TruthTrack – Fake News Prediction

Dr Rakesh Kumar M
Computer Science and Engineering
Rajalakshmi Engineering College,
Chennai, India
rakeshkumar.m@rajalakshmi.edu.in

Reshma S
Computer Science and Engineering
Rajalakshmi Engineering College,
Chennai, India
210701213@rajalakshmi.edu.in

Reshma A
Computer Science and Engineering
Rajalakshmi Engineering College,
Chennai, India
210701211@rajalakshmi.edu.in

Abstract:

In our modern era where the internet is ubiquitous, everyone relies on various online resources for news. Along with the increase in the use of social media platforms like Facebook, Twitter, etc. news spread rapidly among millions of users within a very short span of time. The spread of fake news has far-reaching consequences like the creation of biased opinions to swaying election outcomes for the benefit of certain candidates. Moreover, spammers use appealing news headlines to generate revenue using advertisements via click-baits. In this paper, we aim to perform binary classification of various news articles available online with the help of concepts pertaining to Artificial Intelligence, Natural Language Processing and Machine Learning. We aim to provide the user with the ability to classify the news as fake or real and also provide how much percentage the news is fake.

Keywords:

Fake news , Social media platforms , Binary classification , Natural Language Processing(NLP) , Machine Learning , Artificial Intelligence .

INTRODUCTION:

Fake or false news can be described as news that contains entirely or partially false information. The impact and proliferation of fake news have become a major threat to society and the integrity of information. The short and long-term negative effects on the world through the decrease in trust in mainstream media, the spread of false claims, the incitation of violence, and the growth of conspiracy theories. To circumvent fake news, several fake News detection methodologies have been developed. There have been several comprehensive studies on fake news detection alongside machine learning models trained to identify fake news through various parameters. Social media platforms have also implemented several guidelines to disrupt fake news from spreading through automated and manual methods .As an increasing amount of our lives is spent interacting online through social media platforms, more and more people tend to hunt out and consume news from social media instead of traditional news organizations. It had been also found that social media now outperforms television because the major news source. Despite the benefits provided by social media, the standard of stories on social media is less than traditional news organizations. However, because it's inexpensive to supply news online and far faster and easier to propagate through social media, large volumes of faux news, i.e., those news articles with intentionally false information, are produced online for a spread of purposes, like financial and political gain.The extensive spread of faux news can have a significant negative impact on individuals and society. First, fake news can shatter the authenticity equilibrium of the news ecosystem for instance; it's evident that the most popular fake news was even more outspread on Facebook than the most accepted genuine mainstream news . Second, fake news intentionally persuades consumers to simply accept biased or false beliefs. Fake news is typically manipulated by propagandists to convey political messages or influence for instance, some report shows that Russia has created fake accounts and social bots to spread false stories. Third, fake news changes the way people interpret and answer real news, for instance, some fake news was just created to trigger people's distrust and make them confused; impeding their abilities to differentiate what's true from what's not..Often Internet users can pursue the events of their concern in online form, and increased number of the mobile devices makes this process even easier. But with great possibilities come great challenges. There even exist many websites that produce fake news almost exclusively. They intentionally publish hoaxes, half-truths, propaganda and disinformation asserting to be real news – often using social media to drive web traffic and magnify their effect. The most goals of faux news websites are to affect the general public opinion on certain matters.Thus, fake news may be a global issue also as a worldwide challenge. Many scientists believe that fake news issue could also be addressed by means of machine learning and AI.There’s a reason for that: recently AI algorithms have begun to work far better on many classification problems (image recognition, voice detection then on) because hardware is cheaper and larger datasets are available.The accuracy of the detection achieved by the system is around 90%.This text describes an easy fake news detection method supported one among the synthetic intelligence algorithms – naive Bayes classifier, Random Forest and Logistic Regression. The goal of the research is to look at how these particular methods work for this particular problem given a manually labelled news dataset and to support (or not) the thought of using AI for fake news detection.

RELATED WORK:

Mykhailo Granik et. al. in their paper [3] shows a simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News). They achieved classification accuracy of approximately 74%. Classification accuracy for fake news is slightly worse. This may be caused by the skewness of the dataset: only 4.9% of it is fake news. Himank Gupta et. al. [10] gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of tweets in 1 sec. Firstly, they have collected 400,000 tweets from HSpam14 dataset. Then they further characterize the 150,000 spam tweets and 250,000 non- spam tweets. They also derived some lightweight features along with the Top-30 words that are providing highest information gain from Bag-of-Words model. 4. They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately18%. Marco L. Della Vedova et. al. [11] first proposed a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing its accuracy up to 78.8%. Second, they implemented their method within a Facebook Messenger Chabot and validate it with a real-world application, obtaining a fake news detection accuracy of 81.7%. Their goal was to classify a news item as reliable or fake; they first described the datasets they used for their test, then presented the content-based approach they implemented and the method they proposed to combine it with a social-based approach available in the literature. The resulting dataset is composed of 15,500 posts, coming from 32 pages (14 conspiracy pages, 18 scientific pages), with more than 2, 300, 00 likes by 900,000+ users. 8,923 (57.6%) posts are hoaxes and 6,577 (42.4%) are non-hoaxes. Cody Buntain et. al. [12] develops a method for automating fake news detection on Twitter by learning to predict accuracy assessments in two credibility-focused Twitter datasets: CREDBANK, a crowd sourced dataset of accuracy assessments for events in Twitter, and PHEME, a dataset of potential rumors in Twitter and journalistic assessments of their accuracies. They apply this method to Twitter content sourced from BuzzFeed’s fake news dataset. A feature analysis identifies features that are most predictive for crowd sourced and journalistic accuracy assessments, results of which are consistent with prior work. They rely on identifying highly retweeted threads of conversation and use the features of these threads to classify stories, limiting this work's applicability only to the set of popular tweets. Since the majority of tweets are rarely retweeted, this method therefore is only usable on a minority of Twitter conversation threads. In his paper, Shivam B. Parikh et. al. [13] aims to present an insight of characterization of news story in the modern diaspora combined with the differential content types of news story and its impact on readers. Subsequently, we dive into existing fake news detection approaches that are heavily based on text-based analysis, and also describe popular fake news datasets. We conclude the paper by identifying 4 key open research challenges that can guide future research. It is a theoretical Approach which gives Illustrations of fake news detection by analyzing the psychological factors.

EXISTING SYSTEM

Through the inserted data, the algorithm will first learn to distinguish between bogus and authentic news. After understanding the distinction, the system will learn to make judgments based on the data presented. Fake news tracker programmes monitor the collection, analysis, and visualisation of fake news. [4]The bogus database displays no news channel names, but the genuine dataset displays individual headquarters for each station. Manipulating the concept of dataset fraudulent channels are exploiting an unregistered news portal. As a result, using the original dataset, one may compare and explicitly identify them.The data analysis also involves a number of dangers. The proper usage of data assessment in relation to references must be considered. During data analysis, there are some assessment elements that Python does not recognise, which creates the data clarity difficulty.

PROPOSED SYSTEM:

The approach to developing a fake news prediction model involves several key steps. First, we perform text transformation using either a Count Vectorizer, which converts a collection of text documents into a matrix of token counts, or a TF-IDF Vectorizer, which converts the text into a matrix of TF-IDF features that reflect the importance of words in the document relative to the entire corpus. We then decide whether to use headlines or the full text of news articles for the analysis, as each option may yield different insights and effectiveness in classification. Next, we extract optimal features by selecting the most frequently occurring words and/or phrases, converting all text to lowercase to maintain uniformity, and removing common stop words such as "the," "when," and "there" to focus on more meaningful words. We also ensure that only words appearing at least a certain number of times in the dataset are included, to filter out less informative words. Finally, we implement a Linear Regression model to classify news as fake or real, where the model predicts a continuous score that, when compared against a threshold, determines the classification.

A. *OBJECTIVES*

- Detection and indentification: To accurately detect and identify false or misleading information presented as news.

- Enhancing Trust: To restore and enhance public trust in media and information sources by ensuring that the news consumed is accurate and reliable.

B. *APPROACH*

- Data Collection :

  First, text data is gathered by the system from a variety of sources, such as papers, publications, and websites. The text summarizing technique uses this data, which spans a wide range of subjects and areas, as its input.

- PreProcessing :

  The text goes through preprocessing procedures like tokenization, stopword removal, and sentence splitting after data collection. The text data is ready for additional analysis and summary after these procedures.

- Linear Regression:

  Linear regression can be used for classification by setting a threshold on the predicted output to distinguish between classes. For fake news prediction, the model learns to assign a continuous score to each news article. By selecting an appropriate threshold, scores above it can be classified as real news, while those below are classified as fake news.

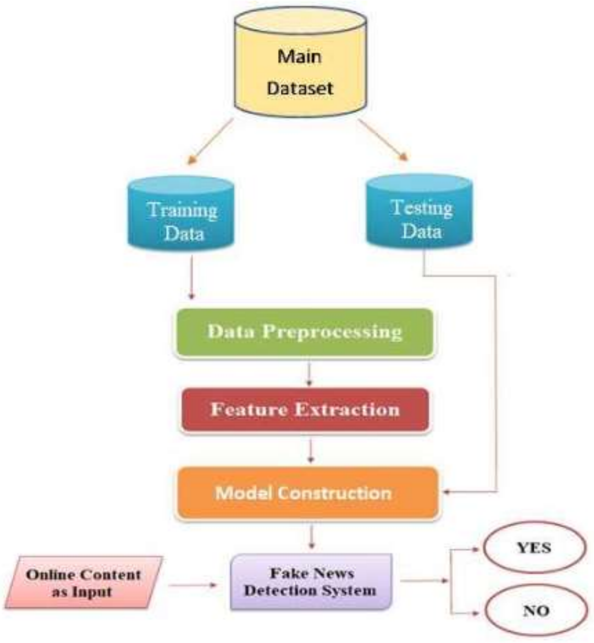
C. *ADVANTAGES*

- Information was very clear and understandable.
- It gives accurate predictions which is very clear to the user.
- User friendly and faster time compatibility.

METHODOLOGY:

1)Data Collection and Preprocessing: The first step requires gathering a diverse dataset of images containing various fruits and vegetables from different sources. It's pivotal to ensure that the dataset encloses a wide range of classes, variations in appearance, and environmental conditions to increase the model's generalization capability. Once collected, the images undergo preprocessing to standardize attributes such as size, color, and orientation. Techniques like resizing. normalization, and augmentation are applied to ensure uniformity and increase the dataset's diversity, which aids in training a more robust model.

2)Model Architecture Design:



This paper explains about the development of the system . The first part is static which works on machine learning classifier. We trained the model with 4 different classifiers and chose the best classifier for final execution. The second part is dynamic which takes the keyword/text from user and searches online for the truth probability of the news. In this paper, we have used Python and its Sci-kit libraries . Python has a huge set of libraries and extensions, which can be easily used in Machine Learning. Sci-Kit Learn library is the best source for machine learning algorithms where nearly all types of machine learning algorithms are readily available for Python, thus easy and quick evaluation of ML algorithms is possible. We have used flask and machine learning for the web based deployment of the model, provides client side implementation using HTML, CSS and Javascript.

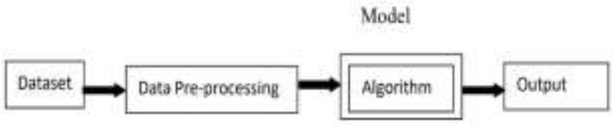


Figure 4.1 : Data Flow Diagram

The DFD takes an input-process-output view of a system i.e., data objects flow into the software, are transformed by processing elements, and resultant data objects flow out of the software. The dataset contains real and fake news information. Then the information is fed to algorithm .Thus news is analyzed as fake or real.

3)Training Procedure:

The designed Linear Regression model for fake news prediction is trained using the preprocessed text dataset. During training, the model learns to map the input text features (extracted using Count Vectorizer or TF-IDF Vectorizer) to their corresponding labels (fake or real) by adjusting its coefficients to minimize the error between the predicted scores and the actual labels . The training process involves iteratively feeding batches of text data into the model, computing the loss (often using mean squared error), and updating the model's parameters to reduce this loss.Techniques such as feature scaling and regularization are employed to optimize convergence and prevent overfitting.

4)Cross-Validation and Validation Strategies:

To ensure the robustness of the trained model, techniques such cross-validation as k-fold cross-validation are employed. The dataset is partitioned into multiple subsets, and the model is trained and evaluated iteratively on different combinations of training and validation sets. This helps in assessing the model's stability and generalization performance across diverse data distributions and mitigates the risk of overfitting.

5) Dataset Selection:

For developing the fake news prediction model, we have selected a dataset from Kaggle, which provides a comprehensive and well-labeled collection of news articles. The dataset includes a diverse range of sources, ensuring that the model can generalize well across different writing styles and perspectives. Each article is accurately labeled as either fake or real, sourced from reliable fact-checking organizations. The dataset is balanced, with a comparable number of fake and real news articles, preventing bias towards any particular class. It includes both headlines and full text, allowing the model to capture nuances in both concise and detailed content. Additionally, the dataset is up-to-date, reflecting current topics and language use, which is crucial for maintaining the relevance and effectiveness of the model in detecting contemporary fake news.

6) Implementation steps:

1)**Pre-processing:** The text goes through preprocessing procedures like tokenization, stopword removal, and sentence splitting after data collection. The text data is ready for additional analysis and summary after these procedures.

2). Cleaning: Cleaning up the text data is necessary to highlight attributes that we’re going to want our machine learning system to pick up on. Cleaning (or pre-processing) the data typically consists of a number of steps.

a) Remove punctuation: Punctuation can provide grammatical context to a sentence which supports our understanding. But for our vectorizer which counts the number of words and not the context, it does not add value, so we remove all special character.

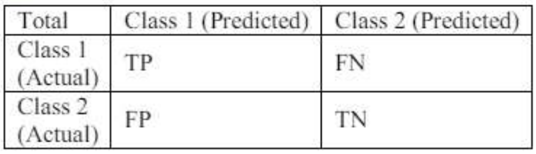
b) Remove stopwords: Stopwords are common words that will likely appear in any text. They don’t tell us much about our data so we remove them

c) Stemming Stemming helps reduce a word to its stem form. It often makes sense to treat related words in the same way. It removes suffices, like “ing”, “ly”, “s”, etc. by a simple rule-based approach. It reduces the corpus of words but often the actual words get neglected. eg: Entitling, Entitled -> Entitle

3) Feature Generation: We can use text data to generate a number of features like word count, frequency of large words, frequency of unique words, n-grams etc. By creating a representation of words that capture their meanings, semantic relationships, and numerous types of context they are used in, we can enable computer to understand text and perform Clustering, Classification etc [19].

4) Vectorizing Data: Vectorizing is the process of encoding text as integers i.e. numeric form to create feature vectors so that machine learning algorithms can understand our data

5) **Confusion Matrix:** A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made .



Total	Class 1 (Predicted)	Class 2 (Predicted)
Class 1 (Actual)	TP	FN
Class 2 (Actual)	FP	TN

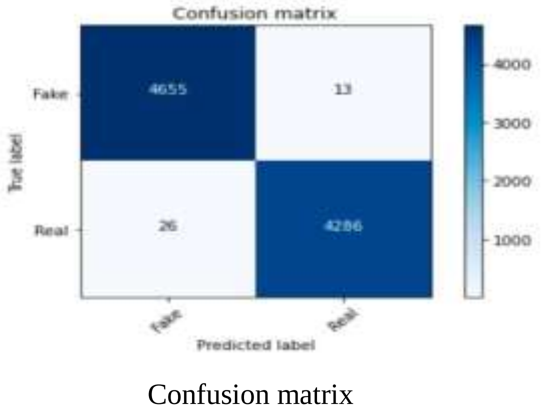
6) Logistic regression used for Classification: Logistic regression can be used for classification by setting a threshold on the predicted output to distinguish between classes. For fake news prediction, the model learns to assign a continuous score to each news article. By selecting an appropriate threshold, scores above it can be classified as real news, while those below are classified as fake news.

Testing Procedure:

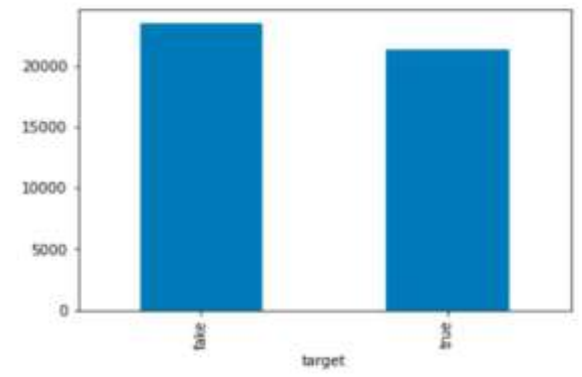
The effectiveness of the Linear Regression model for fake news prediction is evaluated using a separate testing dataset. In this process, the preprocessed text features from the testing dataset are fed into the model. The model then predicts whether each article is fake or real. These predictions are compared to the actual labels to measure how accurate the model is. We use metrics like accuracy, precision to assess performance. This testing helps us understand how well the model works with new, unseen data.

RESULTS:

Implementation was done using the above algorithms with Vector features- Count Vectors and Tf-Idf vectors at Word level . Accuracy was noted for all models. We used K-fold cross validation technique to improve the effectiveness of the models. A. Dataset split using K-fold cross validation This cross-validation technique was used for splitting the dataset randomly into k-folds. (k-1) folds were used for building the model while kth fold was used to check the effectiveness of the model. This was repeated until each of the k-folds served as the test set. I used 3- fold cross validation for this experiment where 67% of the data is used for training the model and remaining 33% for testing. Logistic Regression and Random Forest), their confusion matrix showing actual set and predicted sets are mentioned below:

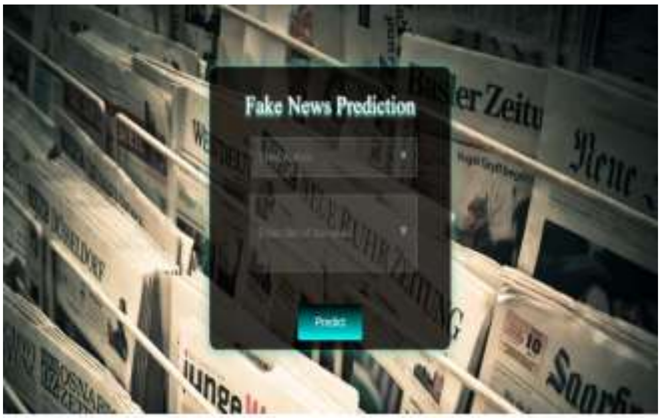


Confusion matrix

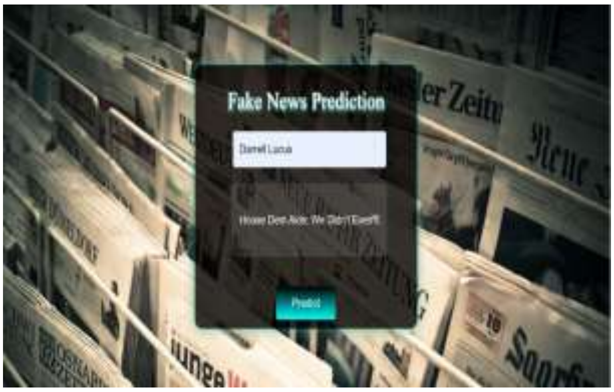


Analysing fake and real news from the dataset.

SCREEN SHOT:



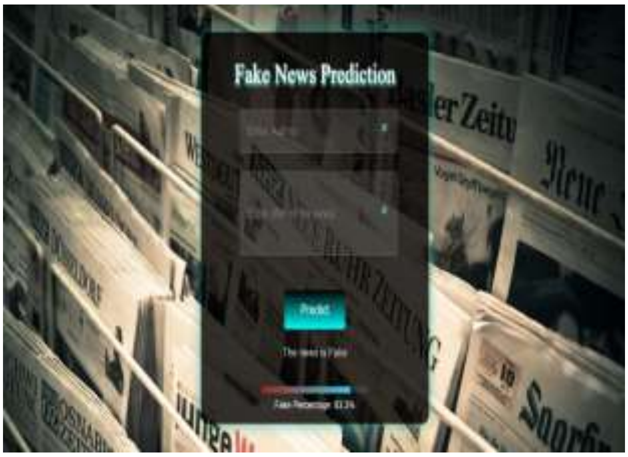
This image shows the front end of our fake news prediction.



Darrell Lucus

House Dem Aide: We Didn't Even S

This image shows the front end of our fake news prediction after providing input from the user.



The news is Fake

Fake Percentage: 83.3%

This is the output after prediction which shows whether the news is fake or real and also the fake percentage of the news.

CONCLUSION

Due to increasing use of internet, it is now easy to spread fake news. A huge number of persons are regularly connected with internet and social media platforms. There is no any restriction while posting any news on these platforms. So some of the people takes the advantage of these platforms and start spreading fake news against the individuals or organizations. This can destroy the repute of an individual or can affect a business. Through fake news, the opinions of the people can also be changed for a political party. There is a need for a way to detect these fake news. Machine learning classifiers are using for different purposes and these can also be used for detecting the fake news. The classifiers are first trained with a data set called training data set. After that, these classifiers can automatically detect fake news. The data we used in our work is collected from the Kaggle.com and contains news articles from various domains to cover most of the news rather than specifically classifying political news. The learning models were trained and parameter-tuned to obtain optimal accuracy. Fake news detection has many open issues that require attention of researchers. For instance, in order to reduce the spread of fake news, identifying key elements involved in the spread of news is an important step. Machine learning techniques can be employed to identify the key sources involved in spread of fake news.

FUTURE WORK:

Fake news is categorized as any kind of cooked-up story with an intention to deceive or to mislead. In this paper we are trying to present the solution for fake news detection task by using Machine Learning techniques. Many events have resulted to a rise in the prominence and spread of phony news. The widespread impacts of the massive onset of fake news can be seen, humans are conflicting if not outright poor detectors of fake news. With this, endeavours are being made to automate the task of fake news detection. The most mainstream of such actions include blacklisting of sources and authors that are unreliable. Even though these tools are useful, but in order to produce a progressive complete end to end solution, we are required to represent for tougher cases where reliable sources and authors are responsible for releasing fake news. Here, the purpose of this project was to build a model that help us to recognize the language patterns that can be used to classify fake and real news with the help of ML (machine learning) techniques. The outcomes of this project shows the capability of ML to be fruitful in this task. We have tried to build a model that helps in catching many intuitive indications of real and fake news as well as in the visualization of the classification decision. Now-a-days fake news is such a big problem that it is affecting our society as well as our facts and opinions. The problem that needs to be solved can be solved using AI and Machine learning techniques.

REFERENCES

1. M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2019 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2019, pp. 900-903.

2. H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2019 11th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2019.

3. M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2019 22nd Conference of Open Innovations Association (FRUCT), Jyvaskyla, 2019, pp. 272- 279.

4. C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," 2019 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2019.

5. Jadhav, S. S., & Thepade, S. D. (2019). Fake News Identification and Classification Using DSSM and Improved Recurrent Neural Network Classifier, Applied Artificial Intelligence, 33(12), 1058-1068, https://doi.org/10.1080/08839514.2019.1661579

6. Kaliyar, R. K., Goswami, A., Narang, P., & Sinha, S. (2020). FNDNet–A deep convolutional neural network for fake news detection. Cognitive Systems Research, 61, 32-44. https://doi.org/10.1016/j.cogsys.2019.12.005

7. Kaur, S., Kumar, P. & Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. Soft Computing, 24(12), 9049–9069. https://doi.org/10.1007/s00500-019-04436-y

8. Kesarwani, A., Chauhan, S. S., & Nair, A. R. (2020). Fake News Detection on Social Media using K-Nearest Neighbor Classifier. 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), Las Vegas, NV, USA, pp.1-4, https://doi.org/10.1109/ICACCE49060.2020.9154997

9. Khan, J. Y., Khondaker, M., Islam, T., Iqbal, A., & Afroz, S. (2019). A benchmark

study on machine learning methods for fake news detection. Computation and Language. https://arxiv.org/abs/1905.04749

10. Rahman, M. M., Chowdhury, M. R. H. K., Islam, M. A., Tohfa, M. U., Kader, M. A. L., Ahmed, A. A. A., & Donepudi, P. K. (2020). Relationship between SocioDemographic Characteristics and Job Satisfaction: Evidence from Private Bank 47 Employees. American Journal of Trade and Policy, 7(2), 65-72. https://doi.org/10.18034/ajtp.v7i2.492

11. Reis, J. C. S., Correia, A., Murai, F., Veloso A., & Benevenuto, F. (2019). Supervised Learning for Fake News Detection. IEEE Intelligent Systems, 34(2), 76-81, https://doi.org/10.1109/MIS.2019.2899143

12. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2020). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600

13. Abdullah-All-Tanvir, Mahir, E. M., Akhter S., & Huq, M. R. (2019). Detecting Fake News using Machine Learning and Deep Learning Algorithms. 7th International Conference on Smart Computing & Communications (ICSCC), Sarawak, Malaysia, Malaysia, 2019, pp.1-5, https://doi.org/10.1109/ICSCC.2019.8843612

14. Ahmed, H., Traore, I., & Saad, S. (2019). Detection of online fake news using n-gram analysis and machine learning techniques. Proceedings of the International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 127–138, Springer, Vancouver, Canada, 2019. https://doi.org/10.1007/978-3-319-69155-8_9

15. Ahmed, H., Traoré, I., & Saad, S. (2020). Detecting opinion spams and fake news using text classification. Secur. Priv., 1(1), 1-15. https://doi.org/10.1002/spy2.9

16. Rampersad G, Althiyabi T 2020 “Fake news: Acceptance by demographics and culture on social media” J. Inf. Technol. Politics 2020, 17, 1–11.

17. NaphapornSirikulviriya; ukreeSinthupinyo. “Integration of Rules from a Random Forest.” International Conference on Information and Electronics Engineering (p. 194 : 198). Singapore: semanticscholar.org. 2011.

18. Jasmin Kevric et el. “An effective combining classifier approach using tree algorithms for network intrusion detection.” Neural Computing and Applications , 1051–1058. 2017.

19. ShivamB.Parikh and PradeepK.Atrey. “Media-RichFake News Detection: A Survey.” IEEE Conference on Multimedia Information. Miami, FL: IEEE. 2018.

20. MykhailoGranik and VolodymyrMesyura. “Fake news detection using naive Bayes classifier.” First Ukraine Conference on Electrical and Computer Engineering (UKRCON). Ukraine : IEEE. 2017.

21. Gilda, S. “Evaluating machine learning algorithms for fake news detection.” 15th Student Conference on Research and Development (SCOReD) (pp. 110-115). IEEE. 2017.

22. Akshay Jain and AmeyKasbe. “Fake News Detection.” 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). Bhopal, India: IEEE. 2018.

23. ArushiGupta and RishabhKaushal. “Improving spam detection in Online Social Networks.” International Conference on Cognitive Computing and Information Processing (CCIP). semanticscholar.org.2015