A Pinnacle Technique for Detection of COVID-19 Fake News in Social Media

R. Sandrilla, M. Savitha Devi

*Abstract***:** *Today the world is gripped with fear of the most infectious disease which was caused by a newly discovered virus namely corona and thus termed as COVID-19. This is a large group of viruses which severely affects humans. The world bears testimony to its contagious nature and rapidity of spreading the illness. 50l people got infected and 30l people died due to this pandemic all around the world. This made a wide impact for people to fear the epidemic around them. The death rate of male is more compared to female. This Pandemic news has caught the attention of the world and gained its momentum in almost all the media platforms. There was an array of creating and spreading of true as well as fake news about COVID-19 in the social media, which has become popular and a major concern to the general public who access it. Spreading such hot news in social media has become a new trend in acquiring familiarity and fan base. At the time it is undeniable that spreading of such fake news in and around creates lots of confusion and fear to the public. To stop all such rumors detection of fake news has become utmost important. To effectively detect the fake news in social media the emerging machine learning classification algorithms can be an appropriate method to frame the model. In the context of the COVID-19 pandemic, we investigated and implemented by collecting the training data and trained a machine learning model by using various machine learning algorithms to automatically detect the fake news about the Corona Virus. The machine learning algorithm used in this investigation is Naïve Bayes classifier and Random forest classification algorithm for the best results. A separate model for each classifier is created after the data preparation and feature extraction Techniques. The results obtained are compared and examined accurately to evaluate the accurate model. Our experiments on a benchmark dataset with random forest classification model showed a promising results with an overall accuracy of 94.06%. This experimental evaluation will prevent the general public to keep themselves out of their fear and to know and understand the impact of fast-spreading as well as misleading fake news.*

Keywords: **Fake News, Corona Virus, COVID-19, Naïve Bayes, Random forest, Machine Learning, Preprocessing.**

I. INTRODUCTION

Social media is becoming very popular now-a-days, especially in the consumption of news due to the fast dissemination of information, free and easy access by the millions of users in the globe.

R. Sandrilla*, Assistant Professor Department of Computer Science at Sacred Heart College (Autonomous) Tirupattur and pursuing her Ph.D. degree in Periyar University, Bangalore, India.
M. Savitha Devi, Department of Computer Science, Periyar University Constituent College of Arts and Science, University, Bangalore, India.

This enables the broad propagation to false information which is known as fake news. One major reason for spreading this type of fake news is mainly because to confuse and mislead the readers. In spite Coronavirus fake news in social media is no exception. Corona virus is a transmissible disease mainly related to SARS virus. This virus was initially found in Wuhan city, in late December 2019.

Later this issue has been raised as a global one and was declared a pandemic to the world. According to the World Health Organization (WHO) the epidemic disease worsens and spreads to about one eighth of the countries in the world. COVID-19 started to affect diverse people in different ways. People get infected and develop mild to moderate illness. So, the public from various places happened to expose their thoughts about covid-19 on various platforms. This paved the way for the creation and development of fake news about COVID-19. When people are in need of access to high quality evidence still, they face strong barriers to take suggested actions. To understand the misleading information WHO has developed to support infodemic to fight against the epidemic disease and to stop the reader from misleading evidence. The term infodemic has been coined to outline the perils of distortion phenomena during the management of virus outbreaks since it could even speed up the epidemic process by inducing and crumbling social response [1]. It is something with a surplus of information which can be either true to the evidence and some may not. The aim of the infodemic is not to eliminate but to manage the situation to prevent and to respond to the misleading information. Because collecting the data from the social media or the other news media platforms without fact checking may severely affect the general public's lifestyle, harmony, the old one's emotional psychology and their mental behaviors. In the fast spread of COVID 19 the globe is in grief and facing the situation of identifying the fake or the real news propagation. The real news helps to mitigate the calamity whereas the fake one amplifies it. This erratic situation inspires the researcher to understand and to develop a model which will help to recognize the changing nature of news in social media related to COVID 19 pandemic. This investigation will help to analyze and detect fake news in social media with less effort. It mainly focuses on social media fake news related to COVID-19 the open source dataset has been collected for detection of misinformation. A machine learning-based classifier is built to detect the misinformation at the time of the pandemic. We also performed extensive testing of our model in different challenging settings, showing that it achieves very high accuracy. The following discussion is divided into four main sections.

Retrieval Number: 100.1/ijitee.A81761110120
DOI: 10.35940/ijitee.A8176.1110120
Journal Website: www.ijitee.org

256

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

A Pinnacle Technique for Detection of COVID-19 Fake News in Social Media

The literature review discusses and summarizes the findings from previous relevant studies in this area. In the methodology section, the data collection, data annotation, data cleaning and prepossessing, data exploration are presented in detail. The classification result section illustrates the major and minor findings of this study.

Lastly, the discussion section concludes with the major findings based on relevant arguments and previous findings. Some limitations and strengths of this study, the future works are also discussed in this section.

II. LITERATURE REVIEW

The Outbreak of COVID 19 exposed to another parallel problem of epidemic led by fake news. This problem of misleading information interrupts the public health communication and also leads to mass fear and anxiety. In the meanwhile, few researchers investigated their studies and brought forward the issues in highly scholar field related to misleading information produced by the social networks on this COVID 19 pandemic. The investigation on fake news propagation deploys various machine learning techniques to help the public in classifying the data that they are viewing is fake or not. This analysis is done using the untruthful content by classifying and comparing the given data with some pre known data that which contains both misleading and fake information [9],[10]. Typically, in order to develop a working model to detect the misleading information practicing the machine learning techniques and methods are very important. The machine learning techniques must undergo certain stages to completely train and develop a detection model starting from data preprocessing to feature detection and extraction. [11],[12] states that the stages starting from data preparation, data selection and data extraction facilitates in handling the huge amount of data required for building a detection model. There exist many misleading information detection websites however most of the websites are human based where the analysis done in this detection model is manual. Though the manual detection is done by highly expertise people it has few drawbacks as high expensive, slow in process, highly subjective, very tuff in handling the large amount of data. Due to this [13] proposed an integration of Machine learning and knowledge engineering that can be helpful in detection of fake news. Hence [14] proposed an automated classification that represents a prolific trend of study. Many researchers worked on detecting fake news detection model [15] surveyed on many automated misleading information systems and proposed a few detection models that leads to detection of mis information. In this context, misinformation related to COVID 19 has been increasing in a fast pace due to pandemic. They were quick to spread deceptive information. [16] reasoned out on Description Logics to detect inconsistencies between the trusted and untrusted news. A novel mathematical model is created to show the spread of COVID 19 in Smart cities. The study proposed an integrated framework with machine learning techniques to prevent the spread of COVID 19 fake news.[17]. Since COVID 19 pandemic and infodemic spread is in parallel it is very essential to propose a framework to fight against the fake news. A multilingual cross domain dataset for COVID 19 has been collected and a machine learning based classifier is built to detect the misinformation at the time of the COVID pandemic [18]. The fake news about the corona virus has become an increasing Fame in the diffusion of shaping the news histories online. The videos posts and articles related to the current pandemic are extensively disseminated throughout the social network platforms. The researcher proposed a novel multi-level voting ensemble model using twelve classifiers combined to predict based on their false prediction ratio [19]. Spreading of this COVID 19 pandemic disease not only changed the healthy behavior among the globe but rapidly given space to spread the false news in directing or misleading the public in numerous rumors.. Inspite of this rumors public not only had very serious impact on their healthy lifestyle but erroneous increase in the spread of virus along with the fake information among the public disturbed their mental health and brought them to serious stress. An article [20] stated that an old man father of three was report to commit suicide upon hearing his diagnosis of COVID 19.

III. PROBLEM STATEMENT

The Dataset used in this paper is the collected tweets about the pandemic diseases COVID-19. As the outbreak of COVID-19 is still existing, radical preventive measures are taken to crack the situation and to lead economically balanced lives. To directly report the issue of false news creation and dissemination, through various online social platforms. To address the issue of fake news generation and dissemination through various online social platforms, an appropriate feature extraction technique is chosen to improve the efficiency of the existing ML classifiers. A unique and efficient approach for the detection of fake news using machine learning model will be proposed to develop an efficient fake news detection system. Mathematically, the problem statement can be represented as- To identify $S = \{fake, real\}$ for a document D where $D = \{d1, d2, ..., dn\}$ and tx represents the text in a news article ax chosen from corpus with series of engagements that is composed of title, body and label of the article as $exyz = (tx , by , lz)$. The task is to evaluate and analyze the best feature extraction method FE where $E = \{Tf\text{-}Idf, CV, HV\}$ using machine learning classifier to compute high efficiency in our project

IV. PROPOSED METHOD

The proposed approach extracts the features of the posted or shared content in a social network without modifying its original context. Hence, it focuses on maintaining the textual words in a sentence in its original structure until given as the input to the learning classifiers. By applying the classification model, the proposed approach learns the represented text model and classifies the textual data into real news and fake news. Then, the proposed approach takes the collaborative decision from the classified data that are generated from the base learners using the weight-based method. Finally, the proposed approach classifies the real and fake news.

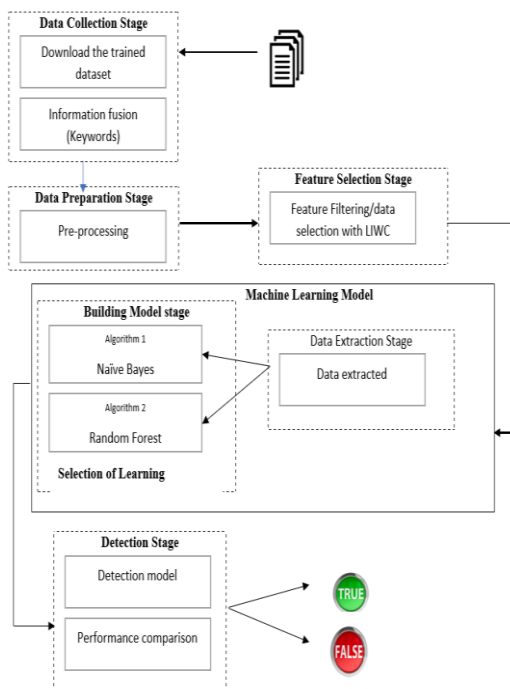The detailed architecture is illustrated in the Fig. 1.

Fig.1. A Proposed Fake news detection framework

A. The Dataset Collection

There are many ways to disseminate the fake news about COVID 19 some such sources are Facebook, twitter, etc., The Dataset used here in this proposed system is news collected from the Kaggle termed as Corona.csv. the dataset has a shape of 1148×4. The first column identifies the news, the second and the third are the title and the fourth column has labels representing the news is either Real or False, respectively.

B. Preprocessing

In order to deliver a precise and better input for more consistent outcomes to detect and prevent COVID-19 cases, data pre-processing is considered as a significant stage. The first step in preprocessing is to extract all the relevant COVID-19 data from the storage. The plenty of unused strings or characters which exists in the label column are removed. The second step is to preform data fusion where the collected data are integrated to produce more reliable, accurate, and valuable information. The third step during pre-processing of COVID-19 data is to preform dimensionality reduction, in which the number of variables is reduced by extracting a set of main variables. Fourth step the extraction is done linguistically using LIWC dictionary, we first extracted verb-noun dependencies from the British National Corpus, where nouns are either direct or prepositional objects to verbs. Here the stop words are removed. Besides fifth steps focus on feature extraction and selection. These two methods are very important because it can be used to filter irrelevant or redundant features from the selected datasets. The redundant textual strings are removed from the corpus using a regular expression (re) in the next step as shown in Figure 1. The re and panda's library has been used to perform the pre-processing task [2]. The re- regular expression library in phyton is used to define a search pattern using a sequence of characters whereas clean method from pandas is used for cleaning the missing values in python Data

Frame. The possible activities in these steps including Wrappers, Filters, and Embedded. The last step is a basic statistical analysis on COVID-19 data in order to interpret the data before knowledge-based algorithms are applied.

C. LIWC Generated Feature

Linguistic Inquiry and Word Count (LIWC) dictionary includes a word classification and count tool. In this study, to improve the classification accuracy of the fake news detection and to reduce the computation time in large scale comments a model is developed with combinations of classifiers using Linguistic Inquiry and Word Count (LIWC) dictionary. LIWC dictionary is used during the preprocessing phase. LIWC reads the texts from a given dataset and its text analysis module then compares each word in the text against a user-defined dictionary. The dictionary identifies which words are associated with which psychologically-relevant categories. Then it calculates the percentage of total words that match each of the dictionary categories. LIWC can be used in computational linguistics as a source of features for deception detection [3],[4].

D. Feature Extraction

The number of new features is analyzed and extracted based on the titles and news articles. In order to extract numerical features from a textual document, tokenization, counting and normalization is done. The features such as capital letter in the text, stop words, Proper Nouns, POS, Type token ratio of each article are analyzed in the corpus. Here the capital letter in the title of the article are extracted and identified that the fake news uses a greater number of capital letters in the title to attract the audience by their headlines. Stop words are removed and the proper noun has been increased in the corpus, and the POS tag keeps the account on counting the number of times that each tag is written in the articles

E. Classification Algorithms

The processed dataset retrieved after pre-processing and feature extraction phase is then fed to the classification phase for the identification of fake news article. In this research two models have been used in in order to predict the accuracy of the classification between the real and fake from the corona news dataset, they are Naïve Bayes Classifier and Random forest.

Naïve Bayes

Naïve Bayes is a renowned classification method. It is a part of simple machine learning algorithms. It is well-known and most popular algorithm used in former as well as later cases to find the accuracy of news to predict the real and true. There is numerous algorithm which follow the same principle, in the classification method. One among those are naïve Bayes. This is a classification algorithm particularly used for text classification. It uses Bayes theorem in order to find the accuracy of the given data by independently accepting the previous predictors. This is an easy build model with no iterative parameter, especially used for very large set of data. Inspite of many machine learning algorithms naïve Bayes are widely used because of its sophisticated classification methods.

258

Normally this algorithm works on the previous or on the posterior probability. It assumes that the that the result of the value of a predictor (x) on a given class () is independent of the values of other predictors. This assumption is called class conditional independence [5].

The following is the formula for naive Bayes classification fake.

$$P(A|B) = P(B|A) \cdot P(A) / P(B), \ldots\ldots\ldots (1)$$

Finding the probability of event, A when event B is TRUE There is multiple naïve Bayes algorithm and, in this study, the used algorithm is multinomial Naive Bayes, since it can be implemented used easily for text classification in the identification of number of occurrences of a given particular word. The news data is divided into train data and the test data in order to predict the test data first. Using the confusion matric the predicted values are compared with the truth test.

Random Forest Model

Random forest is another interesting type of classification algorithm developed by Leo Breiman [6] that uses an ensemble of classification trees [7],[8].

This is also one type of important classification method to clearly classify the text data.

This is considered to be an efficient classification algorithm, since it combines multiple decision trees models results together and forms as ensemble classifier. Here the considered corona datasets are split into subsets and then runs on the data

V. EXPERIMENTS AND RESULTS

In this section, we report on the experimental setup, evaluation results, and the necessary analysis

A. *Experimental Setup*

All the experiments are conducted in a python environment.

The libraries of python are NumPy, Pandas, Sklearn. We have used certain libraries of phyton such as Naïve Bayes as the main model for Multinomial Naive Baye**s,** NumPy, pandas, Sklearn.

Fortunately, Sklearn provides an easy-to-implement object called *MultinomialNB(),* Sklearn ensemble for random forest classifier. Few of the models and their imports are listed in the table 1 which are required to carry out our experiments.

Table- I: Models and their Sub models

Main Model	Imports
sklearn.naive_bayes	MultinomialNB
sklearn.ensemble	RandomForestClassifier
sklearn.model_selection	train_test_split
matplotlib.pyplot	plt
sklearn.metrics	confusion_matrix
seaborn	Sns
sklearn.feature_extraction	CountVectorizer
nltk.tokenize	word_tokenize

The samples taken are prepared for data analysis and the classification algorithms are applied.

After training the model, a confusion matrix is created in order to predict exactly the number of misclassified samples.
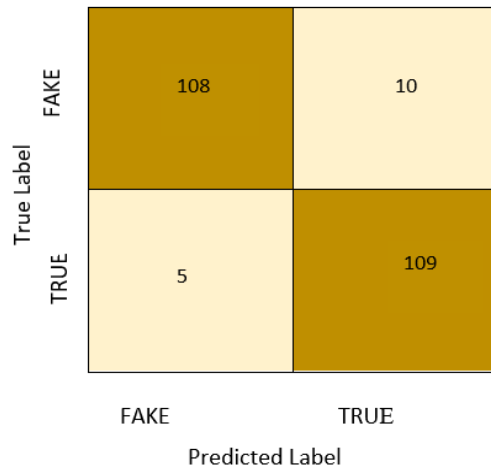

Fig.2. Confusion matrix of Naïve Bayes model on COVID-19 news dataset.

Preparing the model using Random forest classification Algorithm.

After the completion of naïve Bayes classification prediction. The random forest classification algorithm is used to observe the confusion matrix.

This is done in order to predict the highest accurate classification model.
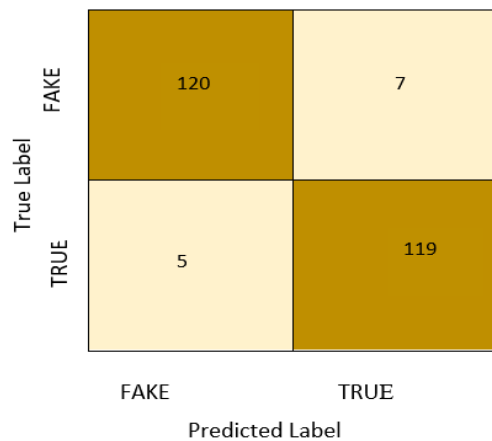
Fig.3. Confusion matrix of Random Forest model on COVID-19 news dataset

VI. EVALUATION AND RESULT

From the above confusion matrix, we evaluate the performance of the classifiers with below mentioned metrics shown in Table II.

The experimental results from the model is actually predicted and the performance of the classifier is tabulated in Table III with the terms of Precision, Recall, Accuracy and F1 Measure is shown below.

Here the TP and TN that denote the number of positive and negative instances that are correctly classified, while FP and FN that denote the number of wrongly classified positive and negative instances, respectively.

259

Table- II: The Evaluation Metrics.

| #	Metrics Used	Formula
1	Precision	TP/TP+FP
2	Recall	TP/TP+FN
3	F1 Measure	2(P*R)/(P+R)
4	Accuracy	TP+TN/TP+TN+FP+FN

VII. RESULT

We evaluate the performance of the system in terms of Precision, Recall and Accuracy and F1 Measure.

Table- III: The Experimental Result

Model Used	Metrics			
	Precision	Recall	Accuracy	F1 measure
Naive Bayes	95.05%	91.05%	95.05%	93.5%
Random Forest	96.6%	92.08%	96.06%	94.06%

After implementing classification algorithms, a comparison table is prepared for comparing both the Machine learning algorithms. Out of which Random forest gives better accuracy compared to naïve Bayes classification algorithm.
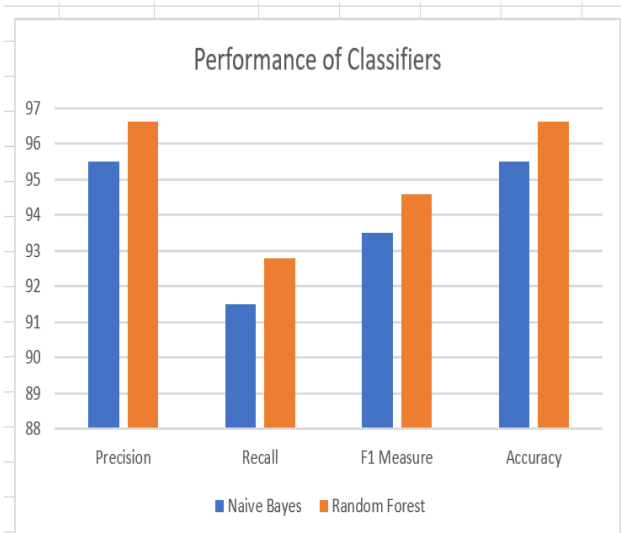
Fig.4. Comparison of Classification Accuracy

VIII. CONCLUSION AND DISCUSSION

In this paper we have investigated and presented a new model for detecting the fake news about the serious COVID 19, through two different machine learning techniques. The reliable dataset is used to build the model with two different classification algorithms. Furthermore, we have investigated the two methods and compared their accuracy. The Model that achieves a highest accuracy is Random Forest and its highest accuracy score is 96.06%. Therefore, we conclude that by using the proposed model any news related to COVID 19 or other news from small to large dataset can be classified as true or fake news with less time which in turn helps the users to be strong enough or not to get confused in this pandemic situation. This Also helps the user to believe in a particular news. Though fake news detection is growing hot topic in the emerging research area it has a scarce in number of datasets. It is quite a difficult task in collecting the real time news or regarding the current affairs. The proposed model is a benchmark dataset that is executed with an existing dataset, showing that the model performs well against it.

In our future work more news articles related to the pandemic can be considered and collected in the corpus of data for more perfection, since Coronavirus is still around up until the time we investigated with this model.

REFERENCES

1. Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carl Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. "The covid-19 social media infodemic". arXiv preprint arXiv:2003.05004, 2020.
2. O. Batchelor, Reference Services Review 45(2), 143 (2017)
3. Angel Hern´andez-Casta neda and Hiram Calvo. Deceptive text detection using continuous semantic space models. ˜ Intelligent Data Analysis, 21(3):679–695, 2017.
4. Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 309–312. Association for Computational Linguistics, 2009.
5. An empirical study of the naive Bayes classifier I. Rish
6. Breiman L: Random forests. Machine Learning 2001, 45: 5–32. View Article Google Scholar
7. Breiman L, Friedman J, Olshen R, Stone C: Classification and regression trees. New York: Chapman & Hall; 1984.Google Scholar.
8. Ripley BD: Pattern recognition and neural networks. Cambridge: Cambridge University Press; 1996.View ArticleGoogle Scholar.
9. J. Y. Khan, M. T. I. Khondaker, A. Iqbal, and S. Afroz, ''A benchmark study on machine learning methods for fake news detection,'' 2019, arXiv:1905.04749. [Online]. Available: http://arxiv.org/abs/1905. 04749
10. M. K. Elhadad, K. F. Li, and F. Gebali, ''A novel approach for selecting hybrid features from online news textual metadata for fake news detection,'' in Proc. 3PGCIC, Antwerp, Belgium, 2019, pp. 914–925.
11. A. K. Tyagi, ''Machine learning with big data,'' in Proc. SUSCOM, Jaipur, India, 2019, pp. 1011–1020.
12. M. K. Elhadad, K. F. Li, and F. Gebali, ''Fake news detection on social media: A systematic survey,'' in Proc. IEEE PACRIM, Victoria, BC, Canada, Aug. 2019, pp. 1–9, doi: 10.1109/PACRIM47961.2019.8985062.
13. Sajjad Ahmed, Knut Hinkelmann, Flavio Corradini," Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks-a survey", In A. Martin, K. Hinkelmann, A. Gerber, D. Lenat, F. van Harmelen, P. Clark (Eds.), Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE 2019). Stanford University, Palo Alto, California, USA, March 25-27, 2019.
14. R. Kumar and R. Verma, ''KDD techniques: A survey,'' Int. J. Electron. Comput. Sci. Engr., vol. 1, no. 4, pp. 2042–2047, Aug. 2008.
15. R. K. Kaliyar and N. Singh, ''Misinformation detection on online social media—A survey,'' in Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT), Jul. 2019, pp. 1–6, doi: 10.1109/ICCCNT45670.2019. 8944587.
16. Detecting fake news for the new coronavirus by reasoning on the Covid-19 ontology Adrian Groza Computer Science Department, Technical University of Cluj-Napoca, Memorandumului 14, Cluj-Napoca, Romania adrian.groza@cs.utcluj.ro April 26, 2020
17. A Machine Learning Solution Framework for Combatting COVID-19 in Smart Cities from Multiple Dimensions Ibrahim Abaker Targio Hashem1 , Absalom E. Ezugwu2, * , Mohammed A. Al-Garadi3 , Idris N. Abdullahi4 , Olumuyiwa Otegbeye2 , Queeneth O. Ahman5 , Godwin C. E. Mbah5 , Amit K. Shukla6 , Haruna Chiroma
18. FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19 Gautam Kishore Shahi 1 and Durgesh Nandini2 University of Duisburg-Essen, Germany1 ,University of Bamberg, Germany2 gautamshahi16@gmail.com 1durgeshnandini16@yahoo.in Copyright c 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved. 1 https://gautamshahi.github.io/FakeCovid/

19. Automating Fake News Detection System using Multi-level Voting Model Sawinder Kaur · Parteek Kumar · Ponnurangam Kumaragur
20. Joe W. Coronavirus: Indian man "died by suicide" after becoming convinced he was infected. The Telegraph. https://www.telegraph.co.uk/global-health/science-and-disease/coronavirus-indian-man-died-suicide-becoming-convinced-infected/. Published 2020.

AUTHORS PROFILE

R. Sandrilla completed her Master of Computer Applications from Mount Carmel College, Bangalore and Master of Philosophy in Computer Science from KMG arts and Science college, Gudiyattam. Currently she is working as Assistant Professor in the Department of Computer Science at Sacred Heart College (Autonomous) Tirupattur and pursuing her Ph.D. degree in Periyar University, Salem. She is having 10 years of Teaching experience and 4 years of Research experience. She has published five research papers in National, International journals and Conference proceedings. She has presented two research papers in National and International conferences. Her research areas include Web Mining and Machine Learning.

M. Savitha Devi completed her doctorate in Mother Teresa Women's University, Kodaikanal. She has completed her M.Phil. from Periyar University, Salem, and M.Sc. Computer Science in Vysya College, Salem where she got University 11th rank in the year 2003. She completed her under graduate course in Government Arts College, Salem. Currently working as an Assistant Professor and Head, Department of Computer Science, Periyar University Constituent College of Arts and Science, Harur. She has 16 years of Teaching experience and 10 years of Research experience. She has published more than 25 research papers in National, International journals and Conference proceedings. She has published books related to network security. Her research areas include E-Content Development, Network Security, Web Mining and Machine Learning. She is currently acting as a University Swayam Mentor.