IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Fake News Detection Using Different Machine Learning Algorithms

Name of Author: Sanket Joshi
Designation of Author: Student
Department: Computer Science and Information Technology
Project Guide: Prof. Vijay Chandode
Head of Department: Prof. Dr. Sushil Kulkarni
College: MBESs College of Engineering, Ambajogai

Abstract: The proliferation of misinformation on digital platforms threatens the trustworthiness of media outlets and online sources. While traditional machine learning models like Naive Bayes, SVM, and Random Forest have shown promise in fake news detection, they often fall short in accuracy and scalability when compared to newer ensemble methods. This paper enhances existing methodologies by incorporating two advanced tree boosting algorithms—XGBoost and LightGBM. These models are evaluated against traditional classifiers using a real-world dataset. The results show that XGBoost and LightGBM significantly outperform traditional models in terms of accuracy, training efficiency, and robustness. The implementation was carried out entirely using Python in a Jupyter Notebook environment, enabling step-by-step analysis and visualization. This work contributes to a modernized and efficient approach to automated misinformation detection.

Keywords- Fake news, Text classification, TF-IDF, Machine Learning, Decision Tree Classifier, Logistic Regression, Gradient Boosting Classifier, Random Forest Classifier, AdaBoosting, SVM, XGBoost and LightGBM.

I. INTRODUCTION

The spread of misinformation poses a major threat to society, especially in an era where news spreads rapidly on digital platforms. Fake news undermines public trust, influences political opinions, and creates confusion. Detecting and mitigating the spread of fake news using machine learning is a crucial area of research.

Previous studies have used traditional classifiers such as Decision Tree, Logistic Regression, Random Forest, Support Vector Machines (SVM), Gradient Boosting and AdaBoosting with various feature extraction techniques like Count Vectorizer and TF-IDF. Although these models show reasonable accuracy, they often suffer from issues like overfitting and slow training when used with large datasets.

This paper introduces two state-of-the-art gradient boosting techniques—XGBoost and LightGBM—for fake news detection. These methods are known for their accuracy, regularization capabilities, and computational efficiency. Our goal is to demonstrate how tree boosting models can outperform traditional classifiers in both accuracy and scalability.

Data- There are two datasets namely “True.csv” and “Fake.csv”, used in our project, sourced from Kaggle. The Fake dataset consists of 23471 rows of data from various news articles available on the internet whereas the True dataset consists of 21407 rows of data. The attributes of both the datasets are –

1. id – Unique ID for the news article.
2. title – Title of the news article.
3. text – The text of the news article. It might be incomplete in a few cases.
4. subject – Type of news article.
5. date – Date of publication of the news article.

II. RELATED WORK

Fake news detection has been a focus of many recent studies due to the rise of misinformation on digital platforms. Traditional machine learning algorithms like Naive Bayes, Logistic Regression, Decision Tree, SVM, and Random Forest have been widely used. For example, Gupta et al. (2019) explored feature extraction methods for classifying misinformation on social media platforms.

In terms of deep learning, Ruchansky et al. (2017) proposed CSI, a hybrid model combining Recurrent Neural Networks (RNNs) and user behavior analysis, which significantly improved the reliability of fake news classification. However, deep learning models require large datasets and are computationally expensive.

Recent advancements in ensemble learning introduced tree boosting methods that have demonstrated strong performance in many text classification problems. Chen and Guestrin (2016) developed XGBoost, a gradient boosting framework that incorporates regularization, tree pruning, and parallel processing. Similarly, Ke et al. (2017) proposed LightGBM, which is optimized for speed and memory usage, especially for large-scale datasets.

However, few studies have applied both XGBoost and LightGBM for fake news detection using real-world news datasets. This research aims to bridge that gap by integrating these models and evaluating their performance against traditional classifiers on a labeled news corpus.

III. PROPOSED SYSTEM

The proposed fake news detection system is designed to classify news articles as either real or fake using a combination of traditional and advanced machine learning models. The key components of the system are as follows and Fig. 1 Shows the block diagram of the model

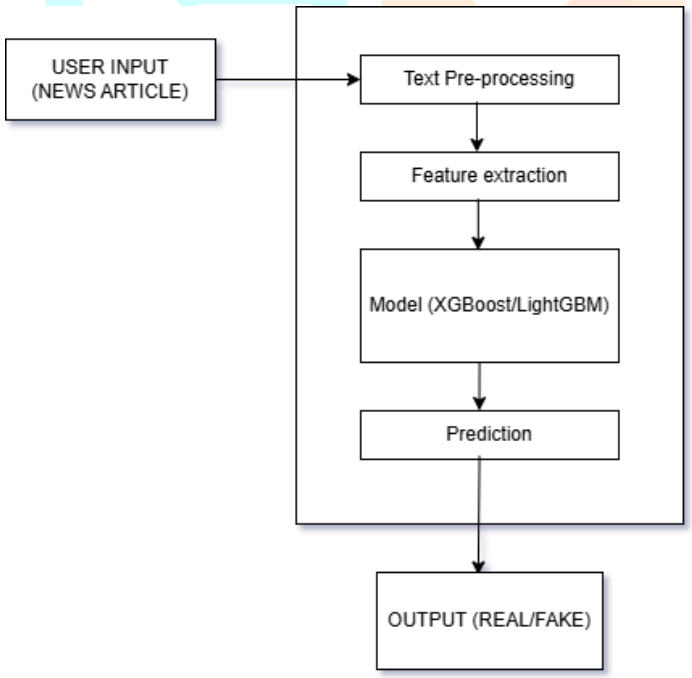

Fig. 1 BLOCK DIAGRAM

The system pipeline consists of:

- *Data Preprocessing*:
  1. Merge and shuffle labeled datasets (e.g., Fake.csv, True.csv)
  2. Remove unnecessary columns (e.g., title, date)
  3. Clean text (remove punctuation and stop words, and convert text to lowercase)

- *Feature Extraction:*
  1. Use TF-IDF Vectorizer for numerical representation of text

- *Model Training:*
  1. Traditional models: Naive Bayes, Logistic Regression, Decision Tree, SVM, Random Forest
  2. New models: XGBoost, LightGBM
  3. Hyperparameter tuning using grid search/cross-validation

- *Evaluation Metrics:*

  1. Accuracy:

  Indicates the overall correctness of the model by calculating the ratio of correctly predicted observations to total observations.

  Formula: (TP + TN) / (TP + TN + FP + FN)

  2. Precision:

  Measures the accuracy of positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positives.

  Formula: TP / (TP + FP)

  3. Recall (Sensitivity or True Positive Rate):

  Indicates the model's ability to correctly identify actual positives. It is the ratio of correctly predicted positives to all actual positives.

  Formula: TP / (TP + FN)

  4. F1-Score:

  The harmonic mean of precision and recall. It provides a balance between the two and is useful when the dataset has imbalanced classes.

  Formula: 2 * (Precision * Recall) / (Precision + Recall)

- *Execution Platform:*

  All experiments, data preprocessing, model training, and evaluation were carried out using Python in Jupyter Notebook.

Traditional definitions of artificial intelligence (AI) are thought to be inadequate. AI can mimic human behavior with consciousness, sensitivity, and spirit because of being more robust. The advent of machine learning (ML) provided the means to help bring this vision closer to reality. Machine learning is an area of artificial intelligence that, by definition, focuses on enabling computers to learn without explicit programming. It is built on the idea of using algorithms that are fed by a lot of data to replicate behavior. The algorithm builds a model and learns which choice to make in a variety of circumstances. So, depending on the circumstances, the machine can automate its tasks.

Supervised Learning: It is a type of artificial intelligence that uses machine learning. The concept is to "guide" the algorithm's learning process using samples of expected results that have already been labeled. After then, artificial intelligence picks up new information from each example and modifies its settings to close the discrepancy between actual and expected results. To generalize, learning with the goal of predicting the outcome of new situations, the margin of error is hence decreased across the training sessions. If the labels resemble discrete classes, the result is referred to as classification, and if continuous quantities are referred to as regression.

TF-IDF vectorizer: One of the most widely used feature extraction techniques is the term inverse document frequency (TF-IDF). This technique is divided into two stages, in which the term frequency (TF) is calculated first, and the inverse document frequency (IDF) is calculated in the second stage. TF(t) = No of times the t term appears in a doc. Total No of terms in the document IDF(t) = Log (total No. of documents no. of documents containing term t)

N-gram level vectorizer: This is a sub-technique of TF-IDF, in which a slice of letters (N), is displayed in a matrix representing TF-IDF scores of N-grams. This technique was used to overcome the problem of selecting the right features and their numerical values. In such cases, the use of a TF-IDF classification associated with unigrams or bigrams has been suggested.

Logistic Regression: The advantages of logistic regression include probability modelling, the capacity to depend on features, and the flexibility to update the model. However, for higher accuracy, logistic regression requires a big data set, but Naive Bayes may function with small datasets as well.

Decision Tree: An algorithm called a decision tree aims to arrange people into groups that are as like one another as feasible in terms of the variable that needs to be predicted. The algorithm's output is a tree that shows the hierarchical connections between the variables. By selecting the explanatory variable that allows for the best individual separation, a subpopulation of people is obtained at each iteration of an iterative procedure. When no more splits are possible, the algorithm terminates.

Gradient Boosting Algorithm: Gradient Boosting is a functional gradient algorithm that repeatedly selects a function that leads in the direction of a weak hypothesis or negative gradient so that it can minimize a loss function. Gradient boosting classifier combines several weak learning models to produce a powerful predictive model.

Random Forest Algorithms: Random Forest is an ensemble technique used for classification and regression predictive models. They are made up of numerous Decision Tree blocks that are utilized as separate predictors. The basic idea behind the method is that multiple predictors are developed, and their various predictions are pooled, rather than attempting to obtain an optimum method all at once. The class receiving the most votes becomes the final prediction.

AdaBoosting Algorithm: The Ada Boost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are reassigned to each instance, with higher weights assigned to incorrectly classified instances.

SVM: Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Extreme Gradient Boosting (XGBoost): It is faster with high accuracy as compared to others. An optimized version of GBM with Regularization, Parallel processing and other improvements. It handles null values automatically.

LightGBM: It is fastest with high accuracy compared to others. It's very efficient when large and high dimensional tabular datasets are available. It uses less memory compared to XGBoost as well. It uses histogram-based learning and supports native categorical features.

IV. EXPERIMENTAL RESULTS

For better implementation and results, we created a separate dataset CSV file for Fake & True news. We created a dataset containing more than 20,000 news articles. Below is the screenshot of the result of making the dataset. Further, a Jupyter Notebook was created to implement the ML program. We have used Logistic Regression, Gradient Boosting, Decision Tree, and Random Forest. After TF-IDF vectorization and data cleaning, we trained and tested the models with these classifiers, we obtained the following accuracies: For Logistic Regression as 98.86%, Decision Tree as 99.69%, Random Forest as 99.06%, SVM as 99.50%, Gradient Boosting as 99.68%, Ada Boosting as 99.70%, XGBoost as 99.87% and LightGBM as 99.90% . The following Table 1 summarizes the accuracy achieved by each model and Fig. 4 Graphical comparison of Machine Learning Techniques.

Model	Accuracy (%)
Logistic Regression	98.86
Decision Tree Classifier	99.69
Random Forest	99.06
SVM	99.50
Gradient Boosting	99.68
AdaBoosting	99.70
Extreme Gradient Boosting (XGBoost)	99.87
LightGBM	99.90

Table 1 Comparison of the models

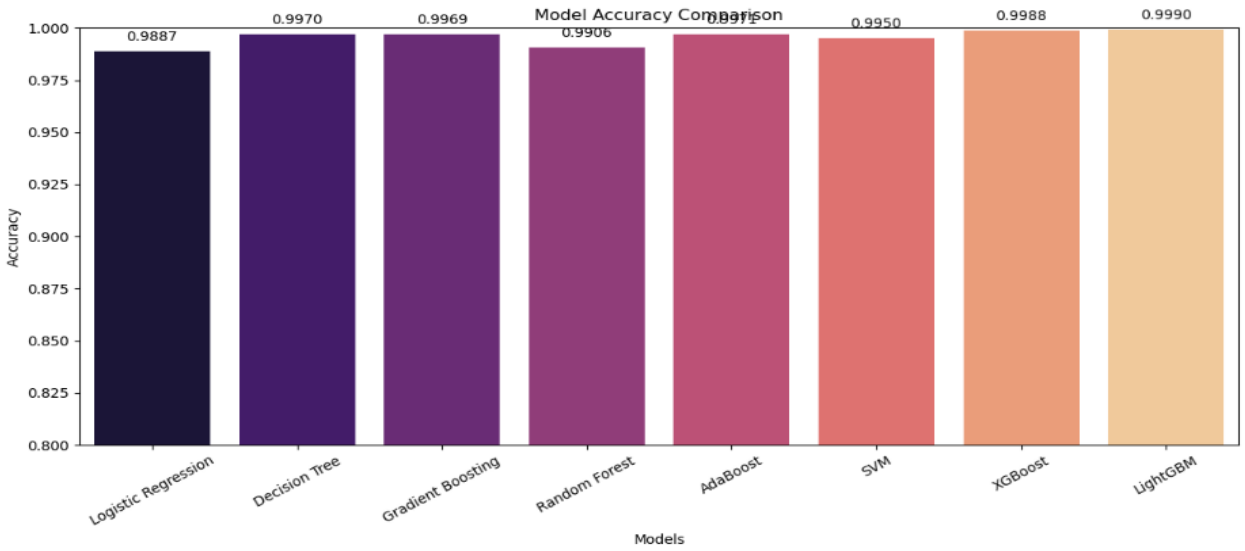

Fig. 2 Comparison of Machine Learning Techniques

```
In [74]: news = str(input())
         manual_testing(news)
```

```
Donald Trump Sends Out Embarrassing New Year’s Eve Message; This is Disturbing,"Donald Trump just couldn t wish all Americans a
Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and  the very dishonest fake ne
ws media.  The former reality show star had just one job to do and he couldn t do it. As our Country rapidly grows stronger and
smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy an
d Healthy New Year,  President Angry Pants tweeted.  2018 will be a great year for America! As our Country rapidly grows strong
er and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a H
appy and Healthy New Year. 2018 will be a great year for America!  Donald J. Trump (@realDonaldTrump) December 31, 2017Trump s
tweet went down about as welll as you d expect.What kind of president sends a New Year s greeting like this despicable, petty,
infantile gibberish? Only Trump! His lack of decency won t even allow him to rise above the gutter long enough to wish the Amer
ican citizens a happy new year!  Bishop Talbert Swan (@TalbertSwan) December 31, 2017no one likes you  Calvin (@calvinstowell)
December 31, 2017Your impeachment would make 2018 a great year for America, but I ll also accept regaining control of Congress.
Miranda Yaver (@mirandayaver) December 31, 2017Do you hear yourself talk? When you have to include that many people that hate y
ou you have to wonder? Why do the they all hate me?  Alan Sandoval (@AlanSandoval13) December 31, 2017Who uses the word Haters
in a New Years wish??  Marlene (@marlene399) December 31, 2017You can t just say happy new year?  Koren pollitt (@Korencarpente
r) December 31, 2017Here s Trump s New Year s Eve tweet from 2016.Happy New Year to all, including to my many enemies and those
who have fought me and lost so badly they just don t know what to do. Love!  Donald J. Trump (@realDonaldTrump) December 31, 20
16This is nothing new for Trump. He s been doing this for years.Trump has directed messages to his  enemies  and  haters  for N
ew Year s, Easter, Thanksgiving, and the anniversary of 9/11. pic.twitter.com/4FPAe2KypA  Daniel Dale (@ddale8) December 31, 20
17Trump s holiday tweets are clearly not presidential.How long did he work at Hallmark before becoming President?  Steven Goodi
ne (@SGoodine) December 31, 2017He s always been like this . . . the only difference is that in the last few years, his filter
has been breaking down.  Roy Schulze (@thbthttt) December 31, 2017Who, apart from a teenager uses the term haters?  Wendy (@Wen
dyWhistles) December 31, 2017he s a fucking 5 year old  Who Knows (@rainyday80) December 31, 2017So, to all the people who vote
d for this a hole thinking he would change once he got into power, you were wrong! 70-year-old men don t change and now he s a
year older.Photo by Andrew Burton/Getty Images.


LogisticRegeression Prdiction: Fake News
DecisionTreeClassifier  Prediction: Fake News
GradientBoostingClassifier Prediction: Fake News
RandomForestClassifier Prediction: Fake News
AdaBoostClassifier Prediction: Fake News
SVMClassifier Prediction: Fake News
XGBoostClassifier Prediction: Fake News
LightGBMClassifier Prediction: Fake News
```

Fig. 3 Article is Fake

```
In [75]: news = str(input())
         manual_testing(news)
```

```
As U.S. budget fight looms, Republicans flip their fiscal script","WASHINGTON (Reuters) - The head of a conservative Republican
faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts, called himsel
f a “fiscal conservative” on Sunday and urged budget restraint in 2018. In keeping with a sharp pivot under way among Republica
ns, U.S. Representative Mark Meadows, speaking on CBS’ “Face the Nation,” drew a hard line on federal spending, which lawmakers
are bracing to do battle over in January. When they return from the holidays on Wednesday, lawmakers will begin trying to pass
a federal budget in a fight likely to be linked to other issues, such as immigration policy, even as the November congressional
election campaigns approach in which Republicans will seek to keep control of Congress. President Donald Trump and his Republic
ans want a big budget increase in military spending, while Democrats also want proportional increases for non-defense “discreti
onary” spending on programs that support education, scientific research, infrastructure, public health and environmental protec
tion. “The (Trump) administration has already been willing to say: ‘We’re going to increase non-defense discretionary spending
... by about 7 percent,’” Meadows, chairman of the small but influential House Freedom Caucus, said on the program. “Now, Democ
rats are saying that’s not enough, we need to give the government a pay raise of 10 to 11 percent. For a fiscal conservative, I
don’t see where the rationale is. ... Eventually you run out of other people’s money,” he said. Meadows was among Republicans w
ho voted in late December for their party’s debt-financed tax overhaul, which is expected to balloon the federal budget deficit
and add about $1.5 trillion over 10 years to the $20 trillion national debt. “It’s interesting to hear Mark talk about fiscal r
esponsibility,” Democratic U.S. Representative Joseph Crowley said on CBS. Crowley said the Republican tax bill would require t
he  United States to borrow $1.5 trillion, to be paid off by future generations, to finance tax cuts for corporations and the r
ich. “This is one of the least ... fiscally responsible bills we’ve ever seen passed in the history of the House of Representat
ives. I think we’re going to be paying for this for many, many years to come,” Crowley said. Republicans insist the tax packag
e, the biggest U.S. tax overhaul in more than 30 years,  will boost the economy and job growth. House Speaker Paul Ryan, who al
so supported the tax bill, recently went further than Meadows, making clear in a radio interview that welfare or “entitlement r
eform,” as the party often calls it, would be a top Republican priority in 2018. In Republican parlance, “entitlement” programs
mean food stamps, housing assistance, Medicare and Medicaid health insurance for the elderly, poor and disabled, as well as oth
er programs created by Washington to assist the needy. Democrats seized on Ryan’s early December remarks, saying they showed Re
publicans would try to pay for their tax overhaul by seeking spending cuts for social programs. But the goals of House Republic
ans may have to take a back seat to the Senate, where the votes of some Democrats will be needed to approve a budget and preven
t a government shutdown. Democrats will use their leverage in the Senate, which Republicans narrowly control, to defend both di
scretionary non-defense programs and social spending, while tackling the issue of the “Dreamers,” people brought illegally to t
he country as children. Trump in September put a March 2018 expiration date on the Deferred Action for Childhood Arrivals, or D
ACA, program, which protects the young immigrants from deportation and provides them with work permits. The president has said
in recent Twitter messages he wants funding for his proposed Mexican border wall and other immigration law changes in exchange
for agreeing to help the Dreamers. Representative Debbie Dingell told CBS she did not favor linking that issue to other policy
objectives, such as wall funding. “We need to do DACA clean,” she said.  On Wednesday, Trump aides will meet with congressional
leaders to discuss those issues. That will be followed by a weekend of strategy sessions for Trump and Republican leaders on Ja
n. 6 and 7, the White House said. Trump was also scheduled to meet on Sunday with Florida Republican Governor Rick Scott, who w
ants more emergency aid. The House has passed an $81 billion aid package after hurricanes in Florida, Texas and Puerto Rico, an
d wildfires in California. The package far exceeded the $44 billion requested by the Trump administration. The Senate has not y
et voted on the aid.


LogisticRegeression Prdiction: Not a Fake News
DecisionTreeClassifier  Prediction: Not a Fake News
GradientBoostingClassifier Prediction: Not a Fake News
RandomForestClassifier Prediction: Not a Fake News
AdaBoostClassifier Prediction: Not a Fake News
SVMClassifier Prediction: Not a Fake News
XGBoostClassifier Prediction: Not a Fake News
LightGBMClassifier Prediction: Not a Fake News
```

Fig. 4 Article is Real (Not Fake)

V. CONCLUSION

In this paper, we looked at a computerized model for verifying news extracted from social media, which provides clear demonstrations for recognizing fake news. This study demonstrates that even basic algorithms can produce reasonable results in detecting fake news. As a result, the findings of this investigation suggest that systems like this could be very useful and effective in dealing with this critical issue. Web scraping is also a key part of this paper as the scraped data will be based on real-time news and will be more reliable than the ready-made datasets available all over the internet. It is an efficient and fast process, and it is relatively easy to maintain. The dataset used in this study is expected to be used in arrangements that use machine learning-based statistical calculations, for example, Logistic Regression (LR), Decision Tree, Gradient Boosting, Random Forest, AdaBoosting, SVM, XGBoost and LightGBM. In the future, the prototype’s efficiency and accuracy can be improved, along with the proposed model’s user interface.

VI. REFERENCES

[1] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 436-441.
[2] Iftikhar Ahmad, Muhammad Yausaf, Suhail Yausaf and Muhhammad Dovais Ahmad, “Fake news detection learning ensemble methods”, Hindawi, vol. 2020, pages 1-11, October.
[3] 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.
[4] C. Buntain 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.
[5]. 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.
[6] Fathima Nada, Bariya Firdous Khan, Aroofa Maryam, Nooruz-Zuha, Zameer Ahmed “Fake news detection using logistic regression”, Vol. 6 Issue 5, May 2019, IRJET.
[7] M. Granik 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.

[8] 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).