Using Association Rules Mining to Analyze Human Rights Violations in Indonesia

Hendro Margono, Xun Yi, and Gitesh K. Raikundalia

***Abstract*—** Human rights are a set of basic rights inherent in humanity. Understanding of human rights is an important part of individual status as human beings who possess dignity and values of mutual respect for each other. Moreover, comprehending human rights violations also significantly enriches our knowledge regarding diversity of violation actions occurring in everyday life including abuse and ignorance of basic human rights. This paper discusses how to detect violation patterns by using association rules mining. These techniques provide powerful tools to identify patterns which occur in a database. Finding human rights violation patterns is one of the challenges in this work. The paper provides an overview of our human rights violations database and describes how data preparation is provided. Moreover, it discusses how data mining could provide solutions for finding frequent patterns human rights violations in Indonesia, and how it could uncover new knowledge about types of violations.

***Keywords*—**Data mining, human rights, human rights violation.

I. Introduction

HUMAN rights are a prominent part of how people interact with others at all levels in society including family, community, educational institutions, workplace, politics and international relations. Human rights violations are typically complex, with some perpetrators engaging in more than one type of violation. Therefore, the violations that occur in life have many causes. Human rights violations may be caused by socio-economic conditions and unstable political situations in a state, which occur in some countries [1, 2]. Geography, poverty, culture, and background also have big roles on committing acts of violence [3]. Moreover, sometimes people who commit one type of violence may well commit other types of violence [4].

Human rights violations are continually occurring across Indonesia and this has increased in the past few years. In 2008, the Indonesia National Human Rights Commission (KomnasHAM) received around 4,000 complaints regarding human rights violations; in 2009 there were about 5,000 complaints [5], and in 2010 KomnasHAM received around 230 cases every month. From January to December of 2011, there were 6,289 cases [6]. This shows that human rights violations are increasing every year, especially at the grassroots level. Elsam [7] reported that "during 2011, as in the previous years, there has not been any significant progress with regard to the settlement of past human rights abuses. Stagnation occurred in two levels, namely the settlement policy and the implementation of past human rights violations cases. The problems underlying the stagnation are still the same with the previous years; it was due to the debates on technicalities, evidence, and the absence of the House's recommendation". On the other hand, serious human rights concerns remain such as a surge in religious violence, particularly against Ahmadiyah, a group that considers itself Muslim but that some Muslims consider heretical. Violence continued at Papua and West Papua provinces, with less investigation by the police to hold perpetrators [7].

There is a major question of why do people commit violations, and how do social and economic conditions, education background, age, ethnicity, religion, or region encourage people to commit violations? To answer this question, these elements should be analyzed more deeply to know the link between them by mining data in a human rights violations database created from reported facts. This study will give contribution to knowledge regarding human rights in Indonesia.

It is an interesting problem to determine the link between violation cases and to what extent people's background, such as social and economic conditions, education background, age, ethnicity, religion, or region, will influence people to commit violations. For analyzing this problem, this research uses data mining techniques, especially association rule mining as a tool to identify data in a human rights violations relational database.

The database has many attributes regarding a person which include name, background, age, types of violation, roles, events, and geographical area. All attributes will have a relationship with each other. Attribute age could have a link to a type of violation, geographical area, name, date or background. Some attributes may have a strong influence on other attributes. In addition, the relational attributes in a human rights database will illustrate that every event has many causes for effects on other events or people.

A tool for analyzing the relationships between attributes in the human rights violations database is data mining. Data mining techniques will be used to determine the relationships. Data mining techniques are powerful tools to

Hendro Margono, Xun Yi, Gitesh K. Raikundalia are at School of Engineering and Science, Faculty of Health, Engineering and Science. Victoria University, PO Box 14428. Melbourne, Victoria 8001. Australia. Phone: +61 3 9919 4702; e-mail: hendro.margono@live.vu.edu.au; Xun.Yi@vu.edu.au; Gitesh.Raikundalia@vu.edu.au).

discover interesting patterns and knowledge in large amounts of data [8].

This paper contributes in identifiying relationships and influences between attributes in the human rights violations database by using data mining techniques especially *association rules mining*. The relationship between attributes will illustrate that every violation has some cause as a reason why the actors or perpetrators committed violations. In association rules mining, there are two measurements to determine how strong the relationship is between attributes. The measures of a strong relationship are *rule support* and *rule confidence*, whereby the rule should comply with both the *minimum support threshold* and the *minimum confidence threshold* [8].

All itemsets which occur in the data set should satisfy the minimum support threshold and minimum confidence threshold as prerequisites in association rules mining. After finding the percentage of minimum support threshold, *Apriori Algorithms* should be used to generate all items set in the dataset.

This paper is organized as follows. Section 2 describes the related work of this paper. In section 3, we describe two association rule techniques that will be used to analyze research problems: Apriori Algorithm and FP-Growth algorithm techniques. In section 4, implementation of Apriori algorithm and FP-Growth to solve research problems will be detailed. Analysing and predicting the relationship amongst attributes will discover new human rights violations patterns. The last section is the conclusion.

II. Related Work

DeRosa [9] reported that data mining techniques and data analysis have been used by the United States to identify accurate information about terrorism. Critical information for determining whether a person is of interest or not to terrorism-related investigation is needed by the United States government. There are two techniques that the government has used. Firstly, "subject-based link analysis uses public records or other large collections of data to find links between a subject, a suspect, an address, or other piece of relevant information and other people, places, or things. Secondly, patterns-based analysis may also have potential counterterrorism uses. Patterns based queries take a predictive model or patterns of behavior and search for that patterns in data sets" [9]. These techniques have been successful and have been of significant benefit to counteract terrorism in the United States when the tragedy on 11 September 2001 occurred.

In similar work but with different techniques, data mining was also applied to national security, privacy and civil liberty as discussed by Thuraisingham [10]. Data mining techniques have been used to handle national security and information security problems such as intrusion detection.

III. Data Preparation for Human Rights Violation Database

Preparing data for the database involves exploring data from some sources such as from the Internet, human rights violations report, evidence of violations, news, human rights annual report or fact finding of human rights violations. Hence, the data that can be taken from some human rights violation reports, news, human rights violations evidences, or data warehouse such as annual report from human rights commission and verified evidence report from some Non-Government Organizations.

The amount of data recorded in a human rights violation database as mentioned above are about 30 cases which will be added continuously.

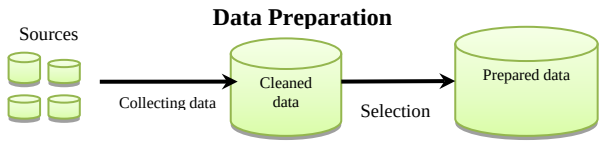

Fig 1: Data preparation

The human rights violations database consists of entities which have various attributes and relations. The entities in this database represent an existing unit or objects which are independent. Two entities in the database are "Event" and "People", and detail about them include:

a) An Event is something that occurs, with a beginning and an end, and evolves until its logical conclusion. The event possibly a single act, a series of related acts, or a combination of related acts happening together.
b) People are an individual or a group who have a part in or in relation to an event.
c) An attribute is a feature that an entity holds in spite of any context, such as the name, age, sex, address or physical appearance of a person. [4].

There are many tables in the human rights violations database and the main tables are people and events. The example table people can be seen in Table I.

TABLE I
VICTIMS OF HUMAN RIGHTS VIOLATIONS

victimsid	Background						
	age	Education	Ethnicity	Religion	Region / Province	Job	Marital Status
01p	28	SHS	Chinese	Buddhist	DKI	U	S
02p	25	SHS	Sunda	Islam	West Java	P	M
03p	33	B	Java	Islam	Centre Java	J	M
04p	32	SHS	Ambon	Christianity	Maluku	CO	S
05p	45	JHS	Papua	Christianity	West Papua	CO	M
06p	24	B	Bali	Islam	DKI	W	S

Education column: SHS = Senior High Schools, JHS = Junior High School, B= Bachelor degree.
Job column: U=unemployment, P=private, J=journalist, CO= chief of Organization, W=writers.
Status column: S=single, M=married.
Sex column: M-male, F=female

Table II describes variables classification for victims of human rights violation using the letters *a* to *l*:

TABLE II
VARIABLES CLASSIFICATION FOR VICTIMS OF HUMAN RIGHTS VIOLATIONS

Age is classified by age groups:
- a1 = Children of ages 0 -12 Years
- a2 = Teenager of ages 13-18 Years
- a3 = Youth of ages 19-25 Years
- a4 = Young people of ages 26-35 Years
- a5 = The Middle of ages 36 – 50 Years
- a6 = The elderly of ages 51 – 100 Years

Education is classified as:
- b1 = Primary School (PS)
- b2 = Junior High School (JHS)
- b3 = Senior High School (SHS)
- b4 = Diploma degree (D)
- b5 = Bachelor degree (B)
- b6 = Master degree (M)
- b7 = PhD degree (P)

Ethnicity is classified as:
- c1 = Jawa
- c2 = Sunda
- c3 = Betawi,…etc.

Religion is classified as:
- d1 = Islam
- d2 = Christianity
- d3 = Hinduism
- d4 = Buddhism
- d5 = No religion

Region / province is classified as:
- e1 = DKI Jakarta
- e2 = West Java
- e3 = Central Java
- e4 = East Java
- e5 = Bali…, e33 = West Papua

Job is classified as:
- f1 = Government officer
- f2 = Private sector employee
- f3 = Journalist, …etc.

Marital status is classified as:
- g1 = Single
- g2 = Married
- g3 = Widower

Gender is classified as:
- h1 = Male
- h2 = female

Role in the violation is classified as:
- i1 = Victim
- i2 = Perpetrator

Table III describes human rights violation events.

TABLE III
HUMAN RIGHTS VIOLATIONS EVENTS

event sid	Event	Place of event: Road / Suburbs	Place of event: Region / Province	Type of violation	Type of act
E01	Asylum	Tangerang	Banten	Violent or coercive acts by state agents	Death in detention or police custody
E02	Religious Discrimination	Tambun Bekasi	Banten	Violent or coercive acts by non-state agents	Torture
E03	Freedom of press		West Java	Violent or coercive acts by non-state agents	Torture
E04	Abduction	West Pasaman	West Papua	Violent or coercive acts by non-state agents	Abduction
E05	Abduction	West Pasaman	West Papua	Violent or coercive acts by non-state agents	Abduction
E06	Freedom of Expression	Pasar Minggu	DKI Jakarta	Acts and instances involving exploitation of individuals or groups	Prohibition on speech

Table IV describes variables classification for human rights violation events using the letters *j* to *l*:

TABLE IV
VARIABLES CLASSIFICATION FOR HUMAN RIGHTS VIOLATIONS EVENTS

Place where the event:
- j1 = DKI Jakarta
- j2 = West Java
- j3 = Central Java
- j4 = East Java
- j5 = Bali…, j33 = West Papua

Types of violations:
- k1 = Violent or coercive acts by state agents
- k2 = Violent or coercive acts by non-state agents
- k3 = Acts and instances involving exploitation of individuals or groups,…etc.

Type of acts:
- l1 = Death in detention or police custody
- l2 = Torture
- l3 = Abduction
- l4 = Prohibition on speech,…etc

Although both entities are independent, there is a prediction that there is a data relationship between people and events. A person is usually a perpetrator or victim in one or more events. It depends on the person's role in every event. Events will occur when a person is involved in the various acts, whether directly or indirectly, that cause or lead to human rights violations, and which on their own, or in combination with related acts, constitute events [4].

Furthermore, a person's background such as their education, age, social status, religion, ethnicity, and job could lead the person to commit a violation. Using techniques in data mining, we expect to find out the relationship between entities and attributes in the human rights violations database and new human rights violations patterns.

Figure 2 illustrates relational data in the human rights violations database.

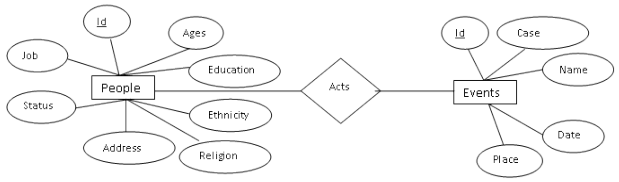

Fig 2: ER diagram of human rights violations database

Table V describes relational human rights violation database.

TABLE V
RELATIONAL HUMAN RIGHTS VIOLATION DATABASE

eventsid	victimsid
E01	01p
E02	02p
E03	03p
E04	04p
E05	05p

Analyzing of Using Data Mining Technique

This research will use one of the data mining techniques to analyze the relationship between attributes and to find out

some new violations patterns in human rights violations database. One of the techniques that can be used is association rule mining. In order to obtain strong relationship between items, minimum support is 2, so the minimum support threshold = 33.33% and the minimum confidence threshold = 67%. For calculating minimum support and minimum confidence, this work uses the formula given by Han, et al. [8].

Support $(A \Rightarrow B) = \mathrm{P}(A \cup B)$

Confidence

$$(A \Rightarrow B) = \mathrm{P}(B/A) = \frac{(support(A \cup B))}{(support(A))} = \frac{(support\_count(A \cup B))}{(support\_count(A))}$$

Since in this work we want to discover human rights violations patterns, hence the minimum confidence should be a high percentage to get the strong relationship between items.

TABLE VI
TRANSACTIONAL DATA

TID	Items
1	a4, b3, c5, d4, e1, f20, g1, h1,i1, j20, k2, l1
2	a4, b3, c2, d1, e2, f3, g2, h1, i1, j20, k2, l2
3	a4, b5, c1, d1, e5, f3, g2, h2, i1, j2, k2, l10
4	a4, b3, c30, d2, e30, f15, g1, h1, i1, j33, k2, l3
5	a5, b2, c30, d2, e33, f15, g2, h1, i1, j33, k2, l3
6	a3, b5, c5, d1, e5, f4, g1, h1, i1, j5, k3, l4.

TID represents set of eviolations events in a database of human rights violations events. Whereas, *I={ a,b,c,…,z}* is a set of items which consist of some information. *D* is a human rights violations evidence database having a set of transactions. Let *A* be a set of items of person background or age or geography. When $A \subseteq T$. An implication form $A \Rightarrow B$ can call an association rule, where $A \subseteq I$, $B \subseteq I$, and $A \cap B \neq \emptyset$.

In Table VI, *TID* of 1 represents of asylum event which contains a4, b3, c5, d4, e1, f20, g1, h1, i1, j20, k2, l1. This means the itemset that occurs in transactional data is an asylum seeker event which contains the victims age of 28 years old (a4), victims' educational background of senior high school (b4), Chinese ethnicity (c4), Buddhist religion (d4), in DKI province (e1), unemployed (f20), single marital status (g1), male (i1), place of events is Banten (j20), violent or coercive acts by state agents (k2) and, type of act about death in detention or police custody (l1).

IV. Association Rules Mining for a Human Rights Violations Database

This research uses association rule mining to address research problem. Using both Apriori Algorithm and FP-Growth in association rules, linking between attributes in the human rights violations database will occur.

Apriori Algorithm in Database

The first step of the Apriori Algorithm is finding frequent patterns. Apriori Algorithm has a prerequisite to fulfill before generating frequent itemsets. The prerequisite is all nonempty subsets must be frequent which call the Apriori Property. For example, in finding types of violation which occur in society, we can generate a table of transaction data for events by using the Apriori Algorithm as shown on Table VII:

TABLE VII
TRANSACTIONAL DATA FOR EVENTS

TID	Items
1	j20, k2, l1
2	j20, k2, l2
3	j2, k2, l10
4	j33, k2, l3
5	j33, k2, l3
6	j5, k3, l4.

Scan D for count for each candidate →

Itemsets	Support count
j20	2
j33	2
j5	1
j2	1
k2	5
k3	1
l1	1
l2	1
l3	2
l4	1
l10	1

1. Generate a table L_{k-1} by joining a set of candidate k-itemsets L_{k-1} with itself. Then the set of candidates is given as C_k Let L_1 and L_2 be itemsets in L_{k-1}. Suppose minimum support required is 2, so the corresponding relative support is 2/6 = 33.33%. Then the result for L_{k-1} is shown in Figure 3.

L_{k-1}

Itemsets	Support count
j20	2
j33	2
k2	5
I3	2

→ L_{k-2}

Itemsets	Support count
j20, k2	2
j33, k2	2
j33,l3	2
k2, l3	2

C_3

Itemsets	Support count
j33, k2, l3	2

→ L_{k-3}

Itemsets	Support count
j33, k2, l3	2

Fig 3: Generation of the candidate itemsets and frequent itemsets, where the minimum support count is 2

2. Then, join $L_{k-1} \bowtie L_{k-1}$ to get candidate itemsets L_{k-2}, where members of L_{k-1} are joinable if their first items are in common. Members l1 and l2 of L_{k-1} are joined if (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧…∧ (l1[k - 2] = l2[k - 2] ∧ l1[k -1] < l2[k -1]). After that, join $L_{k-2} \bowtie L_{k-2}$ to get candidate itemsets C_3 : {{j20, k2}, {j33,k2}, {j33,l3}, {k2,l3} ⋈ {j20,k2}, {j33,k2}, {j33,l3}, {k2, l3}} = {j20, j33, k2}, {j20,j33, l3}, {j20,k2,l3}, {j33,k2,l3}.

3. Prune using the Apriori Property. All nonempty subsets of a frequent itemset must also be frequent. Details explanation regarding the result of joining L_{k-2} itself is as follows:
   - The 2-items subset of {j20, j33, k2} are {j33, k2}, {j20, k2} and {j20, j33}, which are not a member of L_2. Consequently, itemsets {j20, k2, l3} are removed from C_3
   - The 2-items subset of {j20, j33, l3} are {j33, l3} and {j20, l3}, {j20, j33}, which are not a member of L_2. Consequently, itemsets { j20, j33, l3} are removed from C_3
   - The 2-items subset of {j20, k2, l3} are {j20, k2}, {k2, l3} and {j20, l3}, which are not a member of L_2. Consequently, itemsets {j20, k2, l3} are removed from C_3.
   - The 2-items subset of {j33, k2, l3} are {j33, k2}, {j33, l3} and {k2, l3}, which are a member of L_2. Consequently, itemsets {j33, k2, l3} are kept in the Table C_3

The result of finding frequent patterns by using the Apriori algorithm is {j33, k2, l3}, which means the most frequent human rights violation occurring is abduction by non-state agents in West Papua. This technique also can be used to analyze who is usually a victim or perpetrator in violations. The frequent itemsets from transactions in a database D have been found. Next, to find strong association rules from them, the rules should satisfy both minimum support and minimum confidence. The data contains frequent itemsets X = {j33, k2, l3} and the nonempty subsets of X are {j33, k2}, {j33, l3}, {k2, l3}, {j33}, {k2} and {l3}. The resulting association rules are as shown below, each listed with its confidence:

{j33, k2} ⇒ l3, confidence 2/2 = 100%

{j33, l3} ⇒ k2, confidence 2/2 = 100%

{k2, l3} ⇒ j33, confidence 2/2 = 100%

The minimum confidence threshold is 67%, then the result generated above is 'the rules', then all rules are considered strong. Moreover, if we are examining the result, the violation pattern usually occuring is abduction followed by violent or coercive acts by non-state agents in West Papua.

V. Conclusion

This paper covers our work in the important area of human rights violations patterns by applying data mining techniques. The paper describes the application of data mining techniques to a human rights violations database to find the relationships among attributes. We discussed the human rights violations database which has many relationships among attributes. The tool used to analyze the relationships is association rules mining. To our knowledge, no work have been reported on using data mining techniques for mining data in human rights violations databases.

We have seen that in association rules performs well mining finding violation patterns occurring in society. The techniques of Apriori Algorithm is applicable in finding frequent patterns which arise in human rights violations database. This work aims to discover trends of human rights violations which occur in Indonesian by mining data in a human rights violations database. The outcomes of this work is that we found:

1. That the most frequent human rights violations patterns in Indonesia is abduction followed by violent or coercive acts by non-state agents in West Papua.
2. The trends of people who commit human rights violations are found which has relations with type of human rights violations.

Non-Government Organizations, society and government will get benefits regarding kinds of violation information occured within Indonesian society. For the future, these techniques can be used by Non-Government Organizations or the Indonesian government to retrieve information about human rights violations to protect their people for the purpose of law enforcement.

References

[1] N. J. Mitchell and J. M. McCormick, "Economic and political explanations of human rights violations," *World Politics*, vol. 40, pp. 476-498, 1988.

[2] J. C. Kanmony, *Human Rights Violation*: Mittal Publications, 2010.

[3] O. N. T. Thoms and J. Ron, "Do human rights violations cause internal conflict?," *Human Rights Quarterly*, vol. 29, p. 674, 2007.

[4] J. Dueck, M. Guzman, and B. Verstappen, *HURIDOCS events standard formats: a tool for documenting human rights violations*: Huridocs, 2001.

[5] KomnasHAM, "Annual report of human rights 2010," KomnasHAM, Jakarta2010.

[6] KomnasHAM, "Annual report of human rights 2011," KomnasHAM, Jakarta 2011.

[7] Elsam, "Report on the Human Rights Situation of 2011: Towards the Lowest Point of Human Rights Protection," Jakarta 2012.

[8] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 3rd ed. Amsterdam: Elsevier, 2012.

[9] M. DeRosa, *Data mining and data analysis for counterterrorism*: CSIS Press, 2004.

[10] B. Thuraisingham, "Data mining, national security, privacy and civil liberties," *ACM SIGKDD Explorations Newsletter*, vol. 4, pp. 1-5, 2002.

Semantic Web-Based E-Counseling System

Toshika Khandelwal, Garima Joshi, Anish Singhania, Animesh Dutta

Abstract**—** Presently on Internet, it is difficult to find a resource which serves as one-stop destination for students regarding their career-related queries. Students who use Internet to search for help end up with large amount of information, mostly irrelevant. This happens because of current syntax-based web search which just matches words of the query and returns all string-matched results. Thus, this research work proposes a Semantic web-based e-counseling system. It captures the query from the student in Natural Language, understands the logic of the query and returns a reply just as a human counselor would, which helps the student to find a proper career option with less efforts and time. The Natural Language Query is parsed by a NLP Parser to form a data structure which is compared to the Ontology, a hierarchical set of concepts and relations of a domain, and only relevant options are presented to the student.

.

Keywords**—**e-counseling, semantic web, ontology, natural language processing

I. Introduction

A. Counseling System

A counseling system helps a student to select a career depending on the student's choice of subject and field of interest. A counselor assists the student personally and guides the student to the path that it should take.

In a general e-counseling system, the virtual guide of the student follows the syntactic web approach. The approach results in the student having a lot of unwanted and unorganized data. The final outcome is that the student is left to sort a substantial portion of the data by itself, which is not desired. So, the data needs to be organized so that such discrepancies can be avoided.

B. Semantic Web

With all the data that is available in the world today, there is an increasing need for the data to be organized so that it becomes machine processable. Here comes the concept of semantic web, which organizes the data so that it is machine processable and not just human interpretable. Semantic web is the idea of having data on the Web defined and linked in a way such that it is machine processable and not just available for display. For this to become true, underlying support systems and technologies should be designed to make the Web more meaningful and human-friendly.

C. Ontology

Ontology is a formal explicit specification of shared conceptualization. Ontologies form the backbone of the Semantic web; it facilitates machine understanding of the linked information resources. The various annotations on the semantic web form links between resources on the Web and connects them to formal terminologies and these connective structures are called Ontology. Formally, Ontology consists of concepts (classes) and relationships between these concepts (relations) in a hierarchical (or taxonomical) format; thus forming a vocabulary for the domain it is defined in, along with the computerized meaning of the terms in the vocabulary.

II. Related Works

A lot of work has been done earlier and is still in pursue in the field of Semantic Web technologies and Ontology. Work has been done on many layers of the Semantic Web, including content generation, web services and e-connections. [5-6] A Web-based geospatial application using Ontology and Semantic Web Services is designed to provide demonstration of using recent advancement in GIScience in other fields to achieve Semantic based GIS applications over cyber infrastructure which can further contribute to "development of community-guided cyber infrastructure".[10] Semantic Web technologies has been used for an improved exploration and rating of hotels for business customers in order to reduce the search time and costs, which, in turn, results in a huge benefit for the end-users.[9] Ontology has immense applications in the field of medical science and many new systems are being developed which uses semantic technologies as their basic architecture. An ontology based Holonic Diagnostic System(OHDS) has been developed which combines the advantages of holonic paradigm with multi agent system technology, for the research and control of unknown diseases.[8] There are many other applications of Ontology including e-learning systems which provides virtual classrooms, remote courses and distance learning.[7]

In our previous work, the idea of Ontology driven model has been proposed to make e-counseling a better experience for the students and to help them out in selecting an appropriate career option.[11] We expand the previous related

Toshika Khandelwal is with the National Institute of Technology, Durgapur, India, phone: (+91)7679986039; fax: (+91)3432547375; e-mail: toshika28@gmail.com.

Garima Joshi is with the National Institute of Technology, Durgapur, India;e-mail: gjoshi0311@gmail.com.

Anish Singhania is with the National Institute of Technology, Durgapur, India;e-mail: anishsinghania92@gmail.com.

Animesh Dutta is with the National Institute of Technology, Durgapur, India;e-mail: animeshrec@gmail.com