algorithms

Article

Hybrid RFSVM: Hybridization of SVM and Random Forest Models for Detection of Fake News

Deepali Goyal Dev [1],* and Vishal Bhatnagar [2]

[1] GGSIPU, NSUT East Campus (Formerly Ambedkar Institute of Advanced Communication Technologies and Research), New Delhi 110031, India
[2] NSUT East Campus (Formerly Ambedkar Institute of Advanced Communication Technologies and Research), New Delhi 110031, India
* Correspondence: ddepali@gmail.com

Abstract: The creation and spreading of fake information can be carried out very easily through the internet community. This pervasive escalation of fake news and rumors has an extremely adverse effect on the nation and society. Detecting fake news on the social web is an emerging topic in research today. In this research, the authors review various characteristics of fake news and identify research gaps. In this research, the fake news dataset is modeled and tokenized by applying term frequency and inverse document frequency (TFIDF). Several machine-learning classification approaches are used to compute evaluation metrics. The authors proposed hybridizing SVMs and RF classification algorithms for improved accuracy, precision, recall, and F1-score. The authors also show the comparative analysis of different types of news categories using various machine-learning models and compare the performance of the hybrid RFSVM. Comparative studies of hybrid RFSVM with different algorithms such as Random Forest (RF), naïve Bayes (NB), SVMs, and XGBoost have shown better results of around 8% to 16% in terms of accuracy, precision, recall, and F1-score.

Keywords: hoax information; fake news; rumors; social media; random forest; support vector machine

check for updates

Citation: Dev, D.G.; Bhatnagar, V. Hybrid RFSVM: Hybridization of SVM and Random Forest Models for Detection of Fake News. *Algorithms* **2024**, *17*, 459. https://doi.org/10.3390/a17100459

Academic Editor: Ioannis Tsoulos

Received: 11 August 2024
Revised: 3 October 2024
Accepted: 10 October 2024
Published: 16 October 2024

1. Introduction

More people spend time communicating online and consuming news from web media, preferably press agencies. There has been a change in adoption behavior, as it is inexpensive and less time is taken to adopt news or information from media platforms compared to adopting news from press media like newspapers or television. There are various related concepts of information that are distinguished on the strength of various characteristics such as legitimacy or authenticity, intention, and whether it is news or not. Different concepts of information can also be misleading, deceiving, rumor, fake news, or malicious fake news. The authors have presented a different perspective for this study, based on style and on propagation.

1.1. Knowledge-Based Study

This study aims to analyze and detect fake news using a fact-checking process. In the process of fact checking, knowledge is extracted from verified news content to check the news authenticity. Expert-based fact-checking websites involve well-known websites that present statistics on the authenticity of topics, and this information can help in further scrutinization for verification purposes. Table 1 [1] shows a review of expert-based websites. For example, HaoxSlayer focuses on the authenticity of information and categorizes articles and information into hoaxes, junk e-mails, and false news.

Table 1. Expert-based fact-checking websites.

	Topics Covered	**Content Analysed**
PolitiFact	American Politics	Statements
FactCheck	American Politics	TV ads, Debates, Speeches, Interviews and News
Snopes	Politics and other Social Issues	News Articles and Videos
TruthorFiction	Politics, Religion, Nature, Food and Medical	Email Rumours
HoaxSlayer	Ambiguity	Articles and Messages
FullFact	Economy, Health, Education, Crime, Immigration, Law	Articles

Figure 1 shows the automatic fact-checking process. This process is categorized into two parts: knowledge base construction extraction and comparison checking. In the first part, raw facts and data are represented by the knowledge that is extracted from the web. A knowledge graph is constructed with the help of extracted knowledge. Redundancy reduction, invalidity reduction, conflict resolution, credibility improvement, and completeness enhancement are carried out using knowledge extracted from the web. In the fact-checking process, a comparison is carried out between the knowledge extracted from the knowledge base and the news content that is to be verified to check the authenticity of the news [1].

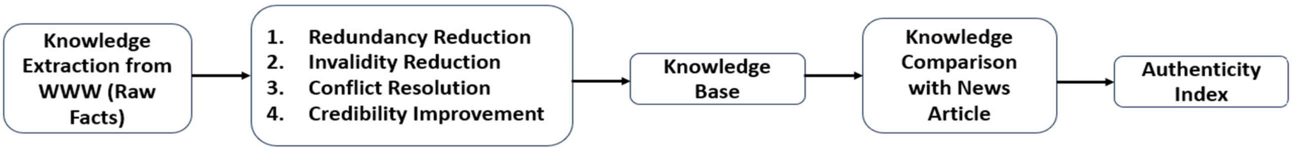
Knowledge Extraction from WWW (Raw Facts)
1. Redundancy Reduction
2. Invalidity Reduction
3. Conflict Resolution
4. Credibility Improvement
Knowledge Base
Knowledge Comparison with News Article
Authenticity Index

Figure 1. Automatic fact-checking process.

1.2. Style-Based Study

This study determines if the intention of news is whether to deceive the public or not. The style is a characteristic that represents and differentiates fake news from veracity. Deception analysis investigates how the style of misleading content is written across various kinds of information. Features or characteristics based on attributes require some supplementary level of calibration or computing, which is time-consuming, but connects to the greater significance of the evaluation of characteristics based on attributes and filtering for the detection of misleading content.

1.3. Propagation-Based Study of Fake News

This study gives information about the spreading of rumors and the process of spreading it. In this propagation-based study, we deal with the following:

a. How are the propagation patterns of false news represented?
b. What will be the measuring parameters for the characterization of dissemination of false news?
c. How do we differentiate the dissemination of fake news from news that is verified?
d. How do we analyze the pattern of fake news in various domains like politics, economy, and education?
e. How does fake news propagate differently for topics like presidential elections and health, for various platforms like Instagram, Facebook, and X (Twitter), in different languages like English, Hindi, Mandarin Chinese, and Spanish?

The sections that make up this document are as follows: The research methods, motives, problem description, and objectives deduced from the literature review and the proposed architecture and proposed algorithm for measuring various parameters are discussed in Section 2. A summary of the process implementation and outcomes is given in Section 3. Conclusions in Section 4 along with the scope of future work.

2. Related Work and Motivation

Due to the dynamicity and extremely complicated and varied online data on media platforms, finding and analyzing reliable information is a significant problem.

Ref. [1] proposed a method for rumor identification at an early stage. For the purpose of early rumor identification, these authors integrated Reinforcement Learning (CM) with a recurrent neural network and took into account datasets from X (Twitter) and Weibo. The increased performance of 93.3% and 85.8% is on par with modern rumor detection techniques. To identify fake news, ref. [2] employed a basic Bag of Words vector, Continuous Bag of Words, and Skip-gram. The authors employed various machine-learning algorithms for classification, and 95.49 percent accuracy was achieved by combining text-based characteristics and stylometric features. Using a deep neural network, ref. [3] was able to find bogus news quickly. The accuracy of the suggested method was roughly 90%. Recurrent neural networks were used by [4] to identify news rumors. The authors evaluated the performance of various evaluation parameters. The datasets were created using real-time tweets from PHEME. Ref. [5] discussed various challenges in the difficulty in obtaining high-quality labeled datasets for online false news detection, and the difficulty in predicting fake news in advance. In order to combat false news, ref. [6] advocated for the identification of fake news based on content, user input, and intervention. The authors also talked about the difficulties in developing quantitative techniques for analyzing fake news. The author of this study also discussed several research topics, including dynamic knowledge bases, fresh intervention techniques, and datasets for intent detection.

Ref. [7] suggested several methods for detecting rumors, as well as determining the right datasets to use for these tasks. For the purpose of identifying false news in online web media, ref. [8] applied 23 supervised classification algorithms to a structured news dataset. Sequential Minimal Optimization surpassed the others in accuracy and F-score. In the future, supervised algorithms may be hybridized and combined for improved outcomes. Ref. [9] suggested propagation-based methods that are resilient to attacks and independent of language. The topic of numerous study gaps, difficulties, and potential future areas of information pollution in social media was covered in [3].

A false news extraction and detection system was put forth by [10] using text present in images. The existing system does not address local news, and the text present in the image is a significant problem because of the shadowing effect. In order to validate the accuracy of the text in the image, the text was extracted from images and the reality parameter was calculated. In 2018, ref. [11] published a survey paper on the identification of rumors. The authors covered a variety of research topics, including timely identification of rumors, investigation of fake news based on deep learning techniques, investigation of fake news across multiple domains, and identification of auditable content. Ref. [12] worked with artificial intelligence to propose four key parameters for accepting fake news and to defeat and protect against these parameters. Fake news detection is a very new area for artificial intelligence research and implementation. Ref. [13] identified the roots and historical patterns of social media misinformation.

Approximately 300,000 tweets can be retrieved from X (Twitter). A linguistic approach can be considered for detection, to distinguish the type of news in different news articles. Ref. [14] introduced two new datasets by crowdsourcing to develop an automatic recognition system. Ref. [15] proposed a new dataset, called LIAR, to detect fake news, and a novel hybrid CNN was developed to integrate metadata with text. This dataset is used for policy research on perspective classification and argument mining. Ref. [16] described a CSI model that was collected, scored, and integrated to detect fake news. The RNN was used to record user temporal patterns in the first module, the next module is used to analyze user behavior, and the other module was used to classify fake news and integrated the first and second modules. The authors also addressed research questions focused on the concepts of reinforcement learning and crowdsourcing in models. Ref. [17] discussed various research gaps and possible research directions in the area of fake news characterization and detection on media platforms.

Ref. [18] conducted the hybridization of RF and SVM models which outperforms the performance of individual models during the categorization of positive or negative sentiment reviews, during the identification of product reviews offered by the Amazon datasets. Considering the datasets of 500 Amazon products, the individual model's accuracy ranges around 80% to 82%, which further increased to 86% during the hybridization. CNN and SVMs were used together as a hybrid approach in [19], in order to detect and classify different kinds of orange diseases. In the blended model, different disorders were identified such as Penicillium, Scab, Melanose, or citrus canker. CNN used for derivation of features and SVMs for the further classification enhances the accuracy of the hybrid model up to 88%, resulting in a significant improvement with respect to using the models separately.

Looking at the continued improvement in results due to the hybridizations of different deep learning models in terms of product reviews or categorization of diseases in oranges, researchers further explored this process in alignment to the specific topic. Ref. [20] proposed the hybridization of CNN and RNN models for the classification of fake news. Further, the model was investigated and validated on different datasets (FA-KES and ISO), achieving excellent results in terms of accuracy, recall, and precision with respect to the baseline model evaluations.

During the analysis behind the need for hybrid models, many researchers came to conclusion that in baseline models, it is very difficult to cater the rapidly changing strategies used by the people who create and share false information, as language and compositional analysis is insufficient and inappropriate. Ref. [21] derived a hybrid approach of using SVM and KNN models to overcome this problem, which successfully identified bogus news. In order to further ensure the robustness and generalization of the results, a cross-validation process must be applied as this prioritizes the complex relationships of models. This holistic approach provides a complex display of different attributes comprising the behavior of users and the dynamics of social media posts.

The continuation, evolution, and impact of cross validation models emerges at large, in terms of different evaluation parameters, in comparison to the baseline models. Hybridized models of CNN and SVMs, RNN and SVMs, or different models, was carried out as mentioned in the literature. During the study, it was observed that hybridization impacted the results at large in different scenarios and on different evaluation parameters in comparison to sole models. Accuracy of fake news classification reached around 88% [20,21]. Our proposed approach derived excellent results in comparison to the mentioned literature and with improvements in different evaluation parameters, varying from 4% up to 10% in different scenarios. Our hybrid approach involved a categorization approach of different outliers at each layer to combat the pervasive spread of misinformation in our rapidly growing digital era. Our model analyzes the highly dimensional data of different attributes, further dividing it into news categories, which results in a tree-based structure that brings optimality, and further enhancing optimality through the binary classification of the datasets, which is very well supported through SVMs.

Fake information can be created and disseminated very easily through media platforms, resulting in real-world impacts. Accurate and trustworthy information detection is a major issue for online communication. Data are very complex and diverse, so a thorough analysis of information pollution is required in the digital world to mitigate harmful societal impacts. So, to summarize the problem from a survey of various papers carried out by different authors: an effective algorithm which can identify fake news on various social media platforms is required, and therefore proposed in this study. The research technique used in this study is displayed in Figure 2.

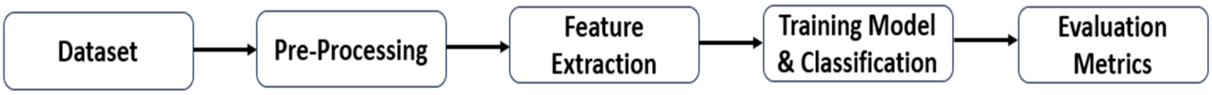

Figure 2. Research methodology.

2.1. Proposed Framework

Figure 3 shows the various steps, such as data acquisition, data labeling, definition of the feature set, classification, and calculation of evaluation parameters.

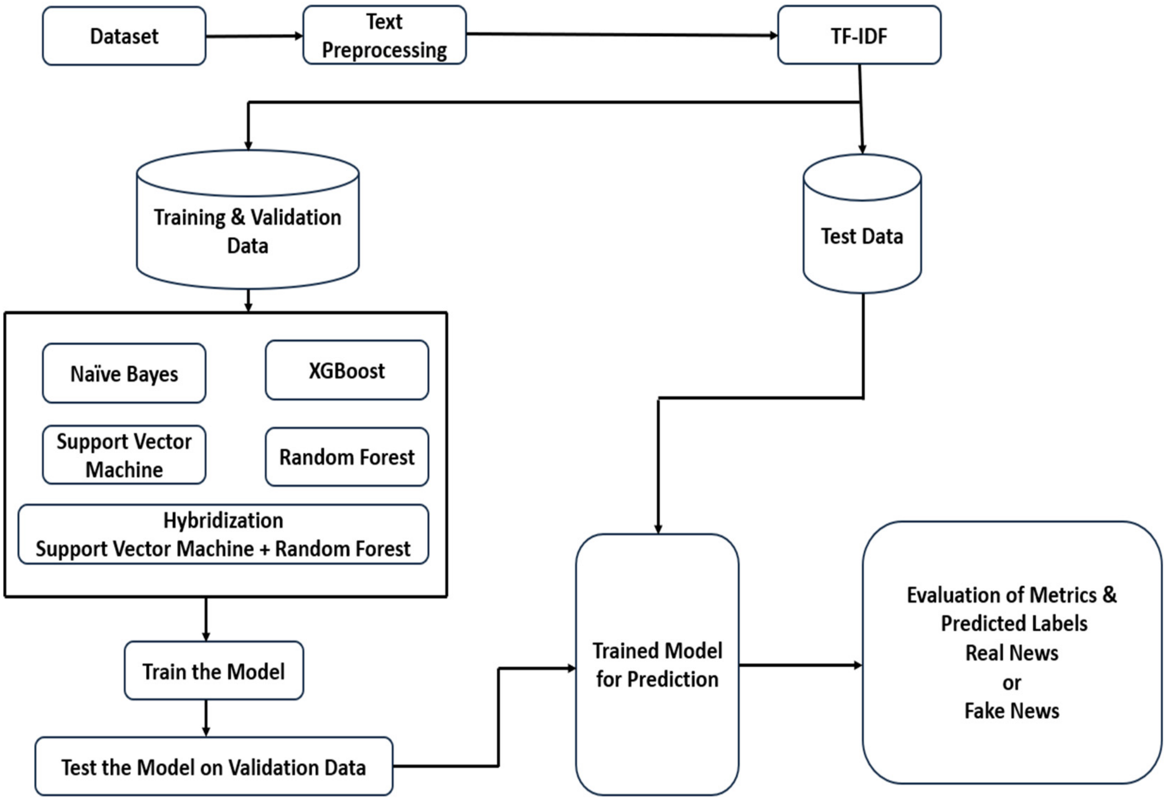
Dataset
Text Preprocessing
TF-IDF
Training & Validation Data
Test Data
Naïve Bayes
XGBoost
Support Vector Machine
Random Forest
Hybridization
Support Vector Machine + Random Forest
Train the Model
Test the Model on Validation Data
Trained Model for Prediction
Evaluation of Metrics & Predicted Labels
Real News
or
Fake News

Figure 3. Proposed framework.

- Data Acquisition: Authors acquired the dataset about fake news on COVID-19 on X (Twitter) from dataworld.com.
- Data Labelling: Data containing tweets from X (Twitter). After cleaning the records, we labeled them 1 (fake news) and 0 (true news). The training set contains 60% data, and the test set contains 40% data containing fake news related to COVID-19.
- Defining the Feature Set: The set of features that authors used consisted of a TFIDF feature vector [22].
- Classifier: The authors considered naïve Bayes [23], Support Vector Machines [24], Random Forest [25], XGBClassifier, and hybrid SVMs and Random Forest.

The naïve Bayes (NB) algorithm, which makes use of the explicit premise that enables a dependency of one another and Bayes's rule, is a simple-to-learn probabilistic algorithm. NB calculates posterior probabilities $P(y|x)$ for each classification for a specific object x based on training data. Applications involving categorization can make use of estimates. NB seems to be a workable solution in many actual applications due to its computational efficiency and many other desirable qualities [26].

The foundation of Support Vector Machines (SVMs) is the notion of learning statistics. Here, SVMs have been used by numerous academics in data categorization and pattern recognition applications. Theoretically, the SVM concept can be described as follows. (1) A notion that indicates the degree of risk or chance of learning error is known as structural risk minimization. The decision-making function is established by the SVM learning process to reduce error rates. The fundamental ideas underlying vector machine technology are core capabilities, in order to generate a nonlinear decision function on the data in the preceding space. (2) SVM learning involves dividing the data into two groups and finding the level with the largest margin that can solve the overfitting problem [27].

One of the strongest algorithms for classification jobs is Random Forest (RF). Large datasets can be accurately and precisely classified using RF. Each tree in the RF system uses random vector values and functions as a classifier. RF creates decision trees during training, forecasts the results of each tree using training data from bootstrap examples, and attributes the selection at random during tree induction. Decision trees are averaged or combined to create predictions using majority voting [28].

A recursive partition of a data space serves as the XGBoost representation of its tree-like model. Highly recognizable nodes that form a rooted tree make up a decision tree. The root node of the tree at the top has no outgoing branches, connections, or edges. For every other node, there is one incoming edge. Intermediate nodes are marked with outgoing edges. Nodes are at the leaves and decision nodes at the lowest level [29].

- Evaluation Parameters for Fake News Detection: Various evaluation parameters are used by the author in the manuscript, as follows.

$$Precision = \frac{|True\ Positive|}{|True\ Positive| + |False\ Positive|} \tag{1}$$

$$Recall = \frac{|True\ Positive|}{|True\ Positive| + |False\ Negative|} \tag{2}$$

$$Accuracy = \frac{|True\ Positive| + |True\ Negative|}{|True\ Positive| + |True\ Negative| + |False\ Positive| + |False\ Negative|} \tag{3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Percision + Recall} \tag{4}$$

2.2. Proposed Algorithm

SVMs are used for binary classification, extending the work to multi-class classifiers by the researchers, as well as working well for the detection of rumors, especially X (Twitter). This algorithm performs preprocessing for the tweets by converting the jargon words used into the standard forms. Further, regular expressions converted the redundant letter words into original words along with the segmentation of words. During the extraction process of the features, structural, user, and other content about the tweet are also considered.

Random Forest is the strongest classification algorithm which classifies large datasets more accurately and precisely. RF is the tree structure, where each tree possesses random vector values and functions as a classifier. This creates various decision trees during the training of datasets, and uses the same for forecasting the data of each individual tree and selection of attributes randomly. Further, RF will result in higher accuracy and robustness by the segregation of multiple trees. Due to the training of multiple trees, variance and standard deviation of each individual tree is reduced, which further enhances the performance. RF removes overfitting and further improves generalizability. For example: consider that the number of trees used is 100, due to a smaller number of trees leading to the over-fitting of the data. In this instance, the random state value is set to 42, which makes the algorithm deterministic, losing its nature of non-determinism. This means that the algorithm will derive the same outcome again and again, and this kind of consistency will further maintain transparency and use in testing.

The basic reason behind the hybridization of RF and SVMs is to predict different types of highly dimensional fake news datasets. In this, we divide the datasets of different news categories into multiple classes where the tree-based approach will bring optimality and further reduce the binary classifications which will be further supported in an optimized way by the SVMs. Random Forest results are exemplary when the news categories provide data with a mixture of numerical and categorical features, along with features on a large scale. SVMs maximize the margins and rely on the concepts of distance between two different points. Furthermore, min-max and different scaling processes are used during the pre-processing step. In the proposed work, the larger the size of the datasets in future research, the more optimized the results of this hybridization model will be, so we started the research with small datasets in order to understand the different types of computational trade-offs which can be resolved in the future. Due to the cross validation of the machine-learning models, there is significant improvement in precision, accuracy, recall, and F1-score in comparison to individual models, as tested on datasets of different types of news categories. The hybridization of the RF and SVM models improves the execution time, that is, the algorithm runs fast in comparison to that of the individual performances, improves interpretability, and also nonlinear dependencies are catered to in the best possible way. In addition to this, as we increase the size of the datasets to greater than 10,000 samples in future work, cross validation of both of the models will produce exemplary results.

The process used for the development of the hybrid RFSVM model is as follows:

Step 1: Identify the spreading of fake news, especially on X (Twitter). In addition to this, try to select the data source containing fake news content, which can further be used for the collection of data and its processing.

Step 2: After collection, preprocessing and cleaning will be carried out. Initially, conversion of the datasets from json format to csv format is carried out, containing user tweets along with the reactions against each tweet. The combined datasets will be loaded into the data frame using pandas.

Step 3: In this step, extraction of text and target variable columns from the dataset will be carried out, which is further used for the removal of tags and symbols. Further, long sentences are split into words. Also, stop words will be removed from the text.

Step 4: This step is used for creating bags of words from the cleaned datasets, and to encode the characters of strings with numerical values, using encoding algorithms.

Step 5: This step is used for the feature selection, extraction, and normalization processes. First, the conversion of the bags of words into vectors is carried out using TF-IDF, which can further be normalized using the min-max algorithm. Then follows the division of the dataset into the training and testing datasets which can further calculate the use of each feature using an entropy-based model of feature selection and selecting the best and optimized feature using the algorithm mentioned.

Step 6: Building the base models such as RF, SVMs, naïve Bayes, and XGBoost individually for the training the datasets. All the parameters are further analyzed individually on different news categories. Further, the hybridization of RF and SVMs is carried out using the mentioned algorithm, in which the analysis of different news categories is carried out along different parameters such as precision, recall, and accuracy.

(a) For each tree in the forest, simply sample n data points. Further, for each node in a tree, we will randomly select m attributes by calculating variance. This leads to the new dataset d′ from the dataset d using the random replacement method along with the assignments of weights against each attribute.

(b) For each subset having a random feature of dataset d′, apply SVMs for each feature subset and generate the output of the classification as one class of SVM, which can further be used to update the weights of all the vectors on the basis of the outcomes of classification. The weights of the vectors can further be increased in case of misclassification and decrease in other cases.

(c) Repeat step (a) and (b), by generating different random datasets till all input vectors are further classified.

(d) The output of the complete dataset is computed using the majority voting process, from the final outputs of each of the random feature subsets Di. Proposed Algorithm is shown in Algorithm 1.

Algorithm 1: Hybrid RFSVM Fake News Detection Algorithm.

```
Input: Dataset
Output: Classification of news as fake/real & evaluation metrics
// Phase 1: Data Set Creation //
                  Download the Dataset
                  For each data belonging to the dataset do
                  Data pre-processing
                  return Processed Data
// Phase 2: Fake Dataset Text Augmentation //
                  For each data in the Dataset
                  FW ← extract (Bag of words)
                  MS ← cosine similarity (FW, FW)
                   If max (MS)
                   AFT ← combine fake text // (Augmented Fake Text)
// Phase 3: Text Classification //
// For Machine Learning:
          tfidf (t, x, X) ← tf (t, x).idf (t, X)
          tf (t, x) ← Log(1+freq (t, x))
          idf (t, X) ← Log (N/count d belongs to D, t belongs to d)
          TF, IDF ← tfidf (Dataset) // Feature Extraction //
         Accuracy ← Random Forest + SVMs
```

Here, FW = fake words, MS = matching score, AFT = augmented fake text, t = term, and D = document.

3. Implementation and Results

The authors implemented the project in Python, and considered data on fake news about COVID-19 from X (Twitter). All packages like XGBClassifier and Random Forest Classifier were imported from the sklearn library in Python. In total, 60% of the data were training data and 40% of the data were testing data. The dataset was read using the read_csv method using the panda package. Authors considered TFIDF [22] as a feature set. TF computes the weight and the frequency of every term occurring in a document.

$$TF(w,j) = \frac{Frequency\ of\ w\ in\ document}{Total\ number\ of\ w\ in\ document} \tag{5}$$

IDF computes the importance of each term w. In sklearn, IDF(t) is

$$IDF\ (t) = Log\,\frac{1+n}{1+df(t)} + 1 \tag{6}$$

The evaluation parameters were determined using naïve Bayes, XGBoost, Support Vector Machines, and Random Forest. For text classification issues, naïve Bayes classifiers are a quick and effective option. They can easily be trained on small datasets. Support Vector Machines tend to give more accurate results on concise datasets and handle high-dimensional spaces efficiently. Random Forest works on large datasets efficiently and produces very accurate results. The XGBoost classifier is scalable to large datasets and produces optimized and efficient computational performance [30]. It also handles sparse data. Table 2 shows the abbreviations used for machine-learning classifiers.

Table 2. Abbreviations used for classifiers.

Support Vector Machine + Random Forest	RFSVM
Support Vector Machine	SVM
Random Forest	RF
XGBoost	XGB
Naïve Bayes	NB

Table 3 signifies the performance of various parameters using classification algorithms like naïve Bayes [26], Support Vector Machines [24], Random Forest [25], XGBClassifier, and hybrid algorithm of SVM and Random Forest. The result predicts that the hybrid RFSVM outperformed in our dataset.

Table 3. Test performance.

Model	Accuracy	Precision	Recall	F1-score
Naïve Bayes	96	85	84.18	84.58
SVM	96.75	86.91	86.63	86.76
Random Forest	96.25	84.88	84.05	84.46
XGBoost	95.75	84.56	83.90	84.63
SVM + Random Forest	97.56	88.21	92.30	93.50

Figures 4 and 5 show the accuracy % and precision % for five different classification algorithms, respectively. Naïve Bayes has the lowest accuracy %, and hybrid SVMs and Random Forest have the highest accuracy %. SVMs have a precision of 86.91%, RF has 84.88%, and XGBoost has 84.56%. Figures 6 and 7 show the recall % and F1-score % of NB, RF, SVMs, XGBoost, and RFSVM, respectively. RFSVM has the highest recall at 92.30% and XGBoost has the lowest recall at 83.90%. RFSVM has the highest F1-score at 93.50%.

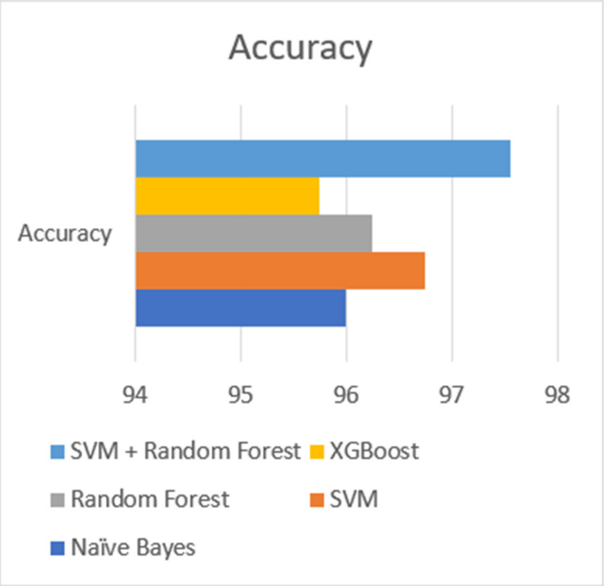

Figure 4. Accuracy % of artificial intelligence algorithms.

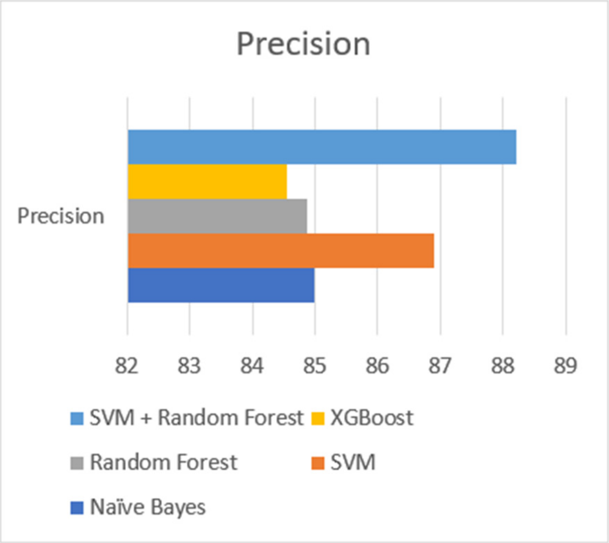

Figure 5. Precision % of artificial intelligence algorithms.

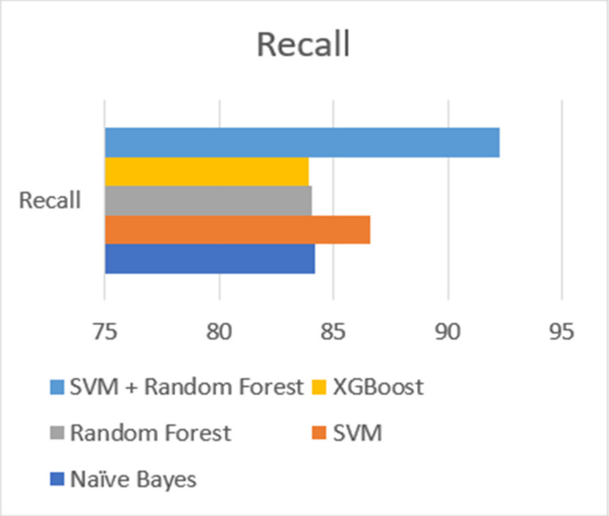

Figure 6. Recall % of artificial intelligence algorithms.

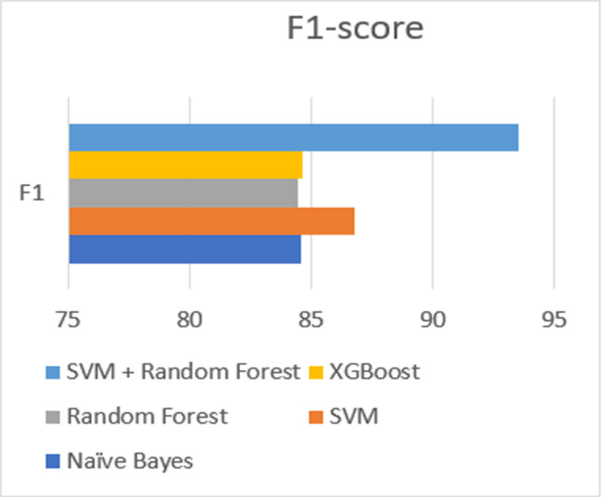

Figure 7. F1-score % of artificial intelligence algorithms.

The results have shown that the RFSVM excels in various parameters with an accuracy of 97.56%, precision of 88.21%, recall of 92.30%, and F1-score of 93.50%, compared to the results of naïve Bayes, Random Forest, SVMs and XGBoost. This means that 92.30% (maximum recall) of fake news was detected successfully with our proposed methodology. Table 4 shows the comparison among various machine-learning classification algorithms in terms of accuracy, precision, recall, and F1-score.

Table 4. Comparison of Performance Results.

Accuracy	RFSVM > SVM > RF > NB > XGB
Precision	RFSVM > SVM > RF > XGB > NB
Recall	RFSVM > SVM > RF > NB > XGB
F1-score	RFSVM > SVM > XGB > NB > RF

SVMs show better accuracy than naïve Bayes, Random Forest, and XGBoost, as naïve Bayes treats features as independent features and SVMs explore the relationship between independent features to a certain degree, and a nonlinear kernel is used as Gaussian. The output therefore varies depending on the features of problem statement interaction and prediction models. SVMs are better than naïve Bayes. The prediction function shows dependencies between variables that naïve Bayes (y (a, b) = ab) does not catch, so it is not a universal approximator. As SVMs utilize kernel trick and maximum margin principle to work better in nonlinear and high-dimensional tasks, SVMs are the best algorithm of all. It also benefits from the correct set of features and extraction/transformation techniques much of the time. Figure 8 shows the graphical comparison among various parameters.

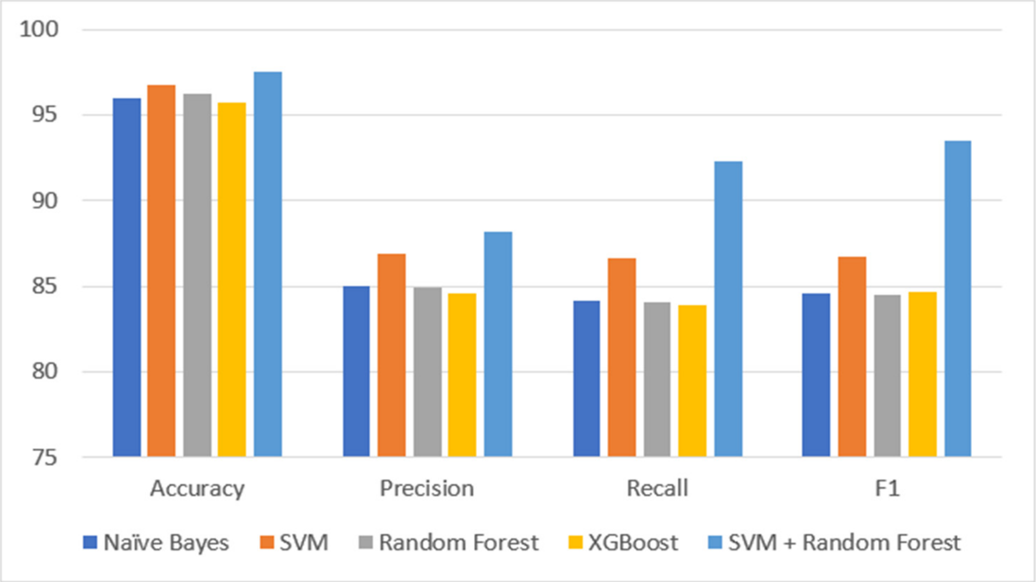

Figure 8. Comparison among evaluation parameters of different classifiers.

Table 5 shows that the hybrid approach of SVMs and Random Forest outperforms others with an accuracy of 0.9756. Different categories of fake news may be evaluated with five different models in terms of accuracy, precision, and F1-score.

Table 5. Comprehensive comparative analysis with baseline studies.

Reference	Classifier Used	Year	Accuracy	Precision	F1-score
[31]	NB	2020	0.60	0.59	0.72
	RF		0.59	0.62	0.67
	LR		0.65	0.69	0.75
	PAC		0.92	0.93	0.9257
[32]	XGBOOST	2020	0.75	-	-
	SVM		0.73	-	-
	RF		0.73	-	-
[33]	SVM	2021	0.8933	-	-
	DT		0.7333	-	-
	NB		0.8689	-	-
	LR		0.9046	-	-
	KNN		0.8998	-	-
Proposed Approach	RFSVM		0.9756	0.8821	0.9350

(a) Analysis of different types of news categories with five different models in terms of accuracy: naïve Bayes (NB), Random Forest (RF), XGBOOST, SVMs, and RFSVM. A comparative analysis of different models in terms of accuracy, with different types of news categories, is presented in Table 6.

Table 6. Accuracy values for different news categories with different ML algorithms.

Category	Avg. Accuracy [NB]	Avg. Accuracy [XGBoost]	Avg. Accuracy [RF]	Avg. Accuracy [SVMs]	Avg. Accuracy [RFSVM]	Reference [34]
Agriculture	0.87	0.89	0.91	0.95	0.97	0.95
Aviation	0.49	0.51	0.53	0.61	0.63	0.68
Sports	0.55	0.57	0.59	0.65	0.67	0.72
Roads	0.64	0.66	0.68	0.72	0.74	0.80
Residential	0.53	0.55	0.57	0.61	0.63	0.61
Forest	0.57	0.59	0.61	0.61	0.64	0.65
Village	0.42	0.44	0.46	0.53	0.58	0.57
Finance	0.87	0.89	0.91	0.95	0.97	0.95
Politics	0.87	0.89	0.91	0.91	0.93	0.91
Technology	0.42	0.44	0.46	0.53	0.55	0.53

RFSVM's combination of Support Vector Machine and Random Forest is around 10% better than the naïve model in terms of accuracy, which is around 6% to 8% better in terms of accuracy compared to other models such as XGBoost, Random Forest, and SVMs.

(b) Analysis of different types of news categories with five different models in terms of precision in Table 7:

RFSVM's combination of Support Vector Machines and Random Forest is around 8% better than the naïve model in terms of precision, which is around 4% to 6% better in terms of precision compared to other models such as XGBoost, Random Forest, and SVMs.

Table 7. Precision values for different news categories with different ML algorithms.

Category	Avg. Precision [NB]	Avg. Precision [XGBoost]	Avg. Precision [RF]	Avg. Precision [SVMs]	Avg. Precision [RFSVM]	Reference [34]
Agriculture	0.90	0.92	0.94	0.98	0.98	0.98
Aviation	0.51	0.53	0.55	0.63	0.70	0.71
Sports	0.57	0.59	0.61	0.67	0.76	0.74
Roads	0.67	0.69	0.71	0.74	0.81	0.82
Residential	0.55	0.57	0.59	0.63	0.65	0.63
Forest	0.59	0.61	0.63	0.63	0.65	0.67
Village	0.43	0.45	0.47	0.55	0.56	0.59
Finance	0.90	0.92	0.94	0.98	0.98	0.98
Politics	0.90	0.92	0.94	0.94	0.96	0.94
Technology	0.43	0.45	0.47	0.55	0.54	0.55

(c) Analysis of different types of news categories with five different models in terms of precision in Table 8:

Table 8. Recall values for different news categories with different ML algorithms.

Category	Avg. Recall [NB]	Avg. Recall [XGBoost]	Avg. Recall [RF]	Avg. Recall [SVMs]	Avg. Recall [RFSVM]	Reference [34]
Agriculture	0.43	0.45	0.47	0.51	0.62	0.61
Aviation	0.23	0.25	0.27	0.35	0.52	0.53
Sports	0.51	0.53	0.55	0.61	0.78	0.78
Roads	0.10	0.12	0.14	0.18	0.33	0.35
Residential	0.25	0.27	0.29	0.33	0.41	0.43
Forest	0.53	0.55	0.57	0.57	0.72	0.71
Village	0.33	0.35	0.37	0.45	0.58	0.59
Finance	0.53	0.55	0.57	0.61	0.69	0.71
Politics	0.53	0.55	0.57	0.57	0.67	0.67
Technology	0.38	0.40	0.42	0.50	0.58	0.60

RFSVM’s combination of Support Vector Machines and Random Forest is around 18% better than the naïve model in terms of recall, which is around 12% to 16% better in terms of recall compared to other models such as XGBoost, Random Forest, and SVMs.

(d) Analysis of different types of news categories with five different models in terms of precision in Table 9:

RFSVM’s combination of Support Vector Machines and Random Forest is around 15% better than the naïve model in terms of F1-scores, which is further around 8% to 12% better in terms of F1-scores compared to other models such as XGBoost, Random Forest, and SVMs.

Table 9. F1-score values for different news category with different ML algorithms.

Category	Avg. F1-Score [NB]	Avg. F1-score [XGBoost]	Avg. F1-Score [RF]	Avg. F1-Score [SVMs]	Avg. F1-Score [RFSVM]	Reference [34]
Agriculture	0.59	0.61	0.63	0.67	0.75	0.75
Aviation	0.32	0.34	0.36	0.45	0.59	0.61
Sports	0.54	0.56	0.58	0.64	0.77	0.76
Roads	0.19	0.21	0.23	0.28	0.34	0.49
Residential	0.35	0.37	0.39	0.43	0.51	0.51
Forest	0.56	0.58	0.60	0.60	0.72	0.69
Village	0.37	0.39	0.41	0.50	0.59	0.59
Finance	0.67	0.69	0.71	0.75	0.85	0.82
Politics	0.67	0.69	0.71	0.71	0.74	0.78
Technology	0.40	0.42	0.44	0.52	0.56	0.57

Cross validation of RF and SVMs provides a significant better result in terms of interpretability, time, and accuracy. Due to the better process of the training and feature extraction, accompanied by the hybridization of RF and SVMs, various improvements in accuracy, precision, recall, and F1-score across different news categories can be achieved. Currently, the hybrid model has been applied on small datasets in order to understand the better visualization of the data across different categories. Applying the hybrid algorithm on small datasets already provides brief insights into how the results are better, for different types of news categories, in comparison to individual machine-learning models. In our approach, currently RF and SVMs have been applied for comparisons within the various classes.

4. Conclusions

In the current social media era, fake news detection is a rapidly emerging topic. The literature surveyed here addresses various research gaps identified on the web and media platforms. The authors have surveyed, summarized, compared, and evaluated the ongoing research on fake news which includes various perspectives on fake news. Various evaluation parameters, such as precision, recall, accuracy, and F1-score have been calculated on different machine-learning classifiers by applying the TFIDF feature set. The authors have presented that hybrid RFSVM outperformed the best among others on the basis of different evaluation parameters. In our approach, currently RF is applied and further SVMs for comparisons within the class. In future, a hybridization of RF and SVMs will be applied on the large datasets along with the integration of more modern techniques such as Transformer based models which can further be used to strengthen the comparisons, which further improve the efficiency and effectiveness of the model. In our proposed model as size of the dataset is not to large, which enable us to provide effective, accurate and fast results using hybridization. Hence, modern techniques such as transformer-based model were not used to strengthen the comparisons. In future research, more focus can be on diversified and labeled datasets. Fake news detection at the initial stage, fake news detection in different languages in cross-platform, and hybridization of various intelligent algorithms can be carried out for better results, dynamic and benchmark datasets are major challenges in domain of false news identification that can be used for potential research opportunities. Images containing fake text is also most promising area as future research.

Author Contributions: Conceptualization, D.G.D. and V.B.; methodology, D.G.D. and V.B.; software, D.G.D. and V.B.; validation, D.G.D. and V.B.; formal analysis, D.G.D. and V.B.; investigation, D.G.D. and V.B.; resources, D.G.D. and V.B.; data curation, D.G.D. and V.B.; writing—original draft preparation, D.G.D. and V.B.; writing—review and editing, D.G.D. and V.B.; visualization, D.G.D. and V.B.; supervision, D.G.D. and V.B.; project administration, D.G.D. and V.B.; funding acquisition, D.G.D. and V.B. All authors have read and agreed to the published version of the manuscript.

Funding: All authors declare that there is no funding available for this article.

Data Availability Statement: Data are available for this article whenever required.

Acknowledgments: The authors are thankful to the learned reviewer for his valuable comments.

Conflicts of Interest: The authors declare that they have no competing interests.

References

1. Zhou, X.; Zafarani, R. Fake news: A survey of research, detection methods, and opportunities. *arXiv* **2018**, arXiv:1812.00315.
2. Reddy, H.; Raj, N.; Gala, M.; Basava, A. Text-mining-based Fake News Detection Using Ensemble Methods. *Int. J. Autom. Comput.* **2020**, *17*, 210–221. [CrossRef]
3. Liu, Y.; Wu, Y.F.B. FNED: A Deep Network for Fake News Early Detection on Social Media. *ACM TOIS* **2020**, *38*, 25. [CrossRef]
4. Alkhodair, S.A.; Ding, S.H.; Fung, B.C.; Liu, J. Detecting breaking news rumors of emerging topics in social media. *Inf. Process. Manag.* **2020**, *57*, 102018. [CrossRef]
5. Meel, P.; Vishwakarma, D.K. Fake News, Rumor, Information Pollution in Social Media and Web: A Contemporary Survey of State-of-the-arts, Challenges and Opportunities. *Expert Syst. Appl.* **2019**, *153*, 112986. [CrossRef]
6. Sharma, K.; Qian, F.; Jiang, H.; Ruchansky, N.; Zhang, M.; Liu, Y. Combating fake news: A survey on identification and mitigation techniques. *ACM TIST* **2019**, *10*, 21. [CrossRef]
7. Bondielli, A.; Marcelloni, F. A survey on fake news and rumour detection techniques. *Inf. Sci.* **2019**, *497*, 38–55. [CrossRef]
8. Ozbay, F.A.; Alatas, B. Fake news detection within online social media using supervised artificial intelligence algorithms. *Phys. A Stat. Mech. Appl* **2020**, *540*, 123174. [CrossRef]
9. Monti, F.; Frasca, F.; Eynard, D.; Mannion, D.; Bronstein, M.M. Fake News Detection on Social Media Using Geometric Deep Learning. *arXiv* **2019**, arXiv:1902.06673.
10. Vishwakarma, D.K.; Varshney, D.; Yadav, A. Detection and veracity analysis of fake news via scrapping and authenticating the web search. *Cogn. Syst. Res.* **2019**, *58*, 217–229. [CrossRef]
11. Alzanin, S.M.; Azmi, A.M. Detecting rumors in social media: A survey. *Procedia Comput. Sci.* **2018**, *142*, 294–300. [CrossRef]
12. Cybenko, A.K.; Cybenko, G. AI and fake news. *IEEE Intell. Syst.* **2018**, *33*, 1–5. [CrossRef]
13. Jang, S.M.; Geng, T.; Li, J.Y.Q.; Xia, R.; Huang, C.T.; Kim, H.; Tang, J. A computational approach for xamining the roots and spreading patterns of fake news: Evolution tree analysis. *Comput. Hum. Behav.* **2018**, *84*, 103–113. [CrossRef]
14. Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic Detection of Fake News. *arXiv* **2017**, arXiv:1708.07104.
15. Wang, W.Y. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv* **2017**, arXiv:1705.0064.
16. Ruchansky, N.; Seo, S.; Liu, Y. CSI: A hybrid deep model for fake news detection. In Proceedings of the 26th ACM International Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 797–806.
17. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
18. Al Amrani, Y.; Lazaar, M.; El Kadiri, K.E. Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Comput. Sci.* **2018**, *127*, 511–520. [CrossRef]
19. Garg, N.; Gupta, R.; Kaur, M.; Ahmed, S.; Shankar, H. Efficient Detection and Classification of Orange Diseases using Hybrid CNN-SVM Model. In Proceedings of the 2023 International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 11–12 May 2023; pp. 721–726.
20. Nasir, J.A.; Khan, O.S.; Varlamis, I. Fake news detection: A hybrid CNN-RNN based deep learning approach. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100007. [CrossRef]
21. Dedeepya, P.; Yarrarapu, M.; Kumar, P.P.; Kaushik, S.K.; Raghavendra, P.N.; Chandu, P. Fake News Detection on Social Media Through a Hybrid SVM-KNN Approach Leveraging Social Capital Variables. In Proceedings of the 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 5–7 June 2024; pp. 1168–1175.
22. Ramos, J. Using TF-IDF to determine word relevance in document queries. In Proceedings of the 1st Instructional Conference on Machine Learning. 2003; Volume 242, pp. 133–142. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1;type=pdf;doi=b3bf6373ff41a115197cb5b30e57830c16130c2c (accessed on 11 August 2024).
23. Rish, I. An empirical study of the naive Bayes classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 4 August 2001; Volume 3, pp. 41–46.
24. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM TIST* **2011**, *2*, 1–27. [CrossRef]
25. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
26. Yager, R.R. An extension of the naive Bayesian classifier. *Inf. Sci* **2006**, *176*, 577–588. [CrossRef]

27. Cortes, V.; Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
28. Ahmad, I.; Yousaf, M.; Yousaf, S.; Ahmad, M.O. Fake news detection using machine learning ensemble methods. *Complexity* **2020**, *1*, 8885861. [CrossRef]
29. Hamsa, H.; Indiradevi, S.; Kizhakkethottam, J.J. Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Proc. Technol.* **2016**, *25*, 326–332. [CrossRef]
30. Malhotra, P.; Malik, S.K. Fake News Detection Using Ensemble Techniques. *Multimed. Tools Appl.* **2024**, *83*, 42037–42062. [CrossRef]
31. Sharma, U.; Saran, S.; Patil, S.M. Fake news detection using machine learning algorithms. *IJCRT* **2020**, *8*, 509–518.
32. Khanam, Z.; Alwasel, B.N.; Sirafi, H.; Rashid, M. Fake news detection using machine learning approaches. *IOP Conf. Ser. Mater. Sci. Eng.* **2021**, *1099*, 012040. [CrossRef]
33. Pandey, S.; Prabhakaran, S.; Reddy, N.S.; Acharya, D. Fake news detection from online media using machine learning classifiers. *J. Phys. Conf. Ser.* **2022**, *2161*, 012027. [CrossRef]
34. Mallick, C.; Mishra, S.; Senapati, M.R. A cooperative deep learning model for fake news detection in online social networks. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 4451–4460. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.