International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-11 Issue-3, January 2022

A Distinctive Approach for Detecting Fake News Using Machine Learning

Md. Rakib Hasan, Ismath Ara Itu

Abstract*: Nowadays, social media platforms have evolved into an esoteric method for audiences to consume information. News spreading through the social network are also a great source of information as well. For the advancement of Internet access, the information consumption pattern is dramatically changing. As a consequence of those, fake news has become one of the prime concerns because of its potentiality to endanger a society in different perspectives as well as has a political and social impact. Because false news causes so much confusion among people, we will train a model to identify all types of fake news in response to public demand. For this respective research, we collect real data from many reliable and reputed online news portals and fake news from unreliable resources. For converting text to vector format Bag of Words is used. Besides TF-IDF is used for extracting the feature and then CNN is used for classification. With an 83.14% accuracy our model can efficiently detect fake and real news. This work paves a path for an easy automatic fake news detection system which will be very helpful for us to prevent spreading the false information and helps to find the truth.

***Keywords*: Bangla Fake News, CNN, Fake News Detection, Natural Language Processing.**

I. INTRODUCTION

Now-a-days number of online news portal, a great source of information is increasing. News spreading through the social network are also a great source of information as well. For the advancement of Internet access, the information consumption pattern is dramatically changing. As a consequence of those, fake news has become one of the prime concerns because of its potentiality to endanger a society in different perspectives as well as has a political and social impact. Automatic detection of fake news a bit complex because it needs the knowledge of political, social context and some common sense which are lacking in present natural language processing algorithms. In this respective research, a benchmark dataset will be presented for fake news detection which can also be used in fact-checking research and an improved way will be engendered based on different machine learning algorithms and deep neural networks. This work will pave a path for an easy automatic fake news detection system which will be very helpful for us to prevent the false information and helps to find the truth.

In today's modern world, we are surrounded with different social media platforms, website because of the advancement of technology. From those sites and platforms, we get different news and we consume this news. Most of them people tend to seek out news from this site and platforms than traditional news organization. This traditional organization provide them real news but most of the time we get false news from different social platform and media. Due to this, different machine learning techniques and natural language processing techniques are used to detect fake news. Here machine learning approach is used to perform the classification of real and fake news. And natural language processing is one of the most well-known fields that allows computer to process and manipulate human language. It is used for easy to learn and readable any language to machine. Fake news detection for bangle is more difficult and challenging than other language because of structure of bangle fake news is very confusing. Though some researchers are improving to detect fake news and use different method and algorithms but very few for bangle. There is still a lot of change to improve more.

II. LITERATURE REVIEW

With the advancement of Information and communication technology more and more online news portal is coming to feed us information. A lot of portals are publishing news without verifying the fact to be the first to publish the news. And off course social media has a big impact for spreading the news. As in social media no one is obliged to share the truth. On the contrary, they sometimes share fake and irrelevant news by choice or just to go with the flow. So, sometimes it's difficult to know which information bears truth and which are not true at all. Many researchers tried to find a way to detect fake news using single or combining multiple machine learning algorithms.

William Yang Wang presents a new, publicly available benchmark dataset named "LIAR" for fake news detection [1]. This database consisting of a huge 12.8k manually labeled short statements is larger than the previously published largest dataset of same type used for fake news detection. They have designed a hybrid CNN to integrate meta data with text that results great in improving a text-only deep learning model.

Manuscript received on January 20, 2022.
Revised Manuscript received on January 25, 2022.
Manuscript published on February 28, 2022.
* Correspondence Author
Md. Rakib Hasan*, Department of Information and Communication Technology, Comilla University, Cumilla, Bangladesh. Email: rakib@cou.ac.bd
Ismath Ara Itu, Department of Management Information Systems, University of Dhaka, Dhaka, Bangladesh. Email: ismathhasan@gmail.com

A fake news detection model is proposed using n-gram analysis and machine learning techniques [2][3]. Two different feature extraction techniques Term Frequency (TF) and Term Frequency-Inverted Document Frequency (TF-IDF) are discussed. Several machine learning techniques are applied to detect fake news and ends with a conclusion of 92% accuracy using Linear Support Vector Machine (LSVM). Some researchers use Count Vector and TF-IDF for extracting feature and then for the detection of fake news they apply machine learning algorithms [3]. Shlok applies TF-IDF of bi-grams and probabilistic context free grammar (CFG) detection on their dataset with different Salgorithms to identify the unreliable sources. They conclude with an accuracy of 77.2% and find a small impact of PCFGs on recall [5].

Lilapati and Rajat work with LIAR dataset from POLITIFACT.COM and concern about the increase of accuracy. They use model ensemble technique to have better accuracy in predicting fake news [6]. Finding gap between how the news is related or unrelated with the real news is the main concern for researchers [7]. Eventually, this approach provides diversity to the field of research with an enviable success rate of 94.21% on test data.

Most of the proposal share their own findings with their accuracy rate for detecting false news. Maximum researchers tried common machine learning algorithms like Support Vector machine, Naïve Bayes, Decision tree, Logistic Regression, Recurrent Neural Network and so on. Some researchers tried their own approaches combining them together but the accuracy is not satisfactory. In these circumstances, an efficient approach is proposed to detect fake news combining deep learning techniques mainly focused on Deep Neural Networks (DNN). This unique approach will provide a new dimension for detecting fake news efficiently with a high accuracy rate as well as lower complexity.

For detecting false news, a supervised system using KNN, Decision Tree, Nave Bayes, Logistic Regression, and SVM works well. However, owing to the large dataset required, development will be time consuming. The amount of the training data determines its performance level. Furthermore, the neural network has done a good job at this, but it is still difficult to improve. It is now a word for a system.

III. METHODOLOGY

The aim of this research is to determine the performance of a system in the identification of false news. The following actions were followed to complete the thesis effectively.

- To build dataset collect data from different reputed and non-reputed news portal by scraping using different scraping tools including (ParseHub, Scrapy etc.)
- Collect the data of different categories.
- Preprocess the data.
- Split the data for train and testing.
- For data read and write panda framework is used.
- Data.shape and data.columns to show dataset effectively.
- Check missing data. For this we have used isnull() function.
- Observe the split data as real data and fake data of dataset.
- Count all the categories.
- Count the data according to categories.
- Fill the missing data with true value.
- Identify, reduce noise and remove all type of noise.
- Target variable encoding to create data shape and partition data.
- Tokenize the data.
- Using Bag of Word, you may transform text documents into numerical characteristics. The simplest is the Count Vectorizer, which counts the number of times a token appears in the text and uses that value as its weight. Finally, to improve accuracy, an unsupervised Bangla POS tagger based on suffix analysis is suggested.
- By using tf-idf convert numeric data into matrix format and
- Finally, construct training and testing sets then model train using CNN.

The mentioned steps are the detailed description of this thesis. But in a nutshell, they can be categorized in some main steps.

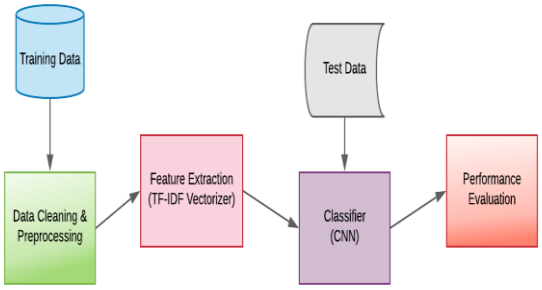

Fig. 1. System flow Diagram of the Proposed System

In Figure 1 the main steps are depicted. Firstly, we need to collect data. Then preprocess the data is a must for better outcome. After preprocessing, the data is ready for feature extraction. Using those feature the model is implemented and then testing the model with test dataset. Finally, it comes to performance analysis. These are main steps in short.

A. Data collection Procedure

Data collection mainly includes data acquisition, data labeling, using existing data. Different types of tools and software are used for different tasks like data discovery, data augmentation, data generation, data labeling, improving the existing data.

For performing those tasks, we can use different scraping tools (Scrapy, Beautiful Soup, Scraper, ParseHub, Octoparse, Mozenda etc.) which will be used for collecting data from web, OpenCV and scikit-image as data augmentation tools. Lionbridge AI will be used for handling image, text, video data and LabelBox will be used for image annotation, text classification.

Our data is split into six columns (Title, Date, Category, Headlines, Description, Class) and classes are classified into two categories Real and Fake.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Our entire dataset is about 2082, with 80% of the data being utilized for training and 20% for testing. Data from reputable Bangladeshi online news portals such as Daily Prothom alo, Bangladesh Pratidin, Ittefaq, Daily KalerKantho, Daily NayaDiganta, bdnews24.com, and others was collected.

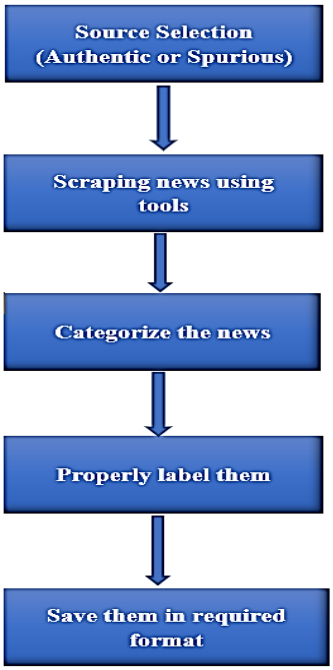

Fig. 2. Data Collection Procedure

Fake class data was gathered from unreliable internet news sources such as Dhaka Channel, baler kontho, moti kontho, Khbor24.com and others. After that, we use Google scraping to verify whether or not the data is false. For web scraping, we must search by news title, create a result for that news, and determine how much of that news is shown in a fake news portal vs a genuine news portal. But this process is very time consuming and hectic. So, we have used some paid scraping tools like ParseHub, OctoParse to web scraping. In ParseHub, at first, we need to download the application from their website. Then need to open a new project. Provide the website url you want to collect data. Then from the user interface one can select the title and description of news as well as any other information. After the "Get Data" and "Run" one can have the option of downloading the extracted news in required format. We divided the news into two categories based on the aforementioned description: false news and genuine news. Figure 2. depicts several major topics as well as a flowchart depicting how we gather and identify data. In Figure 2, we also indicate which sites provide genuine data and which sites provide false data. As a result, this flowchart is split into two parts for news sites, one reputable and the other not. We also observed a news category while gathering data, which indicates which news was produced for a joke and which news was published for an entitlement.

We sometimes get news with an appealing headline, but when we read the item, we notice something else. To effectively identify news, one must analyze all news, which is referred to as reading beyond the headlines. Here are some important aspects we have considered to collect data properly:

- To collect valid data, we consider the reputed source for authentic data and collect fake data from non-reputed source.
- We must consider the full news. Only headline cannot comment actual scenario.
- To identify data that`s the data is real or fake must be check author because some non-reputed author publishes fake news and reputed author always publish real news so it's so important to detect data and collect valeted data.
- To collect news, we focused on news date which claims more authenticity.
- At the time of collecting data, we checked this data is type because it's helpful to arrange data into different categories.
- For data validation, we recheck the news on Google by hold news title by use news title we search Google and find some result if this news is real then show some reputed news site and if this news is fake then show non-reputed site or only show one site because actually fake news published non-reputed site and if is news is fake then it published individual site.

Figure 3 depicts a sample of authentic data. We have arranged 2082 dataset; however, we will only present some datasets as examples. We store news headlines, categories, contents, domain, and classes in our dataset (classes indicate which news is genuine and which is false; we have two classes, fake (for fake news) and real (for true news) (for real news).

Label	1 (Authentic)
Domain	Somoynews.tv
Published Date and Time	9/19/2018 19:37
Category	Sports
Headline	জাতীয় দলের একটু কাছে আশরাফুল
Contents	খুলনায় কাল শুরু হচ্ছে একটি চারদিনের ম্যাচ। এই ম্যাচে থাকছেন আশরাফুল। নিষেধাজ্ঞা কেটে যাওয়ার পর তিনি প্রথমবারের মতো সুযোগ পেয়েছেন বিসিবির কোনো দলে খুলনায় শেখ আবু নাসের স্টেডিয়ামে কাল শুরু হচ্ছে একটি চারদিনের ম্যাচ। এই ম্যাচে খেলছেন মোহাম্মদ আশরাফুল। ম্যাচটা আশরাফুলের কাছে বিশেষ তাৎপর্যময়।

Fig. 3. A Sample Authentic Data

Here we show two different datasets for real and fake news. Real news dataset is shown and leveled as 1. The real news is collected from the reputed news sources on the other hand the fake news is collected from unreliable and non-reputed sources and leveled as 0.

As we mentioned, we have worked with 2082 data. In them 1500 data are collected from reputed sources and considered as real news. Rest of the data are considered non reliable and fake. Figure 4 shows the list of fake news.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Fig. 4. Fake News Dataset

We have shown all category of our dataset. Here we represent ten types of categories and how much times each of category is present like National, International, Sports, Editorial, Technology, Politics, Lifestyle, Entertainment, Finance, Miscellaneous etc in Table 1.

Table-I: Data Category Table

Categories	No. of News
National	970
International	340
Sports	102
Editorial	52
Technology	78
Politics	350
Lifestyle	80
Entertainment	70
Finance	30
Miscellaneous	10

We demonstrated how to process data, we gather text data including false and genuine news from several reputable and non-reputable news platforms. The data is then converted to CSV format, which can be read by machines, using Python's Panda module. Then, using a bag of words method, this text data is converted to numeric format.

We gather the frequency of each word in this procedure and assign the highest feature and document frequency. This function returns text data in numeric format. The TF and IDF values are then calculated. In a particular text, a bag of words is estimated to have a certain frequency.

As a result, we compute the TF and IDF values. Because TF and IDF compute frequency, all documents are calculated. The data is then converted to a matrix and used to generate train and test data. We utilize 80% news for training and 20% news for testing at work. CNN classifier methods are used at. This algorithm uses train data to identify false and genuine news. We obtain our anticipated outcome using this method.

B. Setting up the Required Environment

Bangla is a highly inflectional language, with each word having many meanings depending on inflection. We utilize a library function to identify each word and eliminate noise. This library function removes regular expressions, html tags, and punctuation. For this we have to use some built-in software and library function. We have used Python to build our model. For multidimensional array processing, data manipulation and analysis Numpy and Pandas are used. We utilize different scraping tools and libraries like Sracpy, ParseHub, Octoparse.

We have used SKlearn, Python for building our model. Anaconda is utilized for scientific computing using Python and R distribution that attempts to make package management and deployment easier [12]. PyCharm is an integrated development environment for computer programming, with a focus on the Python programming language.

C. Data Preprocessing

Before sending the raw text input to the classifier, it is essential to perform some preprocessing. A raw text may contain extraneous symbols or other elements marked as noise that are irrelevant to our classification. Preprocessing can be difficult as it depends on how you define "noise" and what causes it. A raw text might contain extraneous symbols as well as various emoticons like :), :(, ;), :D etc. Sometimes emoticons might be helpful for emotion detection and sentiment analysis but in our case, it doesn't work in such way. So, we remove them from our text. Detecting false news is difficult, fake news and genuine news share certain common words, brackets, and tags. As a result, we utilize the normalization function and import a Python library. Using this technique, we first call for text script in Beautiful Soup. Using a special function that returns text and ignores the rest. Built-in function SUB() is used to removes various brackets and returns just text.

Finally, we get noise-free text using this technique. Due to the fact that without decreasing noisy input, the machine would not be able to identify keywords, data reduction is required for training and testing. Before that, we also need to fill the missing data. The missing data is traced and filled with true value.

D. Feature Extraction

Proper feature extraction is very essential for accurate classification. It has a heavy impact on the performance of the machine learning classifier.TF-IDF Vectorizer are used to extract feature from the raw text.

Before that some other processes are required. Tokenization is one of them. In Natural Language Processing, tokenization is a frequent job (NLP). Both conventional NLP techniques like Count Vectorizer and Advanced Deep Learning-based systems rely on it [11]. Before feeding to a model the raw strings need some preprocessing. After finalizing the categories and removing noise it's time to tokenize the raw sentences.

Table II depicts the tokenized version of a raw sentences or string.

Table-II: Tokenized Version of A Sample

Raw sentence	রোহিঙ্গা নির্যাতনে আন্তর্জাতিক অপরাধ আদালতের প্রাথমিক তদন্ত শুরু
Tokenized form	"রোহিঙ্গা" " নির্যাতনে" "আন্তর্জাতিক" "অপরাধ" "আদালতের" "প্রাথমিক" "তদন্ত" "শুরু"

The bag of words model is one of a number of methods for extracting characteristics from text that come from the area of computer science known as Natural Language Processing, or NLP [10]. It does this by counting the number of times words appear in a document. The majority of the time, a bag of words is utilized for numeric format. The same word appears many times in a single line. Numeric format allows us to count their frequency.

In Table III, "রোহিঙ্গা" occurs 3 times from this text script, then "আন্তর্জাতিক" occur 3 times, "মিয়ানমার" occur 2 time, "বাংলাদেশে" occurs 2 times. Finally using of bag of word process, we get frequency of word. And we can simulate this text script and represent numeric format. In this process we have to import pickle library. This library used for vectorization. CountVectorizer() method is used for frequency define with some required arguments (max_features=1500, min_df=4, max_df=0.8) . Here df means that document frequency, max-df remove too much frequency. Here max df =.8 means that more than 80 percent frequency it will remove. And min frequency remove number of low frrquency. Here min_df=4 that means less than 4 frequency it will remove.

Table-III: Frequency Count of A Sample News

	Token	Frequency
রোহিঙ্গাদের বিরুদ্ধে মানবতাবিরোধী অপরাধে অভিযুক্ত মিয়ানমার সরকারের বিচারে প্রাথমিক তদন্ত শুরু করেছে আন্তর্জাতিক অপরাধ আদালত। এর মধ্য দিয়ে রাখাইনে সেনা অভিযান নিয়ে, একটি পূর্ণ তদন্তের পথ সুগম হলো। রোহিঙ্গাদের দেশত্যাগে বাধ্য করা, গণহত্যা, ধর্ষণের পাশাপাশি মৌলিক অধিকার হনন ও লুটপাটের অভিযোগের তদন্তে নেমেছে আন্তর্জাতিক অপরাধ আদালত, আইসিসি…..	"রোহিঙ্গা" "আন্তর্জাতিক" "মিয়ানমার" "বাংলাদেশে"	3 Times 3 Times 2 Times 2 Times . . . .

For translating text to numbers, the bag of words method works well. It does, however, have one flaw.

It gives a word a score depending on how many times it appears in a document. It doesn't account for the fact that the term may appear often in other texts as well. TFIDF addresses this problem by dividing a word's term frequency by its inverse document frequency. The letters TF and IDF stand for "Term Frequency" and "Inverse Document Frequency," respectively. [9]".

The term frequency is calculated as:

$$\text{Termfrequency} = \frac{X}{Z} \quad \text{..... (I)}$$

X =*Number of Occurrences of a word*

Z =*Total words in the document*

And the Inverse Document Frequency is calculated as:

$$IDF = [\frac{V}{N}] \quad \text{..... (II)}$$

V =*Total number of documents*

N =*Number of documents containing the word*

Every word is calculated in a particular way. Using this method, we get tf and idf value then it is converted into array. To prepare our dataset for the model, we first need to extract the feature using this tf-idf method. We have applied some methods to accurately calculate the TD-IDF and then fit the dataset for the model.

E. Classifier

Before Classification we need to perform the tasks including data processing, data cleaning, feature extraction as mentioned earlier. To feed the data into classification algorithms train and test dataset was gained by splitting the main dataset in two different sets. The splitting ratio was 80% for the training dataset and 20% for the test dataset. Text classification or sentence categorization into a set of predefined categories using n-grams, i.e., words or characters, sequences of words or sequence of characters is considered one of the most common tasks in natural language processing. ConvNet are employed to perform the task of classification and make the classification task easier. For this, a series of words w_1, w_2, w_3, ……., w_n is needed to feed to the convolutional network. We get a vector from bag of words that can also be feed to CNN.

Each of the vector is connected with a d-dimensional embedding vector. If a sliding-window of size k moves over the sentence we get a one-dimensional convolution of width k. Every window of the sequence is subjected to the same one-dimensional convolution filter, also known as a kernel. Particularly, a dot-product of the concatenation of the embedding vectors in a given window and a weight vector, followed by a non-linear activation function.

The data is pooled into a single one-dimensional vector in ConceptNet, which is a network of convolution windows.

The most essential elements of the sentence/document should be included in this vector. The notion that ConvNet is merely a feature extractor, leading to a fully connected layer for prediction.

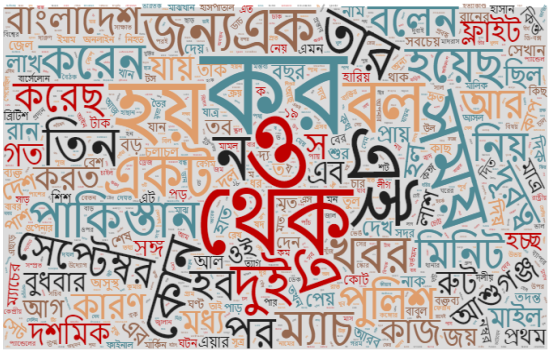

Fig. 5. Real News Word Cloud

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Our model collected 8114 tokens for the fake news and 15205 tokens for the real news. It also extracted some real news keywords and some fake news keywords form the dataset. Here we show the dataset output token words drawing image. Figure 5 depicts several tokens gathered from real-world news by our approach. His important phrases are more highlighted here, and the machine gets the most amount of them.

Fig. 6. Fake news Word Cloud

Figure 6 shows some of the keywords we gathered from false news using our algorithm. His important phrases are more highlighted here, and the machine gets the most amount of them.

IV. RESULT ANALYSIS

The accuracy level of a system may be used to determine its efficiency. Our suggested model is tested on a dataset gathered from a variety of prominent online newspapers. The testing dataset contains almost 417 data points. The ratio of the quantity of properly predicted data to the total amount of data is used to determine accuracy Our system's final(avg) recall and fl-score are both 0.83, which indicates it detects Bengali false news with an 83 percent accuracy rate, which isn't terrible. Our algorithm is extremely good at distinguishing between fake and real news. Table IV displays the outcome.

Table-IV: Experimental Result for CNN

	precision	recall	f1-score
Authentic	0.83	0.85	0.84
Fake	0.82	0.80	0.82
Average	0.82	0.83	0.83

Our model's accuracy level is given in Table IV. By training all datasets, this model will gather certain keywords. The number of false news keywords is 8114, whereas the number of genuine news keywords is 15205. For real news we total collect 15205 keywords. We discover certain limitations by examining the outcome. Firstly, we've gathered some news that doesn't fit into any of the categories. Based on the title and description of the news, we establish a category for it. As a consequence, certain news categories haven't been created properly. Because a news result is defined by her category, the model will provide a variety of outcomes if any news category is incorrectly set. This means that if a news category is altered and the news is tested, the outcome may differ. Second, here's a breakdown of how this masculine model works. The keywords in the description are then matched to the important phrases in the model tales. If any news is tested and given some potion or news, it will yield one result; but, if it is given the maximum potion or the whole potion, it will yield a different result, and many grammatical rules will be disregarded.

V. CONCLUSION

The model announces the outcome based on the description and category, according to the results of the experiment. The model does a category-by-category keyword search from the description to provide a result. To select the most significant phrases, the model compares the stored keywords to those in the dataset. The system then identifies false news and genuine news portal keywords in the input news portal, makes a judgment based on the similarity or dissimilarity of the stored keywords, and determines whether the news is fake or real.

We obtain a fantastic accuracy when compared to the other models. The majority of the models were created for false news identification in English, but we focused on bangle fake news detection. Because of the restrictions, we obtained an 83.14% accuracy, therefore it's a fantastic achievement for us.

The goal of this approach in the future is to increase the capabilities of the model removing the limitations. Accuracy rate will be improved with more data. As this model works based on the keyword, in future all types of news keywords will be recorded in categories. This model will feature a news alert to clearly identify the applicable news. Nonetheless, it will demonstrate how many individuals believe this model is false and how many people believe this story is true. It will also logically describe the customer by comparing or presenting the client with a powerful online news portal that has previously posted this news on their website.

REFERENCES

1. J. K W. Y. Wang, “” liar, liar pants on fire”: A new benchmark dataset for fake news detection,” arXiv preprint arXiv:1705.00648, 2017.
2. A. Hadeer, I. Traore, and S. Saad. "Detection of online fake news using N-gram analysis and machine learning techniques." In *International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pp. 127-138. Springer, Cham, 2017.
3. S. Pal, T. K. S. Kumar, S. Pal.” Applying Machine Learning to Detect Fake News.” Indian Journal of Computer Science, Volume 4, Issue 1, February 2019.
4. F. Monti, F. Frasca, D. Eynard, D. Mannion, M.M. Bronstein. “Fake news detection on social media using geometric deep learning.” arXiv preprint arXiv:1902.06673. February, 2019.
5. S. Gilda, "Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection," 2017 IEEE 15th Student Conference on Research and Development (SCOReD), Putrajaya, 2017, pp. 110-115, doi: 10.1109/SCORED.2017.8305411.
6. L. Waikhom and R. Goswami, “Fake News Detection Using Machine Learning.” Proceedings of International Conference on Advancements in Computing & Management (ICACM) 2019, October 2, 2019.
7. A. Thota, P. Tilak, S. Ahluwalia and N. Lohia,, "Fake News Detection: A Deep Learning Approach," *SMU Data Science Review*: Vol. 1 : No. 3 , Article 10. SMU, 2018.
8. *Machine Learning Project Structure: Stages, Roles, and Tools*, Altexsoft. Accessed on: 04 May, 2020. [Online]. Available: https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

9. Hadeer Ahmed, Issa Traore, SherifSaad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques, Conference paper, First Online: 11 October 2017.
10. "Bag of words model", *Wikipedia*. Accessed on: May 05, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Bag-of-words_model
11. "Tokenization in NLP", *Analytics Vidhya*. Accessed on: April 20, 2021. [Online]. Available: www.https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/.
12. "What is Anaconda and how does it relate to Python", *VENTURE LESSONS*. Accessed on: Jan 15, 2021. [Online]. Available: https://www.venturelessons.com/what-is-anaconda/

AUTHORS PROFILE

Md. Rakib Hasan, has completed his B.Sc. (Honors) and M.Sc. in Computer Science and Engineering from Jahangirnagar University, Dhaka, Bangladesh in 2016 and 2018 respectively. He is now working as a Lecturer at Department of Information and Communication Technology, Comilla University, Cumilla. Previously he worked as a Lecturer at Department of Computer Science and Engineering, Daffodil International University, Dhaka from May 2016 to February 2019. His teaching experience includes different graduate (M.Sc.) and under graduate courses. He is currently working on Artificial Intelligence and Machine Learning. His research interest includes Natural Language Processing, Machine Learning, Cyber Security and Computer Vision. He is actively engaged in educational activities.

Ismath Ara Itu, has completed her B.B.A and M.B.A. in Management Information Systems from university of Dhaka, Dhaka, Bangladesh in 2015 and 2016 respectively. She is currently working on Data Science, Secure Information Systems and Machine Learning. Her research interest includes Machine Learning, Data Science and Information System Security.