iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48172

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: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Detection of fake news using Machine LearningAlgorithms

Tejaswi Gaikwad[1], Bhaskar Rajale[2], Prasad Bhosale[3], Swapnil Vedpathak[4], Mrs. S.S. Adagale[5]
[1, 2, 3, 4]*Computer Department Trinity Academy Of Engineering, Pune, India*
[5]*Faculty of Computer DepartmentTrinity Academy Of Engineering, Pune, India*

Abstract: ***Social media platforms like Facebook, Whatsapp, Twitter, and Telegram are important sources of information diffusion in the modern period, and people believe it with- out questioning its authenticity or source. Social media has fascinated people worldwide in spreading fake news due to its easy availability, cost-effectiveness, and ease of information sharing. The website – pmssgovt online claims that the scheme is applicable to all Indian students studying from class 9 to graduation degree. The website was fake and was charging Rs 450 as the registration fee for students to participate in the said scheme. Fake news can be generated to mislead the community for personal or commercial gains. It can also be used for other personal benefits such as defaming eminent personalities, amendment of government policies, etc. Thus, to mitigate the awful consequences of fake news, several research types have been conducted for its detection with high accuracy to prevent its fatal outcome. This article proposes a framework for detecting fake news based on feature extraction and feature selection algorithms and a set of voting classifiers. The proposed system distinguishes fake news from real news. First, we preprocessed the data taking unnecessary characters and numbers and reducing the words in the dictionary (lemmatization). Second, we extracted some important features by using two types of feature extraction, the term frequency-inverse document frequency technique and bag of words. Third, the extracted characteristics were reduced with the help of the different machine learning algorithm and the analysis of the variance algorithm. At last, challenges and open issues along with future research directions are discussed to facilitate the research in this domain further.***

Index Terms: Naive Bayes, SVM, Logistic Regression

I. INTRODUCTION

Fake news is a manipulated information that resembles newsmedia content in nature but not in management structure or intent . It is continuously exploded via social media, news- papers, online blogs, forums, and magazines, making it hardto identify reliable news sources. The continuous explosionof fake news increases the need for efficient analytical tools capable of providing insight into the reliability of online content. Regular social media users are significantly impacted bythe fake character of news, either negatively or positively. To prevent having a negative impact on the readers, it must be found as soon as feasible. Consequently, the methods and algorithms The paper was reviewed and given the go-ahead forpublishing by the associate editor. The subject of significant research is how to accurately detect fake news.

Fake news sources disregard the mainstream media's editorial policies and standards, which are intended to assure the accuracy and reliability of the material they publish. Fake news primarily draws the attention of the people who are more interested in political talks and stock values andmay affect their mental health, which leads to stress, anxiety, and depression-like issues.

To mitigate the dissemination of fake news, one should focus on the original stories published by the authorized publishers rather than individual articles. It becomes an ideal place for all to create, manipulate, and disseminate fake news. Facebook reported that the malicious actor manipulations accounted for less than one-tenth of 1 percent of public content posted on the site .

Social context-supported techniques leverage other people's social interactions as a secondary assemblage to access and gather information about imitation behaviour. Stance-basedtechniques assess the validity of original news items using users' perspectives from pertinent blog content. Any news is verified by the credibility ratings of pertinent social media posts. A recent area of study is social media's imitation of news discovery. The four scenes that make up the research directions are data-oriented, feature-oriented, model-oriented, and application-oriented. In 2008, the false rumours on Steve Jobs' health (suffering from a heart attack) reported as authentic had great fluctuationsin the stock exchange of Apple Inc. . For instance, research shows that about 19 million bot accounts tweeted in support ofeither Trump or Clinton during the 2016 US presidential elec- tion which perfectly demonstrates how social media greatly contributes to the creation and dissemination of fake news. Fake news is purposefully designed to deceive consumers by playing with the facts and figures. Emulating the fake news as a genuine need to misrepresent reality with various rhetorical forms . There is a possibility that the real news may be citedby fake news in the wrong context to support it . It is quite difficult to detect fake news due to the above factors. Spreading false news is roughly as dangerous as spreading the virus. People are currently encountering fake coronavirus news daily.

This fake information triggers fear and panic among people. Therefore, there is a need for ways to factchecknews.

Fake information can be spread in the form of text, video, pictures, and audio via social media networks such as Face- book and Twitter. The fake news problem has existed for a long time. People used to believe such news even if it was false. Therefore, detecting fake news can be difficult, especially with no supervising body on the internet. The growth of concern regarding the detection of unreliable news is recent. It is difficult for a human to manually detect news, even with the existence of all topics shown on social media. Therefore, thereis a need for an efficient way to help us distinguish false information from true ones posted on social media. One of theefficient ways is to classify the news using machine learning (ML) algorithms.

II. RELATED WORK

The spread of misleading information on social media leadsresearchers to do their best to solve this problem. We present some of the previous work in this direction. Ahmed et al.

[3] proposed using of n-gram model to differentiate between fake and real news. They proposed to generate several setsof n-gram from training data to differentiate between false and true news. They used various features of the n-gram baseline established on words. They preprocessed the dataset by stemming and removing the stop words. They use TF-IDF for text feature extraction. They use six ML algorithms: SGD, k-nearest neighbor (KNN), support vector machines (SVM), LSVM, and decision trees (DT) on three online datasets. They achieved an accuracy (ACC) of 87.0 percnetin differentiating fake from real news using n-gram features and LSVM algorithm. However, trying other algorithms can achieve better performance, such as PA.

Yang et al. [10] proposed a framework for detecting real news and users' credibility using unsupervised learning and probabilistic graphical. Their system achieved an ACC of 75.9 percent. However, incorporating features of news content and user profiles can improve the performance of their unsuper- vised' model. Shu et al. [11] made the fake news trackersystem for fake news problems. First, they collect news and social context automatically to build their dataset. Then, they extract features from the dataset and use ML algorithms to differentiate between false and real news. The experimental results showed that their system achieved an ACC of 74.2 percent. However, using other available features in the datasetlike favorites and re-tweets could enhance the performance.

Ko et al. [15] proposed a cognitive system using backtrack- ing to detect fake news. Their results were an 85.0 percent detection rate. However, they did not clear how to detect fake news and subjective posts. Atodiresei et al. [16] proposed a system for identifying fake users and news on Twitter. Their system will receive a link to a tweet from a user and then compute the tweet credibility, also some statistics such asemotions. However, they didn't clear what are the measure metrics they used to evaluate their work.

Madani et al. [17] focused on fake news that tweeted during the CORONA virus. They proposed a classification approach based on natural language processing, ML, and deeplearning. They used an RF algorithm and achieved 79.0 percentof ACC. However, as compared to other systems, the ACC they achieved is low. Nasir et al. [18] proposed a novel hybrid DL system that gathers convolutional and recurrent neural networks. They used two data sets ISOT and FA-KES. They achieved 99.0 percent of ACC for the ISOT dataset. In comparison to other systems, their system has a good ACC.

III. METHODOLOGY

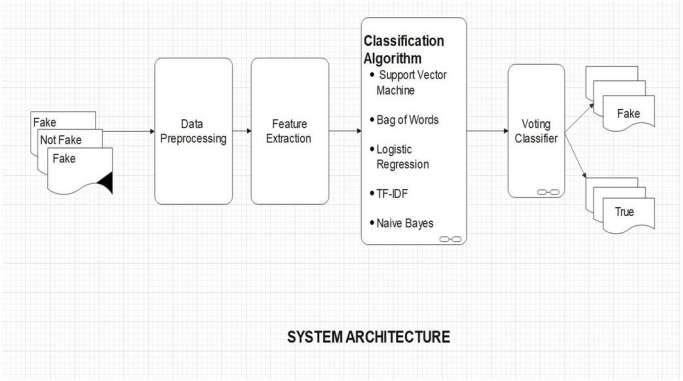


Fig. 1. System Architecture

As u see above, the figure tells us about the architectureof the system the system takes a news as a textual input ,then preprocess the data and classifies using ML algorithmsto figure out the result as fake or real.

A. System consists of the following classifications:

1) *Bag-of-Words:* Bag of words is a Natural Language Pro- cessing technique of text modelling. In technical terms, we cansay that it is a method of feature extraction with text data. This approach is a simple and flexible way of extracting features from documents. A bag of words is a representation of textthat describes the occurrence of words within a document. Wejust keep track of word counts and disregard the grammatical details and the word order. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

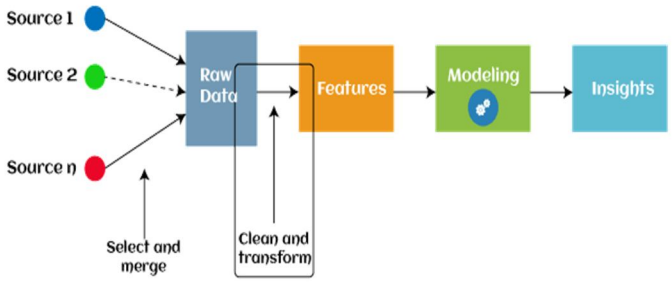


Fig. 2. Data Pre-processing

2) *Feature Engineering:* Feature engineering is the pre-processing step of machine learning, which is used to trans- form raw data into features that can be used for creating a pre- dictive model using Machine learning or statistical Modelling. Feature engineering in machine learning aims to improve the performance of models. It helps to represent an underlying problem to predictive models in a better way, which as aresult, improve the accuracy of the model for unseen data. Thepredictive model contains predictor variables and an outcome variable, and while the feature engineering process selects the most useful predictor variables for the model.

3) *SVM:* 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, pri- marily, 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 datapoint 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, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decisionboundary or hyperplane.

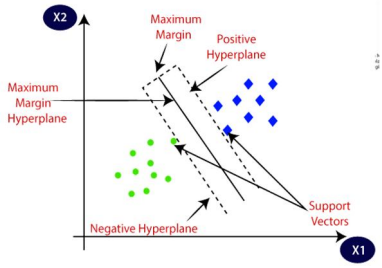


Fig. 3. SVM

4) *Logistic Regression:* Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. 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. In Logistic regression, instead of fitting a regression line,we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). 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. The below image is showing the logisticfunction:

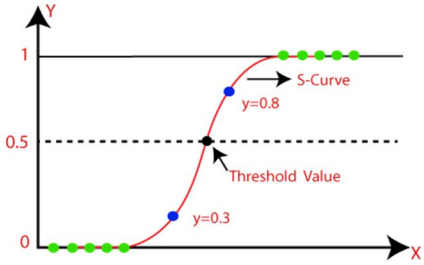


Fig. 4. Logistic Regression

5) *Naive Bayes:* Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naive Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles. Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Fig. 5.
fig: my*label*

IV. ACKNOWLEDGMENT

I would like to acknowledge all the people who have helped and assisted us throughout our project work. First of all I would like to thank our respected guide Prof. S.S. Adagale , for introducing us throughout to the features needed. The time to time guidance, encouragement, and valuable suggestions received from her. I also sincerely thank for the time spent proofreading and correcting many of my mistakes. Further more, we would like to thank respected Dr. N. J. Uke Principal and Dr. M. B. Wagh, Head of the Department of Computer Engineering for the guidance provided by her during our project work. We are also grateful to all the faculty member of Trinity Academy of Engineering, Pune for their support and cooperation.
We would specially like to thank all our friends for their valuable suggestions. We would like to express our gratitude and appreciation to all those who gave us the possibility to complete this project. The acknowledgement would be incomplete without mention of the blessing of the Almighty, which help us in keeping high morale during most difficult situation.

REFERENCES

[1] M. Ghafari, S. Yakhchi, A. Beheshti, and M. Orgun, "Social con- textaware trust prediction: Methods for identifying fake news," in Web Information Systems Engineering—WISE 2018. Cham, Switzer- land:Springer.
[2] F. Ferna´ndez-Reyes and S. Shinde, "Evaluating deep neural networks for automatic fake news detection in political domain," in Proc.
[3] H. Ahmed, I. Traore, and S. Saad, "Intelligent, secure, and dependable systems in distributed and cloud environments," in Proc. 1st Int. Conf. Intell., Secur. Dependable Syst. Distrib. Cloud Environ., vol. 10618, 2017, pp. 169–181, doi: 10.1007/978-3-319-69155-8.
[4] L. Moreira-Matias, J. Mendes-Moreira, J. Gama, and P. Brazdil, "Text categorization using an ensemble classifier based on a mean co- association matrix," in Machine Learning and Data Mining in Pattern Recognition (Lecture Notes in Computer Science), vol. 7376. Berlin, Germany: Springer, May 2014, pp. 525–539, doi: 10.1007/978-3-642- 31537-441.
[5] S. Diab, "Optimizing stochastic gradient descent in text classifica- tion based on finetuning hyper-parameters approach?: A case study on automatic classification of global terrorist attacks," Feb. 2019, arXiv:1902.06542.
[6] A. Bali, M. Fernandes, S. Choubey, and M. Goel, "Comparative perfor- mance of machine learning algorithms for fake news detection," 2019, doi:10.1007/978-981-13-9942-840.
[7] S. T. Indra, L. Wikarsa, and R. Turang, "Using logistic regression method to classify tweets into the selected topics," in Proc. Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS), Oct. 2016, pp. 385–390, doi: 10.1109/ICACSIS.2016.7872727.
[8] D. Keskar, S. Palwe, and A. Gupta, Fake News Classification on Twitter Using Flume, N-Gram Analysis, and Decision Tree Machine Learning Technique. Singapore: Springer, 2020, pp. 139–147,doi: 10.1007/978- 981-15-0790-8-15.
[9] J. T. Pintas, L. A. F. Fernandes, and A. C. B. Garcia, Feature Selection Methods for Text Classification: A Systematic Literature Review, no. 123456789. Cham, Switzerland: Springer, 2021

[10] S. Yang, K. Shu, S. Wang, R. Gu, F. Wu, and H. Liu, "Unsupervised fake news detection on social media: A generative approach," Proc. AAAI Conf. Artif. Intell., vol. 33, Jul. 2019, pp. 5644–5651, doi: 10.1609/aaai.v33i01.33015644.

[11] K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsTracker: A tool for fake news collection, detection, and visualization," Comput. Math. Org. Theory, vol. 25, no. 1, pp. 60–71, Mar. 2019, doi: 10.1007/s10588-018- 09280-3.

[12] H. I. Celenli, S. T. Ozturk, G. Sahin, A. Gerek, and M. C. Ganiz, "Document embedding based supervised methods for Turkish text classi- fication," in Proc. 3rd Int. Conf. Comput. Sci. Eng. (UBMK), Sep. 2018, pp. 477–482, doi: 10.1109/UBMK.2018.8566326.

[13] H. Saleh, A. Alharbi, and S. H. Alsamhi, "OPCNN-FAKE: Optimized convolutional neural network for fake news detection," IEEE Access, vol. 9, pp. 129471–129489, 2021, doi: 10.1109/ACCESS.2021.3112806.

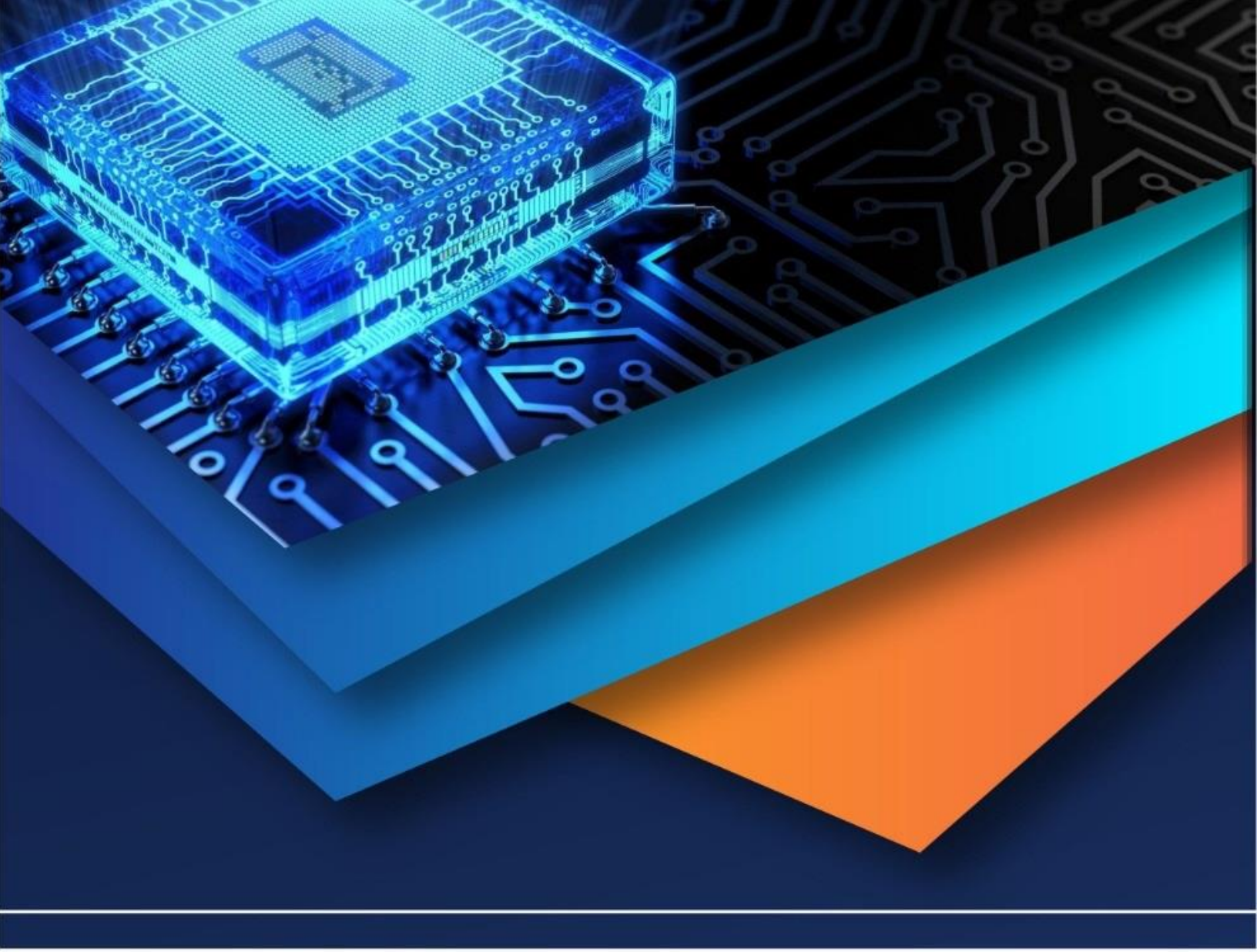
[14] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved chi-square for Arabic text classification," J. King Saud Univ. Comput. Inf. Sci., vol. 32, no. 2, pp. 225–231, Feb. 2020, doi: 10.1016/j.jksuci.2018.05.010.

[15] H. Ko, J. Y. Hong, S. Kim, L. Mesicek, and I. S. Na, "Human-machine interaction: A case study on fake news detection using a backtracking based on a cognitive system," Cognit. Syst. Res., vol. 55, pp. 77–81, Jun. 2019, doi: 10.1016/j.cogsys.2018.12.018.

[16] C.-S. Atodiresei, A. Ta˘na˘selea, and A. Iftene, "Identifying fake news and fake users on Twitter," Proc. Comput. Sci., vol. 126, pp. 451–461, Jan. 2018, doi: 10.1016/j.procS.2018.07.279.

[17] Y. Madani, M. Erritali, and B. Bouikhalene, "Using artificial intelligence techniques for detecting COVID-19 epidemic fake news in Moroccan Tweets," Results Phys., vol. 25, Jun. 2021, Art. no. 104266, doi: 10.1016/j.rinp.2021.104266.

[18] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach," Int. J. Inf. Man- age. Data Insights, vol. 1, no. 1, Apr. 2021, Art. no. 100007, doi: 10.1016/j.jjimei.2020.100007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY