Green University

GUBJSE: ISSN: 2409-0476

Restriction Algorithm for Duplicate Posts on Online Social Networks: Facebook

Babe Sultana, Md. Nasir Hossain Hridoy, Mohammad Mohitul Islam, Ruhul Amin, Farzana Rahman

***Abstract*—** Duplicate content or writing on online social networks is a material that shows up in many more than one location on Online Social Site, Pages etc. Now a days Facebook is an online social networking site that connects people together during the form of expressing personal preferences and opinions as well as communication. In this research paper, we found detecting duplicate material in Facebook groups, pages, and trying to provide a solution for limiting this duplicate content, that is being posted to Facebook and other online social networks. We specified the solution to the issue in the first step and designed an algorithm called Restriction Algorithm for Duplicate Content, which is restricted to posting the copied content in more times on social networks like Facebook. In the second step, we have implemented it to validate our methodology and we have checked the identification of duplicate content of social media writing by using various social media posts as input tests and finally enriching the findings at a satisfactory stage. With optimal computation time, our proposed algorithm can handle large string sizes (more than 10,000 bytes).

***Index Terms*—** Social Media, Online Social Network, Duplicate Contents, Facebook Group

I. Introduction

TODAY, as a medium of information and communication, social media is rising undoubtedly very fast. Social media and particularly online Social Networks (OSN) have grown since the start of this century as sites where audiences create and share information and interact with it rigorously via various actions. Among the online social websites such as Facebook, Instagram Twitter, and LinkedIn draw users more and more. Online Social Networks have billions of users and they have different sets of understanding, competence, and abilities, so it can be seen as a group model of knowledge [1]. This landscape is dominated by Facebook and YouTube, as remarkable U.S. adult majorities use each of these sites. Simultaneously, younger Americans (especially those aged between 18 and 24) stand out for promoting and frequently using a variety of platforms. Approximately 78 percent of 18- to 24-year-olds use Snapchat, and a large majority (71 percent) visit the app repeatedly a day. Furthermore, 71 percent of Americans now use Instagram in this age group and almost half (45 percent) use Facebook [2]. In 2018, a recent survey conducted by the company found that users of the site hated material that was copied and scraped from other outlets. In this article [3], we found that there will be two forms of duplicate content and both could be a concern. The first one is onsite duplication and the second one is offline duplication. Both site duplication is irritating for random users. Onsite replication is when the site duplicates the same content on two or more separate URLs. This is usually something that the site admin and the web development team can handle. And then Offsite replication is when the very same pieces of material are released by two or more pages. It is something that can always not be explicitly monitored but relies on collaborating with third-party companies and the administrators of the violating websites. In figure: 1 we can find an idea that how the percentage of the user is increasing day by day who spend their lots of time in online social networks. So, obviously it will be an important issue to work for the duplicate posts on online social networks. Facebook has confirmed that most of the sites that do

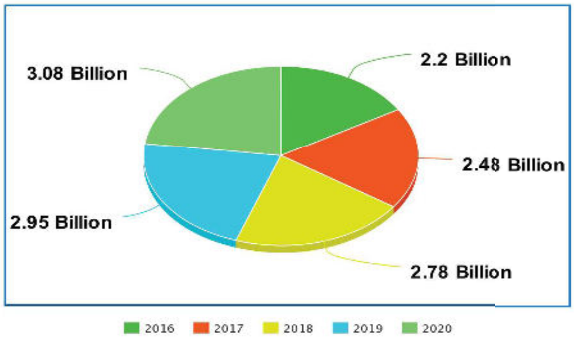


Fig. 1: The worldwide social network users summary from 2016 to 2020 (in billions)

This paper was received on 13 May 2020, revised on 6 October 2020 and accepted on 13 October 2020. This work was financially supported by GUB.

Babe Sultana is with the Computer Science & Engineering, Green University of Bangladesh, Dhaka, Bangladesh. E-mail: *babe@cse.green.edu.bd*

Md. Nasir Hossain Hridoy is with the Computer Science & Engineering, Green University of Bangladesh, Dhaka, Bangladesh. E-mail: *hridoymix123@gmail.com*

Mohammad Mohitul Islam is with the Computer Science & Engineering, Green University of Bangladesh, Dhaka, Bangladesh. E-mail: *mohitul838@gmail.com*

Ruhul Amin is with the Computer Science & Engineering, Green University of Bangladesh, Dhaka, Bangladesh. E-mail: *himel.amin95@gmail.com*

Farzana Rahman is with the Computer Science & Engineering, Green University of Bangladesh, Dhaka, Bangladesh. E-mail: *farzanamumu95@gmail.com*

such things are sites of low quality, rife with cheap advertising, clickbait headlines, and spammy landing pages. Facebook authority think about that and try to make out something new protocol or algorithm. The new protocol would essentially ban any site that appears to be using duplicate content and ads that are irrelevant to the discussion at hand on its News Feed. Also, the algorithm can identify articles that are not relevant to the subject or seem under-exaggerated and intellectually dishonest [3]. Nowadays people are passing their leisure time browsing the online social site and all the time they want unique content, news, funny post which haven't shown yet. By considering this, we investigate that girls and boys passing more time by scrolling various Facebook groups, Facebook pages than scrolling other things. There is a common issue that if one of post gains more popular or gain more like and comment, other members have a common intention that, they copied it and post it more and more time without mentioning the real author or anything else. It is irritating for the audience to see the same post, contents again and again. We have investigated one important point that audience shows their anger in the comment section, why they posted the same status again and again.

In very recent, many researchers have worked with this problem and try to propose new ideas for overcoming this irritating issue. Authors [4] work for Detecting Sockpuppets in Social Media and they used Plagiarism Detection Algorithms. In their research paper, they discuss several possible approaches to classify the user accounts – sockpuppets by adding plagiarism detection algorithms to test and evaluate their success against the various types of threat. In very recent, authors [5] have shown in their paper a plagiarism analysis for social media content and picture using URL as input sample. They use Smith-Waterman Algorithm and Latent Semantic Analysis method for measuring similarities. They only check the similarity measurement but one important thing came out that, there is no restriction analysis is not found on their paper when users are posting often the same post. And also, we study some related research paper who worked with duplicate contents on online social networks. Authors [6-8] work with detecting duplicate contents like finding duplicate web pages, duplicated of partial content detection, duplicate content search etc on online social networks. After concerning the above discussion, we decide to think with some new ideas that the duplicate post is needed sometimes is not. So, we decide that if we give an option than three times can be OK but not more than three times posting the same duplicate post on the same Facebook group, page, or any other social site. After that, by doing these members get an alarming statement for this. So, in this research paper, we propose an algorithm that makes a restriction to post the same status when it has already done three times on the same Facebook group or Facebook page.

II. Background Study

Facebook has become more than a social network, as using it is easy to reach a large number of people. We can see a short overview of Facebook of 2020 from figure 2, where worldwide total active users as of the first quarter of 2020 [9] are shown with monthly approximate comments which are left on Facebook pages [10]. Nowadays, most people use Facebook for their business. We also can see the approximate active small business pages amount on Facebook [11]. Besides pages, another great communication feature of Facebook is a group, which is provided real value with promotional content. There has also an idea of Facebook groups and the number of people uses Facebook groups every month [9]. The research of Facebook with users showed that aged between 18 to 24 are open to brands posting in Facebook groups as long as those posts provide real value rather than just promotion. Another popular Facebook feature is to share posts. In average each user likes 13 posts, also make 5 comments, and share 1 post per month [11]. The like button pressing times in between every 1 minute [10] is given in the figure. Besides, 317,000 status updates; 400 new users; 147,000 photos uploaded; and 54,000 links have been shared in a month [10]. Including good sides, Facebook has some bad sides also. There have a huge number of fake user accounts has and one of the large problems is duplicate posts. Users can get duplicate posts from their Facebook friends. The frequent reason for duplicate posts in the pages can be the item URLs in the news feed have changed. We can see different percentages of Facebook using the year 2020 in figure 3. Need for less data for using Facebook and easily usable, using Facebook is sort of easy to reach most of the people. Generally, people use Facebook for different reasons, like – just for going through the news feed where 45% of people use it for getting news [12]; 88% of user are on there to stay in touch with family and friends [13]; sometimes posting own updates like – 40% people said they would share their health data with Facebook [12]; commenting, liking, and sharing different posts; managing group or page;

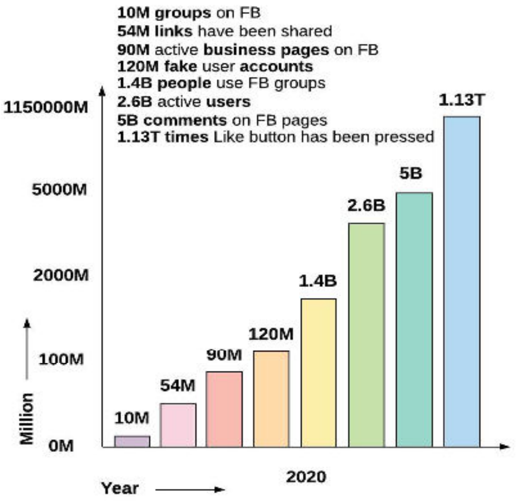


Fig. 2: A short overview of Facebook of the year 2020

business purpose; promoting work; someone uses it for showing their talents, etc. We know about the active small business pages on Facebook in figure 1, here, worldwide 93.7% of businesses use Facebook [14]. There are 49% of users like a Facebook page and 42% of users don't like any pages [12]. Almost 75% of high-income earners use Facebook for its excellent advertising platform [9]. People make the page for many kinds of business purposes, in between them, the average engagement rate of make by videos is 6.13% [11] which is about 11% of Facebook posts, where 89.5% of businesses prefer to share their video content in average Facebook shares. People use Facebook to promote their work also. After video making and sharing, Facebook is famous for selling and buying products. In 2016, the Facebook marketplace was introduced as an online shopping channel for people to buy and sell from each other locally, sort of via Messenger. Product selling business is such an intense that 57% of consumers claim that social media influences them to do the shopping [15] and 78% of consumers say that they have discovered products on Facebook. Among the followers of business pages [15], to receive special offers from different site 39% of users follow the Facebook page. We have known about the Facebook posts sharing rate in the previous figure. After this research, 78% of respondents said, they share posts because it helps them to stay connected to people they may not otherwise stay in touch with. 69% people said that, they share their thought, information for self-fulfillment and it allows them to feel more involved in the outside. The sharing posts on Facebook in the top 500 posts in 2018, 81.8% of videos, 18% of images, and 0.2% of links [14]. People use Facebook for another great reason which is showing their talents like – artists, singers, bloggers, chefs, video content makers, writers, photographers, etc. As it is a great platform, so, users love to share their talents with everyone by Facebook, and most recently there has an intent to get viral via Facebook all over the world which means a position has generated an intense level of attention in the form of a high number of likes, comments, and shares. The median virality rate for Facebook Pages is 1.92%. There are called a Facebook celebrity, it made someone's life!

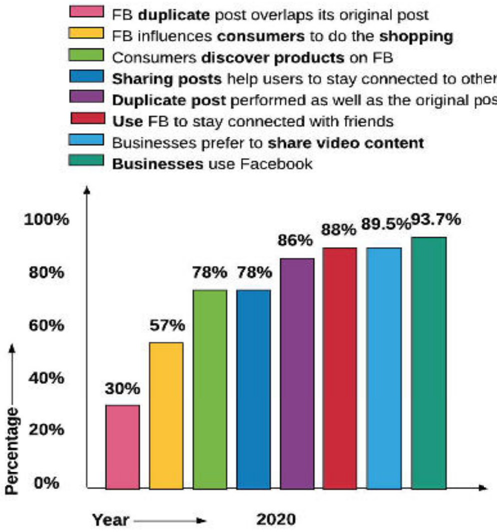


Fig. 3: Analysis of the duplicate post percentages of Facebook until 2020.

The rest of this paper is organized as follows. The details of our proposed algorithm with a working flow diagram are described in section III. In section IV, we have described the outcome of the details of our proposed algorithm with proper direction. And finally, we have included our conclusion with limitations and future direction in Section V.

III. Methodology

A. Problem formulation

Facebook has 2.4 billion members, the world's biggest social networking site. There are also more than one billion users in other social networks, including Twitter and WhatsApp [16-17]. Users want more flexibility and want not to get irritating when they browsing the online social site. Seeing the same post over and over in the same Facebook group is irritating for all. For search engines, duplicate content that poses three key problems [18]:

- Don't even know which version(s) we have to include / prohibit from indexes.
- Don't know how and when to guide the parameters of the links (believe, legitimacy, anchored text, connection equity, etc.) to one site or retain it isolated among different variants.
- Don't know which version(s) should be listed for the results of queries.

Search engines will seldom show several versions of the same content to have the best search experience, and are thus forced to select which version would most likely be the best result. And another important point is that, link equity may be further reduced because other sites often have to choose between the duplicates. Rather than all spammy links referring to one piece of content, they connect to multiple parts, spreading the equity connection between the duplicates. Site owners will experience rankings and traffic failures when duplicate content is present for above those two main issues.

Facebook is one of the leading online social networks where many users use this platform for their business purpose. To do this user create group, page to connect many users. Considering these three key points we can say that, these problems may arise when we use Facebook pages, groups for sharing our unique contents, selling posts, or anything else. By concerning this issue, we have proposed this algorithm to restrict duplicate posts. In figure: -4 we have shown our working flow how this algorithm works. In the proposed algorithm section, we have described the detailed process of our algorithm.

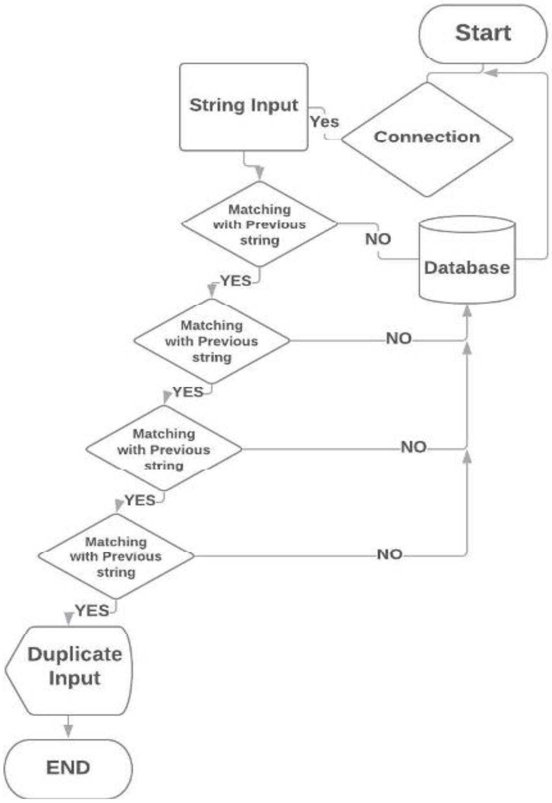

Start
String Input
Yes
Connection
Matching with Previous string
NO
Database
YES
Matching with Previous string
NO
YES
Matching with Previous string
NO
YES
Matching with Previous string
NO
YES
Duplicate Input
END

Fig. 4: Workflow diagram of Restriction Algorithm for Duplicate Posts

.

B. Proposed Algorıthm

The concept of our proposed methodology is based on detecting duplicate contents and making a restriction in most used online social site named Facebook. In our proposed algorithms three functions and database tables have been used. It works in three steps which are shown in our proposed algorithm named Restriction Algorithm for Duplicate Posts.

- **Post () function**: In Post () function, first it will check the Connection between Interface and Database. If the Connection is OK then it will ask the user to input a String. Then initial the input value and store the value in the first table of the Data base. If the input value is equal to the stored value then it will call a sub-function Check (), whose parameter is String. Else the Input String will be stored in the first table of Database.

- **Check () function** : In Sub Function Check (), it will initiate the store value of the second table of Database. If the input value is equal to the stored value then it will call a sub-function Recheck (), whose parameter is also an input String. Else the Input String will be stored in the second table of Database.

- **Recheck () function**: In Sub Function Recheck (), it will initiate the store value of the third table of Database If the Input Value is equal to the stored value then it will stop the process and shows an end statement. Else the input String will be stored in the third table of Database.

Algorithm 1: Restriction Algorithm for Duplicate Content

```
Input: String Input
Output: Restricting same posting after three
        attempts
Post( ) // This is main function
// Connecting to database
if (Connection != Null) then
    I
nitial input variable ← values from users;
Initial connecting variables ← connect to first
  table of database;
if (input variable = connecting variable) then
    Check( $inputvariable$);
    // Check is a sub function
else
    First table ← input variable;
end
Check($stringinputvariable$) // Check is
    a sub function
Initial input variable ← values from users;
Initial connecting variable – ← connect to
  second table of database
if (input variable = connecting variable) then
    Recheck( $inputvariable$);
    // Recheck is a sub function
else
    Second table ← input variable;
end
Recheck($stringinputvariable$)
// Recheck is a sub function
Initial connecting variables ← connect to third
  table of database;
if (input variable = connecting variable) then
    (
print statement(Posting restricted!! Duplicate
  post found!!);
else
    Third table ← input variable;
end
```

IV. Performance Evaluation

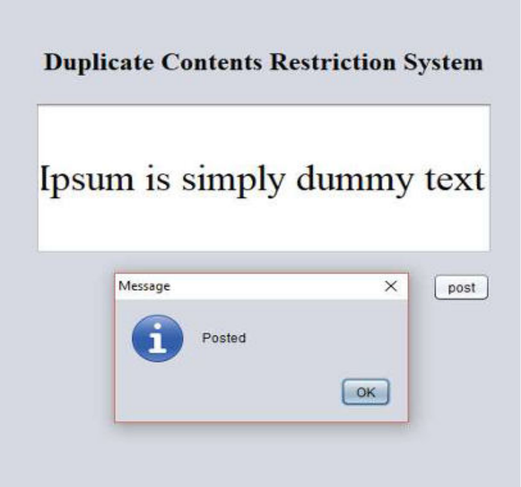
A. Environmental Setup

We have implemented our proposed algorithm which has been developed under the following environment:

- **Operating System:** Windows 10 64bit
- **Processor:** Intel® Core^{TM} i3- 7100 CPU @ 3.90 GHz
- **RAM:** 8 GBytes
- **IDE:** Net beans
- **Programming Language:** Core Java

B. Results and Discussion

1) System Outcome Analysis: We have implemented our proposed algorithm in java to verify our outcomes. We have used a Java text field for taking the



Duplicate Contents Restriction System

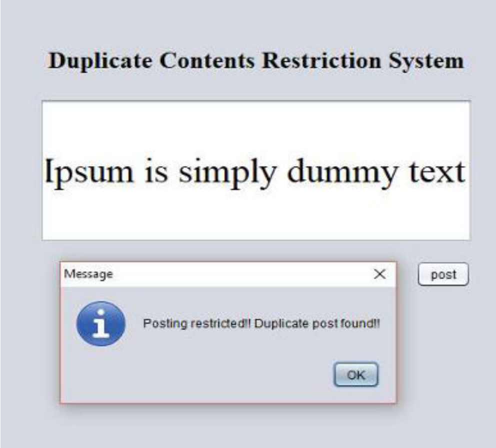
Ipsum is simply dummy text

Message

Posted

Fig. 5: The system will allow the user to post duplicate content three times

input sample. When any user gives input, in the first attempt the system will allow a user to post because of unique content. In the second attempt, the system will allow the user to post duplicate content, and until the third attempt (Figure: -5). In the fourth attempt, the system will restrict the duplicate content posting by giving this alarming message which is shown in Figure: -6.



Duplicate Contents Restriction System

Ipsum is simply dummy text

Message

Posting restricted!! Duplicate post found!!

Fig. 6: In the fourth attempt, the system will restrict the Duplicate content Posting!!

2) Impacts of increasing the size of strings: We have checked our system to see the needed run time for various sizes of string in bytes which are given in Table I. Using these data, we have drawn a bar chart graph where we can see the computation time is increasing with the increase of the size of strings. From the graph, the result says that for 10000 bytes system need only 55 seconds which is much efficient.

3) Computation Complexity: The system will depend on the user inputs. If the user gives input only one time then it will be executed at one time. Second and third times it will be worked for similar or different inputs. When a user gives input the same values in fourth times the system will restrict that for similar posts. If the user gives input to other values then the program will execute smoothly. So, the proposed system will execute n times for n inputs. So, the best, average and worst-case will be Big-oh(n).

TABLE I: Table of Computation Time

Size of String	Run time (s)
1000 Bytes	28
2000 Bytes	30
3000 Bytes	33
4000 Bytes	36
5000 Bytes	40
6000 Bytes	43
7000 Bytes	46
8000 Bytes	49
9000 Bytes	53
10000 Bytes	55

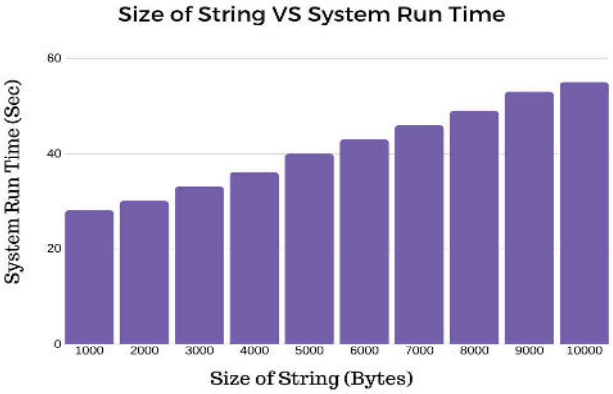


Fig. 7: Impacts of increasing the Size of Strings VS System Run Time

V. Conclusion & Future Work

In this research paper, we have presented an idea with a proposed algorithm which has made a restriction for posting often the same or copied contents of the writing. Here, our attempt is to propose an idea for solving an irritating issue that happened on online social networks. When browsing online sites, users want new content rather than the same content over and over. Our aim is to focus on this irritating issue that happened all over the world on online social networks. In this research paper, we have tried to give an idea for taking a proper step to restrict this occurrence.

There is a limitation of our proposed approach is that, if add or changes in a single byte of strings it couldn't detect any duplicate content. But we will work for it which is under progress and our future goal is to work with the plagiarism measurement analysis.

Acknowledgement

This research was supported and funded by the Green University of Bangladesh.

References

[1] S. Asur, B. A. Huberman : ”Predicting the future with social media”. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology- Volume 01. IEEEComputer Society, 492—499 (2010)

[2] https://www.pewresearch.org/internet/2018/03/01/socialmedia-use-in 2018/.

[3] https://www.socialmedia.biz/no-more-copied-content-on facebook/.

[4] F. Albrektsson, "Detecting sockpuppets in social media with plagiarism detection algorithms," 2017.

[5] M. Kurniawan and K. Surendro, "Similarity measurement algorithms of writing and image for plagiarism on facebook's social media," in IOP Conference Series: Materials Science and Engineering, vol. 403, no. 1. IOP Publishing, 2018, p. 012074.

[6] M. Henzinger, "Finding near-duplicate web pages: a largescale evaluation of algorithms," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 284–291.

[7] C. C. Mysen and J. Chen, "Duplicate content search," Nov. 20 2008, uS Patent App. 11/749,561.

[8] Y. Qingwei, W. Dongxing, Z. Yu, and W. Xiaodong, "The duplicated of partial content detection based on pso," in 2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA). IEEE, 2010, pp. 350–353.

[9] https://zephoria.com/top-15-valuable-facebook statistics/.

[10] https://www.omnicoreagency.com/facebook statistics/.

[11] https://blog.hootsuite.com/facebook statistics/.

[12] https://www.brandwatch.com/blog/facebook statistics/.

[13] https://sproutsocia.com/insights/facebook-stats-for marketers/.

[14] https://buffer.com/resources/facebook-marketing 2019.

[15] https://adespresso.com/blog/facebook statistics/.

[16] E. Djafarova and O. Trofimenko, "'instafamous'–credibility and self-presentation of micro-celebrities on social media," Information, communication & society, vol. 22, no. 10, pp.

[17] R. Enikolopov, A. Makarin, and M. Petrova, "Social media and protest participation: Evidence from russia," Econometrica, vol. 88, no. 4, pp. 1479–1514, 2020.

[18] https://moz.com/learn/seo/duplicate content.

Babe Sultana was born in, Cox's Bazar, Bangladesh, in 1994. She received her B.Sc. Degree in Computer Science and Engineering from Green University of Bangladesh (GUB) in 2018. At present she is working as a Lecturer, Dept. of CSE, Green University of Bangladesh. Also, her publication was about **"Multimode Project Scheduling with Limited Resource and Budget Constraints"** published in **International Conference on Innovation in Engineering and Technology (ICIET) 27-28 December, 2018** and she got **Best Paper Award** and **IEEE Best Paper Award** on this conference. Her research interests include Theory of Optimization, Natural language Processing, Renewable and Sustainable Energy.

Md. Nasir Hossain Hridoy was born in Kaliakair, Gazipur, Bangladesh in 1997. He is currently in 10th semester of his BSc in Computer Science and Engineering degree in Green University of Bangladesh. He is expecting to complete his BSc program within 2021. His publication **"Duplicate Contents Restriction Algorithm for Copied Post on Online Social Network"** has been published in **"International Conference on Sustainable Technologies for Industry 4.0" 24-25 December,2019.** His research interests are Machine Learning, Internet of Things (IoT), Natural Language Processing (NLP), Deep learning, Big Data.

Mohammad Mohitul Islam was born in Dhaka, Bangladesh, in 1994. He is currently in 10th semester of his B.Sc in Computer Science and Engineering degree of Green University of Bangladesh. He is expecting to complete his B.Sc program within 2021. His publication **"Duplicate Contents Restriction Algorithm for Copied Post on Online Social Network"** has been published in **"International Conference on Sustainable Technologies for Industry 4.0" 24-25 December, 2019.** His research interest includes Internet of Things (IoT), Natural Language Processing (NLP), Data Mining, Machine Learning.

Ruhul Amin was born in Araihazar, Narayangonj, Bangladesh in 1995. He is currently in 10th semester of his BSc in Computer Science and Engineering degree of Green University of Bangladesh. He is expecting to complete his BSc program within 2021. His publication **"Duplicate Contents Restriction Algorithm for Copied Post on Online Social Network"** has been published in **"International Conference on Sustainable Technologies for Industry 4.0" 24-25 December,2019.** His research interests are Natural Language Processing, Deep learning, Block chain.

Farzana Rahman was born in Chandpur, Bangladesh in 1996. She received the B.Sc. Degree in Computer Science amp; Engineering from the Green University of Bangladesh (GUB). She is now studying for a Master's degree in Data Science at the Technical University of Dortmund, Germany. She achieved Merit Awards for doing excellent results each semester during her bachelor study period at GUB. She is a trained professional Mobile Application Developer. She has a publication about **"A Framework to Figure Out Breast Cancer Cells Using Ultrasound Images**; at the **11th International Conference on Computing, Communication, and Networking Technologies (ICCCNT) 1-3 July 2020.** Her research interests include Artificial Intelligence, Machine Learning, and Image Processing.