Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 7, No. 4, August, 2023

Stand up Against Bad Intended News: An Approach to Detect Fake News using Machine Learning

Nafiz Fahad [1*], K. O. Michael Goh [2*], Md. Ismail Hossen [1], K. M. Shahriar Shopnil [1], Israt Jahan Mitu [1], Md. A. Hossain Alif [1], Connie Tee [2]

[1] *Faculty of Science & Technology, American International University-Bangladesh, AIUB, 408/1, Kuratoli, Khilkhet, Dhaka 1229, Bangladesh.*

[2] *Faculty of Information Science and Technology (FIST), Multimedia University, Melaka, Malaysia.*

Abstract

The purpose of this approach is to find out the effects and efficiently detect fake news by using a publicly available dataset. However, it is difficult for human beings to judge an article's truthfulness manually, which is why This paper mainly wanted to cure the effect and to found out an automated fake news detection system with benchmark accuracy by using a machine learning classifier, which must be higher than other recent research works. In essence, this work's target is to find out an efficient way to detect fake and real news, and it also the target is to compare with existing work where researchers used machine learning classifiers and deep learning architecture. The proposed approach depended on a systematic literature review and a publicly available dataset where 7796 news data are recorded with 50% real and 50% fake news. The best and benchmark accuracy is 93.61%, achieved by the Support Vector Machine (SVM) among the used Random Forest, Decision Tree, KNN, and Logistics Regression classifiers, and the achieved accuracy is better than the exciting recent research works. Moreover, fake news is detected, people are able to differentiate between fake or real news, and effects are cured when people used SVM.

Keywords:

Fake News; Rumor;
False Information;
Social Platforms;
Machine Learning.

Article History:

Received:	26	January	2023
Revised:	14	June	2023
Accepted:	21	June	2023
Available online:	12	July	2023

1- Introduction

In this era, Artificial intelligence brought the most important changes in the field of information technologies and architectures, like robotic surgery, natural language processing, developing and using intelligent transportation systems, et cetera [1].

Nowadays, due to the internet of things, online platforms like Facebook and YouTube are one of the main sources of news. Human beings are significantly dependent on online news when they are busy with their other daily usual work. Fake news is a human activity; that is not a new thing in our lives. However, true and false news created chaos among each other. Online platforms have a bucket of information, which makes it difficult to differentiate between real and fake news [2]. A bunch of fake news spread throughout the world within a nanosecond through online news. Some misleading news has lost credibility through the social platform. Fake people use chic headlines to attract people easily to click on them. Fake news can have financial and political implications [3]. Lack of evidence of fraud news covers the truth. At present, fake news is spreading online at an alarming rate. Facebook, Twitter, Instagram, et cetera are the most popular social platforms for spreading fake news. An example of how fake news spreads is illustrated below in Figure 1.

* **CONTACT**: fahadnafiz1@gmail.com; michael.goh@mmu.edu.my

DOI: http://dx.doi.org/10.28991/ESJ-2023-07-04-015

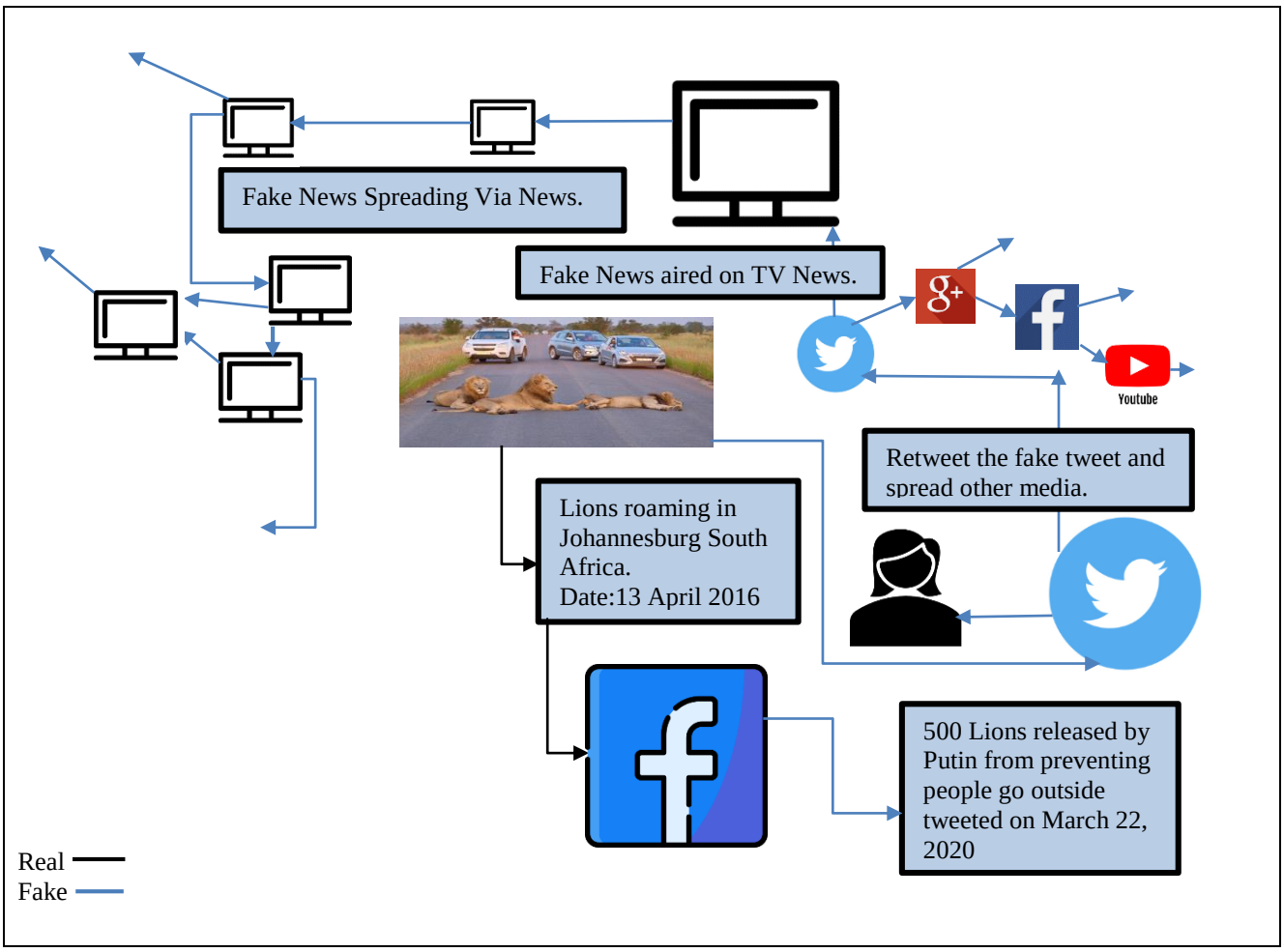


Figure 1. An example of a figure

From this picture, it is seen how the fake news spread. At first, someone posted on Facebook that lions roamed in Johannesburg, South Africa. After that, someone posted on Facebook fake news that a human had released 500 lions, preventing people from going outside. Then someone tweets fake news on Twitter, and this time the fake news is posted on various social platforms and finally shown on television news. However, fake news is used by people's sentiments to mislead and misguide. So, from this picture, it is understood that in this current era, social platforms are the main sources for spreading fake news, which has a bunch of effects on human beings.

Again, in this internet-based, globally connected media, where anyone can participate, intrusions into our daily lives are also now a given. Even though we have constantly had myths and stories that influence our perceptions and attitudes, the sheer number of online print or social platforms ruins the voices of reality [4]. Journalism has an underlying issue as well. Daily, the media disseminates vast amounts of fake information that might elicit anxiety or inspire us to drastically alter our routines. The offensive of fake news, such as the drawbacks of vaccination, the alleged dangers of gluten, sugar, and genetically modified foods, and the risk of developing cancer from acidic diets, can be overwhelming and has a greater chance of becoming widely circulated than actual news [5]. Fake news or false news causes severe health problems and negatively impacts societal trust issues, financial markets, political issues, and economies [6].

The objective of our paper is to find out the effect using the systematic literature review and detect fake news using four supervised machine learning classifiers, including logistic regression (LR), k-nearest neighbor (KNN), decision tree (DT), random forest (RF), and support vector machine (SVM). Our goals are to find out the effects of fake news, the most efficient way to detect fake and real news with the best accuracy, and the most accurate accuracy compared to other recent works.

The rest of the paper is organized as follows: section 2 discusses the existing works; the proposed method is illustrated in section 3; section 4 describes the result and discussion; and lastly, in section 5, a conclusion is drawn.

2- Literature Review

Numerous studies have mostly concentrated on identifying and categorizing bogus news on social platforms and websites like Twitter and Facebook [3]. Fake news or rumors have been conceptually divided into various categories,

and then models of ML have been developed for various domains and sectors [7]. One approach is to extract linguistic information from textual articles, such as n-grams, and then train various ML models, like the support vector machine and stochastic gradient descent [8]. Another strategy involved merging textual elements with data, where the data is auxiliary, like social engagement, to gain improved accuracy with various models. The sociological and psychological theories and how to spot fake information online were also covered by the authors. The authors also covered several data mining algorithms for building models and methods for extracting shared features. These models are built by using information like the writing pose and social environments like status and advocacy [9].

Another way is that the author trained multiple ML models using textual features and metadata. The author primarily utilizes convolutional neural networks. The dependency is captured between the metadata vectors using a convolutional layer, which is followed by a two-directional LSTM layer. In the final prediction, joining maximum text depictions with the metadata presentation using two directional LSTM layers and feeding it with a fully attached layer by using a function named SoftMax. The study uses the political world dataset, which includes remarks from two different parties. Additionally, as a feature set, some metadata is also supplied, like subject, employment, speaker, context, status, party, and history. With the use of a mix of features, including text and speaker, an accuracy of 27.7% was attained [10]. A competing solution, which allocates an article with four labels—"disagree," "agree," "unrelated," or "discuss" relying on how well the headline or article title of the article matches the content of the article, is an attitude identification system. As a landmark set for their multilayer perceptron classifier, linguistic aspects of text like the term frequency are used by the authors, which is a document frequency, and the frequency is inverse, as well as one secret layer, and a function named SoftMax is used on the final layer to get output. In the dataset, each article had a headline, body, content, and labels mentioned as text or numbers. The algorithm performed poorly when labeling test instances as "disagree" but best when labeling them as "agree" [11]. Some of the effects of fake news mentioned in recent research are summarized in Table 1.

Table 1. Some Effects of Fake News

Extracted Information	References
Fake news influences decision-making and distorts one's perceptions.	Zhang et al. (2019) [12]
Pollute the reputation of a well-reputed company by publishing fake news or rumors, creating government harm, making monetary, social, and political losses, using the sentiment of the people and people occurs crimes.	Awan et al. (2021) [13]
Misleading people by using fake or false news.	Meel & Vishwakarma (2021) [14]
Fake news affects societal values, changing opinions, and redefining truths, beliefs, and facts.	Olan et al. (2022) [15]
Threaten the public's confidence and always cause misunderstandings.	Liao and Wang (2021) [16]
Publishers can tarnish reputations, ruin businesses, muddy public discourse, and sway political decisions by publishing fake news.	Taher et al. (2021) [9]
Fake or false news is predominantly viewed as a threat to the rational processes of sense-making and decision-making.	Bastick (2021) [17]
People are unable to assess the reliability of the information, people are tricked by fake news. Thus, opportunity seeker people spreading incorrect information online can "cause harm" and misleading claims have the "terrifying potential to cause actual harm to real people".	Hamdan (2020) [18]
Fake news is a social problem threatening the public's ability to trust legitimate press outlets. Fake news poses such a significant threat to the legitimacy of our press, and thus the democratic legitimacy of our government. Fake news undermines the informing function of the press by eroding the legitimacy and credibility of traditional, reliable news outlets, creating an uninformed public unable to participate effectively in our democracy	Mohseni et al. (2019) [19]
Various intentional harm is debated, and various incentives, such as monetary, social, and political benefits –often drive the fake news to spread.	Shu et al. (2017) [20]

Moreover, recently, Arif et al. [21] applied a Passive Aggressive Classifier (PAC), Bi-directional LSTM (LSTM means long short-term memory), a deep learning algorithm, and ROBERTA, a pre-trained language model, where the author got so much lower accuracies, which are 51% for PAC, 52% for Bi-LSTM, and 47% for ROBERTA for 1st dataset, for 2nd dataset they also got lower accuracies, which are - for PAC, 61% for Bi-LSTM, and 0.28% for ROBERTA. Another author applied DT (Decision Tree), KNN (K-Nearest Neighbors), LR (Logistic Regression), NB (Naïve Base), and SVM (Support Vector Machine) to detect fake news but got lower accuracy [22]. One more author got lower accuracies when applying the NB, LR, MLP (Multilayer perceptron), SVM, and PA (Passive aggressive) [23]. From recent works, it is clear that no one had to find benchmark accuracy, and no one has together done research on finding the effects and detecting fake news with benchmark accuracy.

2-1- Our Contribution

There are numerous examples of machine learning algorithms that are being used to categorize content or text in fake news [15]. However, the majority of the researchers focused on particular domains or datasets, most notably the realm

of politics [7]. As a result, when it is opened to articles from different domains, the algorithm does not produce the best results because it was trained on a specific article's domain. It is challenging to develop an algorithm that is general and performs best across all news domains because each article's textual structure differs across distinct news domains. In this research, we provide an ML ensemble strategy to detect fake news. Our study investigates diverse textual characteristics that could be used to distinguish between authentic and fraudulent content. We train several different algorithms of ML using a variety of ML methods that are not in the existing literature by utilizing those properties [24]. We have carried out thorough experiments on a real-world dataset that is publicly available [25]. The outcomes confirm the enhancement by utilizing the three generally used indicators of performance, namely "accuracy, precision, and recall".

3- Materials and Methods

For this research purpose, systematic literature review (SLR) and machine learning (ML) classifiers are appropriate to fulfill the research objective. SLR helps to find the effect, and ML helps to detect fake news.

3-1- Proposed Method

The diagram in Figure 2 shows the proposed approaches and the steps of the proposed method are illustrated below. The proposed method has seven main steps, which are gathering the data, data pre-processing, model selection, model training, estimation, adjusting the parameter, and prediction.

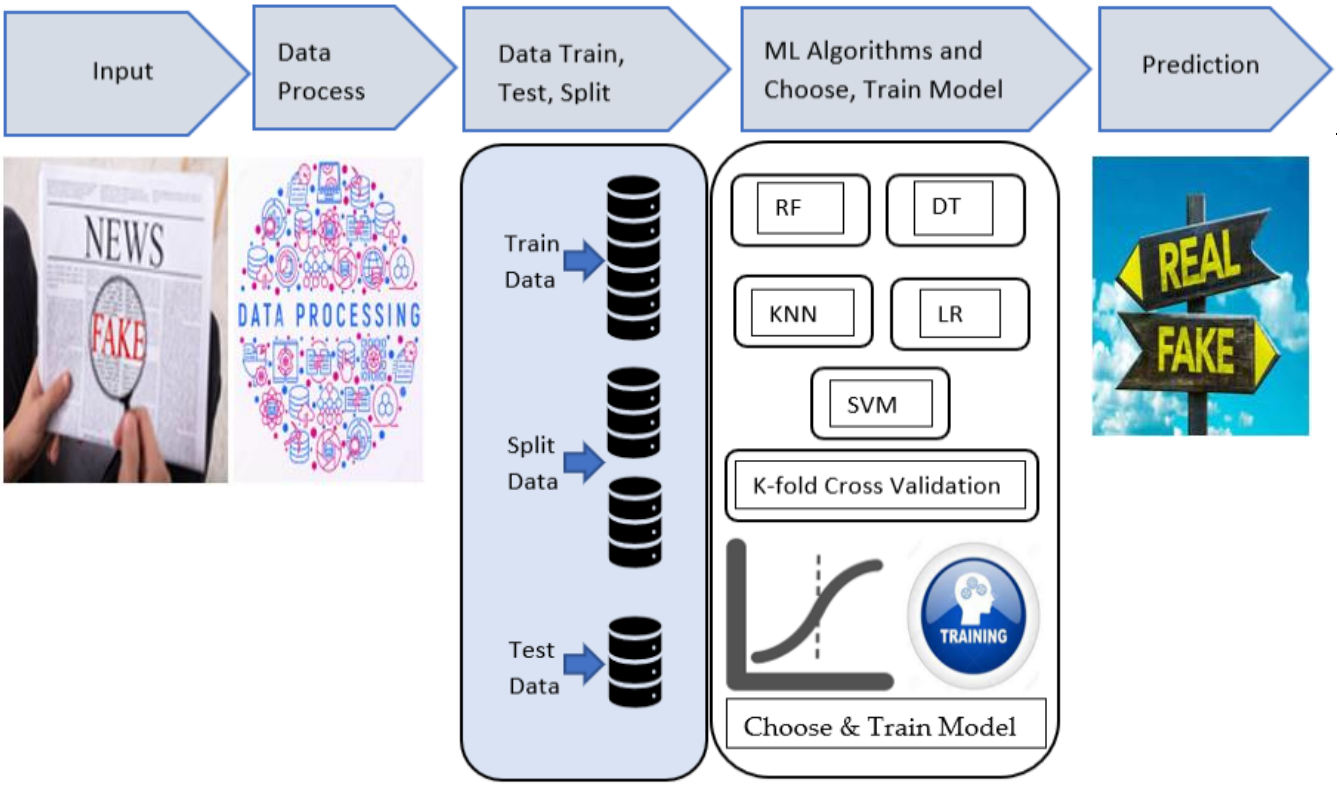


Figure 2. Proposed model diagram

3-2- Data Collection

A standard dataset is used in this research, and the dataset is publicly available on Kaggle. The dataset has a total of 7796 records, with 50% real and 50% fake news [26].

3-3- Data Pre-Processing

The raw dataset consists of several thousand news articles labeled as either real news or fake news. As ML deals with only numeric values, it was needed to format the text of the dataset with meaningful numbers. Besides, missing data were cleaned and reshaped accordingly so that they became suitable for further processing. The interface of the dataset before and after prepossessing is mentioned below (Tables 2 and 3).

Table 2. The feature of the dataset before processing

	Unnamed: 0	Title	text	Label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello…	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol…	Google Pinterest Digg LinkedIn Reddit Stumbleu	FAKE
2	3608	Kerry to go to Paris in a gesture of sympathy	U. S. Secretary of State John F. Kerry said Mon…	REAL
3	10142	Bernie supporters on Twitter erupted in anger at…	-Kaydee (@Kaydeeking) November 9, 2016 T…	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners…	REAL

Table 3. The feature of the dataset after processing

	Unnamed: 0	Title	Text	Label	Content
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello…	0	You Can Smell Hillary's Fear, Daniel Greenfield…
1	10294	Watch The Exact Moment Paul Ryan Committed Pol…	Google Pinterest Digg Linkedin Reddit Stumbleu	0	Watch The Exact Moment Paul Ryan Committed Pol…
2	3608	Kerry to go to Paris in a gesture of sympathy	U. S. Secretary of State John F. Kerry said Mon…	1	Kerry to go to Paris in a gesture of sympathy U…
3	10142	Bernie supporters on Twitter erupted in anger at…	-Kaydee (@Kaydeeking) November 9, 2016 T…	0	Bernie supporters on Twitter erupted in anger at…
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners…	1	The Battle of New York: Why This Primary Matte.

3-4- Model Selection

An important stage in any ML method is model selection. It necessitates the evaluation of the intended result and inputs. The model must make sensible decisions based on the nature of the output. Regression and classification are two components of supervised ML. In both situations, it identifies a certain input structure or relationship to anticipate the precise outcome. Considering the nature of our dataset, we used LR, DT, KNN, RF, and SVM algorithms to examine how the data behaves when subjected to various classifiers. We also used cross-validation (CV). The features we used are represented below in Table 4.

Table 4. Features we used

Classifiers/ technique We Use	Features We Use
Random Forest, Decision tree, KNN, SVM	Use 3 neighbors, train the model, pass the test data, predict the model
K-Fold cross-validation	Shuffle the dataset, Split the dataset, prediction
Logistics Regression	Model fitting, detection, Binary Prediction

3-4-1- LR

Since we are classifying text or content based on a broad feature set with a binary output (true/false or authentic article/fake article), the LR model is used because it provides an equation that is a simple cost function and obtains an equation that categorizes issues between binary or multiple classes. To achieve the best results for a particular dataset, we tuned hyperparameters. Several parameters were evaluated before obtaining the LR model to produce benchmark accuracy [19].

$$h_\theta = \frac{1}{1+e^{-\beta 0 + \beta 1 X}} \quad (1)$$

Here, the meaning of h_0=hypothetical function β_0= intercept, β_1= slope (expected change in the outcome y per unit change in x, x=feature of data).

A sigmoid function is used by the logistic regression to transform the output into a probability value, and the objective is to achieve an optimal probability by minimizing the cost function [24]. The cost function mentioned below:

$$\text{Cost}\,(h_\theta\,(\text{x}),\ \text{y}) = \begin{cases} \log\big(h_\theta\,(x)\big), & y = 1 \\ -\log(1 - h_\theta x)), & y = 0 \end{cases} \quad (2)$$

Here, the meaning of cost=function, $h_\theta\,(\text{x}),\ \text{y}$ = hypothesis function of x and y, y=actual value of training data

3-4-2- DT

An essential tool is DT, which resembles a flow chart by mostly addressing classification problems and functions within a structure. The decision tree branch is determined by the internal node's condition, or "test" on an attribute result

(Figure 3). The leaf node is finally given a class label when all attributes have been calculated. The distance between the root and the leaf serves as a representation of the classification rule. It's fantastic that it can be applied to both a category and a dependent variable. They do a fantastic job of highlighting the most important factors and demonstrating how factors are related. They play a vital role in developing new variables and characteristics that are helpful for data exploration and accurately forecasting the desired variable. Predictive models often use supervised learning techniques to build high accuracy, and this is where tree-based learning algorithms come into play. They excel at mapping relationships that are not linear. They also go by the name CART and do a good job of solving classification or regression problems [14].

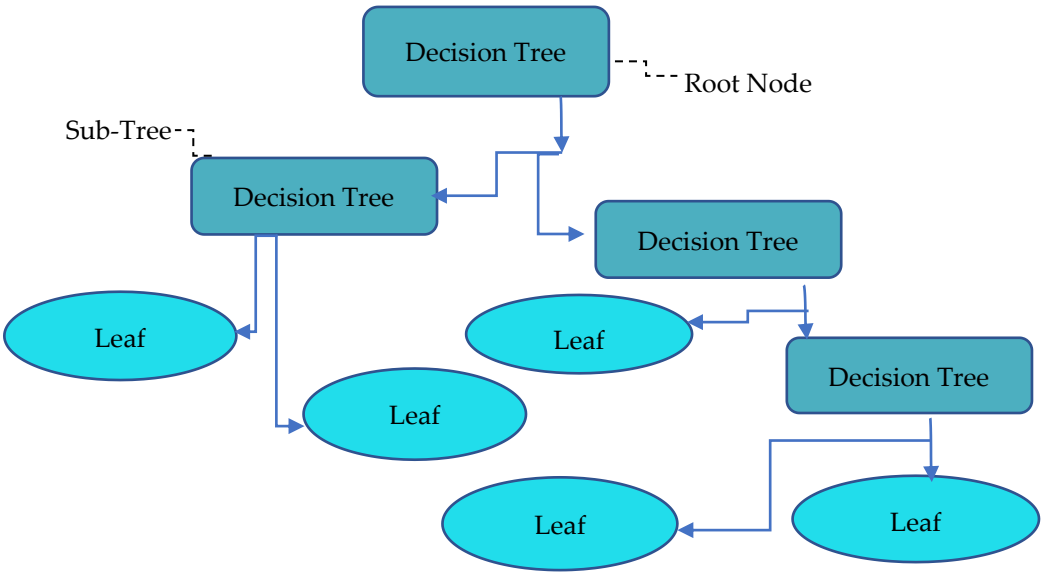


Figure 3. Structure of decision tree

3-4-3- KNN

Without using a dependent variable, KNN is capable of predicting specific data outcomes. We provide enough practice data so that it can identify the precise neighborhood to which a given data item belongs. The number of K measures the majority of new data and the votes of the new data point's neighbors, and the KNN model calculates the distance between a new data point and its nearest neighbors. The new data point is assigned to the class with the shortest distance if K is equal to 1 [27].

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (3)$$

$$\text{Manhattan distance} = \sum_{i=1}^{k} |x_i - y_i| \quad (4)$$

$$\text{Minkowski distance} = \left(\sum_{i=1}^{k} |x_i - y_i|^q\right)^{\frac{1}{q}} \quad (5)$$

Here, $\sum$= summation, k= number of dimensions, i=index, x_i= datapoint from dataset, y_i = new datapoint to be predicted, $q = a$ parameter.

3-4-3- RF

Random Forest derives from the way different decision trees or algorithms of the same type are mixed within a forest of trees. Classification and regression problems can be used to carry out the random forest method [14].

3-4-5- SVM

SVM, or support vector machine, is a machine learning model that is supervised or used by some classification algorithm. This classification held two groups of problems. When we give sets of training data, the SVM model can categorize the next text. In some limited samples, SVM has higher speed and better performance than all other models. The classifier SVM is nothing but a two-dimensional simple line. It takes data points as input and outputs the hyperplane with separate tags. The simple line is the decision boundary. As there are two dimensions divided by a line, one is categorized as blue and another is categorized as red. Whichever tag or point is nearest to the hyperplane is the largest, and vice versa. Whatever the case, several parameters were evaluated before obtaining the SVM model's highest levels of accuracy [22].

3-4-6- CV

In cross-validation, we first divide the data set into k-number of partitions. We divided randomly. This is also called k-fold cross-validation. Then train the ML algorithm with a (k-1)-number of partitions each time from this k-number of partitions and test for the rest. So, what happens is that each part of our k-number of parts will be used as a test at least once. This is done by running the loop once and looping through it a k-number of times and getting different performance values for each different test set (Figure 4). Then finally, the average value of all performance values is taken, which reveals how good or bad the model is [28, 29]. The formula evaluation score mentioned below:

$$Final\ Model\ Evaluation = \frac{1}{k}\sum_{i=1}^{k} S_i \tag{6}$$

$$Cross\ Validation\ MEAN = \frac{Sum\ of\ Total\ k-Folds}{Number\ of\ k-Folds} \tag{7}$$

Here, the meaning of K= Number of Folds, S_i=Performance scores of I^{th} index, $i = index$.

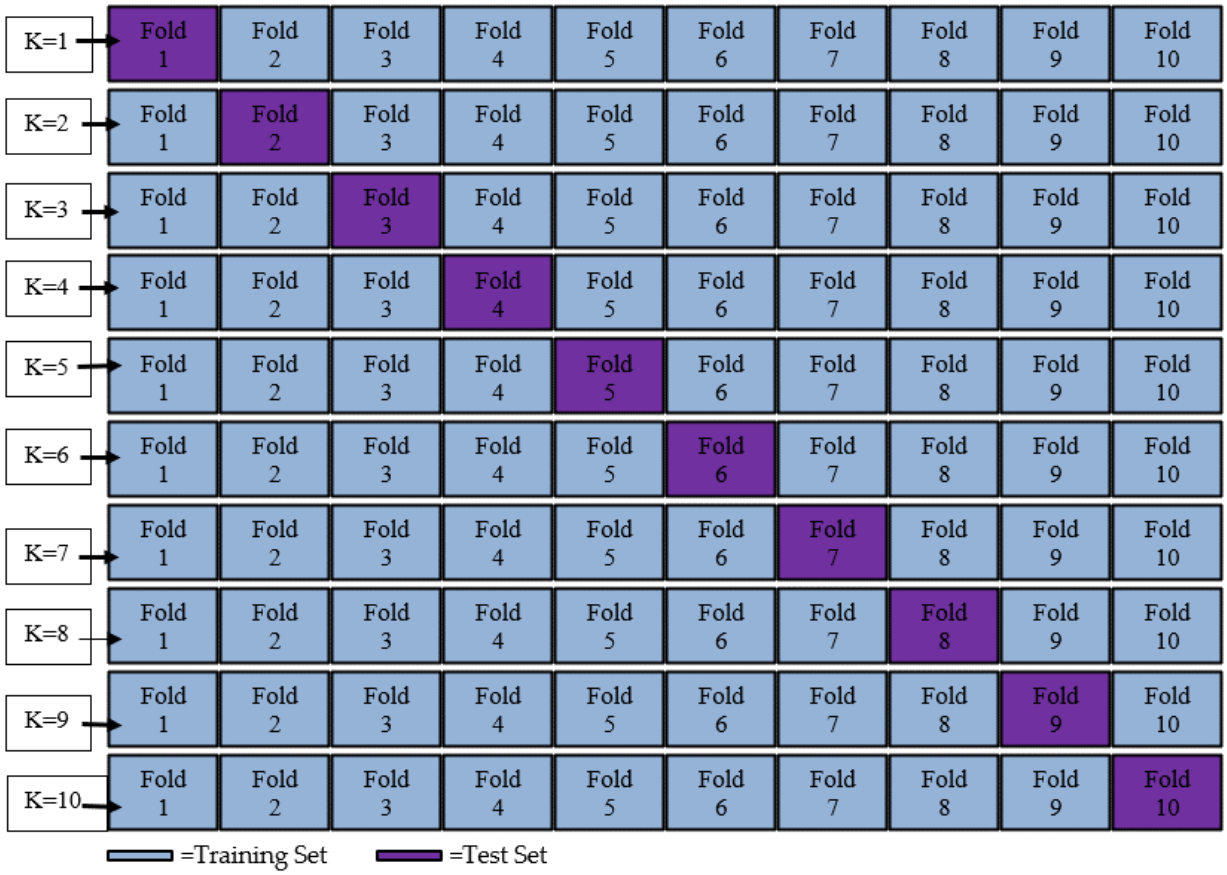


Figure 4. CV diagram for k fold where k=10

3-5- Training Model

For the training model, the pre-processed dataset was split between training and testing, 80% of the data was used for train, and 20% of the data was used for the test.

3-6- Model Evaluation

The model is assessed by a test data set. Several scores, including accuracy, precision, and recall, are calculated. Equations 8 to 10 are used for calculating accuracy, precision, and recall respectively.

3-6-1- Accuracy

For accurately predicting whether it is true or false, accuracy is the most popular measurement [30]. The following equation may be applied to determine a model. Accuracy:

$$Accuracy = \frac{TP+FP}{(TP+FP+TN+FN)} \tag{8}$$

3-6-2- Precision

Precision is a model which made the quality of positive prediction [3]. In our experiment, the number of true positives is divided by the total amount of positive predictions known as precision:

$$Precision = \frac{TP}{(TP+FP)} \tag{9}$$

3-6-3- Recall

Recall which is the sum of correctly classified instances outside of the true class [31]. Our experiment refers to the percentage of articles among all accurately predicted articles that were appropriately expected.

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

Here, TP = true positive, FP = false positive, TN = true negative, and FN = false negative

3-7- Tools Used

Scikit-learn one of the most dependable and useful libraries for ML in Python was used to experiment. Sci-kit-learn provides a range of efficient methods for statistical modeling and ML, including dimensionality, clustering, regression, and classification.

4- Result and Discussion

This section describes the experimental findings that were made during the research. Among the four utilized are the classifiers LR, DT, RF, KNN, and SVM. The variables are taken into account during the experiment to determine recall, accuracy, and precision. Also used to validate the accuracy scores is K-fold cross-validation (K-CV). K-CV divides the dataset into K folds, with each fold acting as the testing set for a subsequent iteration. The dataset in this instance is divided into 10 folds. Thus, k is equal to 10 in this situation. The experimental results of accuracy, precision, recall, and cross-validation are successively reported here. Figure 5 displays the accuracy of the suggested systems' performance for each of the selected classifiers. Figure 5 demonstrates that the LR, DT, RF, KNN, and SVM accuracy scores of test data are 91.00%, 65.47%, 79.73%, 69.44%, and 93.61%, respectively. The SVM classifier achieves a maximum accuracy of 93.61%.

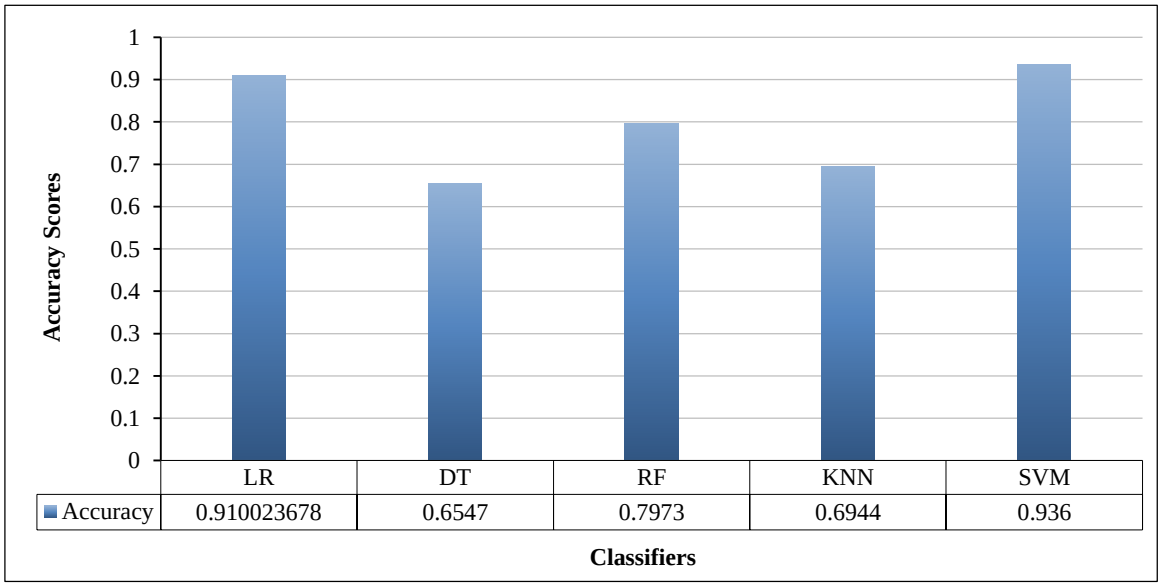


Figure 5. Accuracy vs. Classifiers

Secondly, experiments with precision scores have been conducted to calculate the performance of the suggested strategy. The ratio of total positive observations to accurate positive predictions is determined by the precision score. It illustrates how frequently the favorable forecast comes true. The higher, the better in this situation. For the LR, DT, RF, KNN, and SVM, the experiment yielded precision scores of 88.48%, 72.22%, 76.87%, 77.36%, and 91.32%. Figure 6 displays the results of all classifiers' precision experiments. The data demonstrate that SVM produces the best outcomes.

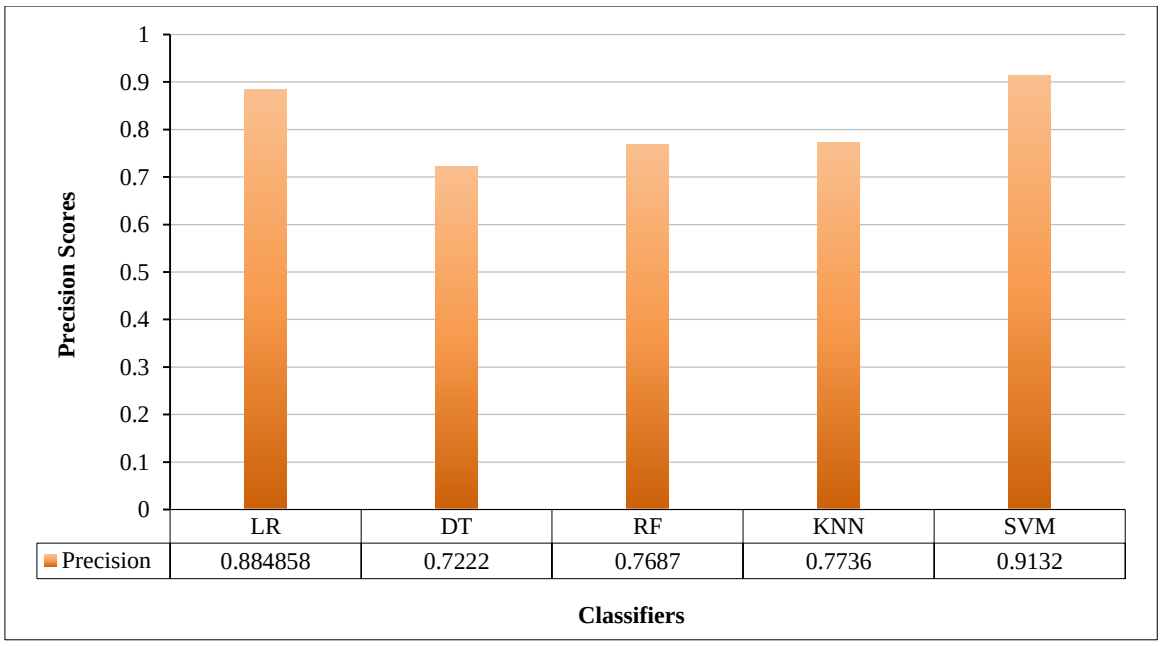


Figure 6. Precision vs. Classifiers

The recall score is also used to compute the observation of the actual positive prediction number over all of the actual label classes. Sensitivity, which shows the percentage of real positive findings, is also known as recall. As the recall score increases, the performance gets better. The corresponding findings for LR, DT, RF, KNN, and SVM are 93.18%, 64.84%, 82.63%, 67.88%, and 95.70%, respectively, based on the attained recall score. The highest rating once again went to SVM performance (95.70%). Figure 7 shows the experimental recall scores.

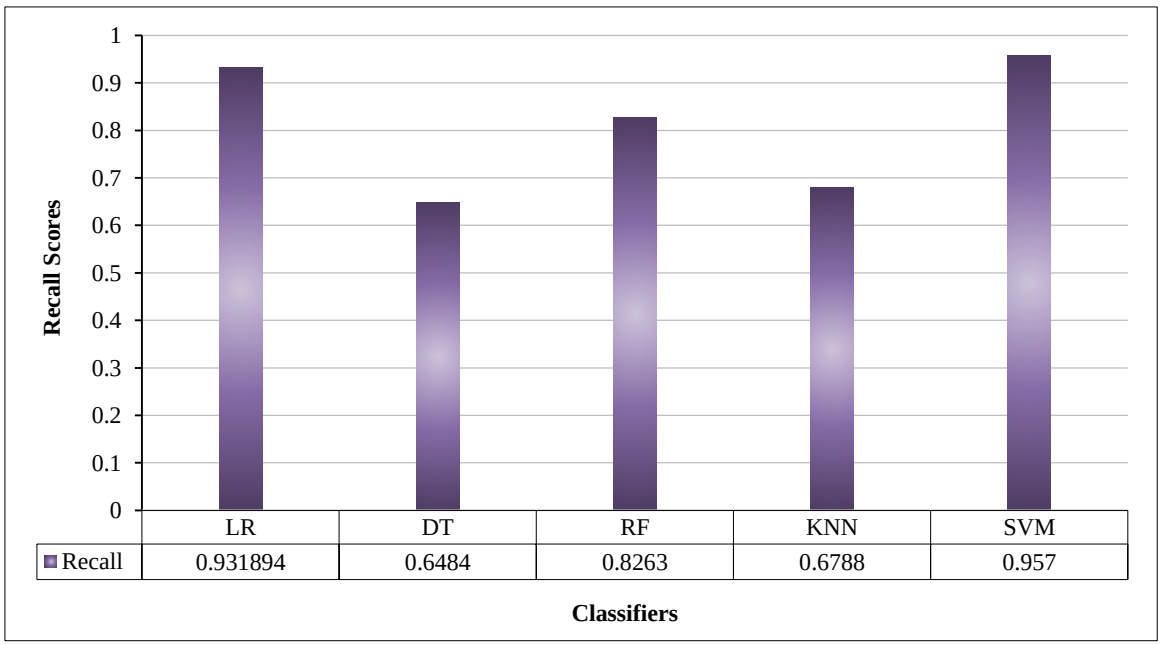
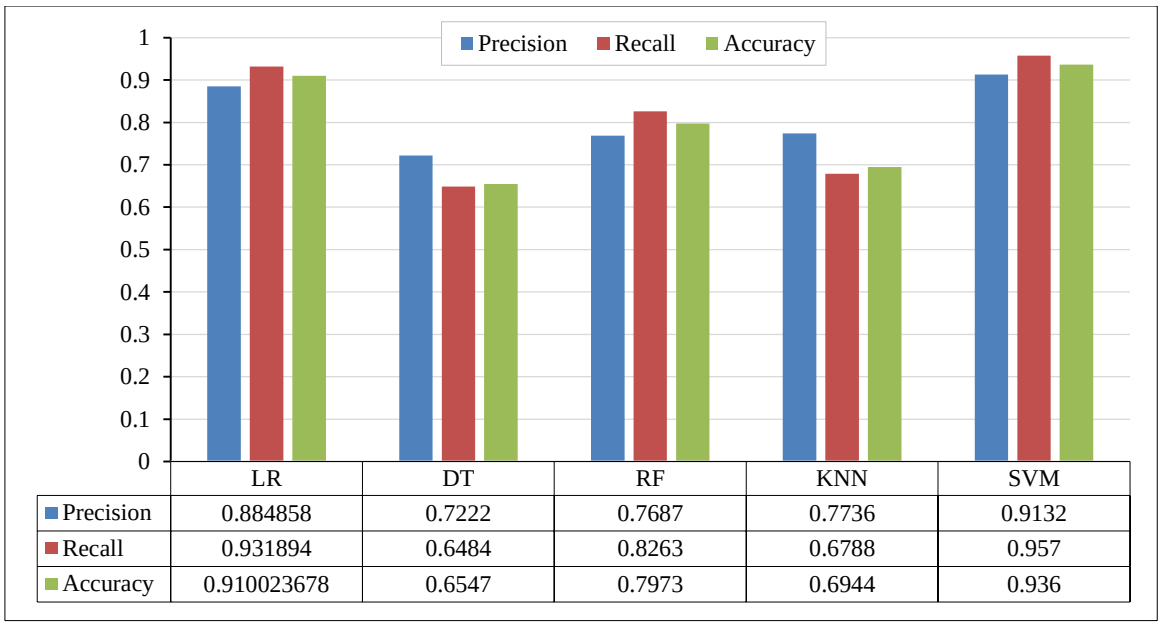


Figure 7. Recall score vs. Classifiers

As an overview of all the classifiers, Figure 8 shows the accuracy, precision, and recall scores for each classifier in a single figure. The chart shows that LR, DT, RF, KNN, and SVM all produce results that are almost identical, with very high scores, in contrast to DT's relatively low scores.

The mentioned Figure 9 confirmed the accuracy of SVM and demonstrated that it offers the maximum accuracy. K-CV is performed, and the outcome is shown in Figure 9. The figure makes it obvious that all K-CV folds yield scores that are essentially the same. The fact that there aren't many variances across the folds suggests that the precision is reliable. Additionally, the 10-fold cross-validation mean is observed to be 93.30%, validating the SVM result since the test's accuracy is likewise the same. These two K-CV points firmly establish that the data utilized for the experiment was adequate and that the accuracy of the experiment was good.



	LR	DT	RF	KNN	SVM
Precision	0.884858	0.7222	0.7687	0.7736	0.9132
Recall	0.931894	0.6484	0.8263	0.6788	0.957
Accuracy	0.910023678	0.6547	0.7973	0.6944	0.936

Figure 8. Summary of accuracy, precision, and recall classifiers

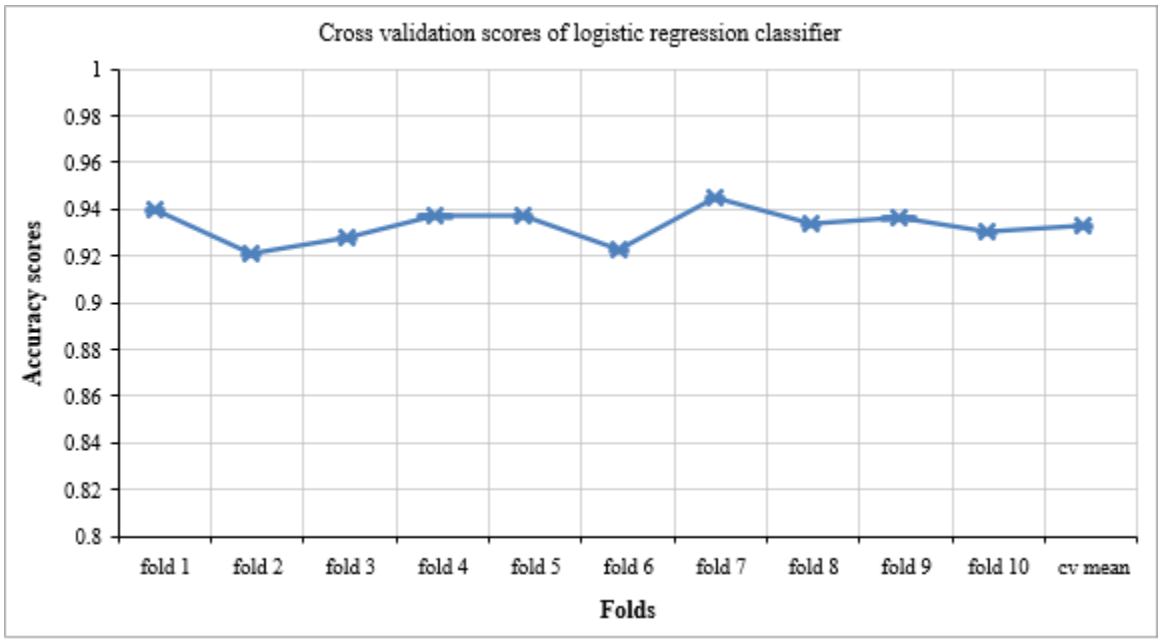


Figure 9. Result from KCV for SVM classifier [32]

4-1- Summarization

The following table is representing all the experiment results' accuracy, precision, and recall mentioned below in Table 5.

Table 5. Summarization of ML Classifiers

ML Classifiers	Accuracy	Precision	Recall
LR	91.00%	88.48%	93.18%
DT	65.47%	72.22%	64.84%
RF	79.73%	76.87%	82.63%
KNN	69.44%	77.36%	67.88%
SVM	93.61%	91.32%	95.70%

This table found that SVM produces the highest accuracy, precision, and recall score in contrast to other classifiers. SVM is the best of our used six classifiers. Whatever SVM efficiently detects fake news with a benchmark accuracy.

4-2- Further Explanation

Accuracy proves whether it is true or false. Our experimental results showed the highest accuracy when producing SVM and found lower accuracy when producing DT. LR accuracy and SVM accuracy are close to 100%. RF, DT, and KNN have an accuracy lower than 90%. Precisions are also applied; precisions provide a positive prediction result, and SVM is again higher than all of the classifiers. All the classifiers except SVM have a lower precision score than 90%. At last, recall is used, which is the sum of correctly classified instances outside of the true class; again, SVM got the highest score. All the classifiers except LR and SVM have more than 90% recall. From accuracy, precision, and recall scores, we saw that SVM has the best result among all the classifiers.

4-3- Comparison

Table 6 mentions a Comparison between recently published papers and our research paper's accuracy. Our research paper's accuracy is attached in the last row.

Table 6. Comparison between our accuracy and others

Authors and Number of models/Classifiers/ Algorithms Used	Dataset	Using Models/ Classifiers	Accuracy
Arif et al. (2022) used 3 models to get accuracy [21].	Accuracies of 1st dataset	PAC, Bi-LSTM, ROBERTA	0.51, 0.52, 0.47
	Accuracies of 2nd dataset	PAC, Bi-LSTM, ROBERTA	-, 0.61, 0.28
Alhkami et al. (2022) used 5 classifiers to get accuracy [22].	Accuracies of COVID-19 Fake News dataset	DT, KNN, LR, NB (Naïve base), SVM	0.74, 0.76, 0.79, 0.75, 0.71, 0.80
	Constraint@AAAI 2021 Dataset	DT, KNN, LR, NB, SVM	0.87, 0.76, 0.91, 0.81, 0.88, 0.65
Zarate and Tovar (2022) Used 5 classifiers to get accuracy [23]	Used their dataset	NB, LR, MLP (Multilayer perceptron), SVM, PA (Passive aggressive)	0.782, 0.756, 0.716, 0.760, 0.796
6 classifiers were used by us to get accuracy.	News dataset	LR, DT, RF, KNN, SVM	0.91, 0.6547, 0.7973, 0.6944, 0.9361

From this table, it can be seen that the present research classifier's SVM result is higher than the recent works written above, and the SVM accuracy is 0.9361, which is a benchmark accuracy.

5- Conclusion

Fake news is spreading within a nanosecond through online platforms with lots of effects, which is why detection is an urgent need to get rid of its effects. From others' work, it is also clear that fake news has lots of effects, and machine learning is very important to detect fake news. These statements are supported by this paper. Therefore, having an automated algorithm to recognize fake news with satisfactory accuracy is a crucial need. Even though there were some existing works in this domain, the level of accuracy was not satisfactory. Hence, we have proposed an advanced approach to detecting real and fake news by choosing the best classifiers among Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors, and Support Vector Machines. Among all the classifiers, this proposed approach achieved benchmark accuracy and was an efficient way to detect fake and real news using a Support vector machine. In essence, those five machine learning algorithms are utilized among the ensemble learners, who have consistently outperformed the former in all performance criteria. However, our experiment used a dataset collected from Kaggle, which helped us achieve 93.61% benchmark accuracy by using a Support Vector Machine. That is why a human can detect fake or real news when using a Support Vector Machine, and then the effect of fake news is cured. Whatever the case, this research has some limitations, which will be addressed in future works. In this research, feature selection has not been applied. Improved performance will be achieved by reducing features. One more limitation is that no deep learning models are used, which gives the best performance. Limitations will be solved in our future work.

6- Declarations

6-1- Author Contributions

Conceptualization, N.F. and I.H.; methodology, N.F.; software, N.F.; validation, N.F., K.M.S.S., and I.J.M.; formal analysis, N.F.; investigation, N.F.; resources, K.O.M.G.; data curation, N.F. and A.H.A.; writing—original draft preparation, N.F.; writing—review and editing, T.C.; visualization, N.F.; supervision, I.H.; project administration, N.F.; funding acquisition, K.O.M.G. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/code/ahmedxmahmoud/fake-news-detection/input

6-3- Funding and Acknowledgements

For providing the dataset, the authors are grateful to Kaggle. However, this project is partially supported by Faculty of Information Science & Technology (FIST), Multimedia University (also known as UNIVERSITI TELEKOM SDN. BHD.) and TM R&D Fund (Grant no. MMUE/220023 or RDTC/221054) in collaboration with the Faculty of Science & Technology, American International University-Bangladesh, AIUB.

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Buzea, M. C., Trausan-Matu, S., & Rebedea, T. (2022). Automatic Fake News Detection for Romanian Online News. Information (Switzerland), 13(3), 151. doi:10.3390/info13030151.

[2] Adiba, F. I., Islam, T., Kaiser, M. S., Mahmud, M., & Rahman, M. A. (2020). Effect of corpora on classification of fake news using naive Bayes classifier. International Journal of Automation, Artificial Intelligence and Machine Learning, 1(1), 80-92.

[3] Stewart, E. (2021). Detecting Fake News: Two Problems for Content Moderation. Philosophy & Technology, 34(4), 923–940. doi:10.1007/s13347-021-00442-x.

[4] Alnazzawi, N., Alsaedi, N., Alharbi, F., & Alaswad, N. (2022). Using Social Media to Detect Fake News Information Related to Product Marketing: The FakeAds Corpus. Data, 7(4), 44. doi:10.3390/data7040044.

[5] Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. Social Science and Medicine, 240, 112552. doi:10.1016/j.socscimed.2019.112552.

[6] Zafarani, R., Zhou, X., Shu, K., & Liu, H. (2019). Fake news research: Theories, detection strategies, and open problems. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 3207–3208. doi:10.1145/3292500.3332287.

[7] Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1), 1–4. doi:10.1002/pra2.2015.145052010082.

[8] Ahmed, H., Traore, I., & Saad, S. (2017). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, ISDDC 2017, Lecture Notes in Computer Science, vol 10618. Springer, Cham, Switzerland. doi:10.1007/978-3-319-69155-8_9.

[9] Thaher, T., Saheb, M., Turabieh, H., & Chantar, H. (2021). Intelligent detection of false information in Arabic tweets utilizing hybrid Harris Hawks based feature selection and machine learning models. Symmetry, 13(4), 556. doi:10.3390/sym13040556.

[10] Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). doi:10.18653/v1/p17-2067.

[11] Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2017). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task. arXiv preprint arXiv:1707.03264. doi:10.48550/arXiv.1707.03264.

[12] Zhang, C., Gupta, A., Kauten, C., Deokar, A. V., & Qin, X. (2019). Detecting fake news for reducing misinformation risks using analytics approaches. European Journal of Operational Research, 279(3), 1036–1052. doi:10.1016/j.ejor.2019.06.022.

[13] Awan, M. J., Yasin, A., Nobanee, H., Ali, A. A., Shahzad, Z., Nabeel, M., Zain, A. M., & Shahzad, H. M. F. (2021). Fake news data exploration and analytics. Electronics (Switzerland), 10(19), 2326. doi:10.3390/electronics10192326.

[14] Meel, P., & Vishwakarma, D. K. (2021). A temporal ensembling based semi-supervised ConvNet for the detection of fake news articles. Expert Systems with Applications, 177(5), 115002. doi:10.1016/j.eswa.2021.115002.

[15] Olan, F., Jayawickrama, U., Arakpogun, E. O., Suklan, J., & Liu, S. (2022). Fake news on Social Media: the Impact on Society. In Information Systems Frontiers. Information Systems Frontiers. doi:10.1007/s10796-022-10242-z.

[16] Liao, H. P., & Wang, J. L. (2023). The impact of epidemic information on the public's worries and attitude toward epidemic prevention measures during the COVID-19 outbreak. Current Psychology, 42(1), 145–153. doi:10.1007/s12144-021-01364-9.

[17] Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. Computers in Human Behavior, 116, 106633. doi:10.1016/j.chb.2020.106633.

[18] Hamdan, Y. B. (2020). Faultless decision making for false information in online: a systematic approach. Journal of Soft Computing Paradigm (JSCP), 2(04), 226-235. doi:10.36548/jscp.2020.4.004.

[19] Mohseni, S., Ragan, E., & Hu, X. (2019). Open issues in combating fake news: Interpretability as an opportunity. arXiv preprint arXiv:1904.03016. doi:10.48550/arXiv.1904.03016.

[20] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. doi:10.1145/3137597.3137600.

[21] Arif, M., Tonja, A. L., Ameer, I., Kolesnikova, O., Gelbukh, A., Sidorov, G., & Meque, A. G. M. (2022). CIC at CheckThat! 2022: multi-class and cross-lingual fake news detection. Working Notes of CLEF, CLEF 2022: Conference and Labs of the Evaluation Forum, 5–8 September, 2022, Bologna, Italy.

[22] Alhakami, H., Alhakami, W., Baz, A., Faizan, M., Khan, M. W., & Agrawal, A. (2022). Evaluating Intelligent Methods for Detecting COVID-19 Fake News on Social Media Platforms. Electronics (Switzerland), 11(15), 2417. doi:10.3390/electronics11152417.

[23] Zarate, C. A. J., & Tovar, L. A. N. (2022). Proposal for a Model for Detecting Fake News on Social Media in Mexico. International Conferences e-Society 2022 and Mobile Learning 2022, 12-14 March, 2022.

[24] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. arXiv preprint arXiv:1708.07104. doi:10.48550/arXiv.1708.07104

[25] Mitchell, T. M. (2006). The discipline of machine learning. Carnegie Mellon University, School of Computer Science, Machine Learning Department, Pittsburgh, United States.

[26] Mahmoud, A. (2022). News.csv. Kaggle, San Francisco, United States. Available online: https://www.kaggle.com/datasets/antonioskokiantonis/newscsv (accessed on April 2023).

[27] Mahmud, M., Kaiser, M. S., McGinnity, T. M., & Hussain, A. (2021). Deep Learning in Mining Biological Data. Cognitive Computation, 13(1), 1–33. doi:10.1007/s12559-020-09773-x.

[28] Ghafoor, H.Y., Jaffar, A., Jahangir, R., Iqbal, M.W., & Abbas, M.Z. (2022). Fake News Identification on Social Media Using Machine Learning Techniques. Proceedings of International Conference on Information Technology and Applications. Lecture Notes in Networks and Systems, vol 350, Springer, Singapore. doi:10.1007/978-981-16-7618-5_8.

[29] Zhang, K., Wang, W., Chen, L., Liu, Y., Hu, J., Guo, F., Tian, W., Wang, Y., & Xue, F. (2020). Cross-validation of genes potentially associated with neoadjuvant chemotherapy and platinum-based chemoresistance in epithelial ovarian carcinoma. Oncology Reports, 44(3), 909–926. doi:10.3892/or.2020.7668.

[30] Jain, A., Shakya, A., Khatter, H., & Gupta, A. K. (2019). A smart System for Fake News Detection Using Machine Learning. 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India. doi:10.1109/icict46931.2019.8977659.

[31] Khan, J. Y., Khondaker, M. T. I., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100032. doi:10.1016/j.mlwa.2021.100032.

[32] Andrade, J. J., Da Fonseca, L. G., Farage, M., & Marques, G. L. de O. (2020). Prediction of the Performance of Bituminous Mixes Using Adaptive Neuro-Fuzzy Inference Systems. Revista Mundi Engenharia, Tecnologia e Gestão, 5(6), 1-14. doi:10.21575/25254782rmetg2020vol5n61367.