tree classifier	Random forest classifier	Gradient Boosting classifier
Accuracy	0.97	0.98	0.99	1.00
Precision	0.95	0.96	0.99	0.99
Recall	0.97	0.99	0.99	1.00
F1-score	0.98	0.99	0.98	0.99

TABLE I: CLASSIFICATION REPORT

VI. CONCLUSION

Manually categorising news involves in-depth knowledge of the area as well as experience in identifying abnormalities in the text. We examined the challenge of classifying false news items using machine learning models and ensemble approaches in this study. This research explores the performance of four machine learning classifiers for detecting false news: logistic regression, decision tree classifier, random forest, and gradient boosting classifier. LIAR is a well-known and freely available dataset that we utilised in our study. We compared the outcomes of each method using a variety of performance indicators. Based on our findings, the gradient boosting classifier performed well when compared to the other techniques.

VII. ACKNOWLEDGMENT

We are grateful to Shri Ramswaroop Memorial College of Engineering and Management for mentoring and supporting us in our endeavour. We are extremely grateful to Er. Shilpi Khanna for assisting us with the research needs and information.

REFERENCES

[1]. An Empirical Comparison of Fake News Detection using different Machine Learning Algorithms. Article in International Journal of Advanced Computer Science and Applications · January 2020.

[2]. Fake News Detection using Machine Learning Algorithms Uma Sharma, Sidarth Saran, Shankar M. Patil Department of Information Technology Bharati Vidyapeeth College of Engineering Navi Mumbai, India.

[3]. Research Article Fake News Detection Using Machine Learning Ensemble Methods.

[4]. Fake News Detection Using Machine Learning Approaches. : Z Khanam et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012040.

[5]. https://www.javatpoint.com/logistic-regression-in-machine-learning

[6]. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[7]. Zhang, X.; and Ghorbani, A. A. (2019). An overview of online fake news: Characterization, detection, and discussion. Information Processing & Management, 57, 102025.

[8]. Horne, B. D.; and Adali, S. (2017). This just in: Fake news packs a lot in title, uses simpler, repeti-tive content in text body, more similar to satire than real news. In Eleventh International AAAIConference on Web and Social Media.

[9]. https://www.javatpoint.com/machine-learning-random-forest-algorithm.

[10]. H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383.

[11]. M. L. Della 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]. https://www.simplilearn.com/gradient-boosting-algorithm-in-python-article#:~:text=Gradient%20Boosting%20is%20a%20functional,produce%20a%20powerful%20predicting%20model.