iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.38561

www.ijraset.com

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

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

Detecting Fake News Using Social Media Platforms

Prof. B. J. Deokate[1], Akansha Gahide[2], Omkar Chauhan[3], Yogita Wani[4], Diptesh Patil[5]
[1, 2, 3, 4, 5]*Deptartment of Information Technology Engineering, SKN Sinhgad Institute of Technology and Science,Kusgoan(BK), Lonavala, Maharashtra, India, 410401*

Abstract: ***Fake news detection is an interesting topic for computer scientists and social science. The recent growth of the online social media fake news has great impact to the society. There is a huge information from disparate sources among various users around the world. Social media platforms like Facebook, WhatsApp and Twitter are one of the most popular applications that are able to deliver appealing data in timely manner. Developing a technique that can detect fake news from these platforms is becoming a necessary and challenging task. This project proposes a machine learning method which can identify the credibility of an article that will be extracted from the Uniform Resource Locator (URL) entered by the user on the front end of a website. The project uses the five widely used machine learning methods: Long Short Term Memory (LSTM), Random Forest (random tree), Random Forest (decision tree), Decision Tree and Neural Network to give a response telling the user about the credibility of that news. Our initial definition of reliable and unreliable will rely on the human-curated data http://opensources.co. OpenSources.co has a list of about 20 credible news websites and a list of over 700 fake news websites. The proposed model is working well and defining the correctness of results upto 87.45% of accuracy.***

Keywords: Data Pre-processing, Fake news datasets, ML algorithms, Prediction.

I. INTRODUCTION

Fake news or junk news or pseudo-news is a type of yellow journalism or propaganda that consists of deliberate disinformation or hoaxes spread via traditional print and broadcast news media or online social media. The news is often reverberated as misinformation in social media but occasionally finds its way to the mainstream media as well. We plan to build a web-based application or browser extension to help users identify if a news source is reliable or fake. Our initial definition of reliable and unreliable will rely on the human-curated data http://opensources.co. OpenSources.co has a list of about 20 credible news websites and a list of over 700 fake news websites. The project begins by building a profile of these sites, crawling both reliable and unreliable sites. The list obtained from this website will as our data set. The crawled information will be stored in the local machine for further data processing including but not limited URL extraction and author analysis. Additionally, external libraries of some machine learning techniques in Recurrent Neural Network(RNN) will be applied for data classification/prediction on the backend server. The figure 1.1 shows a simple representation of the project. In this figure the user open up the front end of the website and enters an URL. This URL will contain an article that needs to be checked by the user. Then web front end responds and tell about the credibility of the news. The propagation of ambiguous information available every day at different platforms such as news blogs, online newspapers and social media. Now a day's young generation is mostly spent his time on social media and internet thus it has become the main source of consuming news or information instead of acquiring from traditional sources. News on social media is more appealing and less expensive compared to other traditional news organization and it is easy to share, like and comment but despite providing the benefit, this class of news from social media is inferior than other traditional news sources. As per survey conducting in 2016, the percentage of consuming news on social media was 62% whereas in 2012 it was 49%. It shows that we are bombarded with information day-by-day and do not have the related resources, knowledge or expertise to verify the information. Fake news intentionally spread for a variety of purpose such as, it can impact their ability to distinguish what is legitimate or what is not legitimate. Detection of news on social media is an interesting problem. Fake news spread in different format like click baits, news blogs and online newspaper. In the recent survey, Facebook referrals account consists 50% fake news sites and 20% genuine websites. Here are some strategies to shield yourself from fake news. First one, Are you familiar with the source? Does it be legitimated? Has it been reliable in the past, if not you may not want to trust it? The second one, if inflammatory headline drew your attention then read one times more before you decide. Many scholars consider that with help of artificial intelligence and machine learning technique we can easily handle fake news detection problem because of recently artificial intelligence algorithms are used in classification problems (voice detection, image recognition) and they work much better. We collect the data of Users' social engagements from different websites but that data is huge, incomplete, unstructured and noisy. Then our main focus on finding the way to extract useful feature, a credible user. Data mining is the way toward taking care of data from a dataset which is undetectable straightforwardly.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

Detection of fake news on social media is the latest evolving research area, which can be solved by different data mining perspective. This research issues divided into four categories.

1) Application Oriented
2) Data Oriented
3) Model Oriented
4) Features Oriented

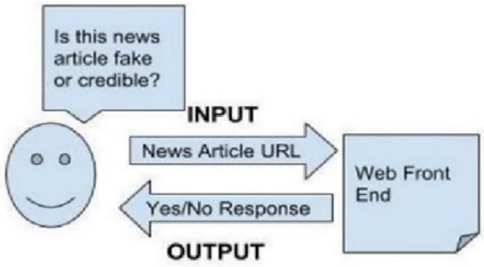


Figure 1 : Basic Layout of Project

II. RELATED WORK

A. Social Media and Fake News

Social media includes websites and programs that are devoted to forums, social websites, microblogging, social bookmarking and wikis. On the other side, some researchers consider the fake news as a result of accidental issues such as educational shock or unwitting actions like what happened in Nepal Earthquake case. In 2020, there was widespread fake news concerning health that had exposed global health at risk. The WHO released a warning during early February 2020 that the COVID-19 outbreak has caused massive ‘infodemic’, or a spurt of real and fake news—which included lots of misinformation.

B. Natural Language Processing

The main reason for utilizing Natural Language Processing is to consider one or more specializations of system or an algorithm. The Natural Language Processing (NLP) rating of an algorithmic system enables the combination of speech understanding and speech generation. In addition, it could be utilized to detect actions with various languages.It suggested a new ideal system for extraction actions from languages of English, Italian and Dutch speeches through utilizing various pipelines of various languages such as Emotion Analyzer and Detection, Named Entity Recognition (NER), Parts of Speech (POS) Taggers, Chunking, and Semantic Role Labeling made NLP good Subject of the search.

The Sentiment analysis extracts emotions on a particular subject. Sentiment analysis is composed of extracting a specific term for a subject, extracting the sentiment, and pairing with connection analysis. The Sentiment analysis uses dual languages Resources for analysis: Glossary of meaning and Sentiment models database. for constructive and Destructive words and attempts to give classifications on a level of -5 to 5. Parts of speech taggers tools for languages such as European languages are being explored to produce parts of language taggers tools in different languages such as Sanskrit, Hindi and Arabic. Can be efficient Mark and categorize words as names, adjectives, verbs, and so on. Most part of speech techniques can be performed effectively in European languages, but not in Asian or Arabic languages. Part of the Sanskrit word "speak" specifically uses the tree-bank method. The Arabic utilizes Vector Machine (SVM) uses a method to automatically identify symbols and parts of speech and automatically expose basic sentences in Arabic text.

C. Data Mining

Data mining techniques are categorized into two main methods, which is; supervised and unsupervised. The supervised method utilizes the training information in order to foresee the hidden activities. Unsupervised Data Mining is a try to recognize hidden data models provided without providing training data for example, pairs of input labels and categories. A model example for unsupervised data mining is aggregate mines and a syndicate base.

D. Machine Learning (ML) Classification

Machine Learning (ML) is a class of algorithms that help software systems achieve more accurate results without having to reprogram them directly. Data scientists characterize changes or characteristics that the model needs to analyze and utilize to develop predictions. When the training is completed, the algorithm splits the learned levels into new data. There are six algorithms that are adopted in this paper for classifying the fake news.

E. Decision Tree

Decision tree is a tree model, the query process corresponds to a path from the root to a leaf. In each inner node one feature value will be examined, by comparing with a pre-calculated value it will decide which subtree to go. When it reach a leaf, the result stored in will be the answer. In our project, Gini impurity is used to determine which feature to examine for every node when building the tree. The decision tree is an important tool that works based on flow chart like structure that is mainly used for classification problems. Each internal node of the decision tree specifies a condition or a "test" on an attribute and the branching is done on the basis of the test conditions and result. Finally the leaf node bears a class label that is obtained after computing all attributes. The distance from the root to leaf represents the classification rule. The amazing thing is that it can work with category and dependent variable. They are good in identifying the most important variables and they also depict the relation between the variables quite aptly. They are significant in creating new variables and features which is useful for data exploration and predicts the target variable quite efficiently.

Tree based learning algorithms are widely with predictive models using supervised learning methods to establish high accuracy. They are good in mapping non-linear relationships. They solve the classification or regression problems quite well and are also referred to as CART.

F. Random Forest

Random Forest are built on the concept of building many decision tree algorithms, after which the decision trees get a separate result. The results, which are predicted by large number of decision tree, are taken up by the random forest. To ensure a variation of the decision trees, the random forest randomly selects a subcategory of properties from each group.

The applicability of Random forest is best when used on uncorrelated decision trees. If applied on similar trees, the overall result will be more or less similar to a single decision tree. Uncorrelated decision trees can be obtained by bootstrapping and feature randomness.

G. Support Vector Machine (SVM)

The SVM algorithm is based on the layout of each data item in the form of a point in a range of dimensions n (the number of available properties), and the value of a given property is the number of specified coordinates. Given a set of n features, SVM algorithm uses n dimensional space to plot the data item with the coordinates representing the value of each feature. The hyper-plane obtained to separate the two classes is used for classifying the data.

H. Naive Bayes

This algorithm works on Bayes theory under the assuming that its free from predictors and is used in multiple machine learning problems. Simply put, Naive Bayes assumes that one function in the category has nothing to do with another. For example, the fruit will be classified as an apple when its of red color, swirls, and the diameter is close to 3 inches. Regardless of whether these functions depend on each other or on different functions, and even if these functions depend on each other or on other functions, Naive Bayes assumes that all these functions share a separate proof of the apples.

I. LSTM

LSTM refers to Long-Short Term Memory, which is a recurrent neural network architecture. It achieved the best known results in natural language text. This model applies binary cross-entropy to the preprocessed content keywords data. The unique characteristic of a recurrent neural network is that it doesn't restart scratches from the beginning. Instead, loops in its chain-like architecture ensure the persistency of the past information, which distinguishes it from traditional neural networks. More precisely, LSTM is capable of learning the long-term dependencies of a given data set. Thus, it often can be successfully applied to many language-related tasks. In the context of our Fake News detection system, we apply LTSM to keywords of an article. To properly use LSTM on content keywords, the model first needs to preprocess the data extracted by Newspaper API. For instance, the preprocessing procedure maps each unique word onto an integer in the defined scope.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

III. METHODOLOGY

This section presents the methodology used for the classification. Using this model, a tool is implemented for detecting the fake articles. In this method supervised machine learning is used for classifying the dataset. The first step in this classification problem is dataset collection phase, followed by preprocessing, implementing features selection, then perform the training and testing of dataset and finally running the classifiers. Below figure describes the proposed system methodology.

The methodology is based on conducting various experiments on dataset using the algorithms described in the previous section named Random forest, SVM and Naïve Bayes, majority voting and other classifiers. The experiments are conducted individually on each algorithm, and on combination among them for the purpose of best accuracy and precision.

The main goal is to apply a set of classification algorithms to obtain a classification model in order to be used as a scanner for a fake news by details of news detection and embed the model in python application to be used as a discovery for the fake news data. Also, appropriate refactoring's have been performed on the Python code to produce an optimized code.

The classification algorithms applied in this model are k-Nearest Neighbors (k-NN), Linear Regression, XGBoost, Naive Bayes, Decision Tree, Random Forests and Support Vector Machine (SVM). All these algorithms get as accurate as possible. Where reliable from the combination of the average of them and compare them.

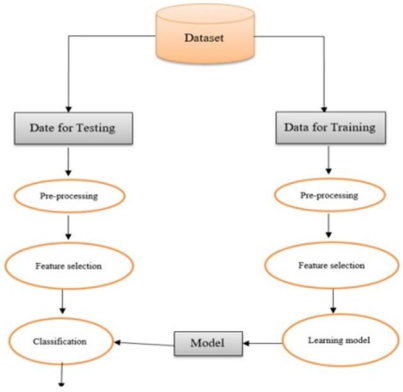
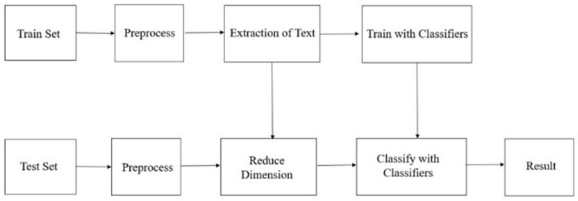


Figure 2 : Describes the Proposed System Methodology

The dataset is applied to different algorithms in order to detect a fake news. The accuracy of the results obtained are analyzed to conclude the final result.

IV. SYSTEM ARCHITECTURE



V. MATHEMATICAL IMPLEMENTATION

A. *Decision Trees*

$$H(X) = -\Sigma (p_i * \log_2 p_i)$$

$$Entropy\ (p) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

B. Logistic Regression

$$P(Y=1|X) \text{ or } P(Y=0|X)$$

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

VI. IMPLEMENTATION AND RESULTS

For the implementation purpose, the four existing approaches are considered. The results of mentioned four models are compared with the proposed model, it is found the accuracy among top 200 results. The demonstration is done using python programming on R studio and some machine learning algorithm. We have performed analysis on various datsets. The results of the analysis of the datasets using the six algorithms have been depicted using the bar plot and charts. The six algorithms used for the detection are as:

A. Random Forests.
B. Naive Bayes.
C. K-Nearest Neighbors (KNN).
D. Decision Tree.
E. SVM

The Bar plot is automatically obtained by Python code using the cognitive learning library when running the algorithm code in jupyter notebook in anaconda platform.

VII. CONCLUSION

It is significant to find the accuracy of news which is available on internet. In the paper, the components for recognizing Fake news are discussed. A mindfulness that not all, the fake news will propagate via web-based networking media. Currently, to test out the proposed method of Naïve Bayes classifier, SVM, Decision trees and NLP are used. In future, ensuing algorithm may provide better results with hybrid approaches for the same purpose fulfilment. The mentioned system detects the fake news on the based on the models applied. Also it had provided some suggested news on that topic which is very useful for any user. In the future, the efficiency and accuracy of the prototype can be enhanced to a certain level, and also enhance the user interface of the proposed model. Hence, this project will make people more informed. It will contribute to start a new revolution against one of the most prevalent hazard i.e. Fake News. It will serve a root and branch eradication of the same.

REFERENCES

[1] 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, published by IEEE.
[2] R. R. Mandical, M. N., Manica R., Krishna A. N. , Shivakumar N, “Identification of Fake news using machine learning”, presented at the 2020 International Conference on Electronics, Computing and Communication Technologies(CONECCT), Bangalore, India, 2nd - 4th July 2020, published by 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, published by 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 Communications (CoCoNet’19), India, 3rd Jan 2020, published by ScienceDirect.
[5] M. A. Panhwar, K. A. Memon, A. Abro, D. Zhongliang, S. A. Khuhro and 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, published by IEEE.
[6] F. C. Akyon, M. E. Kalfaoglu, “Instagram Fake and Automated Account Detection”, presented at 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 31 Oct.-2 Nov. 2019, published by IEEE.
[7] Y. Lahlou, S. E. Fkihi, R. Faizi, “Automatic detection of fake news on online platforms: A survey”, presented at the 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 3-4 Oct. 2019, published by IEEE.

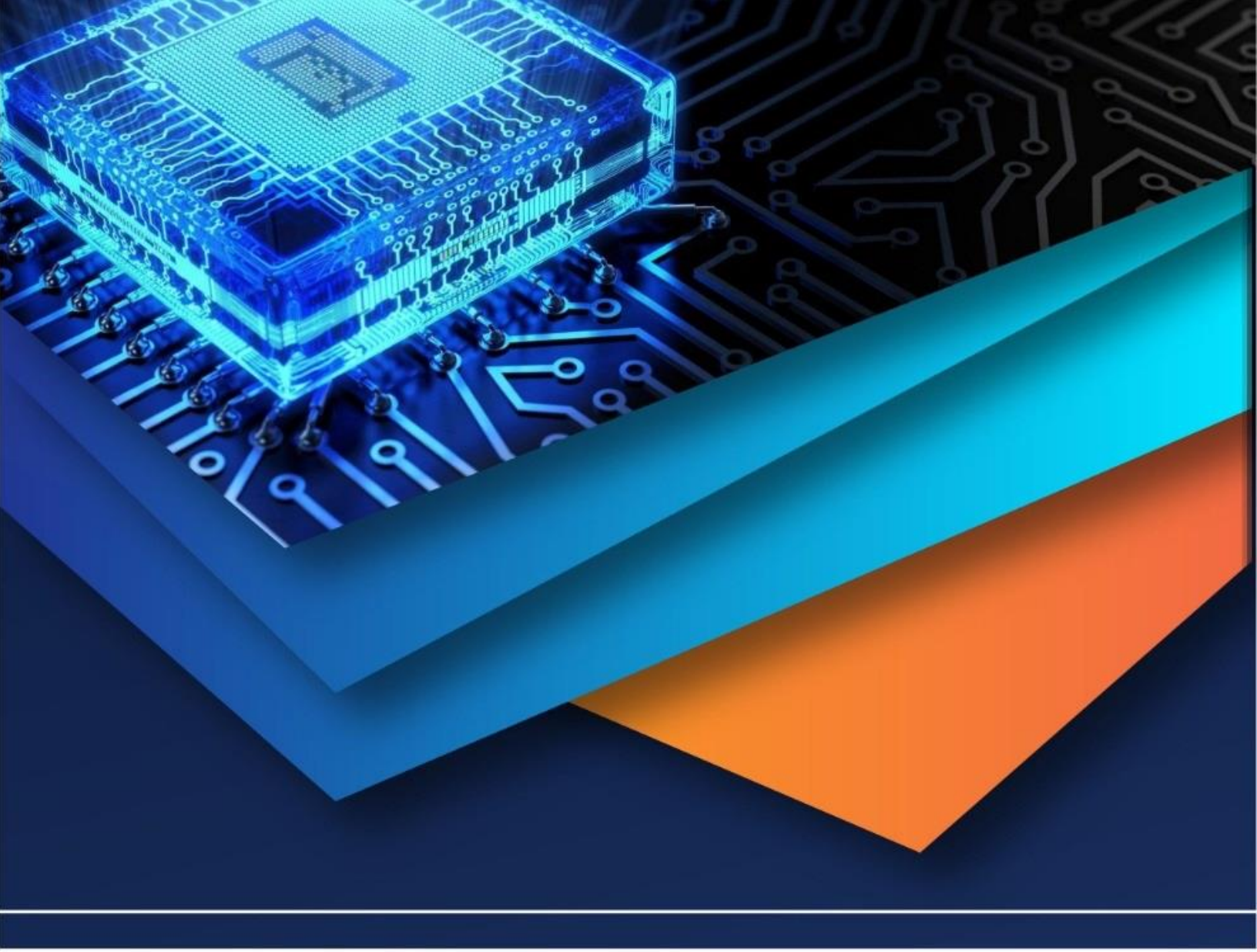
[8] R. Pathar, A. Adivarekar, A. Mishra , A. Deshmukh, "Human Emotion Recognition using Convolutional Neural Network in Real Time", presented at the 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 25-26 April 2019, published by IEEE.

[9] A. Jain, A. Shakya, H. Khatter and A. K. Gupta, "A Smart System For Fake News Detection Using Machine Learning", presented at the 2019 2nd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 28 sep 2019, , published by IEEE.

[10] A. Kesarwani, S. S. Chauhan and A. R. Nair, "Fake News Detection on Social Media using K-Nearest Neighbor Classifier", presented at the 2020 International Conference on Advances in Computing and Communication engineering (ICACCE), Las Vegas, NV , USA, 24 june 2020, published by IEEE.

[11] P. B. P. Reddy, M. P. K. Reddy, G. V. M. Reddy and K. M. Mehata, "Fake Data Analysis and Detection Using Ensembled Hybrid Algorithm", presented by the Third International Conference on Computing Methodologies and Communication (ICCMC 2019), Erode, India, 29 March 2019, published by IEEE

[12] L. Liu, Y. Wang and w. Chi, "Image Recognition Technology Based on Machine Learning", presented at the CACRE 2020 Conference Committee, Dalian, China, 4 sep 2020, published by IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY