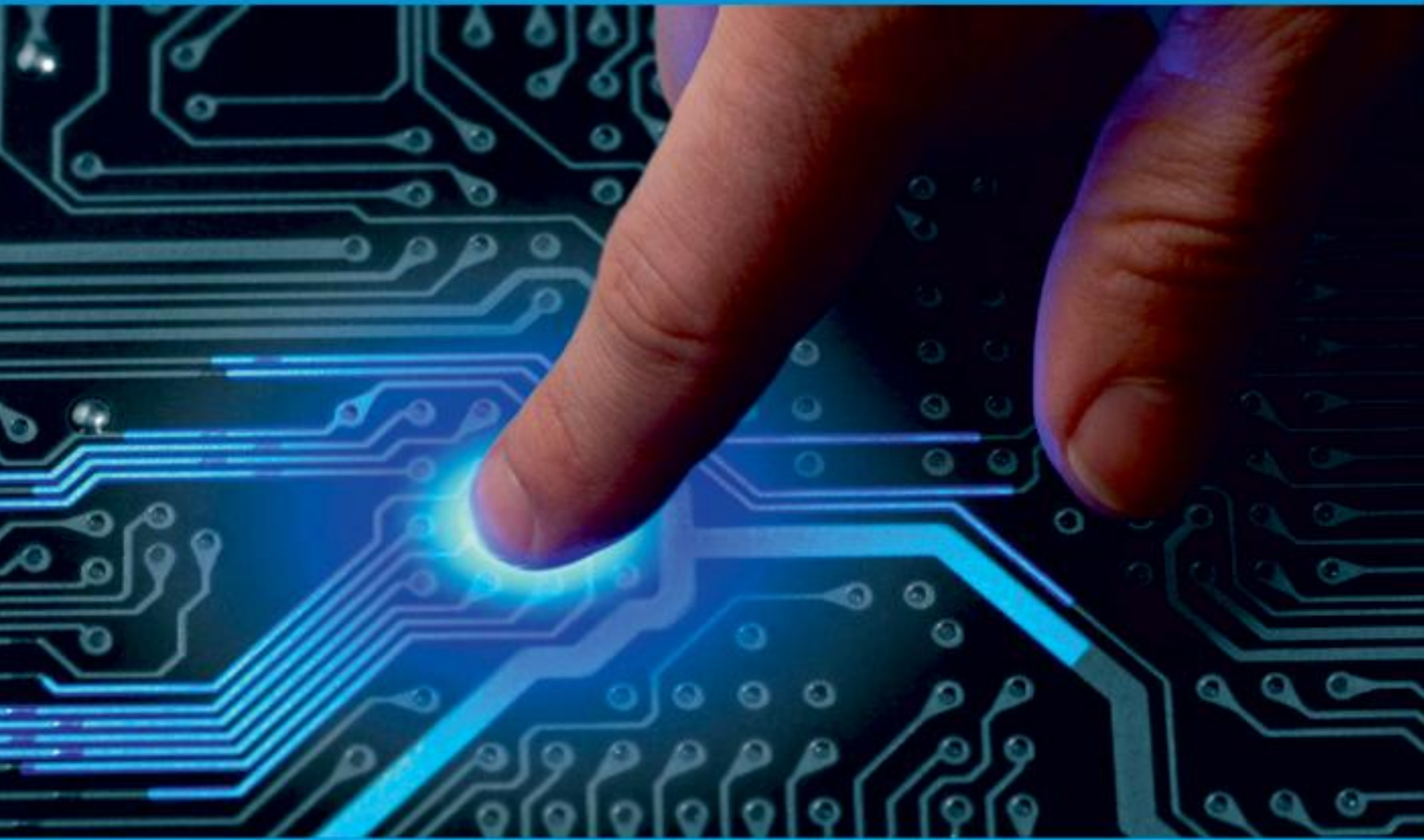
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

| DOI: 10.15680/IJIRCCE.2022.1005108|

Fake News Detection Using Machine Learning and Data Science

Aditya Revadkar, Dhananjay More, Nitin Khot, Shubham Pawar, Prof. Rasika Kachore

Department of Computer Engineering, ISBM College of Engineering, Nande, Pune, India

Department of Computer Engineering, ISBM College of Engineering, Nande, Pune, India

Department of Computer Engineering, ISBM College of Engineering, Nande, Pune, India

Department of Computer Engineering, ISBM College of Engineering, Nande, Pune, India

Department of Computer Engineering, ISBM College of Engineering, Nande, Pune, India

ABSTRACT:

Information preciseness on Internet, especially on social media, is an increasingly important concern, but web scale data hampers, ability to identify, evaluate and correct such data, or so called “fake news,” present in these platforms. In this paper, we propose a method for “fake news” detection and ways to apply it on Facebook, one of the most popular online social media platforms. The results may be improved by applying several techniques that are discussed in the paper. Received results suggest, that fake news detection problem can be addressed with machine learning methods.

Keywords: Decision tree, regression, SVM (support vector machine), and random forest classifier, NLP (natural language processing).

I.INTRODUCTION

Fake news can be simply explained as a piece of article which is usually written for economic, personal or political gains. Fake news is a news designed to deliberately spread hoaxes, propaganda and disinformation. The advent of the World Wide Web and the rapid adoption of social media platforms (such as Facebook and Twitter) paved the way for information dissemination that has never been witnessed in the human history before. Besides other use cases, news outlets benefitted from the widespread use of social media platforms by providing updated news in near real time to its subscribers. The news media evolved from newspapers, tabloids, and magazines to a digital form such as online news platforms, blogs, social media feeds, and other digital media formats. It became easier for consumers to acquire the latest news at their fingertips. Facebook referrals account for 70% of traffic to news websites. These social media platforms in their current state are extremely powerful and useful for their ability to allow users to discuss and share ideas and debate over issues such as democracy, education, and health. However, such platforms are also used with a negative perspective by certain entities commonly for monetary gain [3, 4] and in other cases for creating biased opinions, manipulating mindsets, and spreading satire or absurdity. The phenomenon is commonly known as fake news.

II.DATASET

In our model, we were using true and fake news dataset taken from kaggle. True and fake news dataset is the open source dataset which consist 21417 rows,5 colums and 23581 rows, 5 columns respectively.

Every dataset is mainly divided into three phases such as training, testing and validation. We divide our dataset into training and testing in the ratio of 80:20 that is we divide into 80% of training and 20% of testing. So that, training will contain more data to train and make accurate prediction.

In machine learning the more data we use for train the algorithm,the more accurate our system is.

III.METHODOLOGY

Using twitter dataset in which we are going to pre-process dataset by using natural language processing. (tokenization, removing stop words, steaming). Then this dataset train and test by machine learning algorithms (Decision tree,svm, random forest classifier, logistic regression, etc). It will classify news into fake or true.

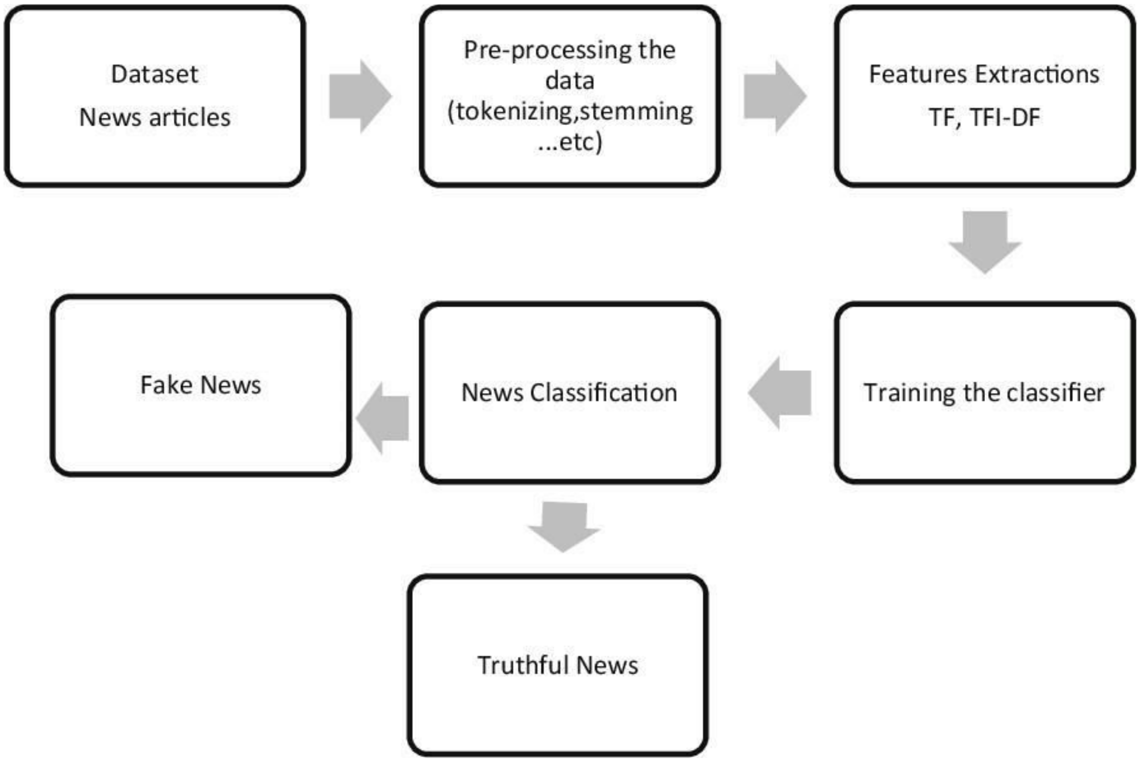


Figure 1. System architecture

PRE-PROCESS:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data pre-processing is a technique that is used to convert the raw data into a clean data set. Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words, sentences called tokens. These tokens help in understanding the context or developing the model for the NLP.

The tokenization helps in interpreting the meaning of the text by analysing the sequence of the words.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005108|

EXTRACTION OF TEXT:

Extracting the text from the article added from the news articles to process. TF- IDF Term Frequency/Inverse Document Frequency) is one of the most popular IR (Information Retrieval) technique to analyze how important a word is in a document, 83% of text-based recommender systems uses TFIDF.TF-IDF weighs the importance of words in a document. For example, “the” is commonly used in any documents so that TF-IDF does not consider “the” important to characterize documents. On contrary, “python” is used in IT relevant topic so that TF-IDF considers “python” as important feature word to recognize topic and category.

TRAIN WITH CLASSIFIERS:

Machine Learning Classifiers can be used to predict the algorithm can predict the class the data. The model is trained using the classifier, so that the model, ultimately, classifies your data. There are both supervised and unsupervised classifiers. Unsupervised machine learning classifiers are fed only unlabelled datasets, which they classify according to pattern recognition or structures and anomalies in the data.

CLASSIFY WITH CLASSIFIER:

Machine Learning algorithm classifier can be used to classify the data.

Result: Getting the result by doing all the steps with average 90%+ accuracy.

Scope of project:

In a future our algorithm tries to recognize fake news or real news on the basis of various dataset and our application will remind to Publisher about their contents.

ALGORITHMS:

A. Random Forest:

It is a combination of decision trees. Here each tree will build a random subset of a training dataset. In each decision tree model, a random subset of variables is used to partition the data set at each node. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

B. Support Vector Machine:

Support Vector Machine(SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. It is mostly used in classification problems.Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors.

C. Logistic Regression:

It is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes and success) or 0 (no and failure).Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

D. Decision Tree:

Decision tree classifier is a supervised learning algorithm and also a very powerful classifier. Decision tree classifier can perform both classification and regression like the support vector machines. All the possible solutions to a decision are graphically represented. It is easy to understand as it uses tree analysis to classify the data. The data is broken into smaller parts and the decision tree is built. Decision trees support both categorical data and numeric data. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.In order to build a tree, we use the CART algorithm**,** which stands for Classification and Regression Tree algorithm.A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

| DOI: 10.15680/IJIRCCE.2022.1005108|

FLOWCHART:

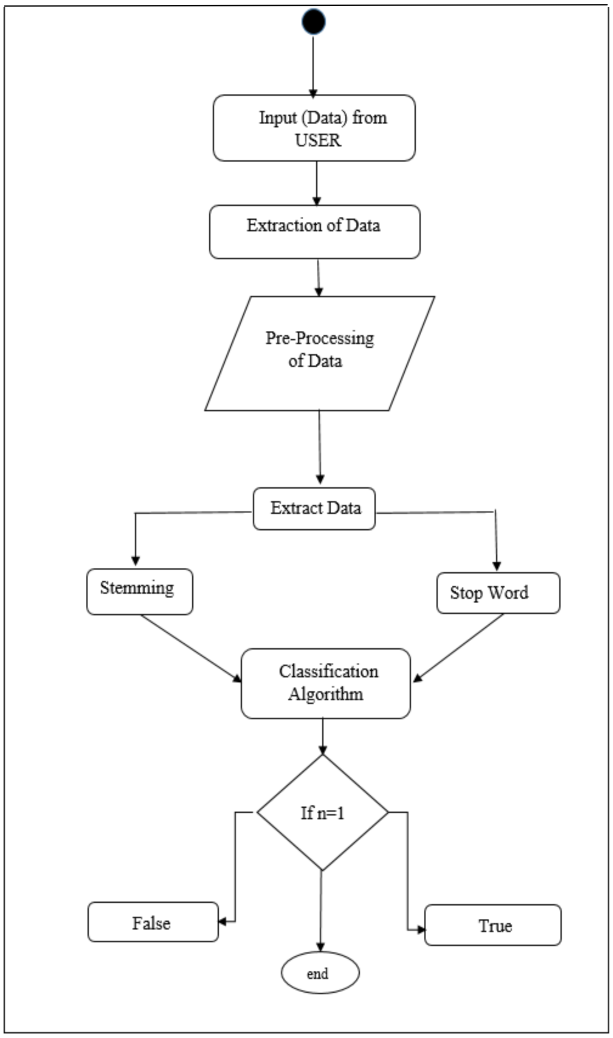


Figure 2.Flowchart

- Out of the four mentioned fields in the data set first we use title as the only source of information followed by text for using SVM classifier, Decision Tree, Random Forest Classifier, Logistic Regression.
- Titles of the news articles were retrieved from the dataset and a data frame was created.
- Every news article labelled as “REAL” and “FAKE” were tagged as 1 and 0 respectively so as to apply SVM classifier, Decision Tree, Random Forest Classifier, Logistic Regression.
- The dataset was randomly shuffled, and then it was divided into two unequal subsets: training dataset, test dataset. Training dataset was used to train the algorithms. Test dataset was used to get the impartial estimation of how well the classifier performs on new data in terms of AUC score. The training dataset contains 80% of the total dataset while rest 20% was allotted to test dataset.

- Firstly we generate our vocabulary by using the bagof-words concept. This approach is simple and is a commonly used way to represent text for use in machine learning, which ignores structure and only counts how often each word occurs. CountVectorizer allows us to use the bag-of-words approach by converting a collection of text documents into a matrix of token counts.

Next, when we transform our training data, fit our model, make predictions on the transform test data, and compute the AUC score for the title.

IV.ANALYSIS AND RESULTS

ANALYSIS

The analysis process involves the following steps:

Step 1: Enter news and Extraction:

News is enter from the user side and then extraction is perform internally.

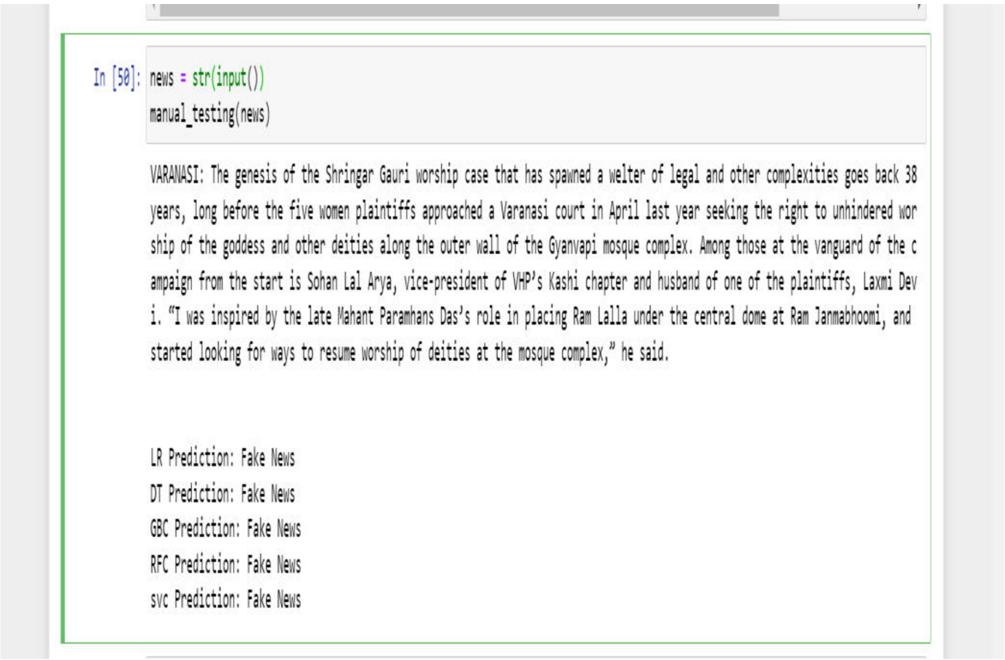
Start:

our algorithms will classify the news as they train and predict the news either it is true or fake

End

RESULTS

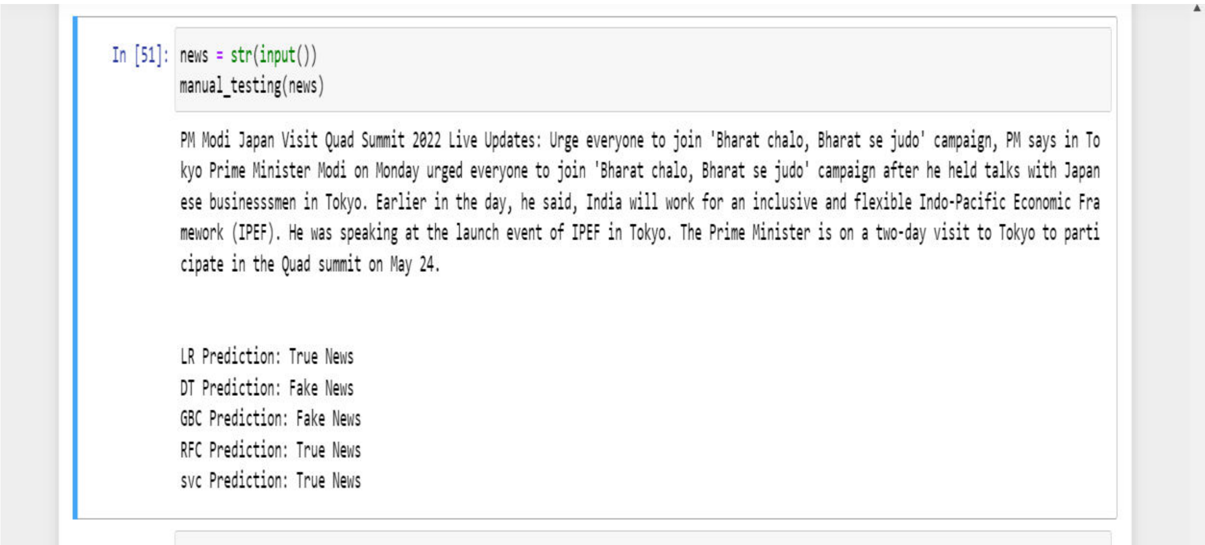
The result of the process is presented here,



```
In [50]: news = str(input())
         manual_testing(news)

         VARANASI: The genesis of the Shringar Gauri worship case that has spawned a welter of legal and other complexities goes back 38
         years, long before the five women plaintiffs approached a Varanasi court in April last year seeking the right to unhindered wor
         ship of the goddess and other deities along the outer wall of the Gyanvapi mosque complex. Among those at the vanguard of the c
         ampaign from the start is Sohan Lal Arya, vice-president of VHP's Kashi chapter and husband of one of the plaintiffs, Laxmi Dev
         i. "I was inspired by the late Mahant Paramhans Das's role in placing Ram Lalla under the central dome at Ram Janmabhoomi, and
         started looking for ways to resume worship of deities at the mosque complex," he said.


         LR Prediction: Fake News
         DT Prediction: Fake News
         GBC Prediction: Fake News
         RFC Prediction: Fake News
         svc Prediction: Fake News
```



```
In [51]: news = str(input())
         manual_testing(news)

         PM Modi Japan Visit Quad Summit 2022 Live Updates: Urge everyone to join 'Bharat chalo, Bharat se judo' campaign, PM says in To
         kyo Prime Minister Modi on Monday urged everyone to join 'Bharat chalo, Bharat se judo' campaign after he held talks with Japan
         ese businesssmen in Tokyo. Earlier in the day, he said, India will work for an inclusive and flexible Indo-Pacific Economic Fra
         mework (IPEF). He was speaking at the launch event of IPEF in Tokyo. The Prime Minister is on a two-day visit to Tokyo to parti
         cipate in the Quad summit on May 24.


         LR Prediction: True News
         DT Prediction: Fake News
         GBC Prediction: Fake News
         RFC Prediction: True News
         svc Prediction: True News
```

V.CONCLUSION

Fake news are false stories in order create propaganda or influence the public on a political or sociological issue with the assistance of social media, internet, and established news sources. Fake news serves various purposes to society such as influencing and persuade people through the means of fake material and damaging a person, event, idea, concept, or people in general. The three main causes are; the social media and internet impact, the rise of unreliable news sources and fall of authoritative news outlets, and lastly the anonymity behind the creation of fake news and misinformation.

VI.FUTURE SCOPE

In a future our algorithm try to recognize fake news or real news on the basis of various dataset and our application will remind to publisher about their contents.

REFERENCES

[1] A. Jain And A. Kasbe, Fake News Detection”, presented at the 2018 International Students,Conference on Electrical, Electronic And Computer Science SCEECS), Bhopal, India, 24th – 25th Feb 2018, IEEE.

[2] 1.R. R. Mandical,. M. Manica R., 4. Krishna, Shivakumar N, Identification of Fake news using machine learning. International Conference on Electronics, Computing and Communication Technologies (CONECCT-2020), Bangalore, India, 2nd - 4th July 2020, IEEE.

[3] S. Deepak and B. Chitturia, Deep neural approach to Fake News identification”, presented at the 2020 International Conference on Computational Intelligence and Data Science (ICCIDS 2019), Amritpuri, India, 26th March 2020, ScienceDirect.

[4] J. Kapusta, P. Hajek, M. Munk and L. Benko.Comparison of fake and real news based on morphological analysis”, presented at the Peer-review under responsibility of the scientific committee of the Third International Conference on Computing and Network Communication (CoCoNet’19), India.

| DOI: 10.15680/IJIRCCE.2022.1005108|

[5] M. A. Panhwar, K. A. Memon, A. Abro, 4.D. Zhongliang, S. A. Khuhro and 6.S. Memon, “Signboard Detection and Text Recognition Using Artificial Neural Networks”, presented at 2019 9th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China, 12-14/ July/ 2019, IEEE.

[6] F. C. Akyon, M. E. Kalfaoglu, “Instagram Fake and Automated Account Detection”, Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 31 Oct.-2 Nov. 2019, IEEE.

[7] Y. Lahlou, S. E. Fkihi, R. Faizi, Automatic detection of fake news on online platforms: A survey,1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 3-4 Oct. 2019, IEEE.

[8] R.Pathar, A.Adivarekar, A.Mishra, A.Deshmukh, Human Emotion Recognition using Convolutional Neural Network in Real Time”, International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 25-26 April 2019, IEEE.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details

www.ijircce.com