electronics

MDPI

Article

Fake News Data Exploration and Analytics

Mazhar Javed Awan [1,*], Awais Yasin [2], Haitham Nobanee [3,4,5,*], Ahmed Abid Ali [1], Zain Shahzad [1], Muhammad Nabeel [1], Azlan Mohd Zain [6] and Hafiz Muhammad Faisal Shahzad [7]

[1] Department of Software Engineering, University of Management and Technology, Lahore 54770, Pakistan; ahmadabidali22@gmail.com (A.A.A.); zainshahzad@gmail.com (Z.S.); muhammad.nabeel@umt.edu.pk (M.N.)
[2] Department of Computer Engineering, National University of Technology, Islamabad 44000, Pakistan; awaisyasin@nutech.edu.pk
[3] College of Business, Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates
[4] Oxford Centre for Islamic Studies, University of Oxford, Marston Rd, Headington, Oxford OX3 0EE, UK
[5] Faculty of Humanities & Social Sciences, University of Liverpool, 12 Abercromby Square, Liverpool L69 7WZ, UK
[6] UTM Big Data Centre, School of Computing, Universiti Teknologi Malaysia, Skudai Johor 81310, Malaysia; azlanmz@utm.my
[7] Department of Computer Science and IT, University of Sargodha, Sargodha 40100, Pakistan; muhammad.faisal@uos.edu.pk
* Correspondence: mazhar.awan@umt.edu.pk (M.J.A.); nobanee@gmail.com (H.N.)

check for updates

Citation: Awan, M.J.; Yasin, A.; Nobanee, H.; Ali, A.A.; Shahzad, Z.; Nabeel, M.; Zain, A.M.; Shahzad, H.M.F. Fake News Data Exploration and Analytics. *Electronics* **2021**, *10*, 2326. https://doi.org/10.3390/electronics10192326

Academic Editor: George A. Tsihrintzis

Received: 23 August 2021
Accepted: 16 September 2021
Published: 23 September 2021

Abstract: Before the internet, people acquired their news from the radio, television, and newspapers. With the internet, the news moved online, and suddenly, anyone could post information on websites such as Facebook and Twitter. The spread of fake news has also increased with social media. It has become one of the most significant issues of this century. People use the method of fake news to pollute the reputation of a well-reputed organization for their benefit. The most important reason for such a project is to frame a device to examine the language designs that describe fake and right news through machine learning. This paper proposes models of machine learning that can successfully detect fake news. These models identify which news is real or fake and specify the accuracy of said news, even in a complex environment. After data-preprocessing and exploration, we applied three machine learning models; random forest classifier, logistic regression, and term frequency-inverse document frequency (TF-IDF) vectorizer. The accuracy of the TFIDF vectorizer, logistic regression, random forest classifier, and decision tree classifier models was approximately 99.52%, 98.63%, 99.63%, and 99.68%, respectively. Machine learning models can be considered a great choice to find reality-based results and applied to other unstructured data for various sentiment analysis applications.

Keywords: detection; fake news; data exploration; analytics; machine learning; random forest; logistic regression; big data; TF-IDF; natural language processing; unstructured data

1. Introduction

Fake news is something that everyone is very fond of and needs no introduction. We have seen that internet use has taken off dramatically in recent years, as social media platforms such as Facebook, Twitter, WhatsApp, etc., have evolved. We also should not forget to mention YouTube, one of the biggest culprits in spreading fake news among the population. These applications have many benefits, such as sharing something useful for the betterment of the population. One biggest disadvantage is fake news, which spreads in the same way that fire spreads in a forest. The reason for spreading fake news would be to achieve financial or political benefits for yourself or your organization [1]. Fake news applies sentiment analysis, the branch of information retrieval and information extraction [2,3].

Over the years, many computer scientists have studied this issue, which arises in our lives every day. They have made several computational algorithms and methods to help solve our daily problems while providing a good solution. Researchers have created many reasonable solutions in the fields of deep learning, neural networks, etc. First of all, one should be checked whether the news is from news channels, newspapers, or social media. It is because news channels sometimes spread a great deal of fake news to their listeners. After this happens, when they realize their mistake of spreading fake news, they come out and apologize publicly [4]. Spreading fake news for the sake of entertainment is a terrible act. One example would be news about the coronavirus. When this deadly virus evolved worldwide, people started to spread the fake word, suggesting that scientists indicated that the world would be free of this virus in the summertime. However, what happened was that it became deadlier than it was in the winter. This type of news should not be shared with the population because when it turns out ot be inaccurate, they become disheartened and depressed. People who are exposed to wrong information are likely to be affected by psychological illness or distress. These researchers indicated that once a person is caught in this loophole, it is challenging to remove themselves from it [5].

The internet has expanded the level of self-assurance in how individuals accumulate information, shape their perspectives, and draw in with subjects of cultural importance [6]. In another report indicated by the Pew Research Center′s Journalism Project, in 2020, 53% of US grown-ups say they acquired news from web-based media "regularly" or "here and there", with 59% of Twitter clients and 54% of Facebook clients routinely devouring information on those sites [7].

As time passes, the amount of fake news that is being spread is also increasing rapidly. This rapid increase could be seen from the last decade due to the evolution of big technology giants such as Facebook, Twitter, YouTube, etc. The issue of fake news was most prominently observed in the 2016 US general election. Such vast sharing of bogus information that is not confirmed affects the reputations of politicians or their political parties and other sectors such as sports, health, and even science [8]. Another heavily influenced sector is the financial market, where we know that a light rumor can bring disastrous changes to the market, ultimately making the owners pay [9].

One of the main reasons that fake news is spreading rapidly worldwide is that we rely heavily on the information we acquire from social media or any other news platform. There has been much evidence that the news that receives the most significant reaction is often proven wrong later [10,11]. One of these pieces of evidence would be the spreading of the coronavirus, where incorrect and fake information was spread around the world [12].

Recently, machine learning models have achieved good performance results in all fields [13–17]. The machine learning techniques that are very useful in detecting news and marking it as fake or real are the random forest classifier technique, TF-IDF vectorizer technique, and logistic regression technique.

Our research determined that we toned to predict a specific news story as fake or genuine from the given dataset, where the amount of news is the given feature and where the response variable will be of two types: fake or real.

The significant contributions of our study are as follows:

- Pre-processed and extensive data exploration are applied in our work to understand fake and real news.
- As per our knowledge, our proposed four machine learning models are more efficient than previous studies reported.
- The proposed approach could help determine fake or real news for various other types of datasets.

The organization of our study is as follows: Section 2 of the paper presents the related work completed for the detection of fake news. Section 3 presents the methods and materials. Section 4 presents the results obtained by applying different machine learning techniques on the given dataset. Section 5 represents the discussion of the results obtained by applying machine learning techniques. Finally, the last section, Section 6, presents the conclusion of the study and future work.

2. Literature Review

Fake news data are pervasive, and it has become an exploration challenge to consistently check the data, content, and distribution to label it as right or wrong. Many researchers have been trying to work on this problem, and they have also somehow been successful. Some have researched the field of machine learning, and some have explored deep learning. Still, no one has ever produced research in the field of sentiment analysis or sentiment information.

Ahmed et al. [18] applied a 4-g model with term frequency and TF-IDF to extract fake contents. The nonlinear machine learning models did not perform well than the linear models for simulated and actual news. A limitation of the study was less accuracy when applied higher n-gram.

Conroy et al. [19] overviewed two significant classes of strategies for discovering fake/false news. The first overviewed class was related to linguistic methodologies, wherein the material of beguiling messages is removed and dissected to relate language designs with double-dealing. The second overviewed type was related to network approaches, in which network data, for example, message metadata or organized information organization inquiries, could be compiled to produce total misdirection measures. We see the guarantee of an imaginative half and half methodology that joins semantic sign and artificial intelligence with network-based social information.

Hussein [20] has produced 41 articles on sentiment analysis (SA) through natural language processing (NLP). The study did not manage wrong/bogus/fake news, but instead, it continued detecting fake websites or inaccurate reviews. Moreover, the more exploration in a feeling challenge, the less the average precision rate is. This paper explains the work that could be completed in the future. The article says that the focus should be on developing a larger examination circle that can explore input consistently in the future.

Bondielli and Marcelloni [21], played with features that were considered to help detect wrong, fake, or even rumored approaches, providing an examination of the different methods used to complete these assignments, and featured how the assortment of applicable information for performing these assignments is challenging. The limitation of the study was that one is to report and examine the different meanings of fake news and bits of gossip/rumors that have not been written correctly. Second, the assortment of important information featured in the study to represent fake news was incorrect, and the performance of the machine learning models was lower.

Bali et al. [22] study on fake news detection was addressed from the standpoint of NLP and ML. Three representative datasets were assessed, each with its own set of features extracted from the headlines and contents. According to the study's results, gradient boosting surpassed all other classifiers. The accuracy and F1 scores of seven alternative maching learning algorithms were investigated, but they all remained under 90%.

Faustini and Covões [23] recommend using one-class classification to detect take news by developing a solely bogus sample in the training dataset (OCC) model. The case study focuses on the Brazilian political scene at the beginning of the 2018 general elections and uses information from Twitter and WhatsApp. The study consumed a great deal of human labour for fact-checking, and the study was quite costly and time-consuming.

Shaikh and Patil's [24] study extracted features from the TF-IDF of news datasets to detect fake news resources, and their datasets were limited. The passive-aggressive classifier and SVM model achieved 95% accuracy. The dataset samples were minimal.

Recent research by Ahmad et al. [25] looks into different linguistic qualities that can differentiate between fake and actual content. They use a variety of ensemble approaches to training a variety of machine learning algorithms. In comparison to individual learners, experimental evaluation reveals the higher performance of the suggested ensemble learner strategy. The KNN model did not perform well for this study. However, the study's implications are only textual data. Other data types are not addressed.

In another study, Hakak et al. [26] developed an ensemble classification model for detecting fake news, which outperformed state-of-the-art models in terms of accuracy. The proposed methodology collects fundamental properties from false news datasets, then categorizes them using an ensemble model that combines three main machine learning methods. However, the study's implications cannot be generalized due to limited dataset considerations.

Abdullah et al. [27] created a deep learning model that was applied to detect fake news. The study was used to detect fake news using a multimodal model. Still, its performance did not produce good results through a convolutional neural network (CNN), and long short-term memory (LSTM) approaches. The model training time was time taken, and the study was biased towards datasets.

A study by Sharma et al. [28] developed a tool for fake news detection. The research took a phony dataset from the general public to determine the basic techniques of how the deep learning models of LSTM and BI-LSTM work. The models achieved high loss rates, and LSTM and BI-LSTM only achieved a performance rat 91.51%.

Nasir et al. [29] determined automatic detection approaches based on deep learning, and machine learning was researched to combat the rise and distribution of fake news. The categorization of fake news, a recent study suggested a novel hybrid deep learning model. The model has effectively verified two fake news datasets, yielding detection results that were much more superior to non-hybrid baseline approaches. Still, in the ISOT dataset machine learning models, the accuracy was less than 90%.

All of the above studies suggested a clear gap in achieving higher performance through machine learning models from datasets based on multiple features such as title and the subject of the fake news.

3. Dataset and Methodology

This section consists of the materials and methods used in this study to detect fake news from the chosen dataset. Furthermore, Section 3.1 explains the datasets and all of the information related to the dataset. Section 3.2 presents the data pre-processing, Section 3.3 is about data exploration, and the last section, Section 3.4, is related to the methods and algorithms essential to solving this problem.

3.1. Dataset Description and Architecture

The dataset used in this study consists of fake news and real news. Each file of the dataset consists of more than twenty thousand examples of fake news and real news. The dataset considers the title, text, subject, and date that the articles were posted, and the dataset comprises information used from the fake and real news datasets used for Ahmed, Traore and Saad [18]. Figure 1 shows an image representing the number of fake and real news samples in the form of a bar chart.

Figure 2 shows the system architecture representing the stages used in our approach. After analyzing the dataset, we pre-processed it, trained and test split it, applied four machine learning classification models to it, and then performed experiments on the test set.

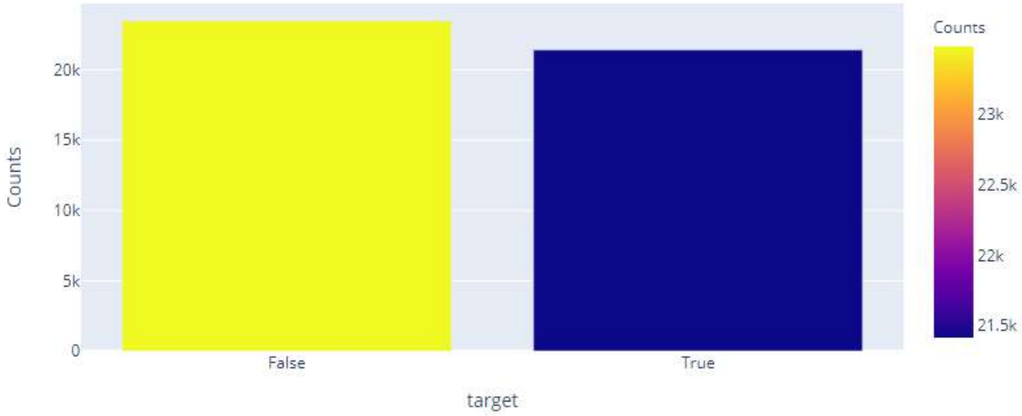


Figure 1. The number of fake and real news articles.

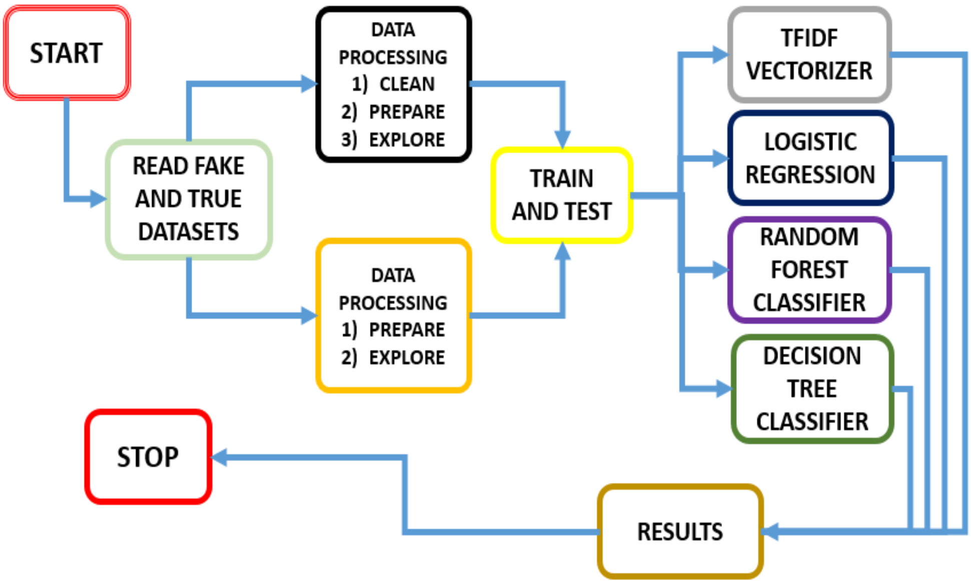


Figure 2. The system architecture of our approach.

3.2. *Data Pre-Processing*

The data needs to be pre-processed before the training, testing, and modeling phases. Before moving to these phases, the real news and fake news are concatenated. In the dataset cleaning process, we removed the columns from the datasets that were not needed for processing. The punctuation and stop words were also removed. Stop-words are those words that frequently occur, such as "I, are, will, Shall, is it, etc. Uppercase letters were converted into lowercase letters. After the dataset was cleaned, it looked good and was ready for the exploration step. However, for the sake of more in-depth research, the dataset exploration was completed on both the cleaned and uncleaned data. For the exploration process, both the fake and real datasets were grouped into a data frame to make the processing easier.

The combined total of fake and real news samples can be seen in Table 1.

Table 1. The number of real and fake news.

Sr. No.	Article Title	Frequency
1	Fake news articles	23,481
2	real news articles	21,417

3.3. Data Exploration

The data exploration stage is used to explore and visualize the data to identify patterns and insights from fake and real news. We plotted various charts using Matplotlib [30] and Seaborn [31] using the Python libraries.

First, we plotted word clouds for the accurate and fake news samples. The word clouds showed all of the essential terms in the datasets. Figure 3a shows the real news keywords in the word clouds for words in the title, showing comments such as Trump, Korea, republican, house, Russia, say, new, leader, white, and senate. Figure 3b shows the word cloud for fake news sample, comprising comments from the titles of the selections, such as Trump, video, watch, Clinton, Obama, Tweet, president, woman, Muslim, democrat.

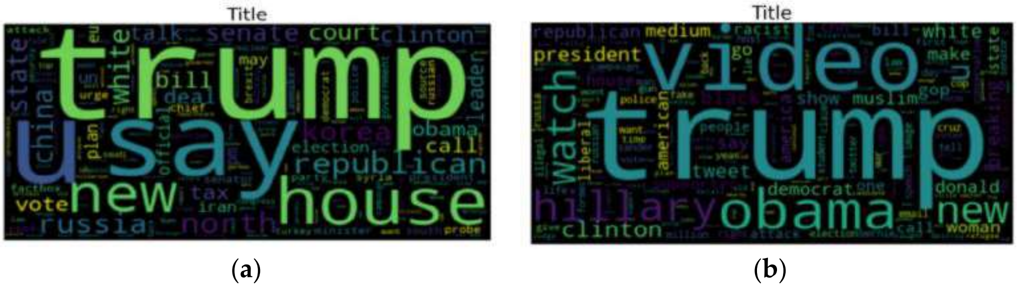


Figure 3. Word clouds comprising title words for the fake and real new samples; (**a**) word cloud of frequently occurring title words from real news samples; (**b**) word cloud of frequently occurring title words from fake news samples.

Figure 4a shows the word clouds of the keywords from the titles from the real news samples, with words such as Trump, state, republican, president, said, Reuters, and party. Figure 4b shows word clouds depicting the keywords from the titles of the fake news samples, with words such as Trump, people, and said.

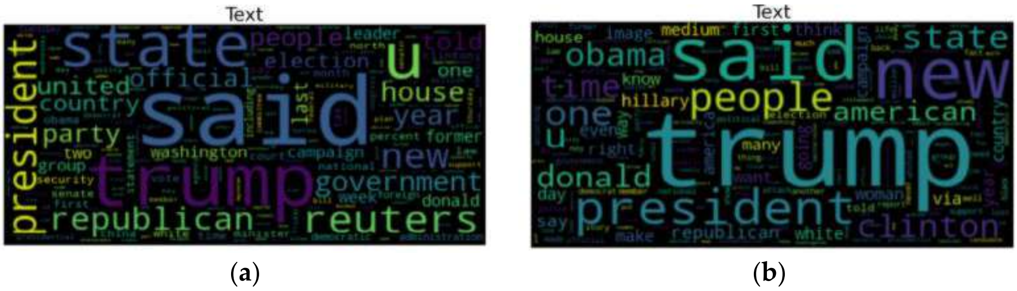


Figure 4. Word cloud of text from real and fake news; (**a**)word cloud representing the frequently occurring words in the real news dataset; (**b**) Word cloud representing the frequently occurring words in the text of the fake news dataset.

Figure 5 shows a line plot of fake and real news against the date. Here, in 2018, the amount of fake news spread is higher than real information.

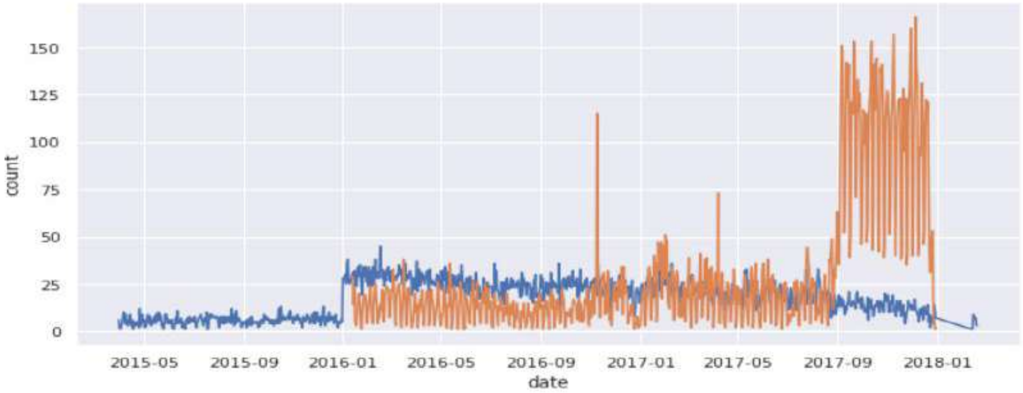
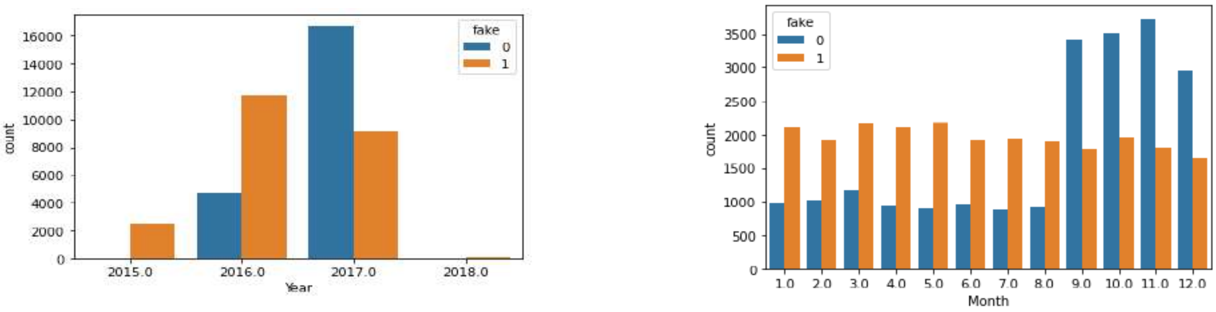


Figure 5. Line chart representing the amount of fake and real news spread based on date.

Further, we created new features, "year," which can be seen in Figure 6a, and "month" in Figure 6b, after using the date column to check which year contained more fake or real news. All of the information for the year 2015 in the dataset is fake news. The amount of fake news is higher until month 8, after which the amount of real news increases drastically. It essentially means that if the month is <=8, then the probability of the news being fake news is higher.



(**a**) (**b**)

Figure 6. Fake and real news counts for year and month; (**a**) year-wise fake and real news; (**b**) month-wise fake and real news.

We plotted a bar chart with counts of various news subjects in Figure 7. Political and world news contained the highest counts after cleaning the dataset.

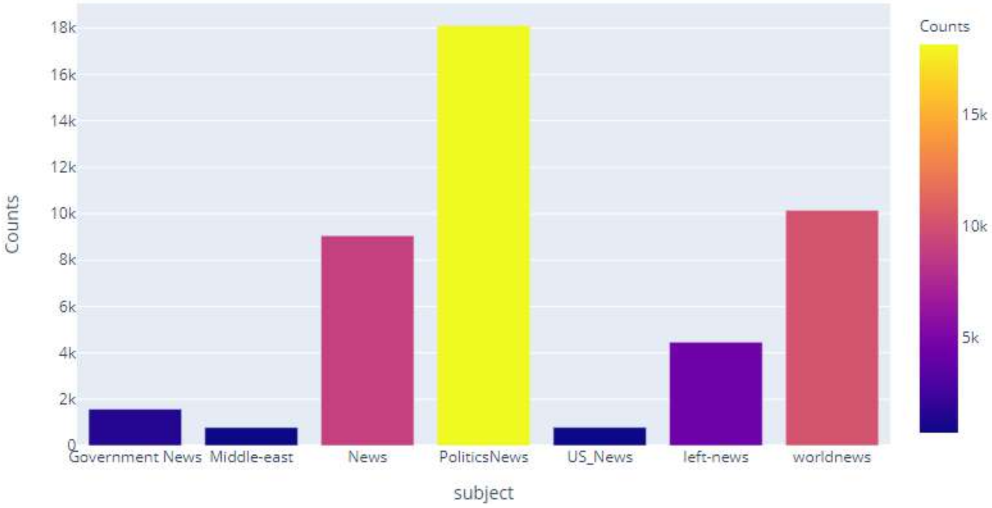


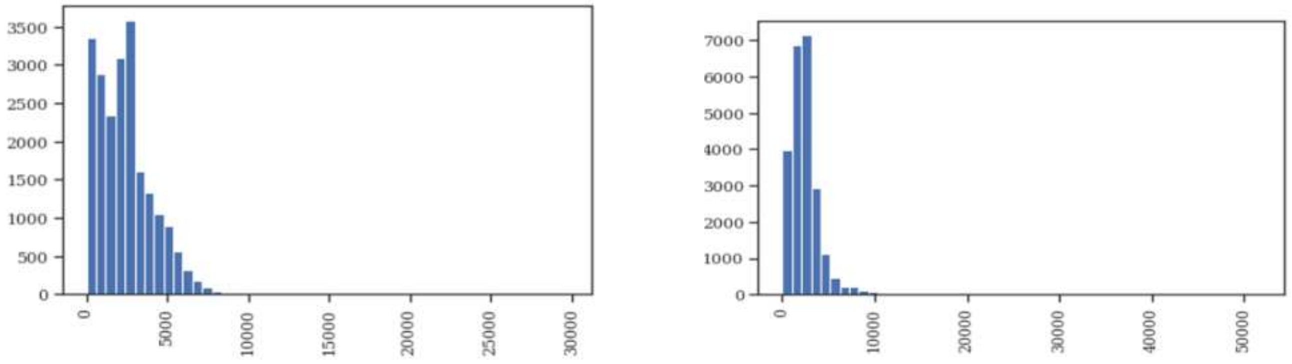
Figure 7. Subject-wise counts of fake and real news.

There are eight different subjects, and their frequencies are seen in Table 2.

Table 2. Frequency of each subject.

S No.	Subject	Frequency
1	Government News	1570
2	Middle East	778
3	News	9050
4	US News	783
5	Left News	4459
6	Politics	6841
7	Politics News	11,272
8	World News	10,145

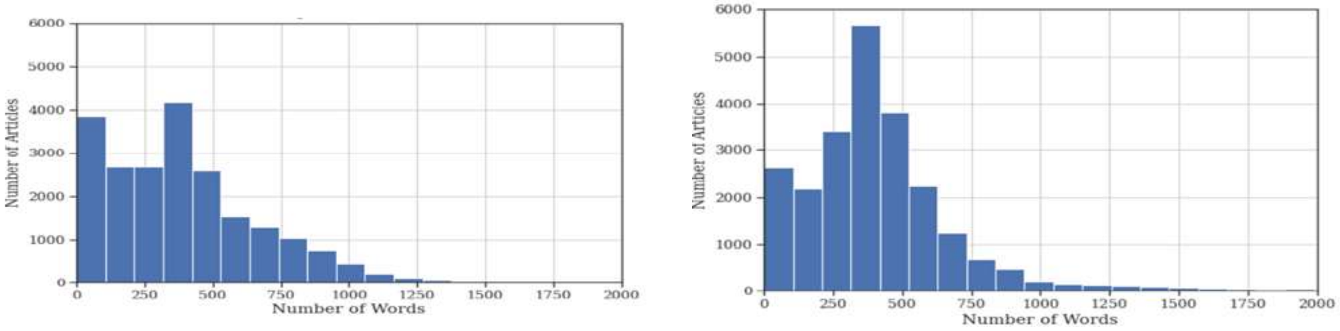
Figure 8a explores the length of the text of real news, and Figure 8b explores text length in fake news. In real news, the longest sentence is 3500, and in fake news, the longest sentence is around 7000.



(**a**) (**b**)

Figure 8. Text length with counts of real and fake news; (**a**) real news text length counts; (**b**) fake news text length counts.

Figure 9a shows the number of articles against the number of words in real news, and Figure 9b shares the same information regarding the number of articles and the number of words for fake news articles.



(**a**) (**b**)

Figure 9. The number of words and articles for real and fake news; (**a**) real news word counts vs. several articles; (**b**) fake news word counts vs. several words.

After the exploration, the data were prepared for modeling, training, and testing, then presented to the machine learning algorithms. The machine learning algorithms were applied to the cleaned and uncleaned datasets. All machine learning algorithms and their explanations are discussed in subsequent sections, along with their confusion matrices and accuracies.

3.4. Our Approach

The methods that we used edict which news was fake and real are discussed in this section.

3.4.1. TF-IDF Vectorizer

A python library known as Scikit learn was used [32]. This library is perfect when performing any task with the TF-IDF vectorizer model. This method includes TF-IDF vectors that represent a term's relative significance in the record or as a whole. The next factor of this method is that term frequency is very important (TF). It represents the frequency of a word occurring in the dataset (we determined the word frequency in an article when undergoing data exploration) [33]. The formula for finding the TF is shown in Equation (1):

$$\mathrm{TF}(t,d) = \frac{Number\ of\ times\ t\ occurs\ in\ a\ document\ 'd'}{Total\ word\ count\ of\ document\ 'd'} \tag{1}$$

The next thing that needs to be determind to ensure that the the model works properly is the IDF, which stands for inverse document frequency. It is used to measure how notable a term is in the entire dataset. The formula for IDF is shown in (2):

$$\mathrm{IDF}(t) = \log_e\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it}\right) \tag{2}$$

The next thing that should be determined is the TF-IDF. The TF-IDF is equal to the inverse document frequency integrated into term frequency, the formula of which is shown in (3):

$$\text{TF-IDF}(t,d) = \mathrm{TF}(t,d) * \mathrm{IDF}(t) \tag{3}$$

The TF-IDF model extracted the feature engineering and counted the most relevant terms from the real and fake news in our dataset. For this reason, it helped to achieve better performance. Second, the technique that we are working with is the TF-IDF vectorizer technique. TF-IDF Vectorizer utilizes an in-memory jargon (a python dictionary) to plan the most successive words to highlight files and process a word event recurrence (scanty) network. The TF-IDF vectorizer is tokenized records and archived recurrent weightings [34].

3.4.2. Logistic Regression

The third technique that we are using to make this model work correctly is the logistic regression technique. Logistic regression in machine learning dictates that logistic regression can discover a connection among the highlights (probability) and likelihood (outcome) of a specific result. A logistic regression classifier is used when the predicting value is categorical. For instance, when predicting the value, it will give either a true or false response. Logistic regression can discover a connection among the highlights (probability) and likelihood (outcome) of a specific result [35]. The logistic regression model can be imported from the sklearn linear_model.

3.4.3. Random Forest Classifier

The random forest has almost the same hyperparameters as a decision tree or a sacking classifier. This technique adds more arbitrariness to the model while developing the trees. First of all, a random forest classifier is a technique that makes different choice trees and consolidates them to produce a more exact and stable prediction. The random forest has hyperparameters that are almost the same as a decision tree or a sacking classifier. This technique adds more arbitrariness to the model while developing the trees [36]. There are diverse arbitrary trees that provide worth, and worth with more votes is the genuine aftereffect of this classifier [37]. It can also be imported from the sklearn, as was the linear model.

3.4.4. Decision Tree Classifier

As we know, this classifier is one of the best classifiers in machine learning. Decision trees are known for their non-parametric supervised learning methods that can be used for processes such as classification and regression tasks. It works in a model way [38]. Tree models where the objective variable can take a discrete arrangement of qualities are called order trees. Decision trees perform with good results and can be made quickly based on Gini index The last machine learning algorithm we will be using is the decision tree classifier. Decision trees are known for their non-parametric supervised learning methods that can be used for both processes, such as classification and regression tasks. Additionally, a decision tree may be suitable for detecting fake news [39]. First of all, it is essential to import the decision tree classifier from the sklearn tree model.

4. Experimental Results

This section has two different sections about the experimental setup in Sections 4.1 and 4.2 is related to the results.

4.1. Experimental Setup

All four models were implemented on Google Colab, which provided a cloud environment. For this, we used python 3.5 and above. The libraries that we used for training and testing were Numpy, Pandas, Scikit learning, Natural language Tool Kit (NLTK), Matplotlib, and Seaborn. We divided the dataset into the training and test set with a ratio of 80:20.

4.2. Results

The results were evaluated through a confusion matrix and a Scikit library classification report of precision, recall, F1-score.

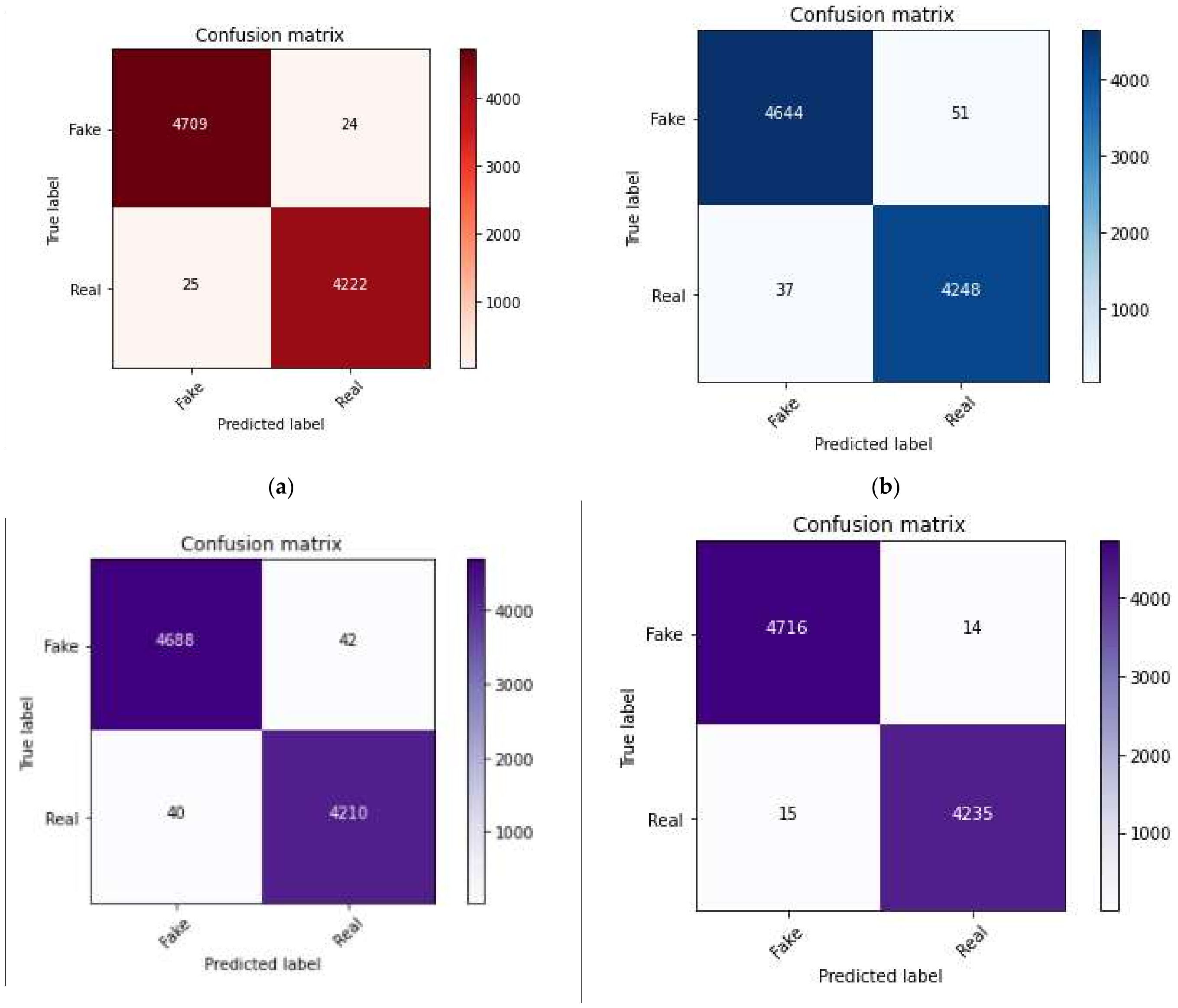
First, the TF-IDF vectorizer was evaluated on the test dataset. The TF-IDF vectorizer achieved an accuracy of 99%, which is almost perfect. The model was able to determine a total of 4709 fake news instances and 4222 real news instances. However, it produced 25 real-fake news and 24 fake-real news, which means that these news samples were somehow real and fake at the same time.

Secondly, the logistic regression model was evaluated based on the test dataset. The model was performed with an accuracy of 98%. The model was able to determine a total of 4644 fake news instances and 4248 real news instances.

Thirdly, the random forest classifier achieved an accuracy of 99%. The model was able to determine a total of 4688 fake news instances and 4210 real news instances.

Lastly, we applied the decision tree classifier, which performed with 99% accuracy. The model determined a total of 4716 fake news instances and 4235 real news instances. The 15 real-fake news and 14 fake-real news instances mean that these news samples were somehow real and fake at the same time.

Figure 10a–d shows the confusion matrix of the fake and real news datasets for the Tf-IDF vectorizer, logistic regression, random forest, and decision tree algorithms.



(**a**)

(**b**)

(**c**)

(**d**)

Figure 10. Confusion Matrix of true and predicted fake and real datasets: (**a**) confusion matrix of the TF-IDF vectorizer model; (**b**) confusion matrix of the logistic regression model; (**c**) confusion matrix of the random forest model; (**d**) confusion matrix of the decision tree model.

A summary of all of the results obtained before cleaning the data, which shows all of the results (accuracies), fake news, and true news inputs in a numeric form, is given in Table 3.

Table 3. Accuracies after applying machine learning models before cleaning.

Sr. No.	Machine Learning Model	Fake News Identified	Real News Identified	Results (Accuracies)
1	TF-IDF Vectorizer	4708	4228	99.51%
2	Logistic Regression	4664	4193	98.63%
3	Random Forest Classifier	4682	4172	98.6%
4	Decision Tree Classifier	4716	4235	99.68%

The summary of all the results obtained after cleaning, which shows all of the results (accuracies), fake news, and true news inputs in a numeric form, is given in Table 4.

Table 4. Accuracies after applying machine learning models after cleaning.

Sr. No.	Machine Learning Model	Fake News Identified	Real News Identified	Results (Accuracies)
1	TF-IDF Vectorizer	4709	4222	99.52%
2	Logistic Regression	4644	4248	98.63%
3	Random Forest Classifier	4678	4173	99.63%
4	Decision Tree Classifier	4716	4235	99.68%

A classification report for all of the machine learning algorithms can also be found. All of the details of the classification report are shown in Table 5. Furthermore, we also calculated the precision, recall, and F1-score of each model.

Table 5. Precision, recall, F1-score after applying machine learning models.

Sr. No.	Machine Learning Model	Precision	Recall	F1-Score
1	TF-IDF Vectorizer	0.99	0.99	0.99
2	Logistic Regression	0.98	0.99	0.98
3	Random Forest Classifier	0.98	0.99	0.98
4	Decision Tree Classifier	0.98	0.99	0.99

5. Discussion

From the result of the current study, we see that all of the classifiers showed exceptional results that would ensure that a research study would be successful. The present study yields more than 90% success rates, which is a feat considering the first time the authors have attempted such a project. This research has shown that fake news can be detected quickly and can be dealt with beautifully. As most research papers are considered successful when results above 80% are achieved, the current study yielding the best possible results that it could is quite an achievement. The recent research showed that fake news did not remain an overwhelming problem in society.

Additionally, this study also determined that the main thing which should be completed in similar studies is dataset cleaning. There are a variety of factors that cause the spread of fake news. Our paper has shown that fake news can be handled. In our opinion, the future work that needs to continue this study would be to make a graphical user interface. GUI is necessary to make an application look attractive, and a good GUI is essential when building an application. Using the GUI, people can just copy-paste any text in the GUI and have its classification results. It shows that technology has made our lives easy as well as challenging. In terms of user requirements, technological options, and support for the decision, we see that if we analyze the user requirements, one main user requirement will be to differentiate between fake and real news. The users will be able to determine what type of news is real and which news is fake. The technologies that are involved in this research study are machine learning techniques. These techniques include the TF-IDF-vectorizer, random forest classifier, logistic regression, and decision tree classifier techniques, which can be used after importing the necessary libraries. We chose this design because these classifiers are capable of producing perfect results in terms of accuracy. A comparison of the different schemes tested within the last three years is shown in Table 6.

Table 6. Comparison of our work with other studies.

Sr. No.	Machine Learning Models	Accuracy	Studies
1	TF-IDF-Vectorizer, Logistic Regression, Random Forest Classifier, Decision Tree Classifier	99.45%	Proposed study
3	Random forest algorithm, Perez-LSVM, Linear SVM, multilayer perceptron, bagging classifiers, boosting classifiers, KNN	99%, 99%,98%, 98%, 98%, 88%	[25]
4	LSTM and BI-LSTM Classifier	91.51%	[28]
5	Term Frequency-Inverted Document Frequency (TF-IDF) and Support Vector Machine (SVM)	95.05%	[24]

From Table 6, we see that the accuracies of other papers are lower than the accuracies of our work. It shows that our results are perfect. One of the limitations of our study is the datasets were not massive. The analysis was only performed on four machine learning models.

6. Conclusions

Our social media is generating every kind of news; mostly, these are fake. Usually, we see clashing realities for a similar point and wonder whether both are valid. We set ourselves in a fix trying to figure out which source to put our confidence. As we have also discussed in the Discussion section, cleaning the dataset is very important. It is essential because it changes the results of the study. As we have seen from determining the frequencies of words as they occur in the dataset, we see that when the data is cleaned, the words such as Trump and said are the most frequently occurring. However, when the dataset has not been cleaned, words such as the, are, and appear the most often. These words on their own have no identity and are considered meaningless until they are used with the other terms. Hence, the datasets should be cleaned to produce accurate results. On a concluding note, the authors want to say that sometimes spreading fake news causes happiness, but for many, it causes sorrow. The spreading of fake news should be stopped as soon as possible. In our research, we used some excellent machine learning algorithms that we're able to show us some splendid results. The algorithms showed an accuracy of more than 99%, which is almost perfect. As a result of this research, people who are pretty addictied to the internet are now not to be afraid of fake news. In the end, there are some limitations and insufficiencies in the presented paper. These occur if the dataset is unbalanced or has not been cleaned, as it will not give accurate results and may be ineffective. The extensive data framework, Spark machine learning, could achieve better results in terms of processing time [40–45]. Furthermore, deep learning-enabled big data models could also be applied to fake news datasets from recently inspired LSTM [46–50].

Author Contributions: All authors contributed equally to this work; conceptualization, M.J.A., A.Y., H.N., A.A.A., Z.S., M.N., A.M.Z., and H.M.F.S.; methodology, M.J.A., A.Y., H.N., A.A.A., Z.S., M.N., A.M.Z. and H.M.F.S.; software, M.J.A., A.Y., H.N., A.A.A., Z.S., M.N., A.M.Z., and H.M.F.S.; validation, M.J.A., A.Y., H.N., A.A.A., Z.S., M.N., A.M.Z., and H.M.F.S.; investigation, M.J.A., A.Y., H.N., A.A.A., Z.S., M.N., A.M.Z., and H.M.F.S.; resources; M.J.A., A.Y., H.N., A.A.A., Z.S., M.N., A.M.Z., and H.M.F.S.; data curation, M.J.A., A.Y., A.A.A., Z.S., M.N., A.M.Z., and H.M.F.S.; writing—original draft preparation, M.J.A., A.Y., A.A.A., Z.S., M.N., A.M.Z., and H.M.F.S.; writing—review and editing, M.J.A., A.Y., H.N., A.A.A., Z.S., M.N., A.M.Z. and H.M.F.S.; visualization, M.J.A., A.Y., H.N., A.A.A., Z.S., M.N., A.M.Z., and H.M.F.S.; project administration, M.J.A., and H.N.; funding acquisition, H.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alonso, M.; Vilares, D.; Gómez-Rodríguez, C.; Vilares, J. Sentiment Analysis for Fake News Detection. *Electronics* **2021**, *10*, 1348. [CrossRef]
2. Rehma, A.A.; Awan, M.J.; Butt, I. Comparison and Evaluation of Information Retrieval Models. *VFAST Trans. Softw. Eng.* **2018**, *13*, 7–14. [CrossRef]
3. Alam, T.M.; Awan, M.J. Domain analysis of information extraction techniques. *Int. J. Multidiscip. Sci. Eng.* **2018**, *9*, 1–9.
4. Kim, H.; Park, J.; Cha, M.; Jeong, J. The Effect of Bad News and CEO Apology of Corporate on User Responses in Social Media. *PLoS ONE* **2015**, *10*, e0126358. [CrossRef]
5. Pulido, C.M.; Ruiz-Eugenio, L.; Redondo-Sama, G.; Villarejo-Carballido, B. A New Application of Social Impact in Social Media for Overcoming Fake News in Health. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2430. [CrossRef]
6. Hamborg, F.; Donnay, K.; Gipp, B. Automated identification of media bias in news articles: An interdisciplinary literature review. *Int. J. Digit. Libr.* **2018**, *20*, 391–415. [CrossRef]
7. Jang, Y.; Park, C.-H.; Seo, Y.-S. Fake News Analysis Modeling Using Quote Retweet. *Electronics* **2019**, *8*, 1377. [CrossRef]
8. Lazer, D.M.J.; Baum, M.A.; Benkler, Y.; Berinsky, A.J.; Greenhill, K.M.; Menczer, F.; Metzger, M.J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. The science of fake news. *Science* **2018**, *359*, 1094–1096. [CrossRef]
9. Kogan, S.; Moskowitz, T.J.; Niessner, M. *Fake News in Financial Markets*; Working Paper; Yale University: New Haven, CT, USA, 2017.
10. Lai, C.-M.; Shiu, H.-J.; Chapman, J. Quantifiable Interactivity of Malicious URLs and the Social Media Ecosystem. *Electronics* **2020**, *9*, 2020. [CrossRef]
11. Wang, Y.; Xia, C.; Si, C.; Zhang, C.; Wang, T. The Graph Reasoning Approach Based on the Dynamic Knowledge Auxiliary for Complex Fact Verification. *Electronics* **2020**, *9*, 1472. [CrossRef]
12. Hua, J.; Shaw, R.J.I. Corona virus (Covid-19) "infodemic" and emerging issues through a data lens: The case of china. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2309. [CrossRef]
13. Anam, M.; Ponnusamy, V.A.; Hussain, M.; Nadeem, M.W.; Javed, M.; Goh, H.G.; Qadeer, S. Osteoporosis Prediction for Trabecular Bone using Machine Learning: A Review. *Comput. Mater. Contin.* **2021**, *67*, 89–105. [CrossRef]
14. Gupta, M.; Jain, R.; Arora, S.; Gupta, A.; Awan, M.J.; Chaudhary, G.; Nobanee, H. AI-enabled COVID-19 outbreak analysis and prediction: Indian states vs. union territories. *Comput. Mater. Contin.* **2021**, *67*, 1–18.
15. Ali, Y.; Farooq, A.; Alam, T.M.; Farooq, M.S.; Awan, M.J.; Baig, T.I. Detection of Schistosomiasis Factors Using Association Rule Mining. *IEEE Access* **2019**, *7*, 186108–186114. [CrossRef]
16. Javed, R.; Saba, T.; Humdullah, S.; Jamail, N.S.M.; Awan, M.J. An Efficient Pattern Recognition Based Method for Drug-Drug Interaction Diagnosis. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 221–226.
17. Nagi, A.T.; Awan, M.J.; Javed, R.; Ayesha, N. A Comparison of Two-Stage Classifier Algorithm with Ensemble Techniques on Detection of Diabetic Retinopathy. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 212–215.
18. Ahmed, H.; Traore, I.; Saad, S. Detecting opinion spams and fake news using text classification. *Secur. Priv.* **2017**, *1*, e9. [CrossRef]
19. Conroy, N.K.; Rubin, V.L.; Chen, Y. Automatic deception detection: Methods for finding fake news. *Proc. Assoc. Inf. Sci. Technol.* **2015**, *52*, 1–4. [CrossRef]
20. Hussein, D.M.E.-D.M. A survey on sentiment analysis challenges. *J. King Saud Univ.-Eng. Sci.* **2018**, *30*, 330–338. [CrossRef]
21. Bondielli, A.; Marcelloni, F. A survey on fake news and rumour detection techniques. *Inf. Sci.* **2019**, *497*, 38–55. [CrossRef]
22. Bali, A.P.S.; Fernandes, M.; Choubey, S.; Goel, M. Comparative performance of machine learning algorithms for fake news detection. In Proceedings of the International Conference on Advances in Computing and Data Sciences, Ghazibad, India, 12–13 April 2019; pp. 420–430.
23. Faustini, P.; Covões, T. Fake news detection using one-class classification. In Proceedings of the 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), Salvador, Brazil, 15–18 October 2019; pp. 592–597.
24. Shaikh, J.; Patil, R. Fake News Detection using Machine Learning. In Proceedings of the 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), San Francisco, CA, USA, 16–17 December 2020; pp. 1–5.
25. Ahmad, I.; Yousaf, M.; Yousaf, S.; Ahmad, M.O. Fake News Detection Using Machine Learning Ensemble Methods. *Complexity* **2020**, *2020*, 1–11. [CrossRef]
26. Hakak, S.; Alazab, M.; Khan, S.; Gadekallu, T.R.; Maddikunta, P.K.R.; Khan, W.Z. An ensemble machine learning approach through effective feature extraction to classify fake news. *Futur. Gener. Comput. Syst.* **2020**, *117*, 47–58. [CrossRef]
27. Abdullah, A.; Awan, M.; Shehzad, M.; Ashraf, M. Fake news classification bimodal using convolutional neural network and long short-term memory. *Int. J. Emerg. Technol.* **2020**, *11*, 209–212.
28. Sharma, D.K.; Garg, S.; Shrivastava, P. Evaluation of Tools and Extension for Fake News Detection. In Proceedings of the 2021 International Conference on Innovative Practices in Technology and Management (ICIPTM), Gautam Buddh Nagar, India, 17–19 February 2021; pp. 227–232.
29. 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]
30. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]

31. Waskom, M.L. seaborn: Statistical data visualization. *J. Open Source Softw.* **2021**, *6*, 3021. [CrossRef]
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Singh, A.K.; Shashi, M. Vectorization of Text Documents for Identifying Unifiable News Articles. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*. [CrossRef]
34. Dey, A.; Jenamani, M.; Thakkar, J.J. Lexical TF-IDF: An n-gram feature space for cross-domain classification of sentiment reviews. In Proceedings of the International Conference on Pattern Recognition and Machine Intelligence, Kolkata, India, 5–8 December 2017; pp. 380–386.
35. Menard, S. *Applied Logistic Regression Analysis*; Sage: London, UK, 2002; Volume 106.
36. Manzoor, S.I.; Singla, J.; Nikita. Fake News Detection Using Machine Learning approaches: A systematic Review. In Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 23–25 April 2019; pp. 230–234.
37. Segal, M.R. *Machine Learning Benchmarks and Random Forest Regression*; Kluwer Academic Publisher: Amsterdam, The Netherlands, 2004.
38. Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man, Cybern.* **1991**, *21*, 660–674. [CrossRef]
39. Lyu, S.; Lo, D.C.T. Fake News Detection by Decision Tree. In Proceedings of the 2020 SoutheastCon, Raleigh, NC, USA, 28–29 March 2020; pp. 1–2.
40. Awan, M.J.; Rahim, M.S.M.; Nobanee, H.; Yasin, A.; Khalaf, O.I.; Ishfaq, U. A Big Data Approach to Black Friday Sales. *Intell. Autom. Soft Comput.* **2021**, *27*, 785–797. [CrossRef]
41. Ahmed, H.M.; Awan, M.J.; Khan, N.S.; Yasin, A.; Faisal Shehzad, H.M. Sentiment Analysis of Online Food Reviews using Big Data Analytics. *Elem. Educ. Online* **2021**, *20*, 827–836.
42. Awan, M.J.; Rahim, M.S.M.; Nobanee, H.; Munawar, A.; Yasin, A.; Azlanmz, A.M.Z. Social Media and Stock Market Prediction: A Big Data Approach. *Comput. Mater. Contin.* **2021**, *67*, 2569–2583. [CrossRef]
43. Awan, M.; Khan, R.; Nobanee, H.; Yasin, A.; Anwar, S.; Naseem, U.; Singh, V. A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach. *Electronics* **2021**, *10*, 1215. [CrossRef]
44. Khalil, A.; Awan, M.J.; Yasin, A.; Singh, V.P.; Shehzad, H.M.F. Flight Web Searches Analytics through Big Data. *Int. J. Comput. Appl. Technol.*. in press.
45. Awan, M.J.; Khan, M.A.; Ansari, Z.K.; Yasin, A.; Shehzad, H.M.F. Fake Profile Recognition using Big Data Analytics in Social Media Platforms. *Interational J. Comput. Appl. Technol.* **2021**, in press.
46. Awan, M.J. Acceleration of Knee MRI Cancellous bone Classification on Google Colaboratory using Convolutional Neural Network. *Int. J. Adv. Trends Comput. Sci. Eng.* **2019**, *8*, 83–88. [CrossRef]
47. Mujahid, A.; Awan, M.; Yasin, A.; Mohammed, M.; Damaševičius, R.; Maskeliūnas, R.; Abdulkareem, K. Real-Time Hand Gesture Recognition Based on Deep Learning YOLOv3 Model. *Appl. Sci.* **2021**, *11*, 4164. [CrossRef]
48. Awan, M.J.; Raza, A.; Yasin, A.; Shehzad, H.M.F.; Butt, I. The Customized Convolutional Neural Network of Face Emotion Expression Classification. *Ann. Rom. Soc. Cell Biol.* **2021**, *25*, 5296–5304.
49. Awan, M.J.; Rahim, M.M.; Salim, N.; Mohammed, M.; Garcia-Zapirain, B.; Abdulkareem, K. Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach. *Diagnostics* **2021**, *11*, 105. [CrossRef]
50. Aftab, M.O.; Awan, M.J.; Khalid, S.; Javed, R.; Shabir, H. Executing Spark BigDL for Leukemia Detection from Microscopic Images using Transfer Learning. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 216–220.