DATA PREPARATION FOR CHATBOT

Richa Patel[1]
Masters of technology in Data Science
Ahemdabad University
Ahemdabad,Gujarat,India
pricha9265@gmail.com

Shakirali Vijapura[2]
Masters of technology in Data Science
Ahemdabad University
Ahemdabad,Gujarat,India
shakiralivajapura97@gmail.com

***Abstract*—Digital documents are available anytime, anyplace due to which the management of documents and retrieval of information have rapidly increased. For information organization and knowledge discovery, it is prominent that the extraction of information from documents becomes automated. One such solution is Text Classification, where given the natural language text will be assigned to one or more predefined categories based on the content. Users can make benefit out of chatbot system through the integration of electronic documents with simulate system by asking and answering questions. Integration of knowledge gathered from popular formats of documents such as digital photos, scanned documents and PDF makes this system. This system firstly extract texts from above mentioned documents using OCR (Optical Character Recognition) followed by applying the filter i.e. preparation of data for machine learning. Algorithms of machine learning learns from the data fed. Hence, right data has been fed to algorithms to get the effective output. The data has to be in correct format, must include meaningful features and is in useful scale.**

***Index Terms*—Chatbot, Data Preprocessing, OCR (Optical Character Recognition), Machine Learning.**

I. Introduction

A Chatbot is a popular agent that is able to communicate with users on a given topic by using natural language [1]. Answering questions from the user, providing comments or bringing a topic to be discussed with the user are the abilities that chatbot beholds normally. For the purpose of education, customer service, site guidance, or even entertainment functions, many chatbots have been deployed on the internet.

A. Architecture of Data Preparation for Chatbot

In Figure.1, chatbot is a representative that talks with the user. User fires the question to chatbot. Chatbot understands the question using Natural language processing then tries to find the appropriate answer to the question from the available documents using machine learning.

There are lots and lots of documents available on the internet in the different formats like documents, images, scanned documents, typed documents, handwritten documents. These documents contain a lots of knowledge. Why not use these documents to provide knowledge to the chatbot? OCR technique read these documents of different formats and gives a plain text document as on output. Now these output text documents are the files from which the machine tries to find an appropriate answer to the user's question. Our system approaches to filter these document before providing it to machine to search the answers using the platform R. Here filter means preparing the data for machine learning.

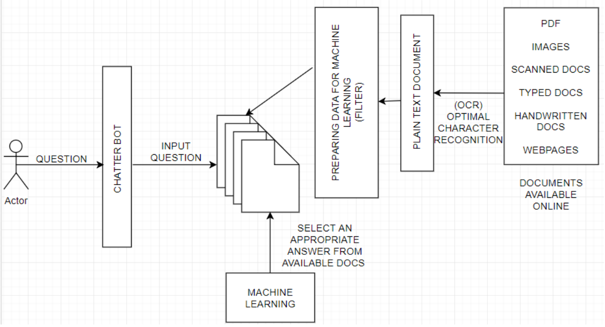

Actor
QUESTION
CHATTER BOT
INPUT QUESTION
PREPARING DATA FOR MACHINE LEARNING (FILTER)
PLAIN TEXT DOCUMENT
(OCR) OPTIMAL CHARACTER RECOGNITION
PDF
IMAGES
SCANNED DOCS
TYPED DOCS
HANDWRITTEN DOCS
WEBPAGES
DOCUMENTS AVAILABLE ONLINE
SELECT AN APPROPRIATE ANSWER FROM AVAILABLE DOCS
MACHINE LEARNING

Fig. 1. Architecture

B. Working Flow of Data Preparation Followed By Machine learning

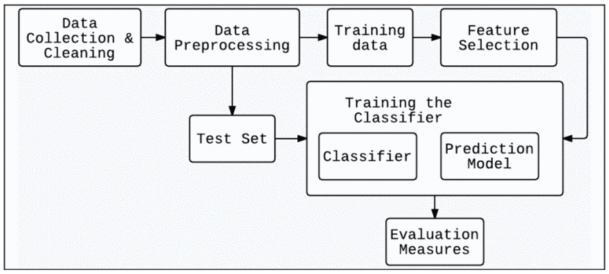


Fig. 2. Working

Multitier classification system's architectural design is explained in this segment, as shown in Figure 1. This design shows that the classification system has seven stages: [2]

1. Data Collection and cleaning
2. Data Preprocessing
3. Training data
4. Feature Selection
5. Training the classifier
6. Test set
7. Evaluation measures

II. Literature Review

A. Pre-processing Data for Machine Learning

Pre-processing refers to the changes applied to our data before supplying it to the algorithm. Data Pre-processing is a method that is used to transform the raw data into a cleaned data set. In other words, whenever the data is assembled from various sources it is assembled in raw format which is not likely for the analysis.

B. Need of Data Pre-processing

For gaining better results from the fitted model in Machine Learning projects the form of the data has to be in a decent manner. Some named Machine Learning model needs information in a particularized format, for example, Random Forest algorithm does not support void values, therefore to perform random forest algorithm void values have to be managed from the initial raw data set. Another phase is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are performed in one data set, and excellent out of them is chosen.

III. Methodology

The more disciplined you are in your handling of data, the more consistent and better results is achieved. The process for summarizing and getting data ready for a machine learning algorithm can be done in three levels:

Level 1: Data Selection
Level 2: Data Pre-processing
Level 3: Data Transformation

It is very likely to be iterative with many loops, but you can follow this process in a linear manner.

A. Data Selection

This step is involved in selecting the data or their subset, that you will be working with. It is a strong recommendation to include as much as data possible but it may not be true.

You actually need to understand your question or problem statement and then you need to find corresponding data that can solve that problem. When selecting data considering certain assumption be careful during the test to record those assumptions.

Below are some questions to help you through this process:

• How much and up to what extent data is available? For example, through time (yearly or monthly), or database table, csv file, connected system. You should be very clear that what you are using from that data.

• The data you wish you had available but it is not available? For example, data that is not recorded or missing value in the dataset.

• Selecting data which will solve or address the problem? Including data is almost always hard than excluding it. You should have a proper reason which data you excluded and why. It is only in a small problem or competition where the data has already been selected but in real life, you need to handle large and dirty data.

B. Data Pre-preprocessing

You need to consider how you are going to use the data, after selecting it. following pre-processing should be followed for getting data into a form that you can work.

1) Formatting Data
2) Cleaning Data
3) Sampling Data

• Formatting the data: We have to convert the data into suitable form because the data selected may not be in a suitable format for you to work with it. The data may be in a combined database or combination of text, audio or video but you would like it in a text file or proper file format.

• Cleaning the data:The removal or fixing of missing data is known as data cleaning. There may be some instances in data that are incomplete or which are not going to address the problem. This data should be removed.

• Sampling data: You have to work with selected data because more data can result in larger computational power, increase runtime for algorithms and also increase memory requirements. So to overcome above problem you can take a sample from data that may be faster for prototyping solution and exploring before considering the whole dataset.

C. Data Transformation

Now the transformation of the processed data is the final step. The problem domain's knowledge and the algorithm with which one is working with will definitely have an impact on this level. As on goes on with the problem it is likely that one has go refers the transformation s of the data that is pre-processed. Scaling of the data, decomposition of attributes and aggregation of attributes are the features of the data transformation as known aa feature engineering.

• Scaling of the data: It is possible that the pre-processed data includes of the attributes which may contains the attributes of different scales such as centimeter, meter or kilometers. Scale of the data attributes should be of the same scale such as between 1 to 2 for smallest and largest numerical values.

• Decomposition of attributes: Sometimes features defines very complex values that may not be useful but when this features are broken into pieces they may give certain constituent parts which proves to be more useful. An example is a birthdate but we need age in our problem so by decomposing the birthdate into age can be more useful.

• Aggregation of attributes: Sometimes by combing some features can be more useful rather than individually. Let's take an example, the session is created whenever a user logs into his/her account then we can simply count the number of sessions created to get the value of currently online users.

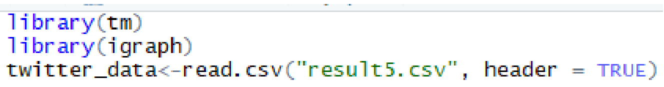
IV. Implementation

The objective of this task is to prepare data to find meaningful insights. We have taken a data from twitter. Around 30 tweets are extracted using R. Then we proposed R script to perform data cleaning and transformation.

A. *Pre-processing and cleaning Data*

Pre-processing refers to the changes applied to our data before supplying it to the algorithm.

For application of machine learning algorithms to the data, information has to be extracted from the raw data by mining process which is done by pre-processing the data. It is evident that one is working with the inconsistent and noisy data if this step is skipped. To find the knowledge from tweets, we don't need the punctuation, numbers, terms and special characters because they don't play a prominent role in finding the knowledge from the data. Hence, the purpose of doing this step is to remove such unwanted things from the data.

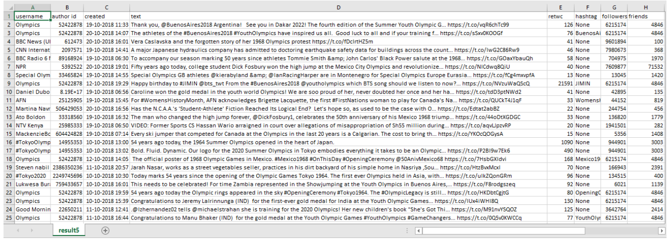


```
library(tm)
library(igraph)
twitter_data<-read.csv("result5.csv", header = TRUE)
```

Fig. 3. Importing Data

Here, the file named 'result5.csv' has twitter data in which Eight columns were extracted named "username"(Username of the user), "author id"(Id generated by system), "created "(created date), "text"(tweet), "retwc"(how many retweets the tweet got), "hashtag"(hashtags in the tweet), "followers" and "friends".

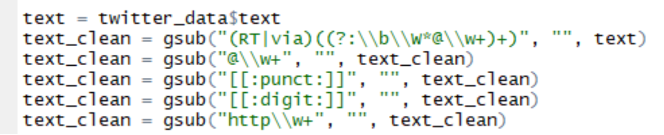
Extracted Data:



	A	B	C	D	E	F	G	H
1	username	author id	created	text	retwc	hashtag	followers	friends
2	Olympics	52422878	19-10-2018 11:33	Thank you, @BuenosAires2018 Argentina! See you in Dakar 2022! The fourth edition of the Summer Youth Olympic G... https://t.co/vqR6chTc99	126	None	6215174	4846
3	Olympics	52422878	20-10-2018 14:07	The athletes of the #BuenosAires2018 #YouthOlympics have inspired us all. Good luck to all and if your training f... https://t.co/sSxv0KOOGf	76	BuenosAi	6215174	4846
4	BBC News (UI	612473	20-10-2018 16:01	Vera Caslavska and the forgotten story of her 1968 Olympics protest https://t.co/fDclrtHZ5m	41	None	9601894	100
5	CNN Internat	2097571	19-10-2018 14:41	A major Japanese hydraulics company has admitted to doctoring earthquake safety data for buildings across the count... https://t.co/IwG2C86Rw9	46	None	7980673	368
6	BBC Radio 6 I	89168924	19-10-2018 06:30	To accompany our season marking 50 years since athletes Tommie Smith &amp; John Carlos' Black Power salute at the 1968... https://t.co/GOaxYbauQh	58	None	704975	1970
7	NPR	5392522	20-10-2018 19:01	Fifty years ago today, college student Dick Fosbury won the high jump at the Mexico City Olympics and revolutionize... https://t.co/NICdwq8QiU	40	None	7609877	71532
8	Special Olym	33465824	19-10-2018 14:55	Special Olympics GB athletes @kierabyland &amp; @IanRacingHarper are in Montenegro for Special Olympics Europe Eurasia... https://t.co/fCg4mxvpfA	13	None	13045	1420
9	Olympics	52422878	12-10-2018 19:29	Happy birthday to #JIMIN @bts_twt From the #BuenosAires2018 @youtholympics which BTS song should we listen to now?... https://t.co/NVzuWaQScQ	21591	JIMIN	6215174	4846
10	Daniel Dubo	8.19E+17	19-10-2018 06:56	Caroline won the gold medal in the youth world Olympics! We are soo proud of her, never doubted her once and her ha... https://t.co/tdD3ptNWd2	41	None	42895	135
11	AFN	25125905	19-10-2018 15:45	For #WomensHistoryMonth, AFN acknowledges Brigette Lacquette, the first #FirstNations woman to play for Canada's Na... https://t.co/QUCkT4J1qF	33	Womensh	44152	819
12	Martina Navr	506429053	20-10-2018 16:56	Has the N.C.A.A.'s 'Student-Athlete' Fiction Reached Its Logical End? Let's hope so, as used to be the case with O... https://t.co/Edtat2abBZ	22	None	244754	456
13	Ato Boldon	33318560	19-10-2018 16:32	The man who changed the high jump forever, @DickFosbury1, celebrates the 50th anniversary of his Mexico 1968 triump... https://t.co/44oDtXGDGC	33	None	136820	1779
14	NTV Kenya	25985333	19-10-2018 06:50	VIDEO: Former Sports CS Hassan Wario arraigned in court over allegations of misappropriation of Sh55 million during... https://t.co/aqvLJpzvRP	20	None	1941501	282
15	MackenzieBc	604424828	19-10-2018 07:14	Every ski jumper that competed for Canada at the Olympics in the last 20 years is a Calgarian. The cost to bring th... https://t.co/YKOcQOGysA	15	None	5356	1408
16	#TokyoOlymp	14955353	10-10-2018 13:00	54 years ago today, the 1964 Summer Olympics opened in the heart of Japan.	1090	None	944901	3003
17	#TokyoOlymp	14955353	10-10-2018 13:02	Bold. Fluid. Dynamic. Our logo for the 2020 Summer Olympics in Tokyo embodies everything it takes to be an Olympic... https://t.co/P2BI9w7Ek6	490	None	944901	3003
18	Olympics	52422878	12-10-2018 14:05	The official poster of 1968 Olympic Games in Mexico. #Mexico1968 #OnThisDay #OpeningCeremony @50AnivMexico68 https://t.co/7HsbGXIdvi	168	Mexico19	6215174	4846
19	Steven nabil	2386350236	11-10-2018 20:57	Jarah Nasar, works as a street vegetables seller, practices in his dirt backyard of his simple home in Nasriya ,Sou... https://t.co/HtzBvxMcxl	70	None	166943	2391
20	#Tokyo2020	2249745696	10-10-2018 10:30	Today marks 54 years since the opening of the Olympic Games Tokyo 1964. The first ever Olympics held in Asia, with... https://t.co/uIkZQonGRm	96	None	134515	400
21	Lukwesa Bur	759433657	10-10-2018 16:01	This needs to be celebrated! For time Zambia represented in the Showjumping at the Youth Olympics in Buenos Aires,... https://t.co/F8rodgszeq	92	None	6021	1139
22	Olympics	52422878	10-10-2018 19:59	54 years ago today the Olympic rings appeared in the sky #OpeningCeremony #Tokyo1964. The #OlympicLegacy is still... https://t.co/HKDbtCgjtG	80	OpeningC	6215174	4846
23	Olympics	52422878	10-10-2018 15:39	Congratulations to Jeremy Lalrinnunga (IND) for the first-ever gold medal for India at the Youth Olympic Games... https://t.co/IUx4iWHI8Q	130	None	6215174	4846
24	Good Mornin	22650211	11-10-2018 12:41	.@lzhernandez02 tells @michaelstrahan she is training for the 2020 Olympics! Her new children's book "She's Got Thi... https://t.co/M91nvYSQOZ	125	None	3642764	2414
25	Olympics	52422878	11-10-2018 16:44	Congratulations to Manu Bhaker (IND) for the gold medal at the Youth Olympic Games #YouthOlympics #GameChangers... https://t.co/0Q5v0KWCCq	77	YouthOlyr	6215174	4846

Fig. 4. Data extracted from twitter

Now, we don't need the punctuations, digits, url links, people tagged and whose tweet is retweeted. So, we are going to remove all that information from the text we have.

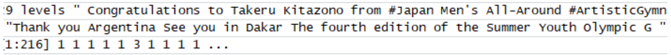


```
text = twitter_data$text
text_clean = gsub("(RT|via)((?:\\b\\w*@\\w+)+)", "", text)
text_clean = gsub("@\\w+", "", text_clean)
text_clean = gsub("[[:punct:]]", "", text_clean)
text_clean = gsub("[[:digit:]]", "", text_clean)
text_clean = gsub("http\\w+", "", text_clean)
```

Fig. 5. Text cleaning code

Here, we have used regular expression to match the pattern we want to find in our data. Input Tweet:" Thank you, @BuenosAires2018 Argentina! See you in Dakar 2022! The fourth edition of the Summer Youth Olympic G... https://t.co/vqR6chTc99"

Snapshot of the data after data Cleaning:



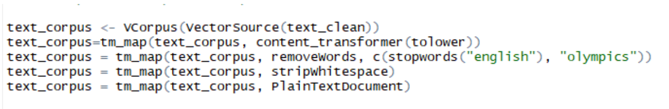
```
9 levels " Congratulations to Takeru Kitazono from #Japan Men's All-Around #ArtisticGymn
"Thank you Argentina See you in Dakar The fourth edition of the Summer Youth Olympic G "
1:216] 1 1 1 1 1 3 1 1 1 1 ...
```

Fig. 6. Cleaned text

We noticed that '@BuenosAires2018', '!', '2022!' and '... https://t.co/vqR6chTc99' is been removed.

B. *Transforming Data*

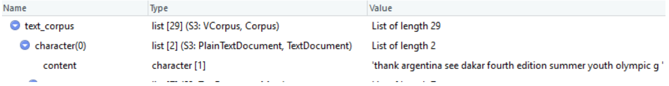
Now, after cleaning of data we want the data free from stopwords, whitespaces, converted to lowercase and formatted to the plain text document extension.



```
text_corpus <- VCorpus(VectorSource(text_clean))
text_corpus=tm_map(text_corpus, content_transformer(tolower))
text_corpus = tm_map(text_corpus, removeWords, c(stopwords("english"), "olympics"))
text_corpus = tm_map(text_corpus, stripWhitespace)
text_corpus = tm_map(text_corpus, PlainTextDocument)
```

Fig. 7. Text Corpus

Output Text:

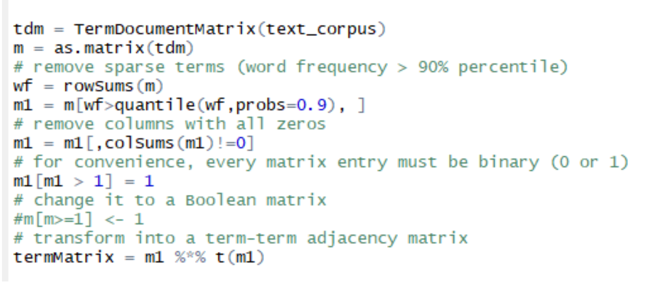


```
Name              Type                                            Value
  text_corpus     list [29] (S3: VCorpus, Corpus)                 List of length 29
    character(0)  list [2] (S3: PlainTextDocument, TextDocument)  List of length 2
      content     character [1]                                   'thank argentina see dakar fourth edition summer youth olympic g '
```

Fig. 8. Text Corpus

Here, we can notice that stopwords like you,in,the and of has been removed, text is converted to lower case and it is in plain text format. We didn't had whitespaces in text so obviously nothing regarding that.

Now, we want to know that what are important words in out text that can be used to extract some meaningful insight. Here, we are going to find out how many times the word has been repeated so that we get the information of how much importance words have in the text.

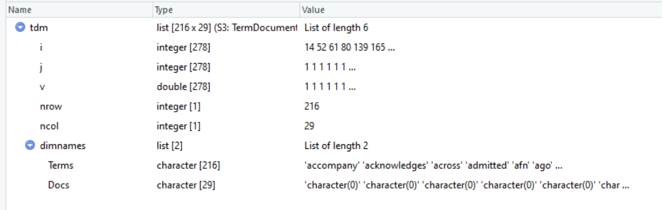


```
tdm = TermDocumentMatrix(text_corpus)
m = as.matrix(tdm)
# remove sparse terms (word frequency > 90% percentile)
wf = rowSums(m)
m1 = m[wf>quantile(wf,probs=0.9), ]
# remove columns with all zeros
m1 = m1[,colSums(m1)!=0]
# for convenience, every matrix entry must be binary (0 or 1)
m1[m1 > 1] = 1
# change it to a Boolean matrix
#m[m>=1] <- 1
# transform into a term-term adjacency matrix
termMatrix = m1 %*% t(m1)
```

Fig. 9. Output of Text Corpus

Output:

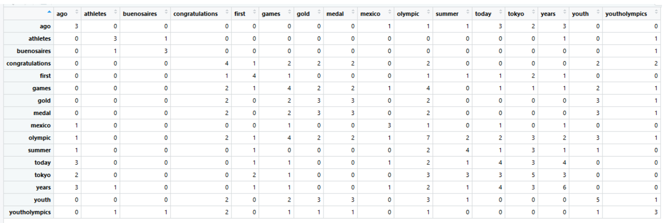
Meta data of term document matrix:



Name	Type	Value
tdm	list [216 x 29] (S3: TermDocument	List of length 6
i	integer [278]	14 52 61 80 139 165 ...
j	integer [278]	1 1 1 1 1 1 ...
v	double [278]	1 1 1 1 1 1 ...
nrow	integer [1]	216
ncol	integer [1]	29
dimnames	list [2]	List of length 2
Terms	character [216]	'accompany' 'acknowledges' 'across' 'admitted' 'afn' 'ago' ...
Docs	character [29]	'character(0)' 'character(0)' 'character(0)' 'character(0)' 'character(0)' 'char ...

Fig. 10. Output of Text Corpus

Term document matrix:



	ago	athletes	buenosaires	congratulations	first	games	gold	medal	mexico	olympic	summer	today	tokyo	years	youth	youtholympics
ago	3	0	0	0	0	0	0	0	1	1	1	3	2	3	0	0
athletes	0	3	1	0	0	0	0	0	0	0	0	0	0	1	0	1
buenosaires	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	1
congratulations	0	0	0	4	1	2	2	2	0	2	0	0	0	0	2	2
first	0	0	0	1	4	1	0	0	0	1	1	1	2	1	0	0
games	0	0	0	2	1	4	2	2	1	4	0	1	1	1	2	1
gold	0	0	0	2	0	2	3	3	0	2	0	0	0	0	3	1
medal	0	0	0	2	0	2	3	3	0	2	0	0	0	0	3	1
mexico	1	0	0	0	0	1	0	0	3	1	0	1	0	1	0	0
olympic	1	0	0	2	1	4	2	2	1	7	2	2	3	2	3	1
summer	1	0	0	0	1	0	0	0	0	2	4	1	3	1	1	0
today	3	0	0	0	1	1	0	0	1	2	1	4	3	4	0	0
tokyo	2	0	0	0	2	1	0	0	0	3	3	3	5	3	0	0
years	3	1	0	0	1	1	0	0	1	2	1	4	3	6	0	0
youth	0	0	0	2	0	2	3	3	0	3	1	0	0	0	5	1
youtholympics	0	1	1	2	0	1	1	1	0	1	0	0	0	0	1	3

Fig. 11. Output of Text Corpus

Above figure shows the term-document matix which defines the number of times particular term is repeated in the document. Columns resemble over terms and rows resembles over documents in the matrix. Now, we have our data ready to feed for machine learning.

Discussion and Future Events

Walking in footsteps of the emerging use of chatbots, it is very important to make the chatbot efficient and productive in terms of answering. This paper proposes to provide the data cleaning and transforming technique so that chatbot can reply more efficiently as the data fed for machine learning doesn't have the unwanted text.

Data reduction of the text file can be further explored which helps in reducing the memory size and the increase the performance. It also maintains the integrity of the data as there is no redundancy.

References

[1] 1Antonius Angga P, 2Edwin Fachri W, 3Elevanita A, 4Suryadi, 5Dewi Agushinta R Gunadarma University Depok, Indonesia 1antoniusangga03@gmail.com,, 2fachrifachrifachri@gmail.com, 3anggarielevanita@gmail.com,4suryadi.guna@gmail.com, 5dewiar@staff.gunadarma.ac.id.

[2] Melody Moh*, Abhiteja Gajjala, Siva Charan Reddy Gangireddy, and Teng-Sheng Moh, Department of Computer Science, San Jose State University,San Jose, CA, USA, *Contact author. Email: melody.moh@sjsu.edu

[3] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73. Ly Pichponreay, Department of MIS, Chungbuk National University, Cheongju, Republic of Korea,ponreay_ly@hotmail.com

[4] AM Rahman1, Abdullah Al Mamun1, Alma Islam2, IEEE1, International Islamic University Chittagong2, amrahman@ieee.org1, almamun@ieee.org1, alma.iiuc@gmail.com2

[5] Godson Michael D'silva1, *, Sanket Thakare2, Sharddha More1, and Jeril Kuriakose1 1Information Technology Department, St. John College of Engineering and Technology, Palghar, India, dsilvagodson@gmail.com