

Fake News Detection system using Decision Tree algorithm and compare textual property with Support Vector Machine algorithm

N. Leela Siva Rama Krishna

Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha University,
Chennai, India
ramaramakrishnakrishna1999@gmail.com

M. Adimoolam

Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha University,
Chennai, India
adimoolamm.sse@saveetha.com

Abstract— To find out precise Fake News Detection on social media, the Decision Tree (DT) machine learning algorithm has to be worked out and also compare textual property accuracy with Support Vector Machine (SVM) machine learning algorithm. **Materials and Methods:** The analysis was done for fake news detection using machine learning algorithms such as the DT algorithm (N=311) and SVM algorithm (N=311) as a proposed research. **Results:** The DT algorithm and SVM algorithm's accuracy for social media fake news were experimented and measured. The DT machine algorithm accuracy measures appears to be 97.67% and it is better than the SVM algorithm accuracy and it appears to be 91.74%. There is a statistically significant difference among the study groups with significance value 0.092 for accuracy and 0.825 for precision for Confidence Interval (CI) 95%. **Conclusion:** The DT machine algorithm is used in examining whether a piece of news from social media is fake or not with accuracy that appears to be better than the SVM algorithm machine learning algorithm.

Keywords— *Decision Tree, Support Vector Machine, Machine Learning, Novel Fake News Detection, Probability, Social media.*

I. INTRODUCTION

News information is easily accessible through an informational database of Facebook news posts as a product context. The format of social media generates massive amounts of information [1]. The DT algorithm is used to detect fake news by combining key learning and man-made technology concepts. Many social media platforms have been focused on spreading fake news, and such fake news can have significant effects. It is extremely useful for people to be able to track their open benefit events online. This work achieved a classification accuracy of around 97 % on the testing area, which is a respectable output for the model's general ease of use with the Internet and online media [2]. Since news becomes more frequent, people become less aware of whether or not the news is true, taking away attention from the real issue. The information list is loaded, which involves the collection of news, after which the data must be merged into the test and train data, and pipelining must be performed to evaluate precision. The software used DT as a fake news detection algorithm [3]. In this research, accuracy identification in fake news was compared using algorithms such as SVM, Linear Regression, Support Vector Machine, and DT. Twitter news was used as the data set in this research. Using machine learning methods, nearly 35 articles on Fake News Detection were published. Frequent interactions have

an important role in influencing public opinion, and as is typical, a few groups aim to profit from it. Wide exchanges balance information in their own unique way from time to time in order to deal with arrival at their destination [4]. This research will discuss and compare various approaches to dealing with this issue, including some standard machine learning approaches such as DT and popular classification algorithms. Fake data can be found in a wide range of areas. They are trying to spread statements, creations, and fake data under the pretext of real news. Their primary purpose is to keep control of information that can be used to determine whether or not they trust it [5 - 6]. These types of sites can be found around the world. As a result, fake news affects people's views. According to researchers, various highly automated idea statistics will help in the detection of fake news [7]. To better understand the characterisation of information in the modern community, as well as the various substance types of reports and their impact on creators, consider current phoney news recognition methods that focus heavily on text-based investigation, as well as usual fake news datasets [8]. Fake news has been around for decades, and with the rise of social media and modern-day journalism, detecting media-rich fake news has become a popular issue in the research community. Given the difficulties associated with detecting fake news research, researchers all over the world are attempting to understand the basic aspects of the problem statement [9]. The best referred paper [10] that implements a productive procedure for detecting fake news. Previously our team had rich experience in working on various research projects across multiple disciplines. [11 - 14]. Now the growing trend in this area motivated us to pursue this project.

There was a gap in the existing systems that approached DT and SVM. Because it may be difficult to detect fake news with accuracy and precision. More values must be added to the dataset in order for the model to be trained to predict accurately. This is due to the fact that man-made cognitive knowledge will eventually become significant, and various instruments will be available to partly verify it. As an author, the machine learning based price prediction solution has been carried out [15 - 19]. The purpose of this research is to use DT and SVM algorithms to detect fake news on social media platforms.

Table 1 Lists the precision and accuracy of DT and SVM algorithms. SVM accuracy appeared to be 90 % to 91.7%

approximately for 10 iterations, and DT accuracy appeared to be 90 % to 97% approximately. The precision of the DT algorithm appears to be higher than that of the SVM algorithm.

Samples	Accuracy (%)		Precision(%)	
	SVM	DT	SVM	DT
1	91.74	97.67	90.12	94.60
2	91.62	96.81	89.92	93.86
3	91.53	96.48	89.54	93.19
4	91.46	96.23	89.37	92.73
5	91.32	95.94	88.96	93.27
6	91.23	95.74	89.68	92.80
7	90.73	95.32	89.35	92.27
8	90.68	95.01	88.73	91.82
9	90.42	94.89	87.43	91.53
10	90.15	94.51	86.18	90.68

Table 2 lists the Accuracy, Precision, Recall, F-measure of DT and SVM algorithms. The DT algorithm appears to perform better than the SVM algorithm.

Model	Accuracy	Precision	Recall	F-Score
SVM	0.91	0.89	0.73	0.75
DT	0.95	0.93	0.78	0.82

Table 3 Statistical Analysis of Mean, Standard Deviation, and Standard Error of Precision and Accuracy of SVM and DT algorithms. DT appeared Accuracy (97.67%) and Precision (94.60%). SVM Accuracy (91.74%) and Precision (90.12%). The standard error is also less in DT compared to SVM.

Algorithm		N	Mean	Standard Deviation	Standard Error Mean
Accuracy	SVM	10	91.1880	0.46852	0.14816
	DT	10	96.9310	0.51954	0.16429
Precision	SVM	10	88.3540	1.23430	0.39032
	DT	10	92.4950	1.20703	0.38170

II. MATERIALS AND METHODS

This research was carried out in the department of Computer Science and Engineering, Saveetha School of Engineering, SIMATS. Two groups taken for research to be carried out. One group is the DT algorithm and the other group is the

SVM algorithm. To perform the Fake News Detection two required datasets were used [20]. There are two news datasets, one with real news and the other with fake news. These can be used to differentiate between real and fake news, as well as to show accuracy based on the news' information. There are 4 attributes and 311 instances in the data set. Using the G power of 80% and the alpha value of 0.005, the sample size is computed to be 23 in each group.

In this study, the DT algorithm is applied to group 1. In both basic and deep machine learning, DT classifiers are required [20]. The Decision Node and the Leaf Node are the two nodes of a Decision tree. Leaf nodes are the output of those decisions and do not contain any more branches, whereas Decision nodes are used to make any decision and have several branches. It is a kind of classification algorithm. In a DT classifier, the use of tokens is compared to news that may or may not be fake, and the accuracy of the news is calculated using Back propagation, which is better than an SVM.

The steps to be carried out for the DT algorithm as follows.

Inputs - Training and Testing data

Outputs - A class of testing data and Accuracy

1. *The data set has been Trained.*
2. *The mean and standard deviation of the predictor variables were calculated.*
3. *Using the gauss density equation, calculate the probability of the predictor variable on a frequent basis. Until all predictor variable values have been calculated.*
4. *For each class, I calculated the probability.*
5. *The highest probability has been obtained.*

In this study, the SVM machine algorithm is kept as group 1. SVM is a machine learning technique that can be used for classification and regression (Reis et al. 2019). The data points are sorted by hyperplane in the SVM method. The SVM algorithm is used to create hyperplanes in an N-dimensional space with N-number of features.

The steps to be carried out for the SVM algorithm as follows.

Inputs - Training and Testing data

Outputs - A class of testing data and Accuracy

1. *The data set has been loaded.*
2. *The dataset was randomly divided into training and testing datasets.*
3. *The target variable has been identified.*
4. *The classifier was created using the training dataset.*
5. *The testing dataset was predicted.*
6. *The classifier was evaluated.*

The testing procedure of the Novel Fake News Detection System is as follows

- 1) *The data from the Dataset is collected and preprocessed.*
- 2) *The data is collected, inserted into the testing data, and vectorized.*
- 3) *The DT algorithm is used for different groups of data that have already been preprocessed.*

Table 4 Comparison - Significance level for SVM and DT algorithms with value $p = 0.05$. Both SVM and DT have a significance level less than 0.05 with a 95 % confidence interval.

		Levene's test for equality of variables		T-test for Equality of means						
				F	Sig	t	df	Sig (2-tailed)	Mean Difference	Std. Error Difference
										Lower
Accuracy	Equal variances assumed	.009	.092	-25.959	18	.000	-5.74300	.22123	-6.20779	-5.27821
	Equal variances not assumed			-25.959	17.811	.000	-5.74300	.22123	-6.20814	-5.27786
Precision	Equal variances assumed	.048	.829	-7.585	18	.000	-4.14100	.54593	-5.28796	-2.99404
	Equal variances not assumed			-7.585	17.991	.000	-4.14100	.54593	-5.28800	-2.99400

- 4) The word count is generated, and the result is displayed based on whether the news is true or false based on the popular keywords.

The proposed system's testing setup is as follows: Hardware-Desktop with 64-bit operating system, 50GB disc space, 8 GB RAM, and Windows 10, Python 3.8, and Google Colab are the software requirements. To determine the accuracy and precision, 10 helps the group observations were held. For our study, we used the IBM SPSS tool as statistical software. In our fake news detection analysis, text and subject are independent variables, while target is the dependent variable. The Independent Sample T-Test was performed using the values obtained from the iterations, and graphs were also plotted.

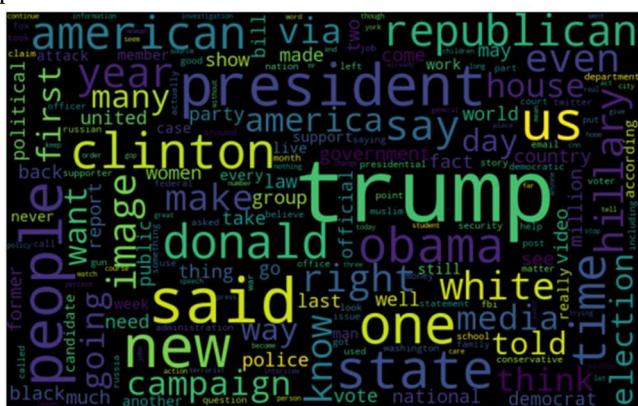


Fig. 1 Result of the word cloud for real news using the DT algorithm. The count of the true words from the dataset appeared to be lesser than the fake words.

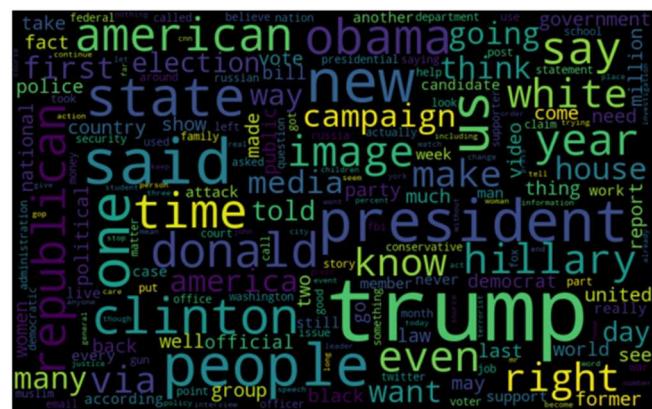


Fig. 2 Result of the word cloud for fake news using the DT algorithm. The count of the fake words from the dataset appeared to be higher than the true words.

III. RESULTS

The DT and SVM algorithms are used to predict fake news detection. The DT algorithm, shown in Fig. 1, is used to predict the true values from the dataset. The count of fake values from the dataset using the DT algorithm is shown in Fig. 2. When comparing Figs. 1 and 2, it can be seen that the fake value count appears to be slightly higher than the true value count. Both algorithms are used to create the confusion matrix. As shown in Figs. 3 and 4, the prediction of real and fake values by the DT algorithm appeared to be higher than that of the SVM algorithm. According to Fig. 5, a comparison of both algorithms has been plotted, including the error rate. From Table 1, as the number of iterations on the sample size increases, so does the accuracy and precision. The DT algorithm outperformed the SVM algorithm in terms of accuracy and precision. Table 2 compares the parameters of the DT and SVM algorithms, such as accuracy, precision,

recall, and F1 score. For ten iterations, the mean accuracy, mean error, and standard deviation of the DT and SVM algorithms were calculated. Table 3 shows that as the number of iterations increases, so does the accuracy and precision. Table 3 shows the comparison of two independent groups (SVM and DT). The outputs produced are effective because the significance value is equal to 0.092 for accuracy and 0.825 for precision. It proves that the DT algorithm has a higher mean accuracy than the SVM algorithm. Table 4 compares the Significance level for SVM and DT algorithms and also it is shown that DT algorithm accuracy and precision is higher than SVM algorithm.

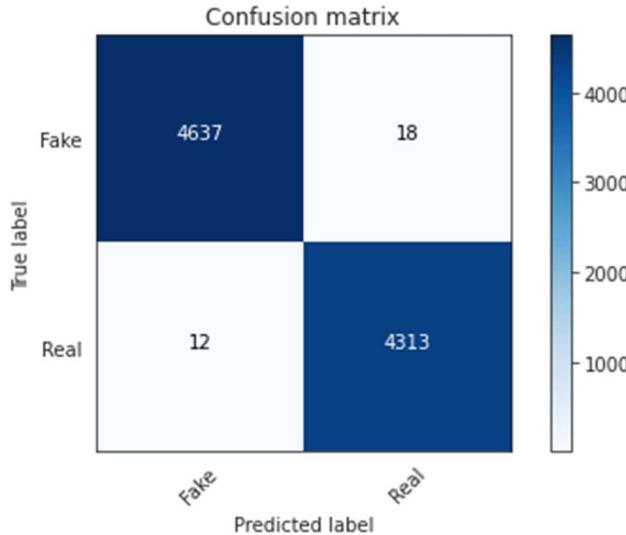


Fig. 3 The SVM algorithm results from the confusion matrix without normalisation. The SVM algorithm predicted 18 real news as fake news and 12 fake news sites as real news, as well as 4637 false news as fake and 4313 real news items as real.

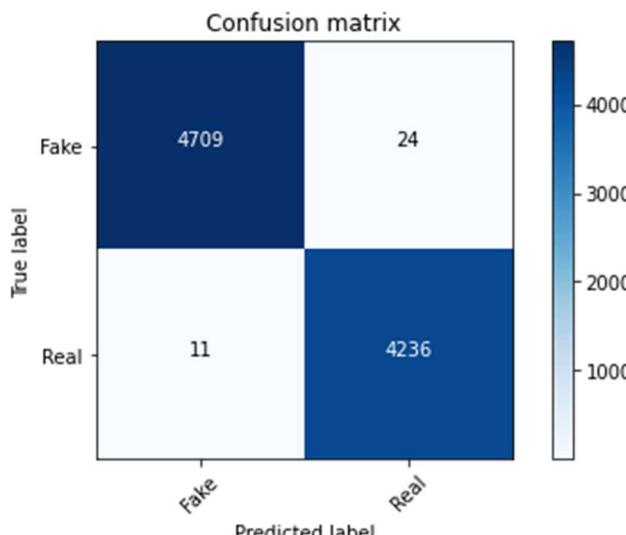


Fig. 4 Outcome of the confusion matrix without normalization using the DT algorithm. 24 real news as fake news and 11 fake news as real news and 4709 fake news as fake and 4236 real news as real.

IV. DISCUSSION

The proposed work analysed the accuracy and precision of DT and SVM algorithms for detecting fake news. On the fake news dataset, the DT algorithm appeared to be more accurate (97.67%) and precise (94.60%), whereas the SVM algorithm

appeared to be less accurate (91.74%) and precise (90.12%). The DT algorithm was found to be significantly more accurate than the SVM algorithm in detecting fake news. The machine's performance is improved by using a higher accuracy algorithm, and the DT algorithm performed better than the SVM algorithm. This paper aims to present an insight into the classification of news articles in the modern community, as well as the various content types of news articles and their impact on readers. Following that, we discuss existing fake news detection approaches that are heavily based on text-based analysis, as well as popular fake news data-sets [21]. Some more similar finding was discussed for the DT algorithm and accurate accuracy was being inferred [20].

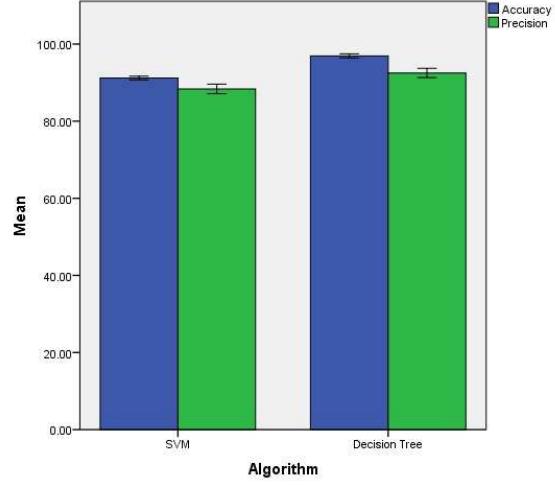


Fig. 5 Comparison of SVM algorithm and DT algorithm in terms of mean accuracy and mean precision. The mean accuracy and mean precision of DT is better than SVM and the standard deviation of DT is slightly better than SVM. X Axis: DT vs SVM algorithm. Y Axis: Mean Accuracy and Mean Precision of detection ± 1 SD.

While many people on social media are real, those who mislead and spread lies may or may not be real. The three major types of fake news contributors are social bots, trolls, and cyborg users. Because of the low cost of creating social media accounts, the creation of misleading accounts is not discouraged [10]. Most actions to detect fake news suggest features that are dependent on data from a specific dataset. Use a newly published dataset, on the other hand, to help us implement the majority of the proposed features discussed in this section [22]. Another technique for high accuracy uses SVM, Random forest, DT, and DT algorithms, with two models proposed. DT appeared to be 98.25 % in this work [23].

Our institution is passionate about high-quality evidence-based research and has excelled in various fields [24 - 27]. This study inspiration and methodology is based on [28-31]. We hope this study adds to this rich legacy.

The fact that the data is unstable is a limitation of this work, which means that any type of prediction model will have errors and make mistakes. To overcome this limitation, concepts such as POS tagging, word2vec, and subject modelling can be used in the future to develop the system. This will give the model a lot more depth in terms of feature extraction and classification.

V. CONCLUSION

For detecting fake news, the DT algorithm performs better than the SVM algorithm, which has an accuracy of 91.74 %. The mean accuracy performance of the DT algorithm is also better than that of the SVM algorithm. The precision performance is also analysed, and it is found that the DT algorithm's mean precision measure of 92.2510 % is better than SVM algorithm mean precision.

REFERENCES

- [1]. Umer, Muhammad, Zainab Imtiaz, Saleem Ullah, Arif Mehmood, Gyu Sang Choi, and Byung-Won On. 2020. "Fake News Stance Detection Using Deep Learning Architecture (CNN-LSTM)." IEEE Access. <https://doi.org/10.1109/access.2020.3019735>.
- [2]. Granik, Mykhailo, and Volodymyr Mesyura. 2017. "Fake News Detection Using Naive Bayes Classifier." 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). <https://doi.org/10.1109/ukrcon.2017.8100379>.
- [3]. Aphiwongsophon, Supanya, and Prabhas Chongstitvatana. 2018. "Detecting Fake News with Machine Learning Method." 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). <https://doi.org/10.1109/ecticon.2018.8620051>.
- [4]. Manzoor, Syed Ishfaq, Jimmy Singla, and Nikita. 2019. "Fake News Detection Using Machine Learning Approaches: A Systematic Review." 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). <https://doi.org/10.1109/icoei.2019.8862770>.
- [5]. Zhou, Xinyi, and Reza Zafarani. 2019. "Network-Based Fake News Detection." ACM SIGKDD Explorations Newsletter. <https://doi.org/10.1145/3373464.3373473>.
- [6]. Cheng, Mingxi, Songli Wang, Xiaofeng Yan, Tianqi Yang, Wenshuo Wang, Zehao Huang, Xiongye Xiao, Shahin Nazarian, and Paul Bogdan. 2021. "A COVID-19 Rumor Dataset." Frontiers in Psychology 12 (May): 644801.
- [7]. Hussain, Md Gulzar, Md Rashidul Hasan, Mahmuda Rahman, Joy Protim, and Sakib Al Hasan. 2020. "Detection of Bangla Fake News Using MNB and SVM Classifier." 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE). <https://doi.org/10.1109/iccece49321.2020.9231167>.
- [8]. Shu, Kai, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities. Springer Nature.
- [9]. Deepak, P., Tanmoy Chakraborty, Cheng Long, and Kumar G. Santhosh. 2021. Data Science for Fake News: Surveys and Perspectives. Springer Nature.
- [10]. Han, Wenlin, and Varshil Mehta. 2019. "Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance Evaluation." 2019 IEEE International Conference on Industrial Internet (ICII). <https://doi.org/10.1109/icii.2019.00070>.
- [11]. Kumar, Muthusamy Senthil, Gelli Vamsi, Ramasamy Srihari, and Praveen Kumar Sehgal. 2006. "Expression of Matrix Metalloproteinases (MMP-8 and -9) in Chronic Periodontitis Patients with and without Diabetes Mellitus." Journal of Periodontology 77 (11): 1803–8.
- [12]. Mehta, Meenu, Deeksha, Devesh Tewari, Gaurav Gupta, Rajendra Awasthi, Harjeet Singh, Parijat Pandey, et al. 2019. "Oligonucleotide Therapy: An Emerging Focus Area for Drug Delivery in Chronic Inflammatory Respiratory Diseases." Chemico-Biological Interactions 308 (August): 206–15.
- [13]. Danda, Anil Kumar, M. R. Muthusekhar, Vinod Narayanan, Mirza F. Baig, and Avinash Siddareddi. 2010. "Open versus Closed Treatment of Unilateral Subcondylar and Condylar Neck Fractures: A Prospective, Randomized Clinical Study." Journal of Oral and Maxillofacial Surgery: Official Journal of the American Association of Oral and Maxillofacial Surgeons 68 (6): 1238–41.
- [14]. Gopalakannan, S., T. Senthilvelan, and S. Ranganathan. 2012. "Modeling and Optimization of EDM Process Parameters on Machining of Al 7075-B4C MMC Using RSM." Procedia Engineering 38: 685–90.
- [15]. John., A., D. Praveen Dominic, M. Adimoolam, and N. M. Balamurugan. 2019. "Share Market Data Prediction Strategies Using Deep Learning Algorithm." Recent Advances in Computer Science and Communications. <https://doi.org/10.2174/2666255813666191209093139>.
- [16]. Adimoolam, M., Raghav Sharma, A. John, M. Suresh Kumar, and K. Ashok Kumar. 2020. "Twitter Data Sentimental Analysis Using Multiple Classifications." Journal of Computational and Theoretical Nanoscience. <https://doi.org/10.1166/jctn.2020.9319>.
- [17]. Balamurugan, N. M., T. K. Rathish babu, M. Adimoolam, and A. John. 2021. "A Novel Efficient Algorithm for Duplicate Video Comparison in Surveillance Video Storage Systems." Journal of Ambient Intelligence and Humanized Computing. <https://doi.org/10.1007/s12652-021-03119-7>.
- [18]. Adimoolam, M., M. Sugumaran, and R. S. Rajesh. 2019. "The Security Challenges, Issues and Countermeasures in Spatiotemporal Data: A Survey." International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. https://doi.org/10.1007/978-3-03146-6_142.
- [19]. Mohan, Senthilkumar, A. John, Ahed Abugabah, M. Adimoolam, Shubham Kumar Singh, Ali Kashif Bashir, and Louis Sanzogni. 2021. "An Approach to Forecast Impact of Covid-19 Using Supervised Machine Learning Model." Software: Practice and Experience. <https://doi.org/10.1002/spe.2969>.
- [20]. Abedalla, Ayat, Aisha Al-Sadi, and Malak Abdullah. 2019. "A Closer Look at Fake News Detection." Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence. <https://doi.org/10.1145/3369114.3369149>.
- [21]. Shu, Kai, and Huan Liu. 2019. Detecting Fake News on Social Media. Morgan & Claypool Publishers.
- [22]. Barbier, Geoffrey, Zhuo Feng, Pritam Gundecha, and Huan Liu. 2013. Provenance Data in Social Media. Morgan & Claypool Publishers.
- [23]. Li, Xin, Peixin Lu, Lianting Hu, Xiaoguang Wang, and Long Lu. 2021. "A Novel Self-Learning Semi-Supervised Deep Learning Network to Detect Fake News on Social Media." Multimedia Tools and Applications, June, 1–9.
- [24]. Stewart, Elizabeth. 2021. "Detecting Fake News: Two Problems for Content Moderation." Philosophy & Technology, February, 1–18.
- [24]. Vijayashree Priyadharsini, Jayaseelan. 2019. "In Silico Validation of the Non-Antibiotic Drugs Acetaminophen and Ibuprofen as Antibacterial Agents against Red Complex Pathogens." Journal of Periodontology 90 (12): 1441–48.
- [26]. Ramesh, Asha, Sheeba Varghese, Nadathur D. Jayakumar, and Sankari Malaiappan. 2018. "Comparative Estimation of Sulfiredoxin Levels between Chronic Periodontitis and Healthy Patients - A Case-Control Study." Journal of Periodontology 89 (10): 1241–48.
- [27]. Mathew, M. G., S. R. Samuel, A. J. Soni, and K. B. Roopa. 2020. "Evaluation of Adhesion of Streptococcus Mutans, Plaque Accumulation on Zirconia and Stainless Steel Crowns, and Surrounding Gingival Inflammation in Primary" Clinical Oral Investigations. <https://link.springer.com/article/10.1007/s00784-020-03204-9>.

- [28] Khan, N. A., Brohi, S. N., & Zaman, N. (2020). Ten deadly cyber security threats amid COVID-19 pandemic.
- [29] Kaun, C., Jhanjhi, N. Z., Goh, W. W., & Sukumaran, S. (2021). Implementation of Decision Tree Algorithm to Classify Knowledge Quality in a Knowledge Intensive System. In MATEC Web of Conferences (Vol. 335, p. 04002). EDP Sciences.
- [30] Kaur, M., Singh, A., Verma, S., Jhanjhi, N. Z., & Talib, M. N. (2021). FANET: Efficient Routing in Flying Ad Hoc Networks (FANETs) Using Firefly Algorithm. In Intelligent Computing and Innovation on Data Science (pp. 483-490). Springer, Singapore.
- [31] Ullah, A., Azeem, M., Ashraf, H., Alaboudi, A. A., Humayun, M., & Jhanjhi, N. Z. (2021). Secure healthcare data aggregation and transmission in IoT—A survey. IEEE Access, 9, 16849-16865.