iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.36604

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Fake News Detection Using Machine Learning Algorithms

Ms. Sarika Tyagi[1], Aditri Rastogi[2], Bhawna Suman[3], Akash Mittal[4], Anushka Saxena[5]
[1]*(Assistant Professor, Dept of Computer Science,Krishna Engineering College,Mohan Nagar Ghaziabad UP-201007 India)*
[2, 3, 4, 5]*(Dept of Computer Science,Krishna Engineering College,Mohan Nagar Ghaziabad UP-201007 India)*

Abstract: ***Fake news always has been a problem. We, too, might have fallen for a false rumor at least once in our lifetime. Moreover, the fight against fake news over social networking media is intricate. Misinformation related to home remedies for COVID 19 that have not been verified, fake news for lockdown extension or release, casualties and damage in any riots, fake consultancies, and conspiracy were prevalent during the lockdown. Many Researchers have implemented several algorithms for the detection of Fake News.***
In this paper, we have used several past published research papers along with our research to compare the performances of three algorithms, i.e., Naive Bayes classifier, Logistic Regression, and Support Vector Machine. This provides an idea of the most practical and efficient algorithm, Support Vector Machine, that can be used for fake news detection.

I. INTRODUCTION

As a considerable amount of our lives is spent interacting online through social media platforms, more and more people tend to seek out and consume news from social media rather than traditional news sources. In Today's world, anybody can post any content over the internet. Individuals get misguided and don't reconsider before sharing such misleading pieces to the most reserved part of the arrangement. Such things are not suitable for a society where some rumors or fake news give rise to negative thoughts. As rapidly the technology is moving, at the same pace, preventive measures are necessary to deal with such things. There are various sites that give false information. They intentionally attempt to bring out purposeful publicity and falsehood under the claim of being true news. Their primary role is to control the data that can cause open to having belief in it. There are lots of cases of such sites everywhere throughout the world. Fake news detection is made to stop the rumors, and fake news spread through social media. This paper provides a comparison between the performances of the three most suitable and mostly used Machine Learning Algorithms, i.e., Naïve Bayes Classifier, Support Vector Machine, and Logistic Regression, for Fake News Detection. This paper also proposes a model that can predict whether the news is real or fake using the Machine Learning Algorithm. We have used the algorithm that has performed the best when comparing the three most suitable algorithms as mentioned above. The model will be using the URL as an input that validates the news, whether it is fake or true. The rest of the paper is organized as follows: Section II outlines the literature review where different research works on fake news detection are discussed briefly. Section III presents the proposed approach where the methodology, i.e., the algorithms compared in the proposed model, is discussed. Section IV defines the dataset used and the implementation part of the paper. Section V concludes the paper and summarizes the results obtained after the implementation of the model.

II. LITERATURE SURVEY

Fake news can be viewed as one of the greatest threats to democracy, public trust, justice, journalism, and the economy, so fake news detection has recently become emerging research attracting tremendous attraction. However, detecting fake news poses several new and challenging research problems.
Many researchers have developed different approaches and applied different algorithms for analyzing and detecting fake news. A computational stylistic analysis based on Linguistic approaches and Natural Language Processing is efficiently applied by Nicollas R. de Oliveria [1]. At the same time, a three-part process using Naïve Bayes Classifier, Support Vector Machine, and Semantic analysis is proposed by Akshay Pratap Singh [2]. Kuai Xu and Fang Wang have used another two perspectives, i.e., domain reputations and content understanding. Their analysis reveals that websites of fake and real news publishers exhibit diverse registration behavior, registration timing, domain rankings, and domain popularity [3]. An innovative hybrid approach that combines two major categories- linguistic cue approaches (with machine learning) and network analysis approaches (using network-based behavioral data) has been provided by Nadia K. Conroy [4]. The existing approaches from a Data Mining perspective, including Feature Extraction and Model Construction, have been reviewed by kai Shu and Amy Sliva [5].

A book has been published by Kai Shu and Huan Liu, which aims to bring researchers and practitioners together for understanding propagation, improving detection, and mitigation of disinformation and fake news on social media [6]. A hybrid Neural Network architecture combining the capabilities of CNN and LSTM is used with two different dimensionality reduction approaches; Principle Component Analysis (PCA) and Chi-Square are being implemented by Muhammed Umer and Saleem Ullah [7]. Traditional machine learning approaches and deep learning approaches are being used by Wenlin Han* and Varshil Mehta[8]. A comparison between the five most suitable machine learning algorithms is made by Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter [9].

III. OBJECTIVES

There is a subtle difference in detecting real and fake news, so it is a daunting challenge. Humans are inefficient at distinguishing between true and false facts, which results in a threat to logical truth and deteriorates democracy, journalism, and governmental institutions' credibility.

The machine learning algorithms that we compared were, Support Vector Machine, Logistic Regression, and Naive Bayes.

These three Machine Learning approaches are excellent in detecting fake news with high confidence, precision, and maximum efficiency. Therefore, after comparing the three machine learning algorithms based on their confusion matrix, efficiency, and accuracy, we needed to conclude which algorithm is the most accurate and efficient in detecting fake news.

IV. METHODOLOGY

A. *Algorithms*

1) *Naive Bayes Classifier:* Naive Bayes algorithm falls under classification in Supervised Machine learning. This works on the principle of conditional probability (given by the Bayes theorem). Google News recognizes whether the news is world news, political, and so on with the help of the Naive Bayes classifier.

Bayes theorem formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

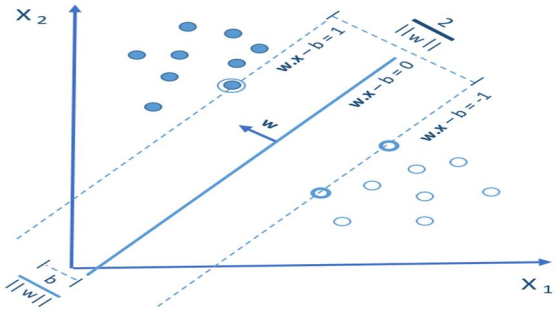
Bayes theorem can be put in simple terms as:

$$posterior = \frac{prior \times likelihood}{evidence}$$

This model is combined with a decision rule by the Naive Bayes classifier. *MAP* decision rule or *maximum a posteriori* is one common rule that picks the most probable hypothesis. The corresponding classifier, a Bayes classifier, is the function that assigns a class label for some *k* as follows:

$$y = argmax_y P(y)\pi_{i=1}^{n} P(x_i|y)$$

2) *Support Vector Machine:* Support Vector Machines (SVM) are an arrangement of related supervised learning techniques operated for grouping and classification. For example, some useless words are present in the textual data. These useless words are called stop words. The features are extracted after these stop words are removed from the text data. After the features extraction, SVM Classifier performs classification on the data; and defines whether the given news is a piece of fake news or real news. A pictorial representation of how the textual data is classified using the SVM Algorithm is shown in the figure given below. The algorithm creates a line or a hyperplane based on which it splits the data into classes.



3) *Logistic Regression:* Logistic regression is an algorithm based on Maximum Likelihood (ML) Estimation. It says coefficients should be chosen to maximize the probability of Y given X (likelihood). With Machine Learning, the computer uses different "iterations" to try other solutions until it gets the maximum likelihood estimates.

The equation to predict:-
The most popular iterative method to estimate the regression parameters is Fisher Scoring.
logit(p) = b0 + b1X1 + b2X2 + —— + bk X
where;
logit(p): ln (p / (1-p))
p: p is the probability of the dependent variable. The value of p is either a "success" or "event".
b: parameters of the model.
X1,X2,X3…Xk: Predictors of the model.

B. Dataset

1) Two CSV files, one containing real news and the other having fake news headlines, are used.
2) Each file has Title, text, subject, and date attributes.
3) There are 21417 true news data and 23481 fake news data given in true and fake CSV files, respectively.

C. Implementation

We first need to process our dataset to apply the algorithms (Pandas, Numpy, and Scikit-learn). We will implement the following vectorizers in order to process the dataset as we are dealing with the text data. The different vectorizers are as follows:

1) *TF-IDF Vectorizer:* It shows how frequent a term is in an entire document. To represent the presence of that term, it tries to assign a metric value. This weight is used to assess how essential a word is to a report in a corpus.
2) *Count Vectorizer:* It represents a record in the form of a matrix data set matrix notation in which a corpus document is represented by each row, each column represents a corpus term, and each cell shows the frequency output of a particular term in a specific document.

We have implemented and shown the comparison between three different classification models. Based on the comparison, we have also considered the accuracy, precision, F1-support, and recall value of the models. A combination of three different models is being trained on the dataset. After comparing the accuracy and evaluating the classification matrix of all the algorithms, we found the most suitable algorithm for our fake news detection. We have implemented that algorithm for the prediction of fake and true news.
.

V. RESULTS

Firstly, Naïve Bayes Model was trained on the labelled dataset on each of the vectorizers, i.e., TF-IDF Vectorizer and Count Vectorizer, and it gives 94.08% accuracy.
Then Logistic Regression model was trained on the labelled dataset on each of the vectorizers, i.e., TF-IDF Vectorizer and Count Vectorizer, and it gives 98.73% accuracy.
Thirdly Support Vector Machine was trained on the labelled dataset on each vectorizer, i.e., TF-IDF Vectorizer and Count Vectorizer. It gives 99.39% accuracy, which was the highest of all the other three models.
The findings mentioned above have been shown in the table (i.e., Table 2) given below.

Classifiers	Accuracy
Logistic Regression	98.73%
Naïve Bayes	94.08%
Support Vector Machine	99.39%

Table 2: Accuracy of Classifiers

Table 3 shows the Precision, Recall, and F1-score for Machine Learning Algorithms such as Naïve Bayes Classifier, Logistic Regression Classifier, and Support Vector Machine Classifier.

Logistic Regression and Support Vector Machine Algorithms have the same level of precision, and Naïve Bayes Classifier has lower precision than both the algorithms.

Support Vector Machine outperforms the other algorithms considering recall as compared to Logistic Regression and Support Vector Machine. The F1-score of both Logistic Regression and Support Vector Machine is the same, but more than that of Naïve Bayes Classifier.

Support Vector Machine outperforms the other algorithms considering Recall, Precision, and F1-score.

1) *Precision:* The ratio between correctly predicted positive observations and the total predicted positive observations is termed as the prediction.
2) *Recall:* The ratio between correctly predicted positive observations and all actual class observations is termed recall.
3) *F1-Score:* The weighted average of recall and precision is the F1-Score.

Classifiers	Precision	Recall	F1-Score
Logistic Regression	0.99	0.99	0.99
Naïve Bayes	0.94	0.95	0.94
Support Vector Machine	0.99	1.00	0.99

Table 3: Precision, Recall and F1-Score for Naïve Bayes, Logistic Regression, SVM

Figure 1 shows the overall comparison among all algorithms. SVM shows the highest accuracy among all the classifiers.

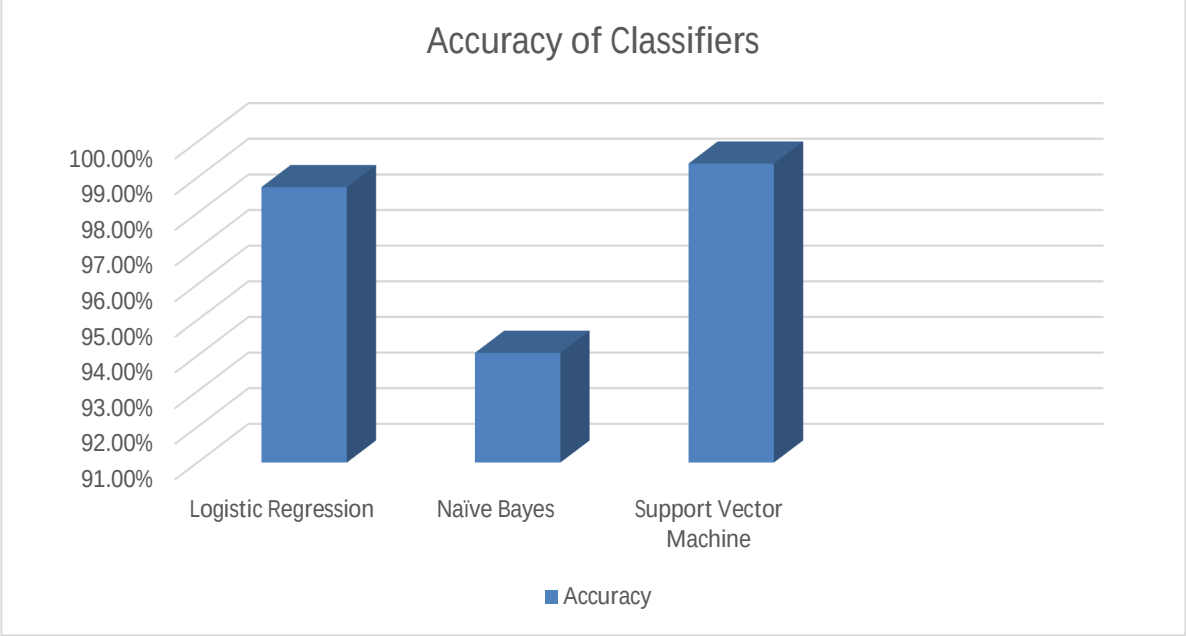


Figure 1: Overall Accuracy Comparison of Algorithms

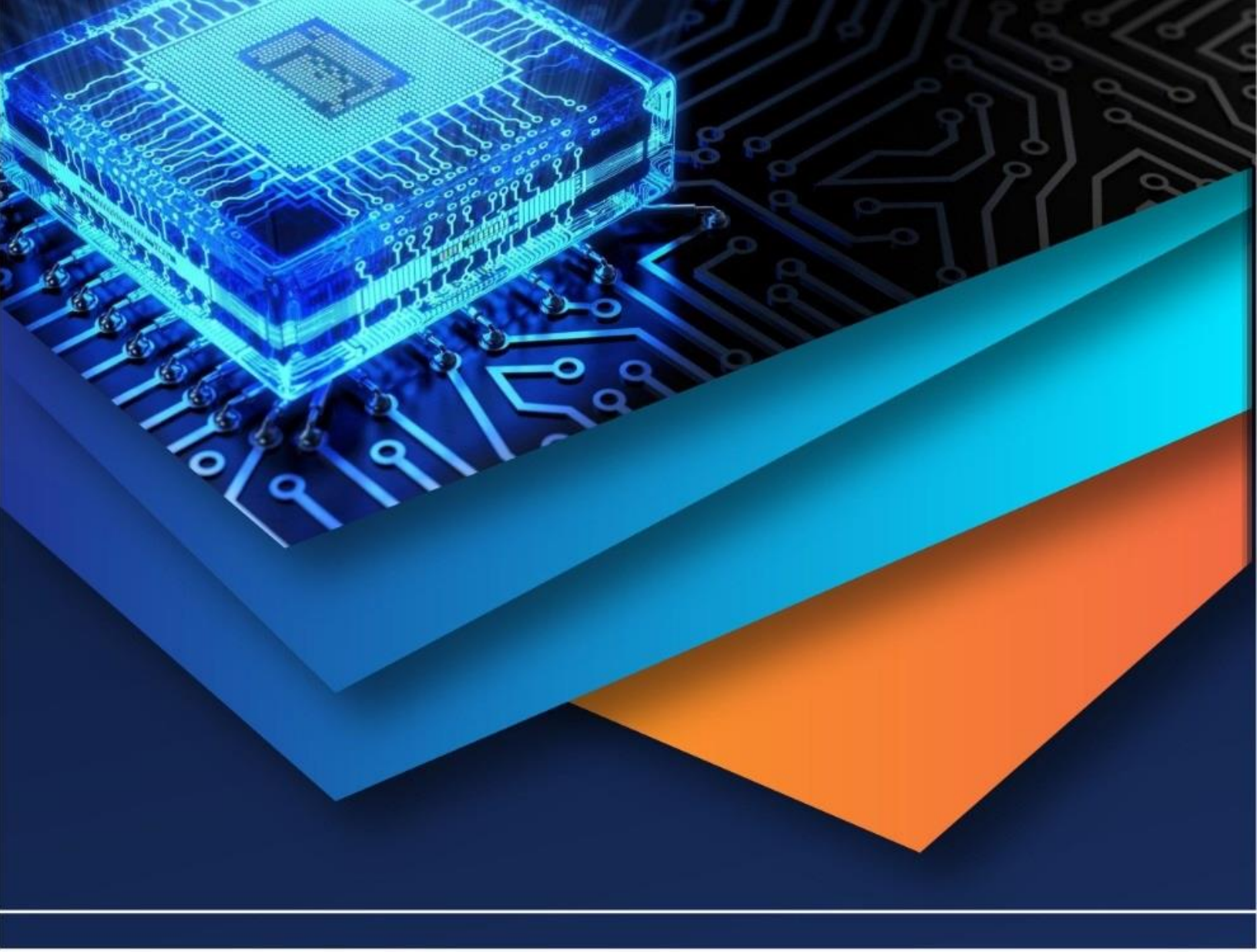
VI. CONCLUSION

We have made a comparison between the performances of the three most suitable machine learning algorithms, i.e., Logistic Regression, Support Vector Machine, and Naïve Bayes for Fake News Detection. These three algorithms are trained on the dataset with TF-IDF Vectorizer and Count Vectorizer.

Lastly, based on the results, the accuracies are compared, and the algorithm, i.e., Support Vector Machine (SVM), which has given the highest accuracy, i.e., 99.39%, is used for the Fake News Detection.

REFERENCES

[1] Nicollas R. de Oliveira, Dianne S. V. Medeiros and Diogo M. F. Mattos "A Sensitive Stylistic Approach to Identify Fake News on Social Networking" Institute of Electrical and Electronics Engineers [IEEE] Journal -2020
[2] Akshay Pratap Singh, Deepanshi Sachdeva and Dheeraj Kumar "Fake News Detection On Social Media" International Journal of Scientific Research in Engineering and Management[IJSREM] –July 2020
[3] Kuai Xu , Feng Wang, Haiyan Wang, and Bo Yang "Detecting Fake News Over Online Social Media via Domain Reputations and Content Understanding" Institute of Electrical and Electronics Engineers [IEEE] Journal -2020
[4] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen "Automatic Deception Detection: Methods for Finding Fake News" Proceedings of the Association for Information Science And Technology -2015
[5] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang and Huan Liu "Fake News Detection on Social Media: A Data Mining Perspective" ACM SIGKDD Explorations Newsletters
-September 2017
[6] Huan Liu and kai Shu "Detecting Fake News on Social Media" Book- July2019
[7] Muhammad Umer, Zainab Imtiaz, Saleem Ullah ,Arif Mehmood, Gyu Sang Choi and Byung-Won On "Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)" – 2016
[8] Wenlin Han* and Varshil Mehta "Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation- 2019
[9] Abdullah-All-Tanvir, Ehesas Mia mahir ,Saima Akhter , Mohammad Rezwanul Huq "Detecting Fake News using Machine Learning and Deep Learning Algorithms -2019



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY