ISSN : 0025-0422

JOURNAL OF THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

सत्यं शिवं सुन्दरम्

Estd. 1949

Accredited Grade 'A' by NAAC

VOL. 55 NO. 1 (Science & Technology)
Vadodara
2021

JOURNAL
OF
THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

JOURNAL
OF
THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

EDITOR
Prof. C. N. MURTHY

THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA
सत्यं शिवं सुन्दरम्
Estd. 1949

Accredited 'A' Grade by NAAC

VOL 55 NO. 1
(SCIENCE & TECHNOLOGY)
VADODARA
2021

ISSN : 0025-0422

Printed by **Jatin H. Somani,** Manager, The Maharaja Sayajirao University of Baroda Press (Sadhana Press), Near Palace Gate, Palace Road, Vadodara and Edited by **Prof. C. N. Murthy,** Editor (Science & Technology) at The Maharaja Sayajirao University of Baroda, Vadodara - 390 002, (India), June 2022.

CONTENTS

AMLA : REVIEW OF A WONDER TREE WITH AMAZING HEALTH BENEFITS

Pallavi Dixit*
Department of Botany,
Mahila Vidyalaya Degree College, Lucknow
*corresponding author email: drpallavidixit80@gmail.com

Abstract

The Indian gooseberry or Amla is a greatest boon to humanity. It is a effective traditional medicine which had been used to treat and to manage diseases since the ancient time. Emblica officinalis is a wonder herb and used as rejuvernator in ayurveda. The eminence of it is so well recognized in Ayurveda that all the famous ancient text have discussed the usefulness and extolled its extraordinary medicinal qualities. It is known for its exceptional medicinal and nutritional properties. It is one of the most extensively studied plant. All parts of Emblica officinalis are useful in the treatment of various diseases.

Keywords: Introduction, Emblica officinalis Plant, Health Benefits

1. Introduction

Amla is botanically known as Emblica officinalis belongs to Phyllanthaceae family. It is native of India and is also grown in different tropical and sub-tropical regions such as China, South East Asia, Malaysia, Pakistan, Uzbekistan and Sri Lanka. It is a effective traditional medicine which had been used to treat and to manage diseases since the ancient time. Emblica officinalis is a wonder herb and used as rejuvenator in ayurveda. The eminence of amla is so well recognized in ayurveda that all the famous ancient text have discussed the usefulness and extolled its extraordinary medicinal qualities. It is known for its exceptional medicinal and nutritional properties. It is full of vitamin C, amino acid and minerals. All parts of Emblica officinalis are useful in the treatment of various diseases.

2. Emblica officinalis Plant

Emblica officinalis is a medium size deciduous plant with height ranging from1-8 metre. The fruit is slightly curved and the branches are scattered around. The bark is grey with hard wood reddish in colour. The Flower and fruits are greenish yellow. A ripened amla fruit is hard weighs appoximately between 60-70 grams. [1]

3. Taxonomic Classification

Kingdom – Plantae

Division – Angiospesmae

Class – Dicotyledonae

Order – Geraniales

Family – Euphorbiaceae

Genus – Emblica

Species – officinalis Gaertn.

4. Health Benefits of Emblica officinalis

All the parts of Emblica officinalis plant are useful in the treatment of various diseases. Fresh fruit of amla is refrigerant,diuertic and laxative. Green fruit is corminative and stomachic. Dried fruit is sour and astringent, flowers are cooling and aperients. The Bark is astringent. [2] Fruit of amla tree are used is various ayurvedic preparations and praised as Dhatri (God of Health) in ayurveda. Some health benefits of Emblica officinalis are as follows-

1. Hair growth - A fixed oil is obtained from the berries that are used to strengthen and promote the growth of hairs. The dried fruits have a good effect on hair hygiene and have long been respected as an ingredient of shampoo and hair oil. [3]

2. Anti-Aging - Emblica officinalis has revitalizing effect as it contains elements which is very valuable in preventing aging and to maintain and protect the body against infections. [4]

3. Skin care - High content of vitamin C boots the collagen cell production in the skin, giving soft supple and youthful skin. [5] The effectiveness of a standardized antioxidant fraction of Phyllathus emblica fruit as a skin lightener and also as an antioxidant was proven. [6]

4. Anti–Bacterial, Anti-Fungal, Anti-viral - Medical studies conducted on Emblica officinalis fruit suggest that it has anti-viral properties and also functions as anti-bacterial and anti-fungal agent. [7]

5. Natural EyeTonic - Fresh Amla Juice or dried amla capsules are good supplement to improve near -sightedness, cataract and glaucoma. It reduce intra ocular tension and corrects the vision. [8]

6. Heart Disease - One teaspoonful of powdered dry amla powder with sugar candy is mixed with a glass of water taken empty stomach may neutralize the blood cholesterol which causes heart disease. [9] Vitamin C present in amla which enlarge the blood vessels and reduce pressure. [10, 11]

7. Anti-cancer Effect - The potential anticancer effects of aqueous fruit extract P. emblica was tested in several different human cancer cell lines such as A 549 (Lungs) HepG2 (Liver) HeLA (cervical), MDA-MB-231 (Brest), SKOV3 (ovarian) and SW620 (Colorectal). Its extract significantly inhibited the growth of several human cancer cells line. [12]

8. Diarrhea - A drink made from Emblica officinalis leaves mixed with lemon juice and misris considered highly beneficial in controlling acute ancillary dysentry. One tablespoonful of Leaves paste mixed with honey or buttermilk is an effective medicare in the treatment of diarrhea and dysentery. [13, 21]

9. Fever - Malays use a decoction of Emblica officinalis leaves to treat fever. The fresh fruit is refrigerant. [14] The seed are given internally as a cooling remedy in bilious affections and nausea and in infusion make a good drink in fevers. [15]

10. Scurvy - Anti-ascorbutic virtues have been attributed to the fruits which are known as the Emblica myrobalans. [16]

11. Digestion - Emblica officinalis is very high in fiber like most fruits, fiber adds bulk to stool and help to move food through the bowels and keep their movement regular. This reduce the chance of constipation. Fiber can also bulk up loose stool and reduce diarrhea. It also stimulates the secretion of gastric and digestive juices so the food is digested efficiently. [7]

12. Antioxidant - As extremely rich source of vitamin C and low molecular weight hydrolysable tannis, which makes Emblica officinalis a good antioxidant. The tannis of amla like emblicanin-A (37%), emblicanin-B (33%), puingluconin and pedunculagin are reported to provide protection against Oxygen radicals included haemolysis of rat peripheral blood erythrocytes. [17] The powerful antioxidant ellagic acid present in amla can inhibit mutation in genes and repairs the chromosomal abnormalities. [18]

13. Diabetes - The fruits are used in the treatment of diabetes [19] and in other reference an infusion of the seeds are used in the treatment of diabetes. [14] Decoction of the leaves are used in the treatment of diabetes mellitus. [7]

14. Spermatotoxicity - Sperm count and viability were increased in mice and in human sperms with ripe Emblica officinalis fruit extract. [20]

15. Respiratory Problems - The juice or extract of the fruit is mixed with honey and pipit added is given to stop hiccough and also in painful respiration. The paste made by 10 gram leaves of Phyllanthus emblica, 5 fruits of Terminalia chebula, 9 seed of Piper nigrum, one garlic are crushed over and mixed with 25 ml. ghee made from cow's milk and clove is very effective is respiratory disorders. [22]

16. Naphroprotective Activity - The study about Emblica officinalis also describes its efficacy against kidney-infection within the body of rats which promote with aging process. [23]

17. Headache - In Indonesia the pulp of the fruit is smeared on the head to dispel headache and dizziness caused by excessive heat.7 Expressed juice of Emblica officinalis along with the other ingredients is to cure fits and insanity. [24]

18. Natural cure for Anaemia - Emblica officinalis is rich in vitamin C or ascorbic acid,an essential ingredient that helps in the absorption of iron supplement. Regular intake of amla removes the risk of anaemia (iron deficiency).

5. Conclusion

Our present generation is very health conscious, to maintain their body fitness they depends upon supplementary vitamins, mineral pills which are sometimes actually fatal for their life. Instead of they can choose the products which are not only natural but also can cut cost to buy those bitter pills to great extent. Such natural herb are freely available in nature like Emblica officinalis, which is a powerhouse of nutrients. Emblica officinalis fruit is full of vitamin C and widely used in the treatment of various ailments. It is considered to be a safe herbal medicine without any adverse effect. So it can considered that Emblica officinalis a traditionally and clinical proven fruit for both its application and efficacy.

References

[1] Gupta Rajni¸ "Amla : A Novel Ayurvedic herb with its Health Benefits",5th International conference on Innovative trends in Science Engineering and Management, Mumbai, Maharashtra.

[2] www.bitterrootrestoration.com

[3] Thakur R.S. Puri H.S. and Husain A. (1989), "Major Medicinal Plants of India Central Institute of Medicinal and Aromatic Plants", Lucknow, India.

[4] Yadav V, Duvey B, Sharma S And B. Deni, "Amla (Emblica officinalis) Medicinal Food and Pharmacolgical Activity". International, Journal of Pharmaceutical and Chemical Sciences Vol. 3, Jul – Sep 2014, 616-619.

[5] www.manushi.indiaorg

[6] Kumar Anil et. Al, International Journal of Pharmaceutical and Chemical Sciences, Vol. 1, Jan-Mar, 2012

[7] Treadway, Linda, "Amla : Traditional food and Medicine; Herbal Gram", The Journal of American Botanical Council Issue - 31, 1994, pp 26

[8] Swetha Dasaroju Krishna Maha Gottumakkala, "Current Trends in Research of Enblica officinalis (Amla) A Pharmacological Perspective" International Journal of Science Rev. Res 24-2, Jan–Feb-2014 ho 25, 150-159

[9] Mirunalini S., Vaithi yanathan V. krisnaveni M, "Amla-A Novel Ayurvedic herb as a functional food for health benefits, A Mini Review" International Journal of Pharmacy and Pharmaceutical Science, Vol. 5 sppl 1, 2013.

[10] Bhattacharya S. K., Bhattacharya A, Sairamm k, Ghosal S, "Effect of Bioactive tannoids Principles of emblica officinalis on ischemia reperfusion induced oxidative stress in rat heart" Phyto medicine, 2012; 9 (2), 171-174.

[11] Anila L., Vijayalakshmi N R, "Flavanoids from Emblica officinalis and mangifera indica effectiveness for dyslipidemia", Journal of Ethno Pharmacology 2002; 79 (1), 81 7.

[12] Singh E, Sharma S, Pareek A, Dwivedi J, Yadav S, and Sharma S, "Phytochemistry, Traditional uses and cancer chemopreventive Activityof Amla phyllanthus emblica The Sustainer"; Journal of Applied Phamaceutical Science 2 (01), 2011; 176-183

[13] Khan KH; Roles of Emblica officinalis in Medicine-A Review, Botany Research International, 2(4), 2009, 218-228.

[14] Nadkarni K. M., Nadkari A. K. "Indian Materia Medica with Ayurvedic Unani Tibbi Siddha Allopathic, Homeopathic and Home Remedies" ISBN-8-7154-142-91, 1999

[15] Drury "Colonel Herb : The Useful Plants of India with notices of their Cheif Medicinal Value in Commerce, Medicines and the Arts" Higginobothamad Co. Madras. 1873

[16] Srivasuki K. P., Nutritional and Health care benefits of Amla. Journal of Pharmacognosy Vol. 3, Issue-2, 2012.

[17] Ghosal S, Triphathi V. K. Chauhan S, "Active Constituents of Empilica officinales Part l The Chemistry and Antioxidant Effect of two new Hydrosable tannis, emblicanin A and B International Journal of Chemistry, 35, 1996, 941-8

[18] Panday Govind "Some Important Anticancer Herbs; A Review" International Research J. of Pharmacy 2(7), 2011, 45-52

[19] Aktar M. S., Ramzan A, Ali A, Ahmad M. "Effect of amla fruit (Emblica officilis Gactn.) on blood glucose and lipid Profile of normal subject and type 2 diabetic patients". Int. J. Food Sc. Nutr. 2011 April 18, (E pub ahead of print)

[20] Chakraborty D, Verma R. spermatotoxic effect of ochratoxin and its amelioration by Emblica officinalis aqueous extract, Acta Pol. Pharm. 2009, 66(6), 689-695.

[21] Sampath Kumar K. P., Bhowmik D, Dutta A, Yadav A, Paswan S, Shweta S, Lokesh D, Recent Trends in Potential traditional Indian Herb Emblica officinalis and its Medicinal Importance, Journal Of Pharmacognosy and Phyto chemistry, 1 (1), 2012, 24-32

[22] Kaushik Vilas Kulkarni, Shrishail M. Ghurghure, Indian Gooseberry (Emblica officinalis): Complete pharmacognosy review. International Journal of Chemistry Studies; Vol. 2, Issue-2, March 2018, pp 05-11

[23] Yokozawa.T, Kim, H. Y., H. J., Okubo, T., Chu, D.-C., and Juneja, L. R. (2007) Amla (Emblica officinalis Gaertn.) prevents dyslipidaemia and oxidative stress in the ageing process. British Journal of nutrition, 97 (6), 1187-1195

[24] Jayaweera D. M. A. (1980) National Science Council of Sri Lanka.

DEVELOPING INTEGRATED MALAPRABHA DIGESTER FOR MANAGEMENT OF KITCHEN WASTE AND HUMAN EXCRETA

Chandrashekhar Parab*, Sameer Shastri, Mrunal Meshram
Civil Engineering Department,
Sinhgad College of Engineering, Vadgaon, Pune, India.
*corresponding author email: chandrashekharparab94@gmail.com

Abstract

Faecal matter of all humans consists of sulfur, nitrogenous and carbonaceous matter. These compounds when undergo reduction, they release many useful compounds in the form of gases. However, when the carbonaceous compounds are subjected to anaerobic digestion, they form organic acids, which when acted by methane forming bacteria forms methane. This project aims to design a integrated system (Integrated Malaprabha), which can harvest energy from human Faecal and kitchen waste in the form of methane, just like we have been doing with the Faecal matter of other animals, but in more efficient way and to also get a treated effluent, from the system, which would significantly decrease the load on a local sewage treatment plant. Malaprabha digester developed by Padmashree late Dr. S. V. Mapuskar is the base for developing this Integrated Malaprabha digester for treating human excreta. Unlike septic tank which releases methane gas into atmosphere, Integrated Malaprabha digester allows to capture this gas and utilizes for cooking purpose. In this report, feasibility of adding kitchen waste along with human excreta is studied and measures to overcome problems due to integration of the same are mentioned. Also benefits, technical feasibility and future scope of Integrated Malaprabha digester is studied.

Keywords: Faecal Matters, Treatment, Anaerobic Digestion, Malaprabha Digester

1. Introduction

Malaprabha digester is one of the component in Decentralized Onsite Waste Management system (DOSIWAM) developed by Padmashree late Dr. S. V. Mapuskar first successfully implemented in 1980 in Dehu village in Pune. Malaprabha Digester technology basically is a 'toilet – linked biogas plant' where human excreta are converted to biogas in a specially designed digester chambers. It has 3 chambers, the first chamber having Hydraulic Retention Time (HRT) of 30 days, and the rest two having together of 15 days of HRT. These final calculations have been estimated by research and development on many systems

installed by Dr. Mapuskar. The HRT of the first chamber is high as compared to the other two as the gas from first chamber is only trapped for reuse. There is negligible amount of gas generation from the other two compartments, which escapes to the atmosphere. The effluent, after treatment, is discharged into the drainage system. Salmonella typhii, a pathogen found in night soil is the most harmful and having longest life span. It can survive for 6 weeks in anaerobic conditions. Thus, to ensure the pathogen free effluent, Hydraulic Retention Time (HRT) for the whole plant is 45 days. The conventional Malaprabha digester was adopted for human excreta or human night soil. It was said by Dr. S. V. Mapuskar that Malaprabha system, with a little modification, can also be used for other feeds. In this project integration of kitchen waste and human excreta was adopted by using modified Malaprabha digester. A small scale model of Malaprabha digester was prepared considering all the standard design aspects. Leftover food, fruits peels along with human excreta was added daily in the model. As the part of the project study flaring time of biogas and reduction in outlet parameters like BOD, COD, TS, TDS, TSS after anaerobic digestion of kitchen and human excreta is studied.

2. Design of Model

Fig. 1: Front view of model

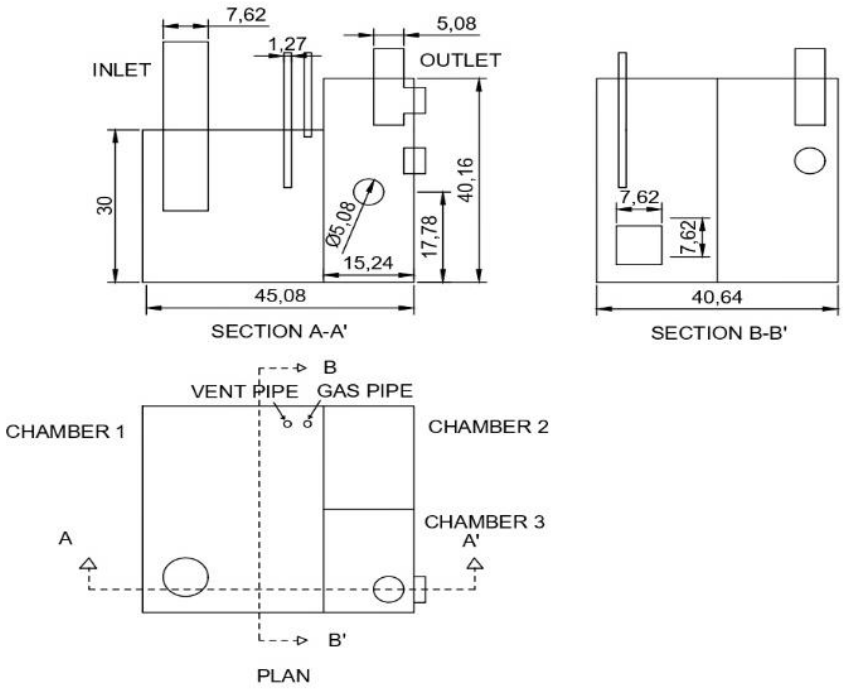


Fig. 2: Plan and section of model (all dimensions are in cm)

Model was designed considering generation rate of kitchen waste and excreta of a single person. Acrylic material was used as its non-corrosive, strong and transparent. It has three compartments having total hydraulic retention time as 45 days. For the lab scale model feed flow rate of kitchen waste slurry and human excreta combined is considered as 1 lit/day. Therefore volume of the model = Q×D.T=1×45=45 litre=45,000 〖cm〗^3. Now the depth of model is considered as 25 cm. therefore surface are of the model = Volume/Depth=45000/25= 1800 〖cm〗^3. Therefore dienssions of the square model will be √1800=42.42 cm×42.42 cm. For the convenience in crafting the model in acrylic sheets, dimenssions are taken in feet and inches. Hence overall dimenssions of the model is taken as 1 ft 5.7 inches×1ft 4 inches=45.08 cm×40.64 cm= 45801.28 〖cm〗^3>45000 〖cm〗^3. Model was prepared by a local craftsmen. Volume of Dimensions of model is shown below.

Sr. No.	Description	Dimension
1	1st chamber	30 cm × 40.64 cm × 30 cm
2	2nd and 3rd chamber	15.24 cm × 20.32 cm × 40.16cm
3	Overall length	45.08 cm

4	Overall width	40.64 cm
5	Inlet pipe diameter	7.62 cm
6	Outlet pipe diameter	5.08 cm
7	Opening b/w 2nd and 3rd chamber	5.08 cm
8	Opening b/w 1st and 2nd chamber	7.62 cm × 7.62 cm
9	Gas pipe and vent pipe diameter	1.27 cm

Table No. 1: Model details

3. Experimental Work

3.1 Sample Collection

Malaprabha digester was designed to run on human excreta and kitchen waste. First model was completely filled with water upto the level of bottom surface of inlet pipe. Human excreta was collected at house and fed in the model manually. Kitchen waste was also collected at house as leftover food like rice, bread, potato, nuts, peels of vegetables and fruits etc. For proper digestion of integrated sample, kitchen waste was grinded before feeding in the model. Grinding helped to decrease size of kitchen waste and thus microorganisms got more surface area to act on. 0.1-0.2 kg of human waste and 0.1-0.2 kg of kitchen waste on wet basis respectively was fed in the model every day having average concentration of 0.830 gBOD/lit/day. Following parameters were checked for the Influent and Effluent sample for Malaprabha digester model 10 days after the first feeding.

1. pH
2. Temperature
3. Chemical oxygen demand
4. Biochemical oxygen demand
5. Total solids
6. Total dissolved solids
7. Total suspended solids
8. Ignition of biogas

A plastic pipe with heat resistive ON-OFF knob was fixed to the gas pipe of the model. Gas was burn with the help of burning candle. Gas was allowed to burn till it get completely finished from the model**.**

Fig. 3: Burning of biogas

4. Result and Discussion

Samples from inlet and outlet from the Malaprabha model were analyzed for pH, temperature, Total BOD (3 days), Total COD, TS, TDS and TSS. Biogas was observed at 10th day from the first feeding. Gas was burnt at alternate day and time of flaring was measured. Following is the analysis and statistical representation of the performed experiments.

1. Date:- 26/09/2020					
Sample collection time:- 9:10 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.8	6.5	-	-
2	Temperature	26	27	o C	-
3	Total BOD	675	48	mg/lit	(-)92.88
4	Total COD	4160	480	mg/lit	(-)88.46
5	TS	2400	200	mg/lit	(-)91.66
6	TDS	461	180	mg/lit	(-)56.61
7	TSS	1939	20	mg/lit	(-)98.96
2. Date:- 02/10/2020					
Sample collection time:- 9:00 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.7	6.9	-	-
2	Temperature	28	29	o C	-
3	Total BOD	625	120	mg/lit	(-)80.80
4	Total COD	7271	390	mg/lit	(-)94.63
5	TS	2907	310	mg/lit	(-)89.33
6	TDS	367	198	mg/lit	(-)46.04
7	TSS	2540	112	mg/lit	(-)95.59
3. Date:- 07/10/2020					
Sample collection time:- 9:00 am					

Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.8	6.9	-	-
2	Temperature	26	27	o C	-
3	Total BOD	715	180	mg/lit	(-)74.82
4	Total COD	6270	350	mg/lit	(-)94.41
5	TS	3100	810	mg/lit	(-)73.87
6	TDS	375	227	mg/lit	(-)39.46
7	TSS	2725	583	mg/lit	(-)78.60
4. Date:- 15/10/2020					
Sample collection time:- 8:55 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.5	7	-	-
2	Temperature	33	33	o C	-
3	Total BOD	825	280	mg/lit	(-)66.06
4	Total COD	7510	480	mg/lit	(-)93.61
5	TS	3210	750	mg/lit	(-)76.65
6	TDS	382	250	mg/lit	(-)34.55
7	TSS	2828	500	mg/lit	(-)82.31
5. Date:- 23/10/2020					
Sample collection time:- 9:20 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.5	7	-	-
2	Temperature	30	32	o C	-
3	Total BOD	1015	511	mg/lit	(-)49.66
4	Total COD	8250	1732	mg/lit	(-)79.00
5	TS	3700	1425	mg/lit	(-)61.48
6	TDS	421	151	mg/lit	(-)64.13
7	TSS	3279	1274	mg/lit	(-)61.14
6. Date:- 30/10/2020					
Sample collection time:- 9:10 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6	7.1	-	-
2	Temperature	32	32	o C	-
3	Total BOD	721	422	mg/lit	(-)41.47
4	Total COD	7150	2711	mg/lit	(-)60.08
5	TS	2705	821	mg/lit	(-)69.64
6	TDS	445	220	mg/lit	(-)50.56
7	TSS	2260	601	mg/lit	(-)73.40
7. Date:- 07/11/2020					
Sample collection time:- 9:00 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.1	7.1	-	-
2	Temperature	30	31	o C	-
3	Total BOD	818	315	mg/lit	(-)61.49
4	Total COD	8567	2759	mg/lit	(-)67.79
5	TS	3150	725	mg/lit	(-)76.98

6	TDS	611	392	mg/lit	(-)35.84
7	TSS	2539	333	mg/lit	(-)86.88
8. Date:- 15/11/2020					
Sample collection time:- 9:10 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.5	6.9	-	-
2	Temperature	29	31	o C	-
3	Total BOD	875	341	mg/lit	(-)61.03
4	Total COD	7421	2114	mg/lit	(-)71.51
5	TS	2985	780	mg/lit	(-)73.86
6	TDS	534	215	mg/lit	(-)59.74
7	TSS	2451	565	mg/lit	(-)76.94
9. Date:- 21/11/2020					
Sample collection time:- 9:20 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.3	7.2	-	-
2	Temperature	30	30	o C	-
3	Total BOD	1014	478	mg/lit	(-)52.86
4	Total COD	8645	3851	mg/lit	(-)55.45
5	TS	2864	632	mg/lit	(-)77.93
6	TDS	596	253	mg/lit	(-)57.55
7	TSS	2268	379	mg/lit	(-)83.28
10. Date:- 29/11/2020					
Sample collection time:- 9:30 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.6	7.1	-	-
2	Temperature	30	32	o C	-
3	Total BOD	987	450	mg/lit	(-)54.41
4	Total COD	8974	3678	mg/lit	(-)59.19
5	TS	3154	810	mg/lit	(-)74.38
6	TDS	440	148	mg/lit	(-)66.36
7	TSS	2714	662	mg/lit	(-)75.60
1. Date:- 26/09/2020					
Sample collection time:- 9:10 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.8	6.5	-	-
2	Temperature	26	27	o C	-
3	Total BOD	675	48	mg/lit	(-)92.88
4	Total COD	4160	480	mg/lit	(-)88.46
5	TS	2400	200	mg/lit	(-)91.66
6	TDS	461	180	mg/lit	(-)56.61

7	TSS	1939	20	mg/lit	(-)98.96
2. Date:- 02/10/2020					
Sample collection time:- 9:00 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.7	6.9	-	-
2	Temperature	28	29	o C	-
3	Total BOD	625	120	mg/lit	(-)80.80
4	Total COD	7271	390	mg/lit	(-)94.63
5	TS	2907	310	mg/lit	(-)89.33
6	TDS	367	198	mg/lit	(-)46.04
7	TSS	2540	112	mg/lit	(-)95.59
3. Date:- 07/10/2020					
Sample collection time:- 9:00 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.8	6.9	-	-
2	Temperature	26	27	o C	-
3	Total BOD	715	180	mg/lit	(-)74.82
4	Total COD	6270	350	mg/lit	(-)94.41
5	TS	3100	810	mg/lit	(-)73.87
6	TDS	375	227	mg/lit	(-)39.46
7	TSS	2725	583	mg/lit	(-)78.60
4. Date:- 15/10/2020					
Sample collection time:- 8:55 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.5	7	-	-
2	Temperature	33	33	o C	-
3	Total BOD	825	280	mg/lit	(-)66.06
4	Total COD	7510	480	mg/lit	(-)93.61
5	TS	3210	750	mg/lit	(-)76.65
6	TDS	382	250	mg/lit	(-)34.55
7	TSS	2828	500	mg/lit	(-)82.31
5. Date:- 23/10/2020					
Sample collection time:- 9:20 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.5	7	-	-
2	Temperature	30	32	o C	-
3	Total BOD	1015	511	mg/lit	(-)49.66
4	Total COD	8250	1732	mg/lit	(-)79.00
5	TS	3700	1425	mg/lit	(-)61.48
6	TDS	421	151	mg/lit	(-)64.13
7	TSS	3279	1274	mg/lit	(-)61.14
6. Date:- 30/10/2020					
Sample collection time:- 9:10 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6	7.1	-	-
2	Temperature	32	32	o C	-

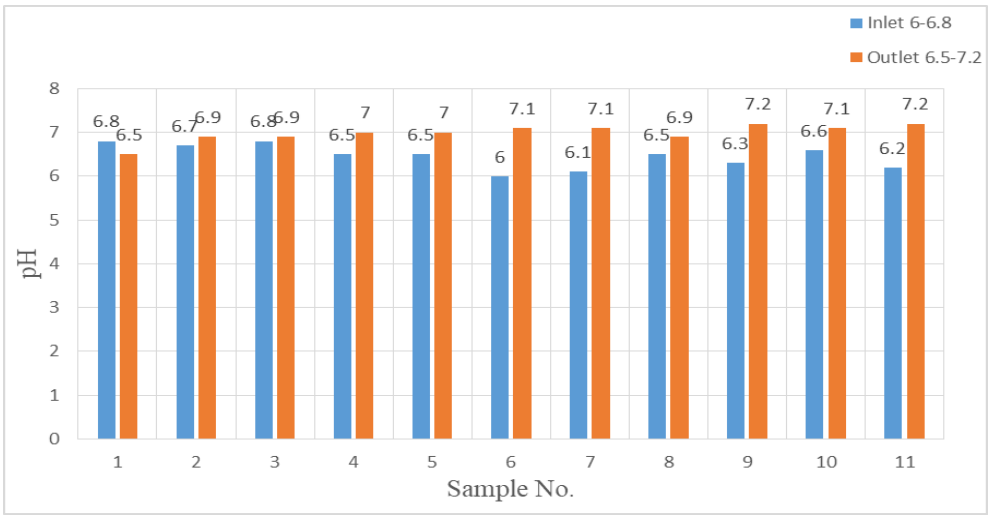
3	Total BOD	721	422	mg/lit	(-)41.47
4	Total COD	7150	2711	mg/lit	(-)60.08
5	TS	2705	821	mg/lit	(-)69.64
6	TDS	445	220	mg/lit	(-)50.56
7	TSS	2260	601	mg/lit	(-)73.40
7. Date:- 07/11/2020					
Sample collection time:- 9:00 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.1	7.1	-	-
2	Temperature	30	31	o C	-
3	Total BOD	818	315	mg/lit	(-)61.49
4	Total COD	8567	2759	mg/lit	(-)67.79
5	TS	3150	725	mg/lit	(-)76.98
6	TDS	611	392	mg/lit	(-)35.84
7	TSS	2539	333	mg/lit	(-)86.88
8. Date:- 15/11/2020					
Sample collection time:- 9:10 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.5	6.9	-	-
2	Temperature	29	31	o C	-
3	Total BOD	875	341	mg/lit	(-)61.03
4	Total COD	7421	2114	mg/lit	(-)71.51
5	TS	2985	780	mg/lit	(-)73.86
6	TDS	534	215	mg/lit	(-)59.74
7	TSS	2451	565	mg/lit	(-)76.94
9. Date:- 21/11/2020					
Sample collection time:- 9:20 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.3	7.2	-	-
2	Temperature	30	30	o C	-
3	Total BOD	1014	478	mg/lit	(-)52.86
4	Total COD	8645	3851	mg/lit	(-)55.45
5	TS	2864	632	mg/lit	(-)77.93
6	TDS	596	253	mg/lit	(-)57.55
7	TSS	2268	379	mg/lit	(-)83.28
10. Date:- 29/11/2020					
Sample collection time:- 9:30 am					
Sr. No.	Description	Inlet	Outlet	Unit	% Reduction
1	pH	6.6	7.1	-	-
2	Temperature	30	32	o C	-
3	Total BOD	987	450	mg/lit	(-)54.41
4	Total COD	8974	3678	mg/lit	(-)59.19
5	TS	3154	810	mg/lit	(-)74.38
6	TDS	440	148	mg/lit	(-)66.36
7	TSS	2714	662	mg/lit	(-)75.60

Table 2: Sample results

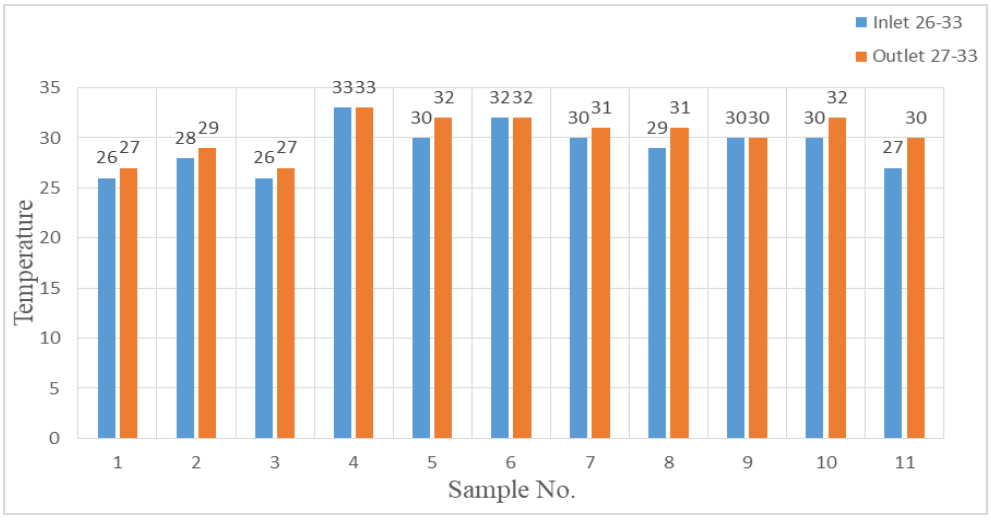
Sr. No.	Date	Duration of flaring		Sr. no.	Date	Duration of flaring	
		min	sec			min	sec
1	26-09-2020	3	30	21	05-11-2020	4	11
2	28-09-2020	3	8	22	07-11-2020	4	25
3	30-09-2020	3	15	23	09-11-2020	5	20
4	02-10-2020	2	15	24	11-11-2020	4	49
5	04-10-2020	2	10	25	13-11-2020	4	30
6	06-10-2020	2	3	26	15-11-2020	4	21
7	08-10-2020	2	49	27	17-11-2020	5	1
8	10-10-2020	2	12	28	19-11-2020	4	25
9	12-10-2020	3	21	29	21-11-2020	3	40
10	14-10-2020	3	41	30	23-11-2020	3	15
11	16-10-2020	2	15	31	25-11-2020	3	22
12	18-10-2020	2	50	32	27-11-2020	3	50
13	20-10-2020	3	46	33	29-11-2020	3	31
14	22-10-2020	4	56	34	01-12-2020	4	10
15	24-10-2020	5	12	35	03-12-2020	3	10
16	26-10-2020	4	58	36	05-12-2020	2	50
17	28-10-2020	4	50	37	07-12-2020	3	7
18	30-10-2020	5	24	38	09-12-2020	2	55
19	01-11-2020	3	46	39	11-12-2020	2	42
20	03-11-2020	3	0	40	13-12-2020	3	15

Table 3: Flaring of biogas

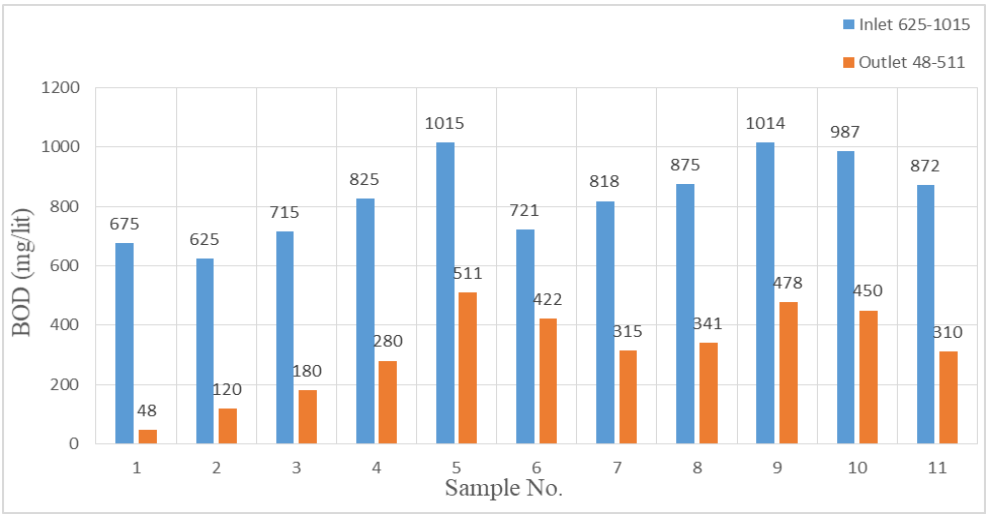
4.1 Graphical Comparision



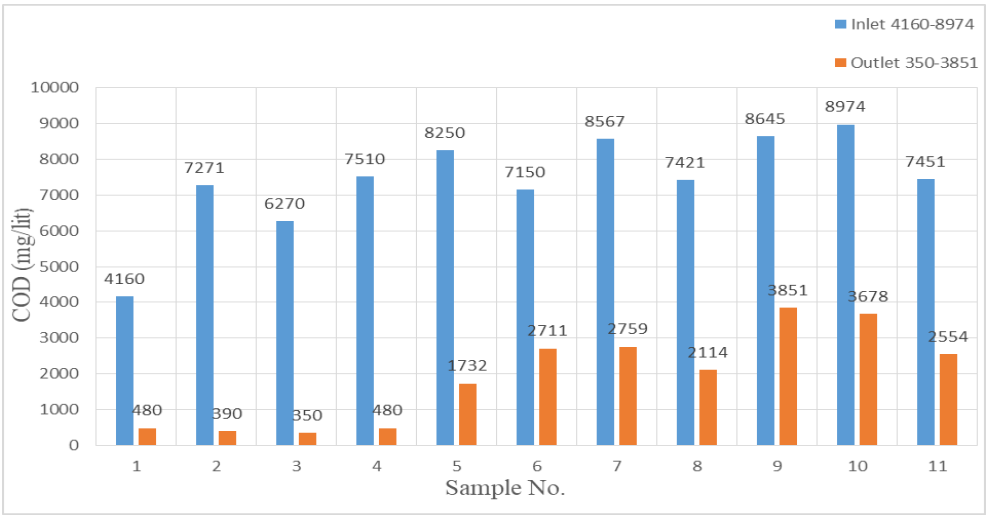
Graph 1: pH results from inlet and outlet samples



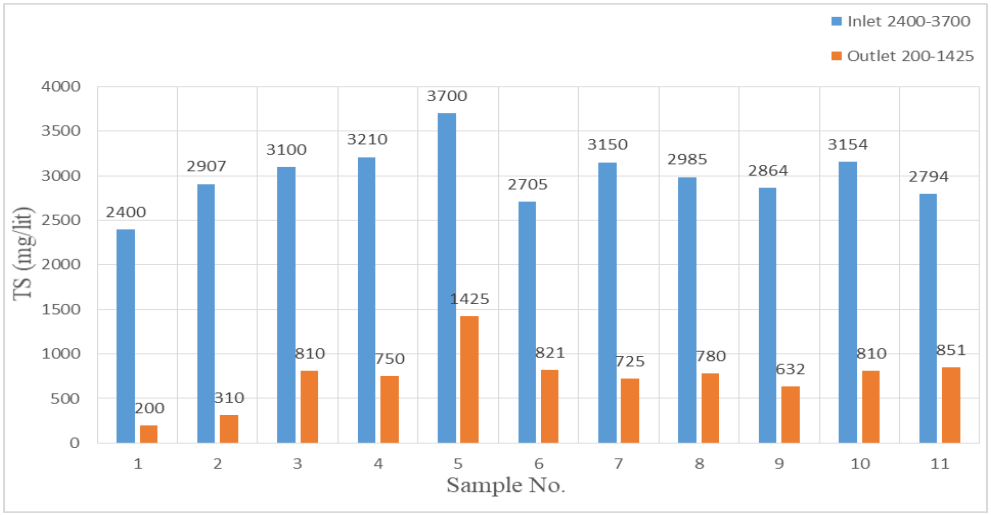
Graph 2: Temperature results from inlet and outlet samples



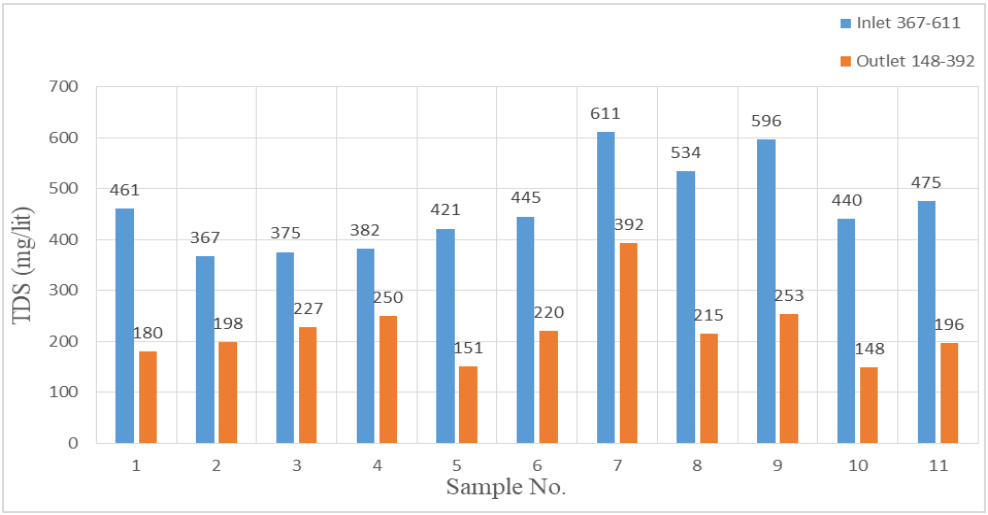
Graph 3: Total BOD results from inlet and outlet samples



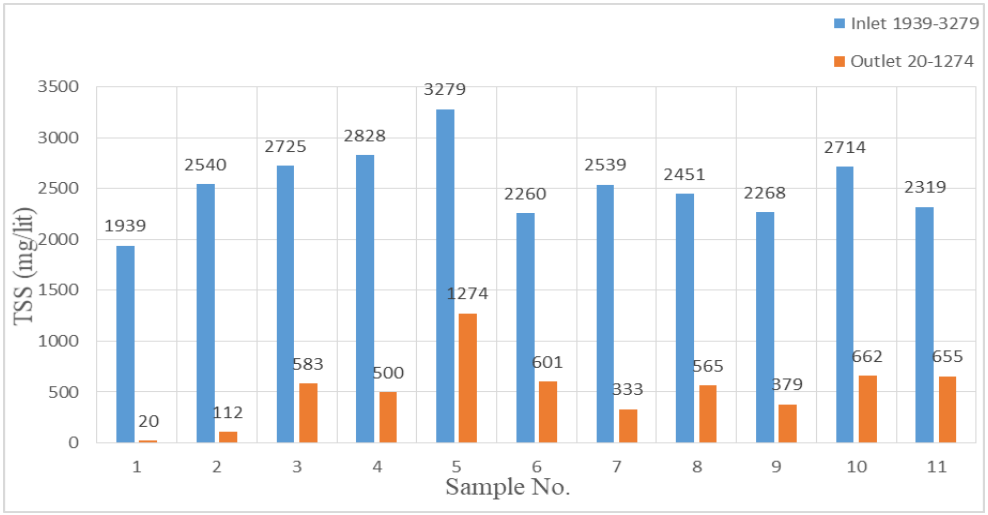
Graph 4: Total COD results from inlet and outlet samples



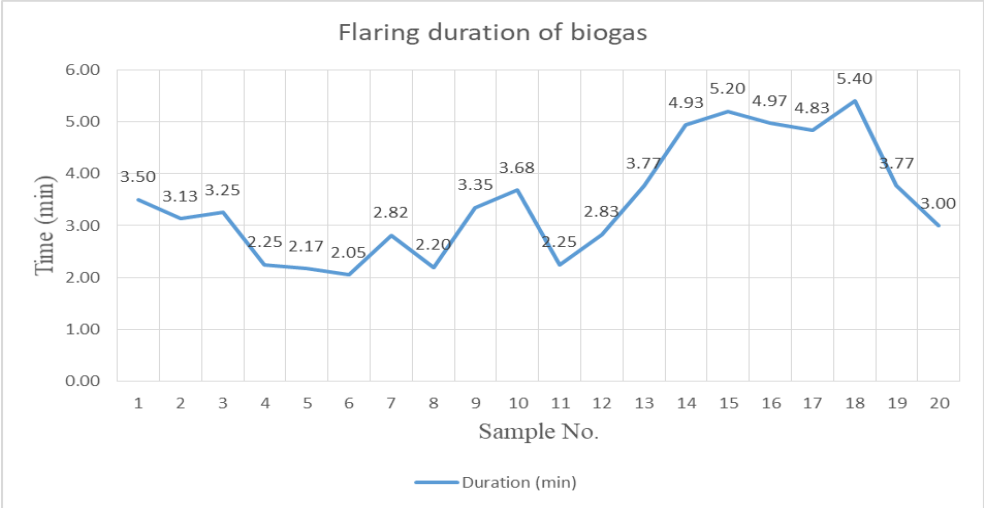
Graph 5: TS results from inlet and outlet samples



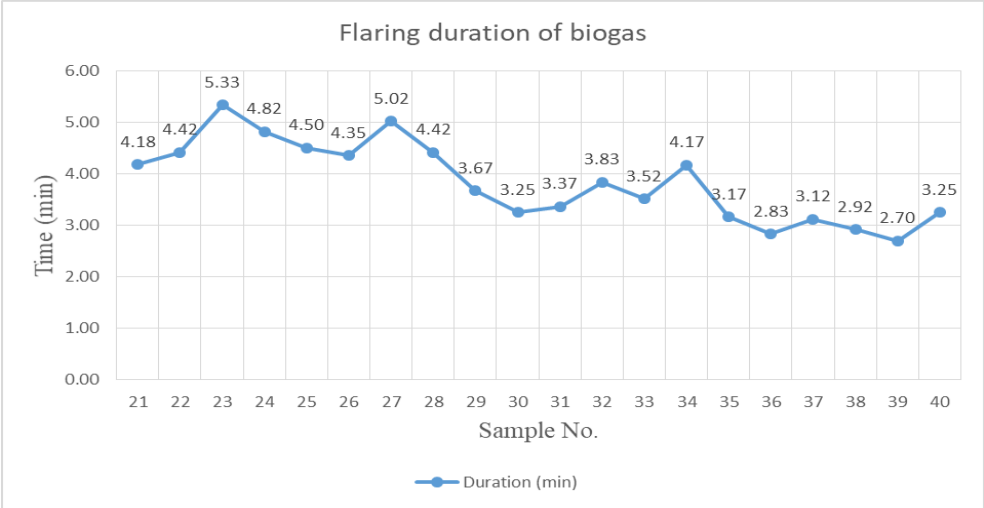
Graph 6: TDS results from inlet and outlet samples



Graph 7: TSS results from inlet and outlet samples



Graph 8: Flaring duration of biogas (1-20)



Graph 9: Flaring duration of biogas (21-40)

4.2 Quantitative Measurement of Biogas

The level of digested organic material drops as there is formation of biogas. Level get drops in the range of 7-10 cm every day. By considering maximum drop i.e. 10 cm, volume of biogas generated in the tank can be calculated as the surface area of 1st chamber is known. i.e. 40.08 cm × 40.64 cm × 10 cm = 16,288.512

cm3 = 0.016288 m3. Assuming density of biogas 1.25 kg/m3 mass of biogas = volume × density = 0.016288 × 1.25 = 0.02036 kg = 20.36 gm.

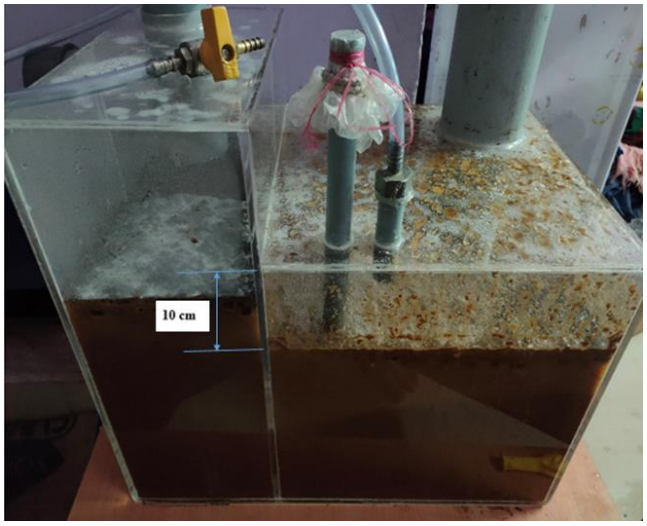


Fig. 4 Drop in level of 1st chamber

5. Cost Estimation of Project

Following calculations are done considering the digester is constructed in a building of 40 families so total families each having a population of 5 people per family, 200 users. For the calculation purpose it is assumed that 30 liters of water will be used for flushing by each person. Therefore, volume of the tank required will be 200×30×45 = 270000 liters of digester volume excluding storage space for gas etc. will be required. By considering space for gas accumulation and the space taken by organic waste, design value for the volume of system required is taken as 294 cubic meter. Volume of 1st chamber is taken as 10 m x14 m x1.5 meter Volume of 2nd and 3rd chamber is equal and taken as 4.95 m x 7 m x 1.5 meter. The plan and section view of such a system is shown below.

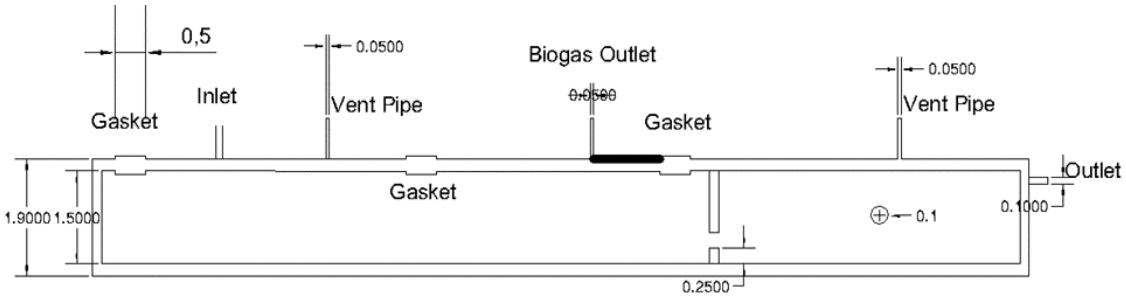


Fig. 5: Section

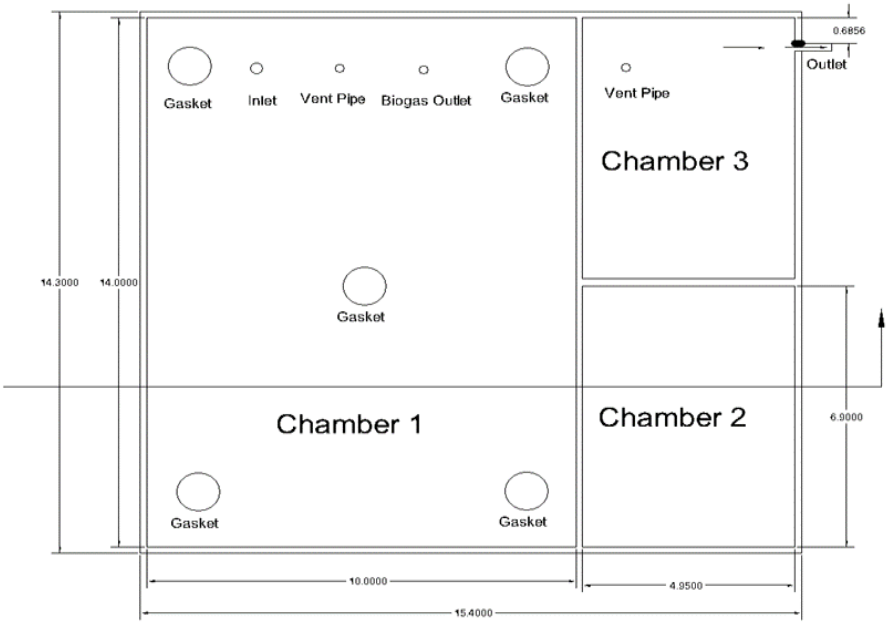


Fig. 6: Plan

The gaskets for the purpose of removal of scum are provided at 4 corners and one at middle for ease of removal of scum. The overall length and breadth of the tank will be taken 15.4 meter and 14 meters respectively. Height of the tank is taken as 1.9 meter. Construction cost of this size of a tank will be approximately equal to Rs. 36,00,000, that is Rs.18000 per user for a life time of energy and their waste management with little to no maintenance. However, since we are planning to construct it in prefabrication, the cost of construction will decrease by a significant amount. (Based on rough estimate, it will be around Rs. 20 lakhs) Also, it will lead to saving of 4.3 gas cylinders per family per year, which would

mean saving of 172 cylinders for the whole building per year. Also, the generated stabilized faucal sludge can be sold in the market as economical and nutritious manure for farmers.

6. Conclusion

From the above results feeding of around 200 gm kitchen waste and 200 gm human excreta, it was possible to produce 5.4 minutes of combustible gas. If this analogy is to be applied to a family of 5, it will be possible to produce approx. 30 minutes of cooking time using biogas. It was observed that there is reduction in BOD, COD, TS, TDS and TSS levels. Reduction of about 93 % to 42 % in BOD, 94 % to 56 % in COD, 91 % to 61 % in TS, 64 % to 35% in TDS and 98 % to 61 % in TSS was observed. Initially the reduction in the parameters observed was high because very less quanity of solids transferred from 1st chamber to 3rd chamber of the model. Outlet parameters are not meeting the standard parameters given by CPCB, therefore additional treatment is required before discharging into the environment.

Anaerobic process is very much dependent upon temperature variation and pH. An optimum pH is necessary i.e. about 6.5-7.5. Regular monitoring of various parameters is required to get best results. Integration of kitchen waste and human excreta in anaerobic digester can produce some amount of scum layer if implemented in large scale. Such problem can be solved with the help of gaskets. If this project if used on actual ground, can help to reduce the pollution parameters to great extent and this can be economical method which will be used to treat human excreta and kitchen waste. Construction cost a brick masonry tank for 200 people will be approximately equal to Rs. 36,00,000. If non-conventional construction material like ferro-cement is used, then cost will be reduced by 50 %. Outlet slurry can be used as excellent manure.

7. Future Scope

1. Further optimization in model to reduce size and cost.
2. Design of technology for mounted shredding machine so that municipal biodegradable waste can be added to the tank.
3. Proper disposal of the outlet slurry
4. Further cost reduction by adopting non-conventional construction materials like ferro-cement, precast cement etc.
5. Detailed gas measurement studies so that provision can be made to store excess gas.

References

[1] Assadawut Khanto and Peerakan Banjerdkij, "Biogas Production from Batch Anaerobic Co-Digestion of Night Soil with Food Waste", www.tshe.org/ea/index.html Environment Asia 9(1) (2016) 77-832016, 2016.

[2] Dr. Raveesh Agarwal, Mona Chaudhary, "Jayveer Singh Waste Management Initiatives In India For Human Well Being" European Scientific Journal June 2015 /SPECIAL/ edition ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431, June (2015).

[3] Alemayehu Gashaw, "Co-digestion of municipal organic wastes with night soil and cow dung for biogas production: A Review", African Journal of Biotechnology Vol. 15(2), pp. 32-44, DOI: 10.5897/AJB2015.14705 Article Number: C7FD78256989, Jan 2016.

[4] Sahassawas Poocheera, Ratchaphon Suntivarakorn, Wasakron Treedet, "Biogas Production from Human Faece and Community Waste Food" Advanced Materials Research Vols. 931-932 (2014) pp 1101-1105 Trans Tech Publications, Switzerland doi:10.4028/www.scientific.net/AMR.931-932.1101, May 2014.

[5] Antony Daisy and S. Kamaraj, "The Impact and Treatment of Night Soil in Anaerobic Digester: A Review", Journal of Microbial & Biochemical Technology, DOI: 10.4172/1948-5948.1000050, 2011.

[6] Ms. Sampada Kulkarni and Mr. RS Arun Kumar, Case study of sustainable sanitation projects Malaprabha Technology Dehu village, Dist. Pune, Maharashtra, INDIA, March 2010.

[7] Jyotsna Mapouskar, "Recycle Human Waste For Health, Wealth and Energy Malaprabha Biogas Plant Developed By Dr. S. V. Mapuskar New Design for Recovery of Biogas from Latrine" Jyotsna Arogya Prabodhan, June 1988,

[8] Central Pollution Control Board Manual

[9] CPCB Standards (cpcb.nic.in)

A BRIEF REVIEW ON SISAL: A RISING CELLULOSIC NATURAL MINOR FIBER OF 21ST CENTURY

T. N. Shaikh, S. B. Chaudhari*, Janki Patel, B. H. Patel[1]
Department of Textile Engineering, Faculty of Technology & Engineering, The Maharaja Sayajirao University of Baroda, Vadodara-390001, Gujarat, India.
[1]Department of Textile Chemistry, Faculty of Technology & Engineering, The Maharaja Sayajirao University of Baroda, Vadodara-390001, Gujarat, India.
*corresponding author email: s.b.chaudhari-ted@msubaroda.ac.in

1. Introduction

Sisal fiber originated in Mexico, is a leaf fiber obtained from the sisal plant of the Agavaceae family and is harder [1-7]. The name "sisal" was initiated from an anchorage urban in Mexico, where sisal stands for "cold water" [4-5]. There are more than 275 kinds of agaves available universally [2, 3, 8]. Agave cantala (Maguey, Cantala/Cantula, Bombay aloe), Agave vera-Cruz (Gray aloe, Elephant aloe, and Railway aloe), Agave fourcroydes (Mexican sisal, Henequen), Agave Americana (Century plant, American aloe), Agave Mexicana and Elephant aloe are the other well-known species, offering fibers of varying nature and properties [1, 3]. However, the most commercialized varieties are Agave Fourcroydes and Agave Sisalana [9-10]. Brazil, Tanzania, Kenya, West Indies, South Africa, Mozambique, Switzerland, Madagascar, Uganda, Haiti, Cuba, China, Venezuela, India, and Indonesia are the main players in the commercial market [1-3, 5-12]. The arid and semiarid regions; Odisha, Andhra Pradesh, Madhya Pradesh, Bihar, Jharkhand, Chhattisgarh, Maharashtra, Tamil Nadu, and Karnataka are the major producers in India for the sisal fiber, many a time is known as Ketki and Raam Baan [2-3,8,11]. The fascinating thing about this natural fiber is unlike others it does not require to be cultivated in an organized way; no pesticides and fertilizer are used in its cultivation. Hence, this plant is mostly grown on banks, bunds, roadsides, and borders of fields with minimum monetary investment and care, categorized as wild fiber. Its extensive root system facilitates reduced soil erosion, as well as more carbon dioxide, is absorbed by this beneficiary plant than it produces. The sisal plant is grown up in all kinds of soil except clay. Thereby a systematic utilization of this fiber in the textile field can result in quite an economical output. The use of this minute technically explored hollow structured natural cellulosic fiber is mainly due to its inherent auspicious characteristics; biodegradability, low density, high specific strength, etc. as well thermal and acoustic insulation like functional properties. Thus a noble but seldom found combination in natural fiber can be visualized in sisal [2-5, 8, 11].

A sisal plant is mainly grown in tropical and subtropical regions. But it can sustain in all types of weather. Sisal plant has a lifespan of 7-12 years, out of which initial 3 to 4

years for the maturity of the plant until leaves can grow maximum to 2 meters. Thereby extracted fiber length will be affected by the length of the fresh leaves, and also by the conditions under which they are processed. Under ideal conditions, the length of extracted fibers can approach the length of the leaves. A fully grown plant can produce 200-250 commercially usable leaves [fig. 1] depending on location, climate, altitude, and variety of plants. There is an average of around 1000-1200 fiber bundles per leaf. The constitution of the leaf includes major stuff 87.5% water and the rest; 8% dry matter, 4% fiber and 0.75% cuticle [1, 3-13].

2. Fibre Extraction Methods

There are three popularly used fiber extraction methods:

1. Water Retting

2. Boiling

3. Scrapping or Mechanical Decortications

1. Water Retting: The fibers are separated from the binding core via a conventional very slow biodegradation process by breaking chemical bonds. These separated fibers are then washed thoroughly before further use. The process is time consuming and usually last for 20 days, also it is unhygienic and non-eco-friendly.

2. Boiling: As the name suggests boiling of sisal leaves followed by their beating to get the fibers are the steps followed in this method. Such extracted fibers need to be cleaned by water and dried in sunlight for a few days to get desired clean usable fibers. The method is restricted to large scale extraction although it is better in all aspects than retting.

3. Scrapping or Mechanical Decortications: The sisal leaves are pulled through pairs of rollers just like sugar cane, as shown in figure 1 to get rid of the juicy matter. The rollers are operated positively by an electric motor. It removes the fleshy pulp from the fibers. However, chlorophyll, leaf juices, and adhesive solids from the extracted fibers can be removed on washing with running water followed by drying in bright sunlight. The process is eco-friendly, hygienic, and fast [2-3, 5, 11, 13].

The fibers obtained on the extraction of leaves are not the same in terms of their physical structure; differing purely based on the location they have been earned.

3. Sisal Fibre Types

Mechanical, ribbon, xylem are the three varieties of the fibers obtained from the sisal leaf (figure 2).

1. Mechanical: The most commercially useful fibers obtained from the peripheral part of the leaf are referred as mechanical fibers, usually they are unevenly thickened with a horseshoe-like shape.

2. Ribbon: This category of the fibers is obtained from the main juicy body of the leaves, ensue in connotation with the conducting tissues and observe a cross-section of a sisal leaf. The mechanical strength of these ribbon fibers arises from the accompanying tissue structure of the ribbon. They are the longest fibers in the group but during processing can be fragmented easily in a longitudinal direction.

3. Xylem: Unlike the ribbon fibers, through the connection of vascular bundles this class of fibers possesses an irregular shape. This thin-walled cells fiber structure can be easily shattered during the extraction process [2, 7, 11, 13].

Low tensile strength and modulus but high breaking strain are expected characteristics for the fiber extracted from the lower part of the leaf. However, the stiffness and strength get added at the middle area of this leaf fiber but the tip has reasonable properties. The fiber toughness, tensile strength, and modulus get descend with an increase in temperature [3]. Although extracted from the same leaf, a wide variation executed in the fiber type can differ their physical and mechanical properties as well as chemical compositions, thus making this natural fiber inherently highly variable. Results of various researchers in these regards are summarized in Tables 1 and 2, which substantiate the expectation [2-5, 11, 13-21].

Sisal fiber is made up of several ultimate cells. This multicellular arrangement of the fiber cells is described by a middle lamella, thickened walls (polygonal shape), and large lumen (rounded corners). Longitudinally the fiber is straight and without crimp, and approximately cylindrical in appearance. There are many stripes and knots on the surface of the fiber, which confirms that the many single cells are arranged in a straight parallel line and make a fiber bundle. The non-uniform cross section of the sisal fiber bundle shows 100-200 single fiber cells with hollow structures which are bonded together by natural gum with a well-defined size of the lumen. A sisal fiber cell is composed of a few fibrillar structures, each consist of fibrillae. The fibrillae are arranged in spiral shape at 40° with respect to the longitudinal axis in the secondary wall as shown in figure 3. The average diameter of its primary wall is a 23μm, and the fibrillae are curved with a slope of 18° to 25° in the inner secondary wall, The central tertiary wall is enclosed with the lumen and the average diameter of the lumen is 11μm [2, 14].

4. Chemical and Physical Characteristics of Sisal Fibre

The Sisal fiber is a group of hollow sub-fibers. The chemical configuration of sisal fiber includes major cellulose stuff followed by hemicellulose, lignin, and pectin. The proportion of these components in the fibers differ as per age, source, measurement methods, etc., causes a large change in the chemical structure of the fiber (Table 1). Spirally oriented cellulose is reinforced in a hemicellulose and lignin matrix of the fiber cell walls. Thus, the exterior surface of the cell wall is consists of lignocellulose and waxy substance layers keeps the cell bound to its neighboring fiber cell wall [2-3, 5, 8, 11-19].

The mechanical and physical parameters of the sisal fiber are affected by its chemical constituents and their concentrations. Major characteristics are given in Table 2 [2-5, 11, 13-21]. The sisal fibers are strong, non-abrasive, non-toxic, and biodegradable. If processing has been carried out carefully, the sisal can acquire creamy white color. Depending on fiber age and location, its length and diameter measures differ [5, 11] and give wide variation in the length of strands of commercial sisal fibers in the range of 0.5 to 1.5 meters and 100 to 300 μm in diameter. Lower micro fibrile angle influenced the mechanical properties to a great extent. The presence of a higher amount of cellulose, glucan polymer in the structure has made it hydrophilic. This hydrophilic, as well as high mechanical strength of the sisal fiber, are attributed to the presence of glucan polymer consists of a linear chain of 1, 4 β bonded anhydroglucose units. Being cellulosic, these fibers on burning behave like cotton; burn quickly and smell like burning paper or grass and it leaves soft gray ash [13, 21]. Similarly, this cellulosic fiber expresses good dye affinity like cotton fiber and especially for acid dye. Hemicellulose in the structure acts as a compatibilizer between lignin and cellulose because it contains many sugar units such as polysaccharides having a degree of chain branching, where lignin provides rigidity and thereby higher stiffness and compressive strength to the fibers. Apart from these lignin also acts as a binder within and between fibers and aid to keep the moisture in the fibers. As a result of this higher moisture regain (10-11%) has been observed than purely cellulosic cotton fibers (8-10%). Wax mainly represents a blend of replaceable long chains of aliphatic hydrocarbons, such as fatty acids, ketone, alkaline, primary and secondary alcohols, and some other constitutes. Pectin in the structure imparts flexibility; thereby this fiber is having high strength, durability, and ability to stretch [13, 22].

5. Brief Summary of Research

Widely varying properties of this natural cellulosic fiber have a limited span of experimentation mainly to fiber and nonwoven/ composite structure. However, some researchers have explored manual spinning and weaving areas but the quantum of work reported is very little. A big research gap has been found due to handling

difficulties faced for very long and intrinsically highly variable stiffer fiber. Their entire zest of research work thereby represented category wise; fiber, yarn, fabric, and nonwoven.

5.1 Fiber

Mukherjee et al [23] had found the variation in the denier with varying fiber diameter, thereby concluding that the fibers were not shaped as cylindrical but ribbon type. The fractured tips of the tensile strength tested sisal fibers were also analyzed through SEM and shown the failure of the fiber was due to the unraveling of microfibrils along with de cohesion, which was finally resulted in tearing of the cell walls. The uncoiling propensity was found to be decreased with the increase in testing speed. Mwaikambo et al [24] had observed SEM images of raw and alkali treated sisal fibers. During this study, they had found irregular and void regions between individual fiber cells. Adjacent ultimate fibers were held together by node like structure. According to them, the tensile strength of sisal fiber bundles depends on the physical characteristics of its interior structure such as the cellulose content, crystallinity index, and microfibrillar angle. Ankita et al [16] had used hemicellulase, pectinase, and cellulase enzymes for treating the sisal fibers at different enzyme concentrations for various time periods. They had observed increased percent weight loss with increased treatment conditions, and this reduced weight was due to the removal of the impurities from the sisal fibers on treatment. The decrease in tensile strength of the fiber with the increased weight loss was also noticed with the progress of enzymatic treatment as more strength imparting noncellulosic materials got removed. They had observed SEM micrographs of raw and enzyme treated sisal fibers. They had concluded that rough and irregular surfaces of untreated fibers get smoothen to some extent while treated. Maria et al [25] had observed from the SEM micrographs of mercerization and acetylation treated fibers that the surface impurities and the separation of ultimate cells occurred due to the extraction of the cementing components such as lignin and hemicelluloses. However, from the SEM results, the noteworthy differences in the fiber surface morphology with respect to the variation of three parameters studied, viz; the temperature, concentration, and time were not seen. Hence, the treatment used during the study to remove cellulose and hemicellulose had caused only changes in the fiber surface without affecting the fiber structure, thus the morphological changes were independent of the treatment conditions used. Oksman et al [26] had observed parenchyma cells surrounded by the technical fibers of varying diameter along with the fiber length which is smallest towards the end. Mukherjee et al [23] observed large variations in fiber diameter at a higher magnification and also found no substantial difference in mechanical properties with the change in fiber diameter. They had testified that with the increase in test fiber length the tensile strength and percent elongation at the break were decreased but Young's modulus was increased. Apart from this, with the increased testing speed, the rise in Young's modulus and tensile

strength were observed, however, no significant change was noticed for elongation. The characteristics of the stress-strain curve for sisal fiber included an initial linear region followed by a curvature indicating the increased rate of strain produced with an increase in stress due to the viscoelastic nature of the fiber. Adriana et al [27] defined on the basis of DSC studies that the degradation of sisal fiber depended on atmospheric conditions similar to other natural stem fibers like jute and hemp. Yang et al [28] had found that the IR spectrum did not change below 200°C treatment whereas, an increase in density and crystallinity were noticed. Mishra et al [29] and Chand et al [30] had treated sisal fibers with acetylation and observed a reduction in moisture absorption on treatment, a valuable outcome for engineering applications. Chand et al [30] had also studied the tensile strength of acetylated sisal fibers and found a reduction in tensile strength due to the loss of hemicellulose in the fiber. Yang et al [31] had done surface modification of the sisal fiber by various treatments like; alkali, benzol/alcohol, dewax, acetylated, thermal, etc., and then worked out effects of the treatments on tensile properties (Table 3). The highest strength was observed with thermal treatment followed by Benzol (alcohol) treatment, because of the increase in crystallinity at 150°C. However, at 200°C the drop in tensile properties was started owing to the degradation of sisal fiber. Other surface treatments enhanced the ductility and decreased the modulus. Significant negative impact on tensile behavior was observed right from acidic, alkaline, or any combination. This behavior was indicative of fiber damage caused along with the removal of strength bearing elements. According to the outcomes of Kalia et al [32] and Hajiha et al [33] acetylation, alkalization/acetylation, and silanization treatments had enhanced the compatibility of natural fibers with matrix. Better thermal stability of sisal fiber was observed on alkali treatment with a 5% concentration, and a change in the morphology of sisal fibers on benzoylation and grafting. However, the surface of sisal fibers became rough in this course in comparison with the clear and smooth surface of the original sisal fibers.

5.2 Yarn

Zwane et al [34] had produced yarn on the traditional spinning wheel and used a liquid binder dissolved in water with a ratio of 1:2. The yarn was made out of optimal concentrated NaOH treated sisal fibers. Decrease in the yarn breaking strength, elongation, linear density, and absorbency time have been noticed with an increased concentration of NaOH. The removal of strengthening layers was the main cause for the reduction in breaking strength, resulting in weaker and thinner fibers. Whereas, increased fragility of the fibers led towards a drop in elongation of yarns. Apart from that, remarkable descent in the yarn linear density was observed with an increase in alkalinity. However, the use of elementary spinning technology has caused higher variation in yarn evenness. Conversely, the yarn stiffness was not influenced by the use of a binding agent in spinning. Higher yarn absorbency observed was accounted for the removal of water hindering elements like; lignin and hemicellulose from the

structure of the sisal fibers. The projected specific surface of the fibrils and thereby absorption was enhanced with the elimination of cementing layers.

5.3 Fabric

Zwane et al [34] had produced fabric using hundred percent 3 ply unwaxed cotton with 80 tex in warp and hundred percent 2 ply sisal yarn (each having seven bundle fibers) with fibers treated in optimal NaOH concentration in weft on a portable weaving loom. The alkali treated fibers were more pliable than control (untreated), as well as the use of cotton warp yarns had added to the fabric hand. The control fabric was rated higher on treated due to its more hardness, roughness, harshness, and more open structure during a subjective assessment. Similarly, flexibility, extensibility, resiliency, and thermal characteristics of fabric from the treated fiber were found better than control fabric as being stiff, non-stretchy, not stiff, and comfortable. They had observed no significant effect of application of polyethylene cationic wax emulsion and other cationic softening finishing agents on fabric hand when compared to control fabric. The suggested potential fabric end use sequence was: macramé, improved mats, and upholstery. The softened fabric was further perceived to have potential for various applications such as carpets, bags, sun hats, etc., important to promote small scale industries.

5.4 Non-woven

Ankita et al [16] had produced non-woven using 100% sisal (softened fibers), parallel laid with 566 and 608 GSM, and cross laid non-woven fabric with 527 and 634 GSM using a manual needle punch machine. They kept 9 mm needle penetration and punch density 150 punches /cm^2 and stroke frequency of 240 stroke/min for all the four non-woven fabrics.

5.5 Composites

Chand et al [35] had developed a sisal-epoxy resin composite and evaluated its mechanical properties; tensile strength, compression strength, and impact strength. The produced sisal-epoxy resin specimen had executed low tensile strength (24KN), high compression strength (29KN), and 18 J impact strength. Calcium phosphate and hydroxyapatite coated composites can be used for both internal and external fixation on the human body for a fractured bone. Sydenstricker et al [36] had noted increased initially but decreased later on trend for the tensile strength of alkali treated (between 0.25 and 2% concentrations) sisal fibers. The best performance was observed with NaOH-treated sisal fibers compared to untreated for sisal/polyester composites. The alkali reinforcement with the polyester matrix had added to tensile strength but lowered moisture absorption. Joseph et al [37] had studied the thermal and crystallization behaviour of sisal/PP composites in relation to fiber content and fiber

treatment by thermogravimetry and differential scanning calorimetry. The treated fiber composites had shown superior properties compared to the untreated as the sisal fiber degraded before the PP matrix. This has happened as cellulose got decomposed earlier; at a temperature of 350°C compared to 398°C decomposition temperature of PP. Apart from these better thermal stability of the sisal/PP composite was realized due to better fiber matrix adhesion. The incorporation of sisal fiber in PP had also enhanced the Tc and % crystallinity values. Sundaresan et al [38] had produced composite boards with 3 ply jute yarn blended with sisal, 100% sisal, and sisal – kenaf blend. They observed that 100% sisal had higher mechanical properties compared to the other two. These Composite boards can be used as a weather shield and false ceilings. Due to higher strength, it can be used in automobiles as fuel tanks, pipelines, and structural materials of Aircraft engineering. Sreekumar et al [39] had studied the tensile and flexural behaviour of the sisal fiber-reinforced polyester composites prepared in two different ways; resin transfer method and compression moulding. Accordingly, the fabrication methods, fibers length, and fibers loading gave a remarkable impact on mechanical properties. Effect of sisal fiber loading on dynamic mechanical analysis and wear properties of jute fiber reinforced epoxy composite was investigated by varying operational parameters, viz; applied load (10–30 N), sliding speed (1–3 m/s), and sliding distance (1000–3000 m) by Gupta and Srivastava [40]. The higher values for storage modulus and glass transition temperature, whereas the minimum value for specific wear rate and coefficient of friction were observed in the case of the hybrid composite produced with Jute (50): sisal (50). Further decrease in the specific wear rate and coefficient of friction was found with alkali treated fibers. Zhong et al [41] had prepared the sisal fiber-reinforced urea-formaldehyde resin composites with varying wt % of fiber content using alkali treated sisal fibers using compression molding. They were evaluated for the effects of sisal loading on mechanical properties; impact strength, flexural strength, and wear resistance. Excellent flexural strength, water absorption, and especially wear resistance were realized with the composite produced with sisal fibre 30 percent by weight due to superior bonding. This finding was substantiated by SEM micrographs of impact fractured and worn surfaces, which had demonstrated the interfacial adhesion between fiber and matrix. Thus, the applicability of these composites in a fiberboard can be expanded. Satyanarayana et al [42] had evaluated the mechanical behavior of chopped sisal fiber-reinforced polyester composites produced by hand laying followed by compression molding. They identified that the specific modulus was near that of glass fiber-reinforced polyester composites. Sisal fiber-reinforced polyester composites have executed three times higher impact strength than a hundred percent polyester composite. The failure mechanism of longitudinally oriented sisal-epoxy composites after four-point bending was studied by Bai et al [43] and found that the damage and failure mechanisms of sisal fiber composites are controlled mainly by the microstructure of sisal fibers. They had also examined interfacial bonding and failure behavior using SEM and found that the debonding of both interfaces between tubular micro-fiber/bonding material and

fiber bundle/matrix under four-point bending tests for all fibers across the beam section. The weaker micro-fiber/ bonding material strength had led to de-cohesion of the cells and also little uncoiling of the micro-fibrils on pullout of the cells at the fast loading rate, caused rapid fracture across the beam. Venkateshwaran et al [44] had evaluated the mechanical and moisture absorption behaviours of hybrid banana-sisal-epoxy composites fabricated using different fiber lengths and different weight percentages after optimizing these parameters by pilot trials. The hybridization of natural fibers did not yield superior properties as expected like those of hybrid synthetic fiber composites. This had restricted its suggested uses for low-cost and low-load bearing applications. The enhanced water absorption, mechanical, and thermal behaviors of hybrid glass-sisal-polypropylene thermoplastic composites were realized by Jarukumjorn et al [45] due to the addition of glass fiber. Aqil et al [46] had suggested that reinforcing the plaster of paris with natural sisal fiber can improve mechanical performance, as well as can prevent sudden failure of some decorative elements due to its weight. Bisanda & Ansell [47] had reported that the incorporation of sisal fibers in an epoxy resin produced stiff and strong composite materials. According to them, the treatment of the sisal fibers with silane, preceded by mercerization, provides improved wettability, mechanical properties, and water resistance. Gopalasetty et al [48] had experimentally evaluated and compared the tensile properties of sisal nanofiber reinforced polymer composites and glass fiber reinforced polymers composites. They had fabricated the composites using the hand lay method. They had treated sisal fibers with sodium hydroxide (NaOH) and hypochlorite (NaClO) to extract the cellulose content present in the fibers. They had used a planetary ball mining machine to reduce the chemically treated fiber size to the nano level. They had observed high tensile strength of sisal nanofiber reinforced polymer composites compared to glass fiber reinforced polymer composites.

6. Applications

Traditionally sisal fiber is used by villagers in making ropes and twine, used in the agricultural field. There are three grades of sisal fiber is available. The lower quality sisal fiber is used in the paper industry to strengthen the recycled paper because of higher cellulose and hemicellulose content, the medium quality is used in the cordage industry for making ropes and twines; which are widely used in the marine and agricultural industry. The Higher grade fibers are used after treatment to convert it into yarn and used by the carpet industry for making carpets, rugs, and bags as shown in figure 4 [3, 8, 11-12, 35]. Sisal composites possess acoustic and thermal insulation properties and based on these, researchers suggested packaging, automobile, construction, and aerospace as their preferable end use criteria. It can be used effectively as internal engine covers, car interiors like seatbacks, door panels, hat racks, etc., and sisal fiber based cement reinforcement plaster for waterproof low cost

roofing. In geotextiles; sisal is used to stabilize the slope, construct road and land reclamation [3, 5, 11, 18, 35].

Unconventional use of the sisal includes dartboards, cat scratching pads made out of it due to higher strength and sustainability against fungal attacks [12, 21]. Due to porosity in its structure, can also be preferred for purification and filtration medium in water purifier, cigarette, tea bags, etc., no doubt till date no such work has been reported. Extraction of sisal generates mainly organic wastes and leaf residue. During the decortication process; sisal juice, fragments of leaves, and crushed particles of parenchymatous tissue (thin cellulose wall)are produced as waste. Biomass left after removal of fibers can be used to generate biogas, can also be used as food for animal and ecological housing materials and fertilizer, etc. Pharmaceutical ingredients like inulin, hecogenin (synthesis of steroid hormones), and inulin can be made from the juice. Intercropping can also be done safely in the initial year of plantation of sisal [1-2, 8, 11, 21].

7. Summary

In the present work, the entire zest of research work has been presented category wise; fiber, yarn, fabric, and non-woven and thereby following conclusions are drawn:

The highly variable natural, strong, minor sisal fiber and its cultivation, fiber extraction, processing, making yarn, and some special product developments are proven sources of employment opportunities and source of income, especially where this plant is grown in abundance. The extent of research and development activities on sisal plants, fibers and connected with engineering applications such as non-woven, composites, automotive, etc. indicates that craftsmen and women can sustain with the income generation through the sisal related activities thus making it a feasible option for rural empowerment.

References

[1] Information from https://en.wikipedia.org/wiki/Sisal_March2020 accessed on 12/09/2019.

[2] Information from https://nptel.ac.in/courses/116102026/_March2020accessed on 12/09/2019.

[3] Mohini Saxena, Ruhi Haque, P.Asokan, Anusha Sharma, Sisal fiber-based polymer composites and their application chapter 22, Cellulose Fibers: Bio- and Nano-Polymer Composites, Springer-Verlag Berlin Heidelberg 2011.

[4] Joseph K, Tolêdo Filho RD, James B, Thomas S, Carvalho LH. A review on sisal fiber reinforced polymer composites. Rev Bras Eng Agrícola Ambient 1999; 3(3):367-379.

[5] Naveen J, M Jawaid, P Amuthakkannan, M Chandrasekar, Mechanical and physical properties of sisal and hybrid sisal fiber-reinforced polymer composites, Woodhead Publishing series in composites science and engineering 2019, 427-440.

[6] J. Gordon Cook, Handbook of Textile Fibers, Vol. 1-Natural Fibers, 2001, Woodhead Publishing Limited

[7] Yan Li, Yiu Wing Mai, Lin Ye, sisal fiber and its composites: a review of recent developments, composites science and technology 60 (2000) 2037-2055.

[8] Lakshmikant Nayak, D. Nag, S. Das, Deb Prasad Ray, Lakshmanan Ammayappan, Utilization of Sisal Fiber (Agave Sisalana L.)-A Review, Agricultural Research Communication Center 2011

[9] Abderrezak Bezazi, Ahmad Belaadi, Mostefa Bourchak, Fabrizio Scarpa, Katarzyna Boda, Novel extraction techniques, Chemical and Mechanical characterization of Agave americana l. natural fibers, composites: part B, 2014, ELSEVIER

[10] P E Zwane, A M Dlamini, N Nkambule, Antimicrobial properties of sisal used as an ingredient in Petroleum Jelly production in Switzerland, Current Research Journal of Biological Sciences 2(6):370-374, 2010

[11] P Sahu and MK Gupta, Sisal fiber and its polymer based composites: A review on current developments, Journal of Reinforced Plastics & Composites 2017, Vol. 36 (24) 1759-1780.

[12] Information from http://www.fao.org/economic/futurefibres/fibres/sisal/en/_March2020 accessed on 14/09/2019.

[13] S.L. Bai, R.K.Y. Li, Y.W. Mai, C.M.L. Wu, Morphological Study of sisal fiber, advanced composites letters, Vol. 11, No. 3, 2002

[14] S.J Eichhorn, J.W.S. Hearle, M. Jaffe and T. Kikutani, Handbook of textile fiber structure, Volume 2: Natural, regenerated, inorganic and specialist fibers.

[15] Information from http://textilefashionstudy.com/chemical-composition-of-sisal-fiber-cellulosic-fiber/_March2020 accessed on 14/09/2019.

[16] Ankita Shroff, Anjali Karolia, Jayendra Shah, Enzyme softening of sisal fiber: A Sustainable Approach for the future, Indian Journal of Applied Research-Vol. 5

[17] Ashish Hulle, Pradyumkumar Kadole, Pooja Katkar, Agave Americana leaf fibers, Fibers 2015, 3, 64-75, MDPI

[18] M Ramesh, K Palanikumar, K Hemachandra Reddy, Comparative evaluation on properties of hybrid glass-sisal/jute Reinforced epoxy composites, sciverse science direct processing Engineering 51 (2013) 745-750.

[19] Tribology of natural fiber composites, chapter 3 sisal reinforced polymer composites 2008, pages 84-107.

[20] Properties of plant fiber yarn polymer composites-An Experimental study-report-Technical university of Denmark-2004, pg 20.

[21] https://textilelearner.blogspot.com/2013/01/sisal-fiber-properties-of-sisal-fiber.html

[22] Robert R. Frank, Bast and other plant fibers, Wood Head Publishing Limited, 2005

[23] Satyanarayana KG, Sukumaran K, Mukherjee P S, Pavithran C, Pillai SG. Natural fiber-polymer composites. Cement & Concrete Composites; p. 117-136, 1990

[24] Mwaikambo LY, Ansell M P, mechanical properties of alkali treated plant fibers and their potential as reinforcement materials II. Sisal fibers. J MATER SCI 41, p. 2497-2508, 2006

[25] Maria Alice Martins, Pedro Kunihiro Kiyohara, Ines Joekes, Scanning Electron Microscopy Study of Raw and chemically modified sisal fibers, Journal of Applied Polymer Science, Vol. 94, 2333-2340, 2004, Wiley Periodicals Inc.

[26] Okasman K, Lennart W, Morphology and mechanical properties of unidirectional sisal-Epoxy composites, Journal of Applied Polymer Science, Vol. 84, p. 2358-2365, 2002

[27] Adriana R. Martin, Maria Alice Martins, Odilon R.R.F. da sila, Luiz H.C. Mattoso, studies on thermal properties of sisal fiber and its constituents, Thermochimica Acta 506, p. 14-19, 2010

[28] Yang GC, Zeng HM, Zhang WB, Thermal treatment and thermal behaviour of sisal fiber. Cellulose Science and Technology; 3: p. 15-19, 1995

[29] Mishra S, Mohanty AK, Drzal LT, Studies on the mechanical performance of biofiber/glass reinforced polyester hybrid composites. Compos Sci Technol 2003; 63: 1377–1385.

[30] Chand N, Verma S, Khazanchi AC. SEM and strength characteristics of acetylated sisal Fibre, Journal of Materials Science Letters; 8: 1307-1316, 1989

[31] Yang GC, Zeng HM, Li JJ, Jian NB, Zhang WB. Relation of modification and tensile properties of sisal fiber. Acta Scientiarum Naturalium Universitatis Sunyatseni; 35: p. 53-7, 1996

[32] Kalia S, Kaushik VK and Sharma RK Effect of benzoylation and graft copolymerization on morphology, thermal stability, and crystallinity of sisal fibers. J Natural Fiber 2011; 8: 27–38.

[33] Hajih H, Sain M and Mei LH Modification and characterization of hemp and sisal fibers. J Natural Fiber; 11: p. 144–168, 2014

[34] Pinkie E. Zwane, Rinn M. Cloud, Development of fabric using chemically treated sisal fibers, AUTEX Research Journal, Vol. 6, No. 2, June 2006

[35] Chand Badshah S B V J, K Arun, B Eswaraiah, Fabrication and Testing of Natural (Sisal) Fiber Reinforced Polymer Composites Material, International Journal of Emerging Trends in Engineering Research, Volume 3, No. 5, May 2015

[36] Sydenstricker TH, Mochnaz S and Amico SC. Pull-out and other evaluations in sisal-reinforced polyester biocomposites. Polym Test; 22: p. 375–380, 2003

[37] Joseph PV, Joseph K, Thomas S, et al. The thermal and crystallization studies of short sisal fiber reinforced polypropylene composites. Compos Part A: Appl Sci Manufact; 34: p. 253–266, 2003

[38] S Sundaresan, R Priyanka, D Lakshmipriya, S Malaiyappasamy, Detailed investigation of properties of Sisal-jute-kenaf yarn composites, IJARIIE (2395-4396) Vol. 2, 2016

[39] Sreekumar PA, Thomas SP, marc Saiter J, et al. Effect of fiber surface modification on the mechanical and water absorption characteristics of sisal/polyester composites fabricated by resin transfer molding. Compos Part A: Appl Sci Manufac; 40: p. 1777-1784, 2009

[40] Gupta MK and Srivastava RK, Tribological and dynamic mechanical analysis of epoxy based hybrid sisal/jute composite. Indian J Eng & Material Science; 23: p. 37–44, 2016

[41] Zhong JB, Lv J and Wei C. Mechanical properties of sisal fiber reinforced urea-formaldehyde resin composites. Express Polym Lett; 1: p. 681–687, 2007

[42] Mukherjee P, Satyanarayana K, Structure and properties of some vegetable fibers. J Mater Sci; 19(12): p. 3925-3934, 1984

[43] Bai S, Wu CM, Mai Y, Zeng H, Li RK, Failure mechanisms of sisal fibers in composites. Adv Compos Lett; 8(1): 13-7, 1999

[44] Venkateshwaran N, ElayaPerumal A, Alavudeen A, Thiruchitrambalam M. Mechanical and water absorption behaviour of banana/sisal reinforced hybrid composites. Mater Des; 32(7): p. 4017-4021, 2011.

[45] Jarukumjorn K, Suppakarn N. Effect of glass fiber hybridization on properties of sisal fiber polypropylene composites. Compos B Eng, 40(7): p. 623-627, 2009.

[46] Aquil M. Almusawi, Thulfiquar S. Hussain and Muhaned A. Shallal, Effect of temperature and sisal fiber, content on the properties of Plaster of Paris, International Journal of Engineering and Technology, p. 205-208, 2018

[47] Bisanda E.T.N., Ansell M.P., The effect of silane treatment on the mechanical and physical properties of sisal-epoxy composites, composites science and technology, Oxford, Vol. 41, p. 165-178, 1991

[48] Gopalasetty Saranya, Mr. M. Karteek Naidu, Mr. P. Kripa Rao, Characterization and Synthesis of nano sisal fiber reinforced composites, International Journal of Engineering Sciences & Research Technology, Vol. 7(12), 2018.

Table 1: Chemical compositions of sisal fiber

Cellulose (%)	43-88
Hemicelluloses (%)	10-24
Lignin (%)	4-20
Pectin (%)	0.8-2.3
Wax (%)	0.15-0.5

Table 2: Properties of sisal fiber

Characteristics	**Measures**
Length (m)	0.5-1.5
Diameter (μm)	100-300
Moisture Content (%)	10-22
Micro fibril angle (θ°)	10-25
Density (g/cc)	1.03-1.5
Tensile strength (gpd)	4.52-5.53
Elongation at break (%)	2-14
Stiffness (KN/mm)	30-38

Table 3: Tensile properties of treated sisal fibers by Yang et al[33]

Sr. No.	Treatment methods	Tensile strength (g/tex)	Elongation at break (%)
1	Untreated	30.7	2.5
2	Benzol/alcohol	38.8	3.7
3	Acetic acid + alkali	9.3	2.6
4	Alkali	31.7	7.5
5	Acetylated	33.2	8.3
6	Thermal	42.0	3.5
7	Alkali thermal	27.6	4.7
8	Thermal-alkali	25.7	4.4

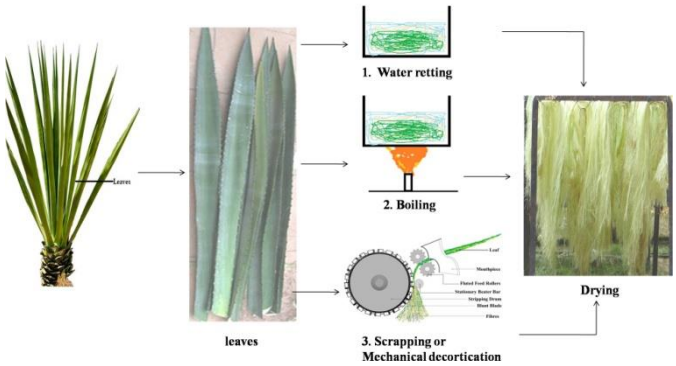


Figure 1: Extraction of sisal fibers

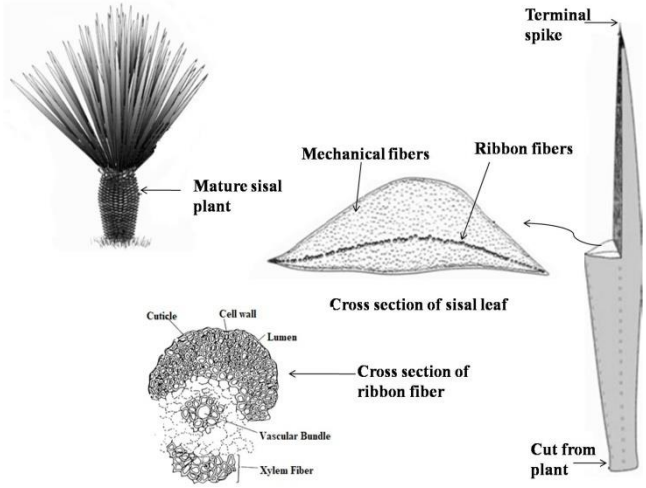


Figure 2: Longitudinal and Cross-section of sisal leaf and fiber

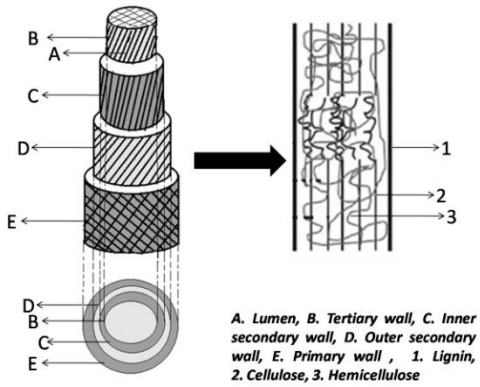


Figure 3: Sisal fiber cell

Figure 4: Application of sisal

A SHORT COMMUNICATION OF E-LEARNING AND ITS CHALLENGES IN INDIA

T. Sumallika*[1], P.V.M. Raju[2]

[1]Department of Information Technology, Gudlavalleru Engineering College, Gudlavalleru

[2]Department of Business and Management Studies, Gudlavalleru Engineering College, Gudlavalleru

*corresponding author email: sumallika.p@gmail.com

Abstract

Nowadays, technology is emerging at a greater pace, with the widespread diffusion of technology, knowledge access becomes easier and remote learning has become a common practice. It enhances the individual's learning ability also improves skills in different dimensions in new subject areas. The best part is, it made individuals acquire knowledge without meeting the teacher physically. One of the prominent modes of distance learning is E-Learning. Online Education has to go long way with rapid technological advancement. India is also one of the countries that evolving around advanced technologies like many other nations. In India, with a heavy population i.e.; 1.3 billion, and the adoption of smart technology is happening at an exponential pace. Especially with the usage of smart mobiles and with 4G and 5G technologies, high-speed internet, and other latest electronic gadgets. People are becoming more and more technology-oriented day by day. The emergence of the World Wide Web has influenced the lifestyle of the people in many ways in India. Ex. doing online shopping, social networks, online food deliveries, online money transactions, and online learning, etc. Even though e-commerce is the most influential industry online, E-Learning is right next to it. In this present scenario, this paper through an adequate light on the overview of e-learning and the challenges facing and future trends in India

Keywords: e-learning, challenges, Technology, Blended classroom

1. Introduction

The rapid diffusion of technology has created a greater influence on the educational system. In the recent past, students have become more technology-oriented learners and relying more on advancedinternet technologies and networks for their better learning experience. Thus the modern educational tools and methodologies have taken a new leap against traditional learning approaches (paper–based).

E-learning is nothing but using advanced electronic technology for accessing various educational curriculums and knowledge content without having any traditional classroom setup.In this ICT (Information and Communication Technology) Era, People are fascinated to learn through online courses with growing knowledge of internet technologies.

In General, E-learning can be defined as the use of information technology connected through the internet for learning. It comprises of

a) Accessing and sharing the learning material.

b) Speaking with the tutors (teachers) and fellow students.

c) Seeking assistancewhile learning process.

d) Obtaining greater knowledge and gaining better experience.

By witnessing greater success in various technologies in India, Our ambitious Prime Minister (PM) intends to transform India by introducing digital technologies through the Digital India Initiative that provide greater opportunities for all kinds of people in the country. This Programme majorly focusses on important sectors like Health, Education, Employment, etc. Majority of the colleges and Universities conducting online classes for being the part of modern India initiate. Which is an indication of acceptanceof the new technologies in the education system.

2. Research Methodology

The present paper is a conceptual study on E-Learning. Thus adopted a qualitative research strategy. 1(Saunders, M, et al 2003), He explained that in conceptual research, it is essential to do a literature survey on the selected topic thereby analyze in-depth to conclude. In researching of qualitative approach, it is always possible to make the necessary changes subjected to the market and integrate the survey. (Ader et al., 2008), He says that, in qualitative research, no independent and dependent variables are involved, it is experimental in nature, Therefore the present study based completely on the qualitative method.

3. Objectives of the Study

i. To analyze various components of E-learning.

ii. To examine the various challenges of E-learning in India.

iii. To evaluate the future trends of E-learning.

4. Online Education Market in India

- In 2016, the online education market in India has a worth of $0.25 billion, and by 2021 it is going to become $1.96 billion withacompound annual growth rate of 52 percent.
- In 2016, 1.6 million users enrolled in different online learning courses. And is estimated to grow up to 9.6 million by the end of the year 2021
- The Classroom education cost is 175% higher is estimated,so online education is more cost-effective than traditional classroom cost.
- In India, 48 percent of target customers are of the 15- 40 age group with higher aspirations and expectations but with lesser income. Noticed that online channel acceptability in the younger demographic is extremely high.
- Below figure 1. Represents India's online education market

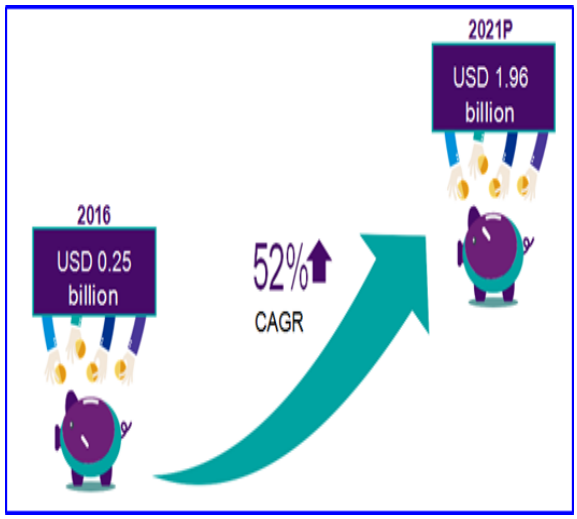


Fig. 1 Online Education Market in India

Source: KPMG in India's research and analysis 2017

The above-said facts reveal that the future potential of online education in India. It has both advantages and disadvantages over traditional education.

The below table: 1 shows advantages and disadvantages of online learning (E-learning)

5. Advantages/Disadvantages of E-learning

Advantages	Disadvantages
Learn from anywhere, at any time	Chances of distraction are very high
Save Money and Time	FraudulentOnline courses
Learn at your own pace	Cannot do courses that require Labs/Workshops
Recognition of online degrees	Lack of transformational power

Table: 1 Advantages /Disadvantages of E-learning

6. An Overview of E-learning

In a nutshell, the E-learning concept must be understood in different ways, one must analyze various definitions with different perspectives, Characteristics that can describean e-learning course/module/program, different types of e-learning methods at different levels, and the teaching approaches adopted for instruction to extract the best performance from both the ends.

The below, picture: 2 represents an overview of E-learning

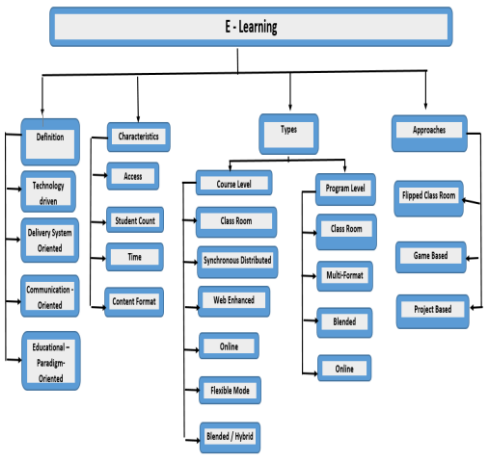


Fig. 2 An Overview of E-learning

A. DEFINITIONS

• **Technology-Driven Definitions:** As per the Technology Dimension, E-learning is all about delivering the content and conducting teaching and learning Programmes by using appropriate technology.

• **Delivery-System-Oriented Definitions:** This type majorly focuses upon the usage patterns of the services rather than the achievement of the outcome. Itdefines E-learning as a process of educational content delivery through an electronic system that includes all activities related to teaching and learning.

• **Communication-oriented Definitions:** This concept includes the usage of various communication and interactive tools for the effective transfer of information between the two parties. It defines as pedagogical interaction between instructors and students, students and students among the students through the World Wide Web by communication tools.

• **Educational-Paradigm-Oriented Definitions:** This views e-learning as a new side of education, it focuses on new developments and innovative breakthroughs in the existing educational system. In this perspective, E-learning is defined as the usage of smart technologies to facilitate remote access to resources and services for better collaboration and exchange of ideas.

B. CHARACTERISTICS

Some characteristics can describe an e-learning course/module/program, which is listed below

• **Access:** This characteristic feature focuses on how students can access and use the program/Course material. They are of two types 1) Online: By using smartphones, laptops, and tablets connected through the internet 2) Offline: By using hard disks, Pen drives DVDs; CDs Learners/students can access various offline content. Furthermore, this mode is more effective as it is free from all net connectivity distractions as sometimes they landed up with unrelated sites**.**

• **Student Count:** This function check whether proper communication established among the students is not?

i) Individual: Interaction between Students/Learners will not be established for interaction, Learner will interact with the tutor, and the task will also be completed on a one-one basis only.

(ii) Group: A Group of learners/Students will communicate among themselves through discussion boards, open forums, e-mails, and chat rooms. Peer interaction is prominent between the learners, whenever they feel the Programme /module /course doesn't suit them, they may drop.

- **Time:** It defines the actual time of information accessed by the students.

i) Synchronous: In this process, the content will be delivered in real-time to the students through virtual platforms like teleconferencing and conference calls.The main limitation of this platform is differences in time zones because the learners are geographically scattered across the world.

ii) Asynchronous: In this method, accessibility of the content has no time restriction, it can be accessed at any time .open forums and discussion boards are the main means for communication. It is helpful to get rid of the incompatibility of different time zones.

- **Content Format:** It defines which type of data and its format is being delivered.

i) Static: It means that the delivered content is through Programme/course/module is the same all the time. Here the continuity of the content can be well maintained. But still few learners /students withdraw as their needs may not be fulfilled through it.

ii) Dynamic: In this process, it is possible to respond more specifically to student/learner requirement, this overcomes the problem static nature, data or information can be modified and delivered to the respondents as and when it is required effectively, But it poses a challenge in creating new content every time and to sustain it for the long run.

C. TYPES

There are various means to segregate E-learning courses based on the level of the course.

- **Classroom Course:** This is a regular way of teaching using computers and simulation designs t in a traditional classroom.

- **Synchronous Distributed Course:** Programme/modules are been taught in the conventional class and streamed through online-based conferences to the learners off-campus.

- **Web-enhanced Course:** This sort of Courses are mostly executed by using online tools, mostly face-face interaction is initiated.

• **Blended/Hybrid Course**: This is a mixture of both online and offline instruction, these will be initiated when the students are from far places not in a position to attend face –to face. There are two types of blended/hybrid courses are given below.

i) Blended Classroom Course: Large portion of the course is conducted through a conventional classroom.

ii) Blended Online Course: a major portion of the course will be conducted online.

• **Online Course:** All requirements will be done to conduct learning sessions online, face-to-face interactions are not be encouraged, those students who have the difficulty in attending the classes physically can take part in it.

D. APPROACHES

Three different teaching approaches adopted mostly in the online learning process, they are briefly discussed below.

• **Flipped Classroom:** This is an innovative approach to execute, where all assignments will be done in the classrooms and all pre-recorded videos and learning material can be made available to the students at home to watch before they come to the class, videos are not only be watched at home but they will be discussed in the class said by Jonathan Bergmann, the founder of this approach, Also said, this will be more useful for the slow learners by completing the assignments in the class with discussing with others, this is a blended method of teaching.

• **Game-based:** The name itself is indicating that it is based on games, students work towards a specified goal in the form of a game. This allows them to experiment and gain points of achievement along the way. This type of learning method helps in improving cognitive and psychomotor processes. It is a useful online teaching mode without interacting with the teacher.

• **Project-based:** This approach helps students to work on a project by exploring real-world challenges. It ensures students solve issues by developing innovative solutions. This, in turn, enhances creativity and practical orientation in comparison with traditional approaches. It is also helpful in both blended courses and online courses. In this approach, instructors play the role of facilitators that guide the learners in their projects.

7. Key Challenges for Online Education in India

Even though online education has large growth potential in India, but it has many future challenges. For instance, a new education policy has posed many challenges to the field of online education. The major challenges are mentioned below.

• In India's population, more than 30% of people are not computer literates, they don't even how a computer can be started.

• Most of the Indian Citizens belong to the communities like farmers, cleaners, sweepers, housemaids, waiters, etc., their financial position may not support them to have a computer or a laptop.

• Teacher's familiarity is also one of the challenges for the new format of online education. They are not trained, some teachers are not familiar with the new education format. Furthermore there no guarantee that a good traditional class teacher to be a good online tutor.

• Nonavailability of required resources for conduction of electronic-based reviews is also a problem, moreover, the question pattern of the exam and number of questions to be asked is also a challenging task.

• It is difficult to teach some practical-oriented subjects through digital education because it involves performing arts and experimentation using chemicals and other machine tools and equipment.

• It is basically a screen-based learning system that many times may not encourages students to practice it.

• Internet connectivity is not proper across the country, there are some people in villages still struggling with 2G or 3G even these days.

• Students must be self-disciplined and well-focused, especially in the online learning process, a survey says that below 17 years of age group people lagging in these skills.

8. Future Trends of E-learning

In the nearest future, India is going to witness new trends in the E-learning market, they are as follows:

i) Hybrid Model: There will be an amalgamation of both offline and online education designs. There are certain additional education activities like e-tuitions, internship programs, after-school coaching,and live projects that will be organizedonline; some of

them also reachthe students through offline touchpoints like labs and community meeting halls. Virtual classrooms and offline teaching pedagogy bothwill help the students to be more interactive and gain practical knowledge about the subject and helps in learning soft skills.

(ii) Addition of new and offbeat subjects: Along with the regular subjects like digital marketing, cloud computing, and data science online education curriculum also offers peculiar subjects like cyber law, culinary management, and forensics, etc.

(iii) Gamification: For making learning activities more effective, engaging, competitive, and rewarding for both professionals and students, the courses must add different features like discounts, badges, and leader boards. Educational institutes and corporates must coordinate with each other to develop learning content.

(iv) Peer-to-peer learning and profile mapping: E-learning enhances better peer-to-peer learning and develops collaborative learning among the students, it helps to share their ideas and exchange notes, and also post their comments on a common forum. Emerged technologies like artificial intelligence (AI), data mining, big data, and facial recognition, etc. are vividly used to teach personalized profile-based courses.

(v) Investor interest will grow: In India, in the last three years, a Considerable number of big deals have been taken place in the E-learning market. For instance, Chan Zuckerberg has invested fifty million dollars in Byju's; Bertelsmann India has invested 8.2 million dollars in Eruditus, and Kaizen Management Advisors and DeVry Inc. have invested 10 million dollars in EduPristine. The Khan Academy, which has earned billions of rupees from the Bill and Melinda Gates Foundation, Google, and Netflix founder Reed Hestings, also exists in the list. This E-learning sector will therefore be creating more enthusiasm in the potential investors to attract funds.

9. Conclusion

In developing countries such as India, the face of education has changed with the emergence of e-learning, made education available even to remote places thereby the literacy rate has drastically increased day by day which further leads to economic growth. This is the fact in the case of nations where technical education is costlier, economic inequalities exist, and scarcity of resources. For such countries E-learning industry is becoming a sunrise industry. It may not totally replace traditional learning, but a hybrid model (both online and offline) will become popular in the coming days. Undoubtedly quality content, distribution, and access (three-fold process) are going to be the better design for e-education.The prices have come down to a greater extent for the network access through various electronic gadgets so that the students can access the information or subject content remotely and explore worldwide opportunities to seek better careers, thanks to satellite technology. Even though the Indian Online

market is at the nascent stage it has to go a long way by carefully focusing on require technology and infrastructure that further improves opportunities in various fields of business.

In the present Business scenario,the online education market is booming with many online platforms like Swayam, Edureka, Coursera, and Udemy, etc, and will positively record exponential growth in the nearest future. However, it cannot be agreed by most people that it is a total replacement for traditional classroom education. So a model which comprises of both modern and conventional learning system will gain momentum in the coming days. More digital learning apps will take birth to cater to the needs of learners, In the future, Digital classrooms will gain their importance by systematically engaging conventional classroom experience. To meet the future requirements of education, the Education Industry must adopt hybrid models and virtual classrooms that are more viable for better implementation of online education.

References

[1] Adèr, H. J., Mellenbergh G. J., & Hand, D. J. (2008). Advising on research methods: A consultant's companion 9

[2] A. Sangrà, D. Vlachopoulos, and N. Cabrera, ‘‘Building an inclusive definition of E-learning: An approach to the conceptual framework,’’ Int. Rev. Res. Open Distrib. Learn., Vol. 13, No. 2, pp. 145–159, 2012. [Online]. Available: http://www.irrodl.org/index.php/irrodl/article/view/1161

[3] B. Tucker, ‘‘The flipped classroom,’’ Educ. Next, Vol. 12, No. 1, pp. 1–6, 2012.

[4] F. Mayadas, G. Miller, and J. Sener, ‘‘Definitions of E-learning courses and programs, developed for discussion within the online learning community,’’ Online Learning Consortium, Newburyport, MA, USA, Tech. Rep. Version 2.0, Apr. 2015.

[6] J. S. Krajcik and P. C. Blumenfeld, Project-Based Learning. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[7] Saunders, M., Lewis, P., & Thornhill, A. (2003) Research method for business students, 3rd edition. New York: Prentice-Hall. 10. P

[8] S. Tobias, J. D. Fletcher, and A. P. Wind, ‘‘Game-based learning,’’ in Handbook of Research on Educational Communications and Technology. Springer, 2014, pp. 485–503.

[9] https://cms.iamai.in/Content/ResearchPapers/d3654bcc-002f-4fc7-ab39-e1fbeb00005d.pdf

[10] https://assets.kpmg/content/dam/kpmg/in/pdf/2017/05/Online-Education-in-India-2021.pdf

[11] https://www.ugc.ac.in/pdfnews/7553683_Online-Courses-or-ProgrammesRegulations_2018.pdf

[12] https://www.india.com/education/government-announces-12-new-swayam-prabha-dth-channels-all-you-need-to-know-about-this-initiative-4031468/

[13] https://www.livemint.com/companies/start-ups/pm-e-vidya-initiative-to-bring-more-opportunities-for-ed-tech-startups-11589774602446.html

[14] https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/ed-tech-firms-go-on-hiring-spree-look-at-creating-3000-new-jobs-next-year/articleshow/74560293.cms?from=mdr

---

NV-LDA: A NOVEL APPROACH TO CLASSIFY THE EMAIL CONTENT USING TOPIC MODELING

Namrata Shroff[1*], Amisha Shingala[2]
[1]Gujarat Technological University, Ahmedabad,Gujarat
[2]Sardar Vallabhbhai Patel Institute of Technology,Vasad,Gujarat
*corresponding author email: nam.shroff@gmail.com

Abstract

In our day to day life, due to over increase emails in our inbox many of the important emails remain unattended. The challenge in Email Classification is to derive the subject intent words from limited number of text and to extract the relevant words. It calls the document clustering so as to create the most representative content. This paper proposes an email classification system that can cluster emails into the meaningful class. Emails with similar content will fall under one cluster. The proposed system extracts representative keywords from the subject and body of each email and topics by Latent Dirichlet allocation (LDA) scheme. The keywords extracted from LDA are compared with knowledge corpus and then appropriate label is predicted for the emails. Term clustering and its relationship with document clustering is considered for interactive expansion of knowledge base. The similar occurring emails are kept in repository which can be further labeled. NV-LDA extracts keywords from subject and body of the emails, maximum term frequency is calculated. The output of LDA is compared with the corpus and the labels are predicted. The keywords are extracted by three methods LDA, TF-IDF and LDA combined and NV-LDA methods for clustering. It is found that using NV-LDA approach, the results are encouraging.

Keywords: LDA, TF-IDF, email classification, Term clustering, Document classification.

I. Introduction

Numerous emails hit our inbox daily with the continuous advancement of digital communications, this makes it difficult for users to search and classify emails that are of interest or important to them on specific topics. Therefore, it is expected to systematically classify these mass emails into similar content so that users can easily and conveniently find the emails that interest them. Typically, for finding important emails on specific topics or subjects user need to classify the emails into appropriate labels. Gmail provides the promotion and social tabs in which the emails are classified accordingly. For the users over time, these important emails, which have been increasing in number, are difficult to manage and process effectively. Because the relationship between important emails to be analyzed and classified is very

complicated, it is difficult to quickly understand the subject and body of each email and it is difficult to accurately classify emails with similar topics in content. Therefore, it is necessary to use automated processing methods for such a large number of emails in order to classify them quickly and accurately.

So here we propose an algorithm in which the important email is classified and a label is assigned to the email so that user will notice that this email is important and he/she will not miss out reading an important email. To classify important emails with similar subjects, we propose the email classification system based on term frequency-inverse document frequency (TF-IDF) and Noun Verb-Latent Dirichlet allocation (NV-LDA) schemes. The proposed system firstly analyzes the top senders from where the emails are coming, also the read emails and threaded emails are considered. Secondly, it uses the TF-IDF scheme to extract subject words from the subject and body of email corpus. Thirdly constructs a representative keyword dictionary with the keywords that are entered by user, and with the topics extracted by the NV-LDA. Then, comparison of the output of NV-LDA and knowledge corpus is done to classify the emails with similar contents. To extract subject and body words from a set of massive emails efficiently, in this paper, we use the Python Gensim that can process data rapidly and stably with high scalability. Furthermore, in order to demonstrate the effectiveness and applicability of the proposed system, this paper evaluates the performance of the proposed system, based on actual email data. As the experimental data for performance evaluation, we use the email corpus of my personal mailbox. The experimental results indicate that the proposed system can well classify the whole email corpus with emails with similar subjects according to the relationship of the keywords extracted from the subject and body of email. The remainder of the paper is organized as follows: In "Related work" section, we provide related work on email classification. "System flow diagram" section presents a system flow diagram for our email classification system. "Email classification system" section explains the email classification system based on TF-IDF and NV-LDA schemes in detail. In "Experimental Result" section, we carry out experiments to evaluate the performance of the proposed email classification system. Finally, "Conclusion" section concludes the paper.

II. Related Work

This section briefly reviews the literature on email classification methods. Email classification is being explored since long. [15] Research in this area is carried out using temporal feature analysis, graph theory, machine learning algorithm [16], neural network algorithm and topic modeling etc. [14] Some of the researchers did email classification with Latent Dirichlet allocation, a powerful topic modeling algorithm which discover the latent topic in the corpus. LDA [6] is used in many applications like Document classification, News article classification, Opinion mining, Research paper classification and Email classification [17].

In this paper, LDA topic model is used; LDA topic model [11] is applied to email topic classification, which solves the problem of reading many emails. In view of the powerful text representation ability and dimensionality reduction effect of LDA model, the corpus is modeled and the topic classification is carried out. Then the accuracy and statistics of keyword matching are obtained by TF-IDF. Value words ratio verifies the reliability of the classification, the results show that the method has a certain effect, can be applied to mail classification. However, the problem of subjective factors in statistic value words and the trial scope of evaluation system are not considered. [1]

The authors proposed Research Paper classification system incorporates TF-IDF and LDA schemes to calculate an importance of each paper and groups the papers with similar subjects by the K-means clustering algorithm. It can thereby achieve correct classification results for users' interesting papers. [2]

In this work, author has approached the problem of automatic email foldering using supervised classification techniques. Author has proposed architecture with a multi-class classifier combination approach, by a majority voting technique. Author has analyzed the entire pipeline of the system, ranging from different techniques for feature weighting, feature selection and classification. They also studied the influence of adding participant fields in the common textual representation (body and subject). [3]

A cascaded-SOM like architecture is presented to deal with the overlapping classes present in a multi-class classification system. SOM (Self Organizing Map) at a particular level is capable of handling two classes simultaneously. In order to generate email vectors, several representation schemas including word2vec, TF-IDF are explored. A large set of features are also exploited for making the classification task more accurate. [4]

Different from the above mentioned methods, our method uses three kinds of keywords:

keywords from the knowledge corpus, keywords extracted from subject and body of email, and topics extracted by LDA scheme. These keywords are used to calculate the TF-IDF of each email, with an aim to considering an importance of email. Then, comparison of LDA topics with the knowledge corpus is done to classify the emails with similar contents, based on the TF-IDF values of each email. Meanwhile, our classification method is designed and implemented for all users, i.e. our system will work in dynamic environment giving the recommendation to user regarding the labels they can be applied in their inbox so that the important mail is attended. To our best knowledge, our work is the first to use the analysis of email subject and body based on TF-IDF and NV-LDA schemes for email classification.

III. Framework for NV-LDA system

The email classification system proposed in this paper consists of four main processes (Figure 1): (1) Preprocessing, (2) Data Management and Topic Modeling, (3) Comparison with knowledge corpus, and (4) Prediction of label. This section describes a system flow diagram for our email classification system.

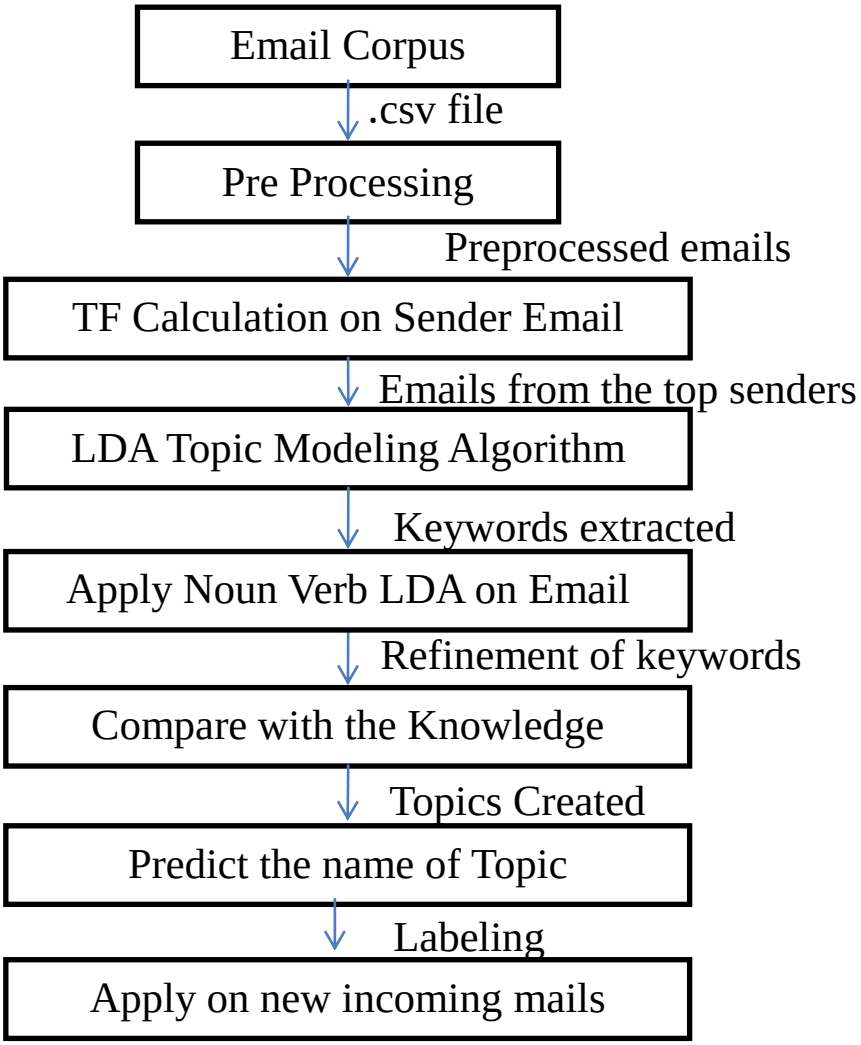


Figure 1: Email Classification system flow diagram

Detailed flows for the system flow diagram shown in Figure 1 are as follows:

Step 1: It automatically collects keywords from email subject and body during a given period. It also executes preprocessing for these data, such as the removal of stop words, the extraction of only nouns and verbs.

Step 2: It constructs a keyword dictionary based on crawled keywords. Because total keywords of whole emails are huge, this paper uses only top-N keywords with high frequency among the whole keywords

Step 3: It extracts topics from the crawled emails containing nouns and verbs by LDA topic modeling

Step 4: It compares the keywords fetch by NV-LDA and keyword dictionary

Step 5: It classifies the emails into appropriate labels.

Step 6: It groups the whole inbox into emails with a similar subject, based on the step 4.

In the next section, we provide a detailed description for the above mentioned steps.

IV. Algorithm for NV-LDA system

I. Preprocessing of Emails

The sender from where the email has arrived is one of important parts in an email as it describes the gist of the email. Typically, email subject, the next most part of emails that users are likely to read and then decides whether to read the email or ignore email. That is, users tend to read firstly a sender address and email subject in order to catch the priority of an email. Because of this, this paper classifies similar emails that are important based on sender's address and subject fast and correct.

As you can see in preprocessing step of Figure 1, the emails are imported and converted into .csv file. It also removes stop words in the email subject and body and then extracts only nouns and verbs from the data. After the preprocessing (i.e., the removal of stop words and the extraction of only Nouns and verbs), the amount of email data should be greatly reduced. This will result to enhancing the processing efficiency of the proposed email classification system.

II. Data Management and Topic Modeling

In order to process lots of keywords simply and efficiently, this paper categorizes several keywords with similar meanings into one representative topic. In this paper, we

construct representative keywords from total keywords of all emails and make a keyword dictionary of these representative keywords. However, even these representative keywords cause much computational time if they are used for email classification without a way of reducing computation. To alleviate this suffering, we use the keyword sets of top frequency 5 among these representative keywords [18].

Topic Modeling: Topic modeling [12] is a type of statistical modeling for discovering the abstract topics that occur in a collection of documents. Latent Dirichlet Allocation (LDA) is an example of topic model and is used to classify text in a document to a particular topic. It is an unsupervised algorithm used to spot the semantic relationship between words a group with the help of associated indicators [9]. When a document needs modeling by LDA [13], the following steps are carried out initially:

- The number of words in the document are determined.
- A topic mixture for the document over a fixed set of topics is chosen.
- A topic is selected based on the document's multinomial distribution.
- Now a word is picked based on the topic's multinomial distribution.

In this paper, after extracting nouns and verbs from the email corpus using LDA scheme, we extract topic sets from the emails.

III. Comparison with Knowledge Corpus

The knowledge corpus is collection of similar keywords grouped together under one topic. The keywords extracted from above process are match with the keywords present in knowledge corpus and the appropriate topic is assigned to it. E.g. if the email contains the keyword like payment, bill, transaction, finance etc. then that email will be labeled as finance.

IV. Predicting the Label

After above process the predicted label will be assigned to the incoming emails. So that the user come to know about the type of email just by looking the label and not looking at senders address and subject. So the important emails will be attended.

V. Experimental Results

In this section, we present our experimental setup in Section I, discuss the optimal parameters of the proposed method in Section II discuss the counter methods used Section III discuss the effectiveness analysis of topic keywords respectively.

I. Experimental Setup

DATASETS: Our personal mail box emails over 7000 in numbers are considered, which have been imported into CSV file, making it convenient to use in Python. We implemented the NV-LDA calculation module using nltk library [7]. The experimental environment is shown in Table 1.

Table 1: Experimental Environment

CPU	Intel ® Core™ i5-7200U
Memory	4 GB
Programming Language	Python Language
Edition	Pyhton 3.6.5
IDE	Jupyter Notebook

II. Counter Methods

We have compared the proposed NV-LDA with the following methods for clustering emails. The keyword dictionaries used for performance evaluation in this paper are constructed with the three methods shown in Table 2.

Table 2: Three methods to construct keywords

Method 1	Using only LDA
Method 2	Using TFIDF and LDA
Method 3	Using NV-LDA

We have compared the proposed NV-LDA with the following methods for email clustering.

Method 1: LDA [5] is a topic model based on assumptions that documents are generated by mixture latent topics in the corpus. Each document can be represented with a topic distribution vector which is inferred based on the words occurrence patterns.

Method 2: TF-IDF [8] can extract only the frequently occurring keywords in emails and LDA can extract only the topics which are latent in emails. On the other hand, the combination of TF-IDF and LDA can lead to the more detailed classification of emails because frequently occurring keywords and the correlation between latent topics are simultaneously used to classify the emails [10].

Method 3: NV will extract only the nouns and verbs keywords in emails and LDA can extract only the topics which are latent in emails. On the other hand, the combination of NV and LDA can lead to the more relevant keyword extraction which will decide the topic in which the email falls clearly. The execution of NV-LDA will be fast and accurate as the only the useful keywords are considered which is going to decide the topic.

III. Effectiveness Analysis in Topic Term Discovery

In this subsection, we analyze the effectiveness of topics by LDA, TFIDF and LDA combined, NVLDA using personal email dataset. We choose 5 large topics ("Personal", "Official", "Finance", "Booking", and "Reminder"), which are supported by large numbers of emails. We adopt the top 10 most frequent terms in each cluster to represent the topic. Table 3 shows the ground truth and Tables 4-6 show the results of topic representative term groups discovered by NV-LDA, TF-IDF and LDA combined and only with LDA respectively. The topic representative terms discovered by NV-LDA. The boldface terms are wrongly detected.

Table 3: Ground Truth / Knowledge Corpus

Topics	Representative Words
Booking	Movies, bus, hotel, car, book, event
Bills	Electricity, item, phone, bill, gas, DTH, credit card, rent
Meeting	Date, time, venue, meeting, schedule, agenda, discussion
Official	Urgent, important, reply, head, principal, gtu, gecg28, gandhinagar
Shopping	Offer discount, fashion, shopping, clothing, off, summer, winter, brands, myntra

Table 4: The topic representative terms discovered by NV-LDA. The boldface terms are wrongly detected.

Topics	Representative Words
Topic 0	Bec owner, grade, myntra, project, computer, may, men women, lifestyle products, ecommerce store, personal care
Topic 1	Mutual fund, rs, courseera, reliance, ‘s, nav, enroll, sms, ananth, folio number
Topic 2	Kindly, thank, dear, sector Gandhinagar, gandhinagar, thanking, technical, pm, faculty members, block

Topic 3	Bec owner principal, be, govt, poly, myntra, spit piludra, hjd kera, botad, merchant visnagar, sigma vadodara
Topic 4	Me, rs, gtu, engg, acty, mec owner principal, sect Gandhinagar, fax, url, ext

Table 5: The topic representative terms discovered by TFIDF and LDA. The boldface terms are wrongly detected.

Topics	Representative Words
Topic 0	Cloud, institute, train, card, project, service, detail, message, live, sender
Topic 1	Fashion, myntra, product, shop, lifestyle, govt, destin, brand, company, ecommerc
Topic 2	Office, ananth, univers, fashion, develop, company, myntra, oper, busi, countri
Topic 3	Owner, princip, Rajkot, ahmedabad, Gujarat, surat, exam, form, vvnagar, vadodara
Topic 4	Patel, faculty, coordi, govern, sector, Gujarat, princip, institut, engg, professor

Table 6: The topic representative terms discovered by LDA. The boldface terms are wrongly detected.

Topics	Representative Words
Topic 0	Patel, style, fontsiz, coordin, shah, professor, sector, good, nahi, leav
Topic 1	Import, office, book, copi, update, offer, bank, nation, life, grade
Topic 2	Owner, principal, Rajkot, ahmedabad, Gujarat, surat, universe, vvnagar, vadodara, exam
Topic 3	Fashion, product, myntra, company, fund, shop, account, reliance, service, statement
Topic 4	Govt, polytechnic, universe, poli, course, science, ahmedabad, paper, research, special

VI. Discussion

Compared with the ground truth in Table 3, we can observe that NV-LDA (as shown in Table 4) can discover more accurate top frequent topic terms than other methods (as shown in Tables 5-6). Taking the topic "Topic 0" as an example, NV-LDA can

discover 6 out of 10 most representative words for the cluster "shopping", while TFIDF and LDA misses 5 and LDA misses 5 keywords. NV-LDA discover the topics as “Official”, “Shopping” and “Finance” properly with the keywords extracted but TFIDF and LDA combined approach and only LDA approach can discover only 2 topics like “Shopping” and “Official” but fails to discover the other topic: "finance". Both the other two methods had extracted the keywords of locations which is not useful in clustering. The proposed NV-LDA focuses on discovering the most significant term groups for email clustering by considering the closeness relations of two terms in the word network. Therefore, NV-LDA can discover more accurate top frequent terms as the representative terms and filters trivial terms at the same time.

VII. Conclusion

We proposed a Noun Verb Latent Dirichlet Allocation (NV-LDA) method for email clustering in this paper. NV-LDA extracts the noun and verbs from the email dataset and then by using LDA find out the hidden topics. We can found groups of significant terms that are closely bounded with each other as a topic representative group. It resolves the noisy and insufficient keywords discovery problem in the previous methods. NV-LDA also addresses the issues of sparsity and noisy in email clustering. Extensive experiments on real-world datasets show that our approach outperforms counterpart methods in terms of relevant keyword extraction. Some future directions can be evaluating the model in terms of accuracy, effectiveness and efficiency. Applying the model in other domain like sms clustering, tweet clustering etcs. and evaluating it.

References

[1] Gong, H., You, F., Guan, X., Cao, Y., & Lai, S. (2018). Application of LDA Topic Model in E-Mail Subject Classification. Advances in Intelligent Systems Research, Volume 161: 144-150.

[2] Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. Human-Centric Computing and Information Sciences, 9(1).

[3] Tony Tam, Artur Ferreira, and André Lourenço. (2012). Automatic foldering of email messages: a combination approach. In Proceedings of the 34th European Conference on Advances in Information Retrieval (ECIR'12). Springer-Verlag, Berlin, Heidelberg: 232–243.

[4] Naveen Saini, Sriparna Saha and Pushpak Bhattacharyya (2018) IEEE Computational Intelligence Society, International Neural Network Society, & Institute of Electrical and Electronics Engineers.

[5] Yang, S., Huang, G., & Cai, B. (2019). Discovering Topic Representative Terms for Short Text Clustering. IEEE Access, 7: 92037–92047.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan,(2003) "Latent Dirichlet allocation," J. Mach. Learn. Res., Vol. 3: 993_1022.

[7] NLTK- http://www.nltk.org/book/ch07.html Accessed 15 June 2021

[8] Bafna P, Pramod D, Vaidya A (2016) Document clustering: TF-IDF approach. In: IEEE int. conf. on electrical, electronics, and optimization techniques (ICEEOT):pp 61–66

[9] Shi J, Li W L. (2009) Topic analysis based on LDA model. Automatic newspaper, Vol. 36: 1586-1593.

[10] Havrlant L, Kreinovich V (2017) A simple probabilistic explanation of term frequency-inverse document frequency (TF-IDF) heuristic (and variations motivated by this explanation). Int J Gen Syst 46(1): 27–36

[11] Yau C-K et al (2014) Clustering scientific documents with topic modeling. Scientometrics 100(3): 767–786

[12] H. Hong and T. Moh, (2015) Effective topic modeling for email, 2015 International Conference on High Performance Computing & Simulation (HPCS): 342-349

[13] Wang Jingru, Chen Zhen (2018) A Comparative Study of Text Subject Extraction Based on Implicit Dirichlet Distribution (Information Science): 102-107

[14] Shah K., Shah N., Shah S., Patel D. (2021) Email User Classification and Topic Modeling. In: Arai K., Kapoor S., Bhatia R. (eds) Proceedings of the Future Technologies Conference (FTC) 2020, Volume 1. FTC 2020. Advances in Intelligent Systems and Computing, Vol. 1288. Springer, Cham

[15] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed and M. A. Al-Garadi, (2017) Email Classification Research Trends: Review and Open Issues, in IEEE Access, Vol. 5: 9044-9064

[16] Sharaff A., Gupta H. (2019) Extra-Tree Classifier with Metaheuristics Approach for Email Classification. In: Bhatia S., Tiwari S., Mishra K., Trivedi M. (eds) Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing, Vol. 924. Springer, Singapore

[17] Alsmadi I, Alhami I(2015) Clustering and classification of email contents. J King Saud Univ Compt Inf Sci. 27(1): 46-57

[18] Xuan J et al. (2017) Automatic bug triage using semi-supervised text classifcation. arXiv preprint arXiv: 1704.04769

CLASSIFICATION OF CLASS IMBALANCE IN SOFTWARE PREDICTION MODELS USING MACHINE LEARNING TECHNIQUES

K. J. Eldho
Department of Computer Science, Mary Matha Arts and Science College, Mananthavady, Kerala, India.
*corresponding author email: eldhorvs@gmail.com

Abstract

In a binary classification problem with data samples from two groups, class imbalance occurs when one class, the minority group, contains significantly fewer samples than the other class, the majority group. In many problems, the minority group is the class of interest, i.e., the positive class. In the software defect prediction problem, the cases of defective software modules are less as compared to non-defective software modules. For such type of problem, software developers take more interest in the correct identification of defective software modules. The failure to identify defective software modules can degrade the software quality. In this paper, machine learning based classifiers are analysed for class imbalance problems in software fault prediction models. This paper achieved its objective in two steps. First a model is proposed to split the dataset into two new datasets. As next step prediction models are tested on the created imbalanced dataset. In this paper, promise repository datasets are used for the experiments. The outcome of this research reveals that applied machine leaning models improve the performance.

Keywords: Software fault Prediction Model, Artificial Intelligence, Smart Debugging, Unbalanced Classification.

1. Introduction

In recent years, with the continuous development of machine learning and data mining, classification imbalance has gradually become a current research hotspot. Generally, unbalanced classification refers to the phenomenon that the distribution of sample numbers among different categories is unbalanced. For example, in the binary classification problem, when the sample sizes of the two categories differ greatly, the classification imbalance problem appears. In practical applications, classification imbalance problems are also common, such as text classification, fraud detection, and medical diagnosis. However, when dealing with the problem of unbalanced classification, the traditional classification model may become inefficient.Software defect prediction is one of the research hotspots in the field of software engineering, and it is also an important work to ensure software quality [1]. Software defect prediction is a typical two-category problem. Its purpose is to divide software modules into defective modules and non-defective modules [2]. It mainly conducts statistical

analysis on historical defect information, excavates the distribution law of historical defects and builds models, so as to compare new software module makes predictions.

Generally, in the defect data set, defective modules belong to the minority category, while non-defective modules belong to the majority category. In this case, it is more important to correctly identify the minority classes than to correctly identify the majority classes. However, traditional classification models often aim at maximizing the overall classification accuracy, but ignore the classification of the minority classes. For example, a defect data set contains 100 samples, of which there are only 1 defective sample and 99 non-defective samples. If a certain classification model predicts all samples as non-defective samples, an overall classification accuracy of 99% can be achieved.But the classification accuracy of defective samples is 0, and such prediction results have no value. Therefore, the imbalance of classification will have a certain impact on the defect prediction results, and it also poses new challenges to the effectiveness of traditional classification models.

2. Literature Survey

At present, there are many methods to solve the problem of classification imbalance, which are mainly divided into three categories: The first category is sampling methods, including over-sampling and under-sampling. Reduce the majority of samples to obtain a new data set with relatively balanced classification. The second category is cost-sensitive learning [3-4]. In the problem of imbalanced classification, it is more valuable to correctly identify a minority class than to correctly identify the majority class. That is, it is more expensive to misclassify a minority class than to misclassify the majority class. However, traditional The classification model believes that the misclassification cost of all categories is the same [5]. Therefore, cost-sensitive learning improves the classification performance of a few classes by assigning different misclassification costs to different classes. The third category is integrated learning [6], which improves classification performance by gathering the prediction results of multiple models. Generally, the performance of the integrated model is better than that of a single model. Although ensemble learning is not proposed to solve the problem of unbalanced classification, it can achieve better results when dealing with the problem of unbalanced classification.

However, when the above three methods solve the problem of imbalance in classification, they often need to combine specific prediction models (ie classification models) or verify them under certain prediction models [7]. How to choose a reasonable prediction model? In addition, the prediction model itself may also be affected by the imbalance of classification. Which prediction models have more stable performance? If we can grasp the performance stability of different prediction models

when the classification is unbalanced, then we can choose a reasonable prediction model in practical application to better guide the software defect prediction work.

Based on this, this paper proposes a classification imbalance impact analysis method to evaluate the impact of classification imbalance on the performance of the software defect prediction model. This method constructs a new data set with increasing imbalance rate on the original unbalanced data set, and then selects 8 typical classification models as defect prediction models, respectively predicts the constructed new data set, and uses ROC (Receiver The area under the operating characteristic curve---Area Underthe Curve [8-10] index is used to evaluate the classification performance of different prediction models, and at the same time, different coefficients of variation are used to predict the different coefficients of the prediction model. The degree of performance stability when the classification is unbalanced. The experimental results show that the performance of the three prediction models, BPN [11] SVM [12] and ELM [13], decreases with the increase of the imbalance rate, indicating that the performance of these three models is very susceptible to classification imbalance. The impact of cost-sensitive learning and ensemble learning can effectively improve their performance and performance stability when the classification is imbalanced.

The main contributions of this article are as follows:

(1) Propose a classification imbalance impact analysis method, transform the original imbalance data set into a new data set with increasing imbalance rate, and select different prediction models to predict the new data set to evaluate different predictions. The degree of stability of the model's performance when the classification is unbalanced.

(2) Experiments were carried out on 5 typical prediction models, and the results showed that the performance of prediction models KNN, NB, BPN are very susceptible to the imbalance of classification, while SVM and ELM is more stable.

(3) Through the work of this paper, we can grasp the degree of performance stability of different prediction models when the classification is unbalanced, so as to select a reasonable prediction model in practical applications, which has a certain guiding role for the research of software defect prediction.

3. Software Defect Prediction Models

In software defect prediction, researchers often need to select some predictive models for experimentation, such as decision trees, logistic regression, and naive Bayes. In order to make the prediction model to be evaluated more representative, statistics are made on the commonly used prediction models. The preliminary statistical results are

shown in Table 1, including the name (abbreviation) of the prediction model, introduction and quotations. Among them, the cited literature means that these literatures have adopted the corresponding prediction model to carry out the experiment. The following is a further introduction to these prediction models.

1. KNN K Nearest Neighbour

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

2. Naïve Bayes

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. In this article, we will understand the Naïve Bayes algorithm and all essential concepts so that there is no room for doubts in understanding.

3. BPN – Back Propagation Neural Network

Backpropagation is the essence of neural network training. It is the method of fine-tuning the weights of a neural network based on the error rate obtained in the previous epoch (i.e., iteration). Proper tuning of the weights allows you to reduce error rates and make the model reliable by increasing its generalization.

4. SVM – Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate.

5. ELM – Extreme Learning Machine

Extreme learning machines are feedforward neural networks for classification, regression, clustering, sparse approximation, compression and feature learning with a single layer or multiple layers of hidden nodes, where the parameters of hidden nodes (not just the weights connecting inputs to hidden nodes) need not be tuned. These hidden nodes can be randomly assigned and never updated (i.e. they are random projection but with nonlinear transforms), or can be inherited from their ancestors

without being changed. In most cases, the output weights of hidden nodes are usually learned in a single step, which essentially amounts to learning a linear model.

3.1 The Impact of Unbalanced Classification on the Performance of Software Defect Prediction Models

In software defect prediction, unbalanced classification means that the number of non-defective samples in the data set is much higher than the number of defective samples. Suppose a data set D = {x1, x2,..., xn}, xi∈ Rd (i = 1, 2, ..., n), that is, the data set includes a number of samples, and each sample contains a In addition, it also includes a category feature to mark the category of the sample, that is, defective or non-defective. According to the characteristics of the category, the data set can be divided into defective type C1 and non-defective type C2. The number of samples is n1 and n2, respectively, and n = n1 + n2. Thus, the imbalance rate of the data set D (ImbalanceRatio)[12] is the ratio of the number of non-defective samples, n2 to the number of defective samples, n1, that is, n2/n1, which is rounded down as IR=(n2/n1). Generally, n2> n1, that is, IR>1. The larger the value of IR, the higher the degree of imbalance in the classification of the data set, and vice versa.

Algorithm 1. New Data Set Construction Algorithm

Input: DataSet - The original unbalanced data set
Output: NewDataSet - New data set
1. According to category and characteristics DataSet is divided into DefectSet and NonDefectSet;
2. Number of defective samples n1=DefectSet.Size();
3. Number of non-defective samples n2=Non NonDefectSet.Size();
4. Imbalance rate r=Math.floor(n2/n1);
5. Form newDataSet=DefectSet;
6. Form restNonDefectSet=NonDefectSet;
7. WHILE restNonDefectSet=NonDefectSet;
8. Randomize the restNonDefectSet
9. IF restNonDefectSet.Size() >=2n1
10. Choose n1 samples randomly from restNonDefectSet
    Save as newDataSet
11. Remove the selected sample from the restNonDefectSet;
12. ELSE
13. save the remaining sample in restNonDefectSet as new Data Set;
14. restNonDefectSet = null;
15. END IF
16. Save new Data Set;
17. END WHILE
18. RETURN all new datasets as new Data Set;

In order to explore the impact of classification imbalance on the performance of prediction models, that is, the changes in the performance of each prediction model in the case of classification imbalance, a set of data sets with different imbalance rates are needed. Therefore, this paper designs a new data set construction algorithm to transform the original unbalanced data set into a new data set with successively increasing unbalance rates. The specific process is shown in Algorithm 1.

4. Experimental Results

In the case of imbalanced classification, the impact analysis method of imbalanced classification proposed in this paper is used to evaluate the performance stability of different prediction models. The experiment was performed using Weka tool under windows platform.

In the forecasting process, it is first necessary to select reasonable indicators to evaluate the performance of each forecasting model. In this paper, the area under the ROC curve (AUC) [13] is used for evaluation. Here we first introduce the most basic performance evaluation indicators. Software defect prediction is a two-class classification problem, and its prediction process will produce 2 different results, as shown in Table 2, where defects are positive cases, and no defects are negative cases. The row represents the actual category, and the column represents the predicted category.

4.1 Test Dataset

The experiment selected 2 unbalanced classification data sets. The basic information is shown in Table 1. These data sets are all defect data sets in the PROMISE library. The first column indicates the name of the dataset, the second column indicates the development language of the dataset program; the third column indicates the number of features contained in the dataset, that is, the feature dimension; the fourth column indicates the samples in the dataset Total, which describes the size of the data set, including small-scale (a few hundred), medium-scale (thousands), and large-scale (tens of thousands); the fifth to seventh columns represent the number of defective samples and non-defective samples in the data set, respectively The number of samples and the defect rate; the eighth column represents the unbalance rate of the data set, which is calculated as IR=(n2/n1), that is, the ratio of the number of samples with no defects to the number of samples with defects is an integer. The larger the value, the more unbalanced the classification of the data set. Jedit-4.3 and Tomcat are open source data sets, and the features they contain are class-level CK metrics, which comprehensively consider the inheritance, coupling, and cohesion in object-oriented programs. So, as to more effectively measure the correlation between software features and defects.

Table 1. Experimental Dataset

Dataset Name	Language	Number of Features	Number of Samples	Number of Defective Samples	Number of non-defective samples	Defect rate (%)	Imbalance rate
Jedit-4.3	Java	20	492	11	481	2.24	43
Tomcat	Java	20	858	77	781	8.97	10

In software defect prediction, the prediction result is jointly determined by the data set and the prediction model. For a certain data set, which includes many software features, all features are used to train the prediction model when feature selection is not performed or the prediction model itself does not have feature selection capabilities; High feature subset, and then train the prediction model based on the selected feature subset. In particular, in this experiment, all the features in the above data set are used to train the prediction model. In order to explore the difference in performance stability of different prediction models in the case of imbalanced classification, the method of this paper is used to evaluate the performance stability of each prediction model in Table 2.

4.2 Comparison of Performance Stability of Different Prediction Models

A binary classification problem, such as fault-prone (positive) and not fault-prone (negative), has four possible prediction outcomes: True Positive (TP) (i.e., correctly classified positive instance), False Positive (FP) (i.e., negative instance classified as positive), True Negative (TN) (i.e., correctly classified negative instance), and False Negative (FN) (i.e., positive instance classified as negative). The four values form the basis for several other performance measures that are well known and commonly used for classifier evaluation. The Overall Accuracy (OA) provides a single value that ranges from 0 to 1. It can be calculated by the equation, $OA = (|T\ P| + |T\ N|)/N$, where N represents the total number of instances in a dataset. While the overall accuracy allows for easier comparisons of model performance, it is often not considered to be a reliable performance metric, especially in the presence of class imbalance. The Area Under the ROC (receiver operating characteristic) curve (i.e., AUC) is a single-value measurement that originated from the field of signal detection. The value of the AUC ranges from 0 to 1. The ROC curve is used to characterize the trade-off between True Positive Rate (TPR) $TPR=TP/(TP+FN)$ and False Positive Rate (FPR) $FPR=FP/(FP+TN)$. A classifier that provides a large area under the curve is preferable over a classifier with a smaller area under the curve. The TPR (True Positive Rate) refers to the ratio of the number of correctly predicted positive cases to the actual number of positive cases, that is, the ratio of the number of correctly predicted

defective samples to the actual number of defective samples. The FPR (False Positive Rate) refers to the ratio of the number of false positive cases to the actual number of negative cases, that is, the ratio of the number of falsely predicted as defective samples to the actual number of non-defective samples.

For a specific prediction model and training data set, the prediction result corresponds to a point on the ROC curve. By adjusting the threshold of the model, a curve passing through (0, 0) and (1, 1) can be obtained below the curve. The area of A is the value of A. In particular, the value range of AT is 0 to 1. When AT is 0.5, it represents the performance of the random guessing model. The larger the value of A, the better the performance of the model. Therefore, a good prediction model should be as close as possible to the upper left corner of the coordinate system. Use the prediction models KNN, Naïve Bayes, BPN, SVM, ELM and the data set Jedit and Tomcat to conduct combined experiments. First, select a data set, and use Algorithm 1 to transform the data set into a new data set with increasing imbalance rate (ie, IR = 1, 2, ..., r); then, use the prediction models to predict the new data set separately to obtain a set of AT values under different imbalance rates, which are recorded as the set S={AUC1, AUC2, ..., AUCr} ; finally , Through the coefficient of variation CV of this group of AT values to evaluate the performance stability of different prediction models under different imbalance rates.

Table 2 shows the experimental results of each prediction model on different data sets, including the mean μ, standard deviation σ, and coefficient of variation CV. The larger the coefficient of variation CV, the more unstable the performance of the prediction model, that is, the greater the impact of imbalance in classification on the performance of the prediction model.

Table 2. Evaluation Results

Prediction Models	Jedit			Tomcat		
	Mean(μ)	Std(σ)	CV	Mean(μ)	Std(σ)	CV
KNN	0.513	0.021	1.710	0.582	0.024	2.135
NAIVEBAYES	0.524	0.021	1.902	0.621	0.026	2.102
BPN	0.581	0.022	1.903	0.624	0.053	2.821
SVM	0.612	0.012	2.102	0.734	0.021	6.031
ELM	0.721	0.020	2.204	0.768	0.032	7.875

It can be seen from Table 2 that the BPN, SVM, and ELM three prediction models have relatively high CV values on both data sets, indicating that the performance of these three models is very susceptible to the imbalance of classification. The performance of models such as KNN, NB and BPN remains relatively stable. In

addition, the sample distribution differences of different data sets will also affect the performance of the prediction model to a certain extent. Therefore, the performance of the same model on individual data sets may be different from the performance on other data sets. For example, the KNN model only shows a slight instability on the Jedit data set (CV value is 1.71%), while the BPN model only shows obvious instability on the Jedit data set the value of CV is 1.903%, because the number of defective samples in this data sets is too small. This makes the initially constructed new data set too small, which affects the performance stability of the prediction model.

Finally, there are external factors, such as the quality of the data set, which may affect the evaluation of the stability of the predictive model's performance by the methods used in this paper. Therefore, this work has conducted more sufficient experiments on data sets of different scales, different defect rates, and different imbalance rates, and all experimental data sets are real data sets in defect prediction, and are also the most commonly used defect data sets. Ensure the validity and reliability of the prediction results.

5. Conclusion and Future Work

In software defect prediction, the impact of unbalanced classification problems has become increasingly prominent. In order to explore the impact of unbalanced classification on software defect prediction and related technologies, researchers have also carried out a large number of empirical studies. In particular, in the experiments in this paper, all the parameters of the prediction model are set to default values. Therefore, the following will explore the impact of different operating parameters on the performance of the prediction model, especially the impact of the used predictive models with unstable performance. In addition, this article explores the impact of unbalanced classification on software defect prediction, that is, the impact on binary classification problems. The imbalance of classification also exists in multi-classification problems. Therefore, the impact of unbalanced classification on multi-classification problems still needs explore further.

References

[1] Yu Qiao, Jiang Shuxuan, Zhang Yanmei, Wang Xingya, Gao Pengfei, and Qian Junyan. (2018). Study on the effect of classification imbalance on the performance of software defect prediction model. Journal of Computer Science , 4, 809-824.

[2] Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. Expert systems with Applications, 36(1), 690-701.

[3] Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. Acmsigkdd explorations newsletter, 6(1), 50-59.

[4] Mohammed, A., Podila, P. S., Davis, R. L., Ataga, K. I., Hankins, J. S., &Kamaleswaran, R. (2019). Machine learning predicts early-onset acute organ failure in critically ill patients with sickle cell disease. bioRxiv, 614941.

[5] Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., ... & Zhu, F. (2018). Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic acids research, 46(D1), D1121-D1127.

[6] Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. Pattern Recognition, 48(5), 1623-1637.

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[8] Tahir, M. A., Kittler, J., & Yan, F. (2012). Inverse random under sampling for class imbalance problem and its application to multi-label classification. Pattern Recognition, 45(10), 3738-3750.

[9] Zhou, Z. H., & Liu, X. Y. (2005). Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on knowledge and data engineering, 18(1), 63-77.

[10] Siers, M. J., & Islam, M. Z. (2015). Software defect prediction using a cost sensitive decision forest and voting, and a potential solution to the class imbalance problem. Information Systems, 51, 62-71.

[11] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., &Folleco, A. (2014). An empirical study of the classification performance of learners on imbalanced and noisy software quality data. Information Sciences, 259, 571-595.

[12] Grbac, T. G., Mausa, G., & Basic, B. D. (2013, September). Stability of Software Defect Prediction in Relation to Levels of Data Imbalance. In Sqamia (pp. 1-10).

[13] Tomar, D., & Agarwal, S. (2016). Prediction of defective software modules using class imbalance learning. Applied Computational Intelligence and Soft Computing, 2016.

PREDICTION OF BREAST CANCER USING MACHINE LEARNING AND DATA MINING APPROACH

Anupam Sen
Department of Computer Science, Government General Degree College, Singur, Hooghly, West Bengal, India
*corresponding author email: anupam.sen2006@gmail.com

Abstract

Machine Learning techniques are playing an important role within the medical field. Machine learning algorithms can be applied to develop model for better prediction of breast cancer. In the proposed study, WEKA tool is used to implement different classification algorithms such as Ada Boosting, Cvglmnet, Glmboost, Glmnet, Kknn, Ksvm, rFerns, Logreg for evaluation of breast cancer prediction. Feature selection (FS) method CfsSub-setEval is used to improve the performance of algorithms for better prediction. The experiments are carried out by using Wisconsin Diagnostic Carcinoma (WDBC) dataset with thirty-two predictors are taken from the UCI repository. Each algorithm's performance is measured using accuracy, Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE). This approach can be further embedded into IoT based breast cancer prediction support system. Proposed method can be helpful for the medical expert to diagnose breast cancer competently.

Keywords: IoT, Kappa statistic, MAE, RMSE, Logreg.

1. Introduction

According to the World Health Organization (WHO) more than 2.1 million deaths are occurring worldwide due to breast cancer which is the most predominant cancer among women. Generally, in case of breast cancer the time when the patient becomes aware of their condition the chances of survival become bleak because there are no pain sensations or symptoms in the early stages. It is true that accurate and timely diagnosis can greatly increase survival chances and aid in reducing treatment costs. Today modern diagnosis involves precise evaluation of patient data and expert decision coupled with varied machine learning approaches and pattern recognition which are proposed to provide supportive aid to experts in their decision- making process. The dominant role of these approaches is extraction of informative knowledge about patient's data and reduction in the time and cost of diagnosis. In this proposed work seven machine learning algorithms are used to predict breast cancer. Feature

selection (FS) algorithm remove unrelated data from the data set to increase the performance of the classifier and also reduce high computational cost. A comparative performance analysis of different classifiers is used to make better prediction of breast cancer.

2. Related Works

L. Gao, M. Ye, and C. Wu [1] have experimented on 9 cancer datasets using SVM (Support Vector Machine) optimized by PSO (Particle Swarm Optimization) combined with ABC (Artificial Bee Colony) to predict breast cancer. Swesi, I. M. A. O and Bakar [2] proposed feature clustering algorithm on high dimensional data for improving the accuracy of the approach, and lower the computational cost. T. Advancements et al. [3] have carried out research work on Elitism Particle Swarm Optimization (EPSO) based and Recursive Feature Reduction (RFR) for selecting genes yield better classification performance and are biologically relevant to cancer. M. R. Mohebian et al. [4] have carried out resaerch work and have built a Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning to identify breast cancer recurrence within 5 years after diagnosis. D. A. Utami and Z. Rustam [5] presented a comparison of PSO–SVM and ABC–SVM machine learning algorithms to detect symptoms of

breast cancer. ABC–SVM method is performed better with an accuracy rate of 88% cancer classification when compared with the PSO–SVM method, which has an accuracy rate of 87 %. S. B. Sakri et al. [6] have implemented particle swarm optimization (PSO) based three popular classifying algorithms, namely, naïve Bayes, IBK, and REPTree, with and without the feature selection algorithm for breast cancer prediction. Naive Bayes performed better output with and without PSO. M. Mahajan et al. [7] proposed a particle swarm optimization method to enhance the performance of the kNN classifier. S. Jeyasingh and M. Veluchamy [8] proposed Modified Bat Algorithm (MBA) for feature optimization with Random Forest (RF) classifier. M. A. Rahman and R. C. Muniyandi [9] implemented a two-step feature selection method based on Artificial Neural Networks with 15 Neurons were used to increase the performance of the classifier. B. Al-Shargabi et al. [10] implemented Multilayer perceptron with feature selection to predict breast cancer which obtained an accuracy rate of 97.70%.

3. Proposed Methodology

In the proposed work, Wisconsin Diagnosis BreastCcancer (WDBC) dataset is obtained from the UCI ML Repository [11]. Dataset has 569 instances with 31 features and a class variable, i.e. (M = malignant, B = benign). Table 1 contains the 11 features selected by CfsSubsetEval with the best first method attribute selection method. The

experiment is carried out using WEKA free software tool with 10-fold cross validation method. In the proposed methodology, seven supervised machine learning algorithms such as Ada Boosting, Cvglmnet, Glmboost, Glmnet, Kknn, Ksvm, rFerns, Logreg are used to build models using all thirty one attributes and only eleven important attributes selected by CfsSubsetEval method. Proposed layout of the model is depicted in fig. 1.

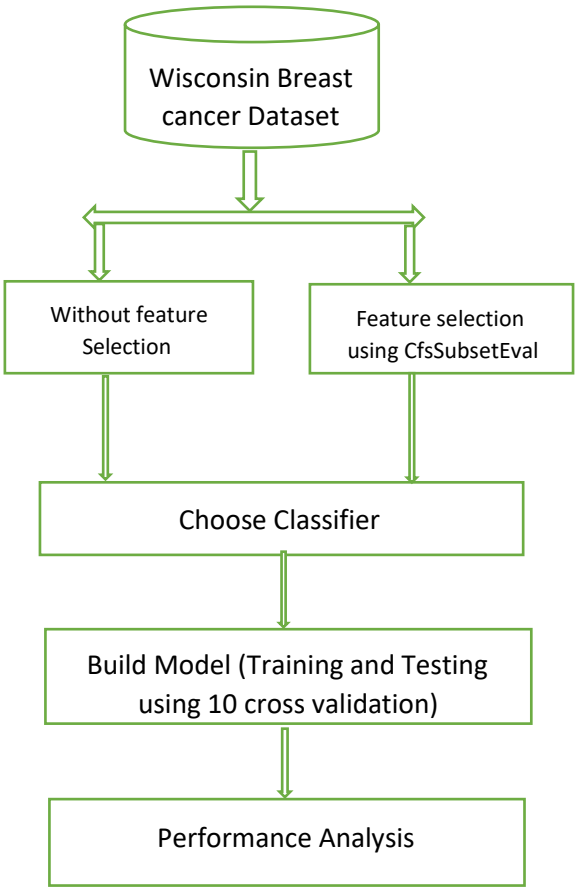


Fig. 1. Layout of proposed model

Table 1. Features selected by CfsSubsetEval method

Sl. No	Attribute Name
1.	texture_mean
2.	concavity_mean
3.	concave points_mean
4.	area_se
5.	symmetry_se
6.	radius_worst
7.	perimeter_worst
8.	area_worst
9.	smoothness_worst
10.	concavity_worst
11.	concave points_worst

Table 2. Performance analysis without feature selection (WFS)

Classification Algorithm	**Accuracy**	**Kappa Statistic**	**Mean Absolute Error (MAE)**	**Root Mean Square Error (RMSE)**
Ada Boosting	96.66 %	0.9281	0.0416	0.1636
Cvglmnet	97.18 %	0.9394	0.0686	0.1608
Glmboost	97.01 %	0.9353	0.1132	0.1853
Glmnet	97.01 %	0.9354	0.0722	0.1596
Kknn	96.83 %	0.9316	0.0511	0.1628
Ksvm	97.53 %	0.9474	0.0468	0.1436
rFerns	95.25 %	0.899	0.0475	0.2178
Logreg	**93.49 %**	**0.8618**	**0.065**	**0.255**

Table 3. Performance analysis with feature selection (FS).

Classification Algorithm	Accuracy	Kappa statistic	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)
Ada Boosting	95.78%	0.9096	0.0461	0.1751
Cvglmnet	97.53 %	0.9468	0.0764	0.1661
Glmboost	96.83 %	0.9317	0.115	0.189
Glmnet	97.36 %	0.943	0.0755	0.1641
Kknn	96.48 %	0.9244	0.0496	0.1633
Ksvm	96.83 %	0.9322	0.0532	0.1587
rFerns	96.30 %	0.9214	0.0369	0.1921
Logreg	96.83 %	0.9323	0.0475	0.1675

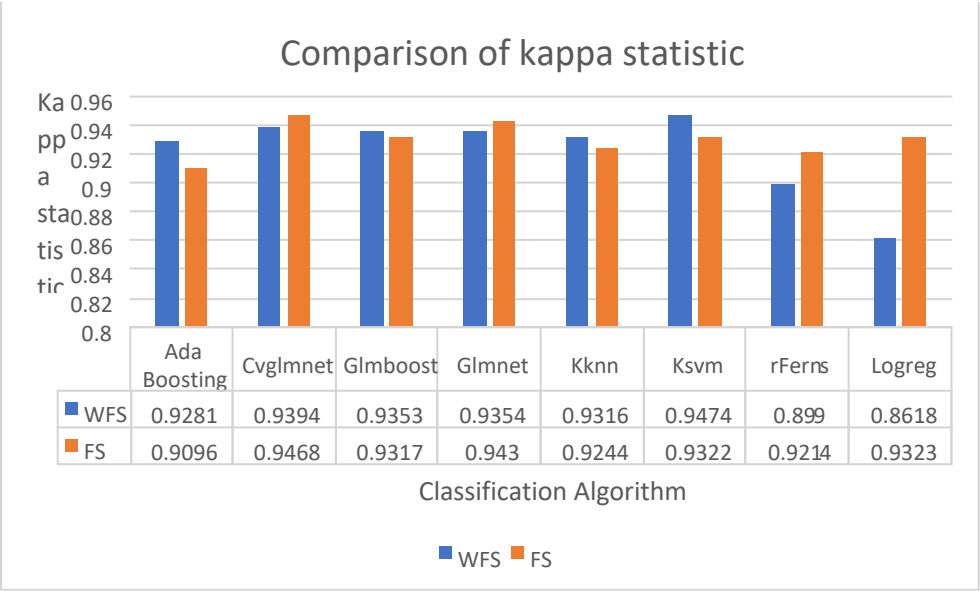


Fig. 3. Comparison of Kappa statistic between without feature selection (WFS) and feature selection (FS)

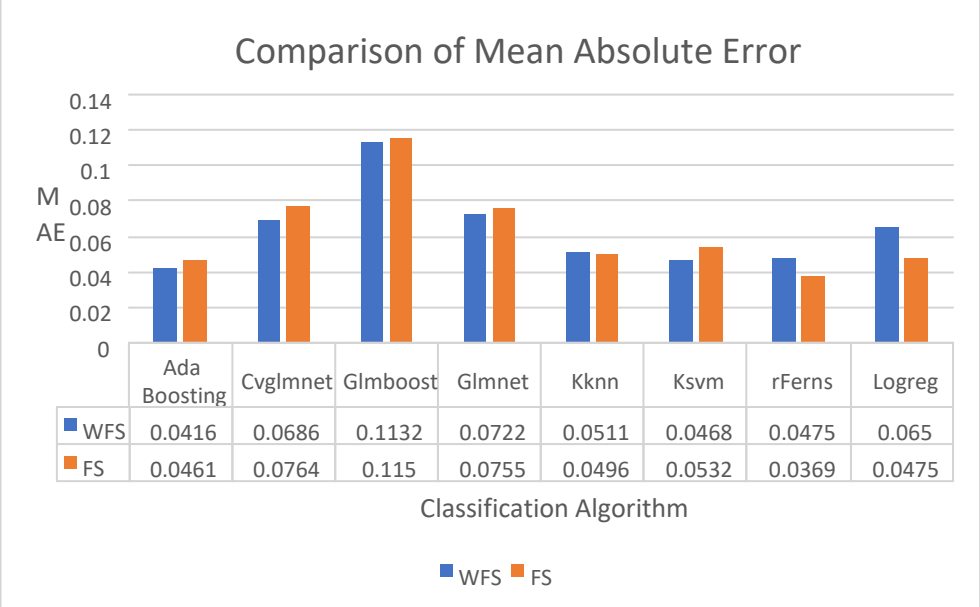


Fig. 4. Comparison of Mean Absolute Error between without feature selection (WFS) and feature selection (FS)

Fig. 4. Comparison of Mean Absolute Error between without feature selection (WFS) and feature selection (FS)

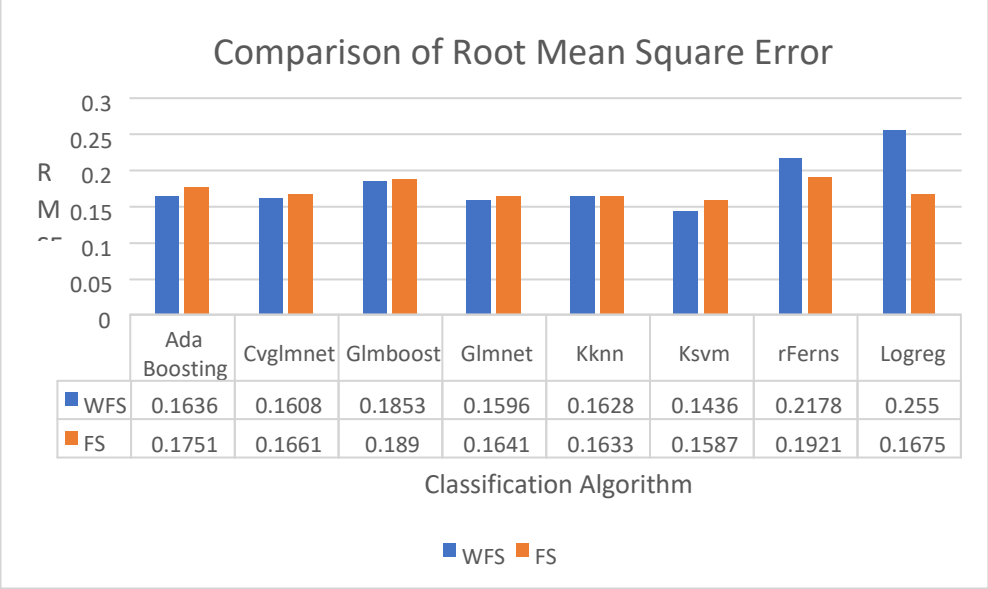


Fig. 5. Comparison of Root Mean Square Error between without feature selection (WFS) and feature selection (FS)

4. Experimental Results

Table 2 represents the experimental results of the different classification algorithms without selecting features elaborating on its accuracy, Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE). In case of Ksvm classification algorithm the accuracy is 97.53% and its performance is better than other algorithms indicating higher Kappa statistic, Minimized Root Square Error compared to any other algorithms with only Mean Absolute Error minimized for Ada boosting. Illustration about the accuracy, Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) of the different classifier algorithms based on attribute selection method for selecting features is presented in Table 3, which shows accuracy value of 97.53 % for Cvglmnet classification algorithm, which reveals that it performs better among other classification algorithms used for selecting features and higher Kappa statistic value. Comparing the different dimensions, find that the Mean Absolute Error is minimum for rFerns classification algorithm whereas Root Square Error is minimum for Ksvm. A Comparative picture is presented in fig. 2, fig. 3, fig. 4 and fig. 5 respectively for without selecting feature (WFS) and with feature Selection (FS) of different classification algorithms with respect to their accuracy, Kappa statistic, Mean Absolute Error, Root Mean square Error. Fig. 2 shows that the accuracy of the classification algorithms Cvglmnet, Glmnet, rFerns and Logreg lead to better results in the proposed work, whereas in Fig. 3 reveals that the Kappa Statistics of the Cvglmnet, Glmnet, rFerns and Logreg classification algorithms perform better results in the proposed work. Fig. 4 depicts that in case of Kknn, rFerns and Logreg classification algorithms the Mean Absolute Error (MAE) is minimized and Fig. 5 provides information about Root Mean Square Error (RMSE), which is minimized for the rFerns and Logreg classification algorithms in the suggested work.

5. Conclusion and Future Work

The main objective of the proposed work is to improve the performance of the different classifiers so that they can more accurately identify the early diagnosis of breast cancer. In this proposed model, good performance is obtained for Cvglmnet, Glmnet, rFerns and Logreg classification algorithms. Drawing from the findings it can be concluded that feature extraction and machine learning algorithms play an essential role in identifying the early diagnosis of breast cancer to reduce cost and time. The future work of the research work is to improve the accuracy of breast cancer prediction by applying newer algorithms and various feature selection methods. The breast cancer prediction can be automated using real time data. Early diagnosis and reduced cost can improve healthcare facility in future.

References

[1] L. Gao, M. Ye, and C. Wu, "Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony," Molecules, Vol. 22, No. 12, 2017, doi: 10.3390/molecules22122086.

[2] C. Technology, "How to cite this article: Swesi, I. M. A. O., & Bakar, A. A. (2019). Feature clustering for pso-based feature construction on high-dimensional data.," Vol. 4, No. 4, pp. 439–472, 2019.

[3] T. Advancements, R. Nagpal, and R. Shrivas, "Cancer Classification Using Elitism PSO Based Lezy IBK on Gene Expression Data," Vol. 1, No. 4, pp. 19–23, 2015.

[4] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning," Comput. Struct. Biotechnol. J., Vol. 15, pp. 75–85, 2017, doi: 10.1016/j.csbj.2016.11.004.

[5] D. A. Utami and Z. Rustam, "Gene selection in cancer classification using hybrid method based on Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) feature selection and support vector machine," AIP Conf. Proc., Vol. 2168, 2019, doi: 10.1063/1.5132474.

[6] S. B. Sakri, N. B. Abdul Rashid, and Z. Muhammad Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," IEEE Access, Vol. 6, pp. 29637–29647, 2018, doi: 10.1109/ACCESS.2018.2843443.

[7] M. Mahajan, S. Kumar, B. Pant, K. Joshi, and V. Tripathi, "PSO Optimized Nearest Neighbor Algorithm," Int. J. Eng. Adv. Technol., Vol. 9, No. 2, pp. 1508–1513, 2019, doi: 10.35940/ijeat.b3574.129219.

[8] S. Jeyasingh and M. Veluchamy, "Modified bat algorithm for feature selection with the Wisconsin Diagnosis Breast Cancer (WDBC) dataset," Asian Pacific J. Cancer Prev., Vol. 18, No. 5, pp. 1257–1264, 2017, doi: 10.22034/APJCP.2017.18.5.1257.

[9] M. A. Rahman and R. C. Muniyandi, "An enhancement in cancer classification accuracy using a two-step feature selection method based on artificial neural networks with 15 neurons," Symmetry (Basel), Vol. 12, No. 2, 2020, doi: 10.3390/sym12020271.

[10] B. Al-Shargabi, F. Al-Shami, and R. S. Alkhawaldeh, "Enhancing Multi-Layer Perceptron for Breast Cancer Prediction," Int. J. Adv. Sci. Technol., Vol. 130, No. September, pp. 11–20, 2019, doi: 10.33832/ijast.2019.130.02.

[11] UCI "Machine Learning Repository" https://archive.ics.uci.edu/ml/index.php

---

A STUDY ON BIG DATA ANALYTICS AND VISUALIZATION TOOLS WITH SPECIAL REFERENCE TO DATA ON COVID 19

Ravitha Sudhakaran, Remya Raveendran*
SCMS School of Technology and Management, Kalamassery, Kerala, India
*corresponding author email: remyaraveendran@scmsgroup.org

Abstract

In this digital era enormous data are generated from various sources and the transition from digital technology has lead to the growth of Big data. It is not the size but it is the value inside the data makes it Big Data. A huge effort is required for the analysis and interpretation of these data and their conversion to knowledge for decision making. Hence it is a potential area for research. An example is health care and epidemiological data such as data related to patients who suffered epidemic diseases like the corona virus disease COVID 19, which is helpful to researchers, epidemiologist and policy makers to handle the disease effectively. Big data tools like Sisense, Tableau, Datawrapper, Power bi, Qlik Sense, Apache Spark etc are used and among these most popular data visualization tools like Power BI and Tableau are discussed in this paper to find our which one suits the best for Covid big data. These tools helps the users to get a better understanding about the data. As pictures can communicate the ideas better than words visualization and visual analytics plays a major role in big data analytics.

Index Terms: Analysis, Big Data, COVID 19, DataWrapper, Infogram, Power BI, Qlik Sense, Sisense Tableau, Visualization

I. Introduction

Big Data refers back to the big extent of information that cannot be stored in a single laptop. It is the facts with so massive length and complexity that none of the traditional facts management equipment may be used to keep it or technique it efficiently. Big data technology can store a massive amount of information about the human beings infected with COVID-19 virus. This records may be successfully used for case identification and helping to allocate the sources for higher protection of public health. These information may be used to tune the virus on a international basis constantly and to create innovation in clinical fields. There are many large statistics visualization gear that can offer public health officers the capacity to peer how COVID-19 progress over the time. This study is to find the nice big data visualization tools suitable for the visualization of COVID-19 statistics by using comparing two maximum popular tools Tableau and Power BI.

II. Types of Big Data

Understanding in which the raw statistics comes from and how it must be treated earlier than analyzing it turns into vital depending at the quantity of large data. The structure of huge facts is not best a key to apprehend its running but also it indicates what insights it could produce.

Structured, Unstructured and Semi-structured

a. Structured

Any facts which will be stored, accessed and processed within the shape of fixed format in termed as structured.

b. Unstructured

Any records with unknown shape or the structure is classified as unstructured records. In addition to its massive length, unstructured information poses a couple of demanding situations in phrases of its processing for deriving price out of it.

c. Semi-structured

Semi-structured facts can include both the types of information. It may be dependent shape however it isn't always honestly described with a desk definition in relational DBMS.

III. 3 V' S OF BIG DATA

Big data can be defined in term of different characteristics.

a. Volume

As the name indicates the size of big data is enormous. The size plays an important role in determining value out of data. To determine a data as big data or not, is dependent upon the volume of data.

b. Variety

The term variety refers to the different sources and the nature to which the data belongs i.e, both structured and unstructured.

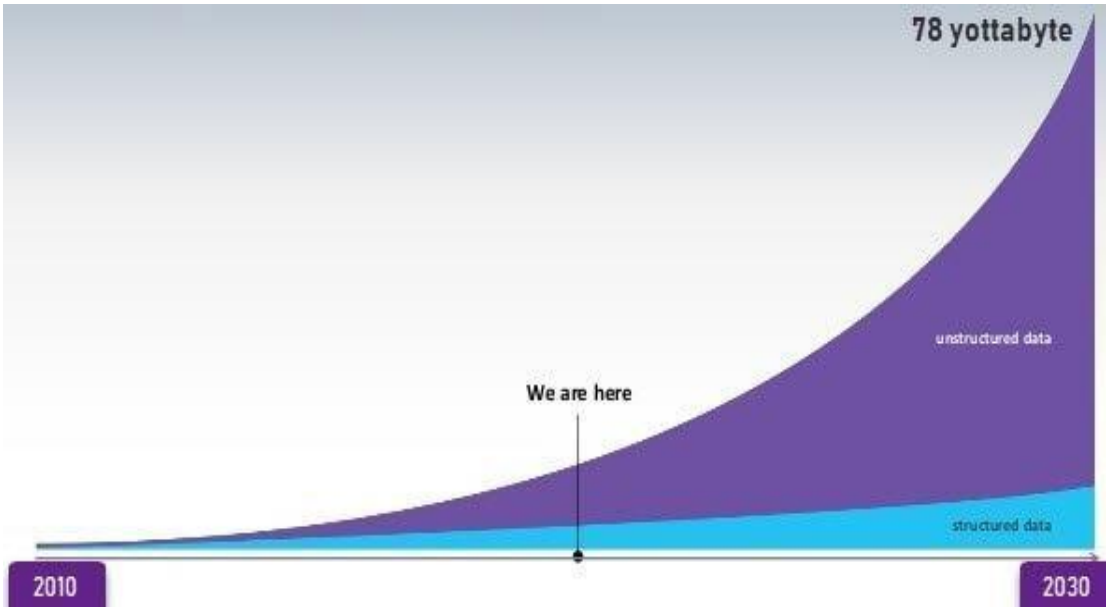


Figure 1. Growth of structured and unstructured data

Nowadays various forms of data like emails, photos, audio, pdfs etc are being considered in the explicit analysis.

c. Velocity

Velocity refers to the speed of generation of knowledge. How fast the info is generated and processed to satisfy the stress, determines real potential within the data.

IV. Big Data Visualization and Analytics

Big data analytics examines large amount of statistics to find hidden styles, correlations and other insights. With today's generation, it's possible to analyze the facts and get solution from it nearly without delay- an attempt that's slower and much less efficient with greater conventional commercial enterprise intelligence solutions. Huge quantity of records requires making use of big data analytical tools to make sense of pandemic and manipulate its spread in a timely manner. It provides real time monitoring of records. As photos can speak the ideas better than words visualization and visible analytics plays a chief role in large data analytics. Data can be visualized in 5 unique categories, it consists of temporal, hierarchical, community, multidimensional and geospatial.

V. Big Data and COVID-19 Visualization

As COVID data is received from one of a kind sources in a huge quantity in various amount, huge data may be used to symbolize these data. The on the spot outbreak of this sickness have created a critical source of data and knowledge. These data are wont to undertake studies and improvement approximately the virus, pandemic and measures to fight this virus and after results. Big data in this modern era may digitally keep a huge amount of records of those patients. It helps to computationally analyze to reveal styles, trends, associations and differences. It can also assist in revealing the insights into the unfold and manage of this virus.

With the unique shooting capability, big statistics may be used to minimize the chance of spreading this virus. A lot of faux records has additionally been generated regarding this pandemic. Hence the choice of dependable statistics from this statistics pool changed into one the various fundamental challenges faced by means of data analytics. In order to conquer this assignment authorities has standardized the source which data can be collected and utilized. Hence WHO and worldmeter acts as the two dependable supply of COVID-19 data. Proper evaluation of this collected data with a view to depict the desired statistics from those massive data is crucial (data analysis techniques).

Data visualization has been so important in communicating these data to the end customers. If the amount of information is small then it is able to be communicated through graphs or charts. But because the COVID-19 is handling quantitative records in a huge amount, data visualization must be accomplished in a comforting way of making sense and comprehensible in a simple manner. Basically data visualization as name indicates its visualizing data. The data is simply words, if we do no longer convey it in a comprehensible manner. Hence data visualization is a way to communicate and make feel of these data. In case of COVID-19, data visualization does now not simply performs a position of verbal exchange, it allows convince the humans to change their behavior. As the virus is spreading in a rapid manner, visualized data is cent percentage better than that of mere words.

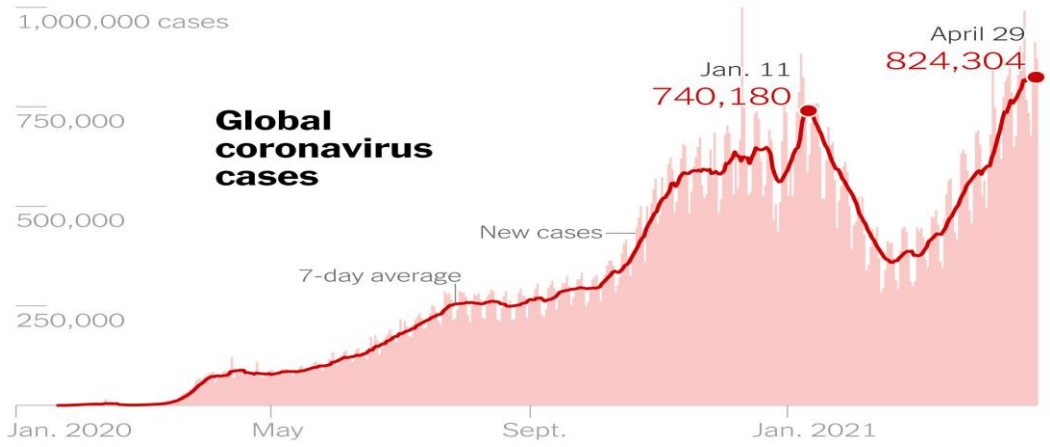


Figure 2. Graph of COVID 19 cases

VI. Big Data Visualization Tools for COVID-19

Data visualization tools enables the visualization designers to create visual illustration of large data units. While managing this data units which incorporates loads of hundreds or hundreds of thousands of data points for the system of creating a visualization, to an awesome extend make a designer's activity less difficult. These data visualization can be used for purposes like dashboards, facts reviews, income and advertising and marketing substances, investor slide decks and anyplace assist the information desires to be interpreted. There are positive not unusual functions for the great visualization gear. It consists of the ease in their use, highquality documentation, required tutorials, and also should be designed in this kind of manner that it feels intuitive to the user. The best visualization tool can manage more than one facts data in a single visual. The output of the tools may be specific charts, graphs and map types. Also the price connected to a tool have to be justifiable in terms of better support, better functions and with better values.

Visualization tools defers for every instance. The choice of the perfect visualization tool relies upon on what the requirement is. There are a whole lot of visualization tool to be had which can be Tableau, Google chart, Sisense, Fussioncharts, Qlik, Power BI, Domo, Polymaps and so forth. The data acquired from the analysis of COVID-19 are giant in different types. These data wishes to be carefully analysed, interpreted and visualized to be able to derive a proper conclusion from the data. Though there is quite a few visualization tools available the most famous tools are Sisense, Qlik Sense, Tableau and Power BI. In my study evaluating these tools to find the fine one appropriate for visualization of COVID data.

VII.Sisense

Sisense is the fastest software when compared to other as there is in chip memory available. It works well with larger volume of data. It helps in combining data from different sources. But while considering the covid data visualization it has some drawbacks.

- Gets slow when the data starts to grow very big.
- It is used mainly for data analysis rather than data visualization.
- Visualizations are difficult for common people to understand.

VIII. Qlik Sense

Qlik is a simple and interactive data visualisation tool which enable users to import and aggregate data from varied big data sources. It has an in-memory data storage of about 500MB. It is a self service Analytics. Qlik can manage small or big data within a single environment. But while considering the COVID data visualization it has some drawbacks. • Inflexible data extraction capabilities.

- Data solution is generally sluggish when working with large data sets.
- Very limited number of visualization types and each one has limited configuration (charts, styles, colours).

IX. Tableau

Tableau is a effective and rapid developing data visualization tool used within the Business Intelligence Industry. It enables in simplifying raw data in a very effortlessly comprehensible layout which will be understood by using professionals at any level in a commercial enterprise. It also allows non-technical users to create customized dashboards. Data analysis could be very rapid with Tableau and the visualization created are within the form of dashboards and worksheets. Data blending, real time evaluation and collaboration of data are the main capabilities of tableau. The tool has garnered interest among the people from all sectors along with commercial enterprise, researchers, different industries, and many others.

a. Pros

- Remarkable Visualization Capabilities.
- Ease of use

- High performance
- Multiple data source connections
- Thriving community and forum
- Mobile friendliness

b. Cons

- High cost
- Inflexible pricing
- Poor after sales support
- Security issues
- IT assistance for proper use
- Poor versioning

X. Current COVID-19 Report Visualized Using Tableau

Name	Casescumulative total	Cases newly reported in last 2 hours	Deaths-cumulative total	Deaths -newly reported in last 2 hours
Glob al	174,520,6 86	424,78 5	3,770,38 1	11,47 9
India	29,274,82 3	91,702	363,079	3,403

Table 1. COVID 19 report

The current covid report (June 11, 2021) as per WHO is given in the table format. As visuals can convey the ideas more clearly it can be represented using tableau. It shows the COVID cases from the starting till the date. It also provides the graphical representation of Death rate.

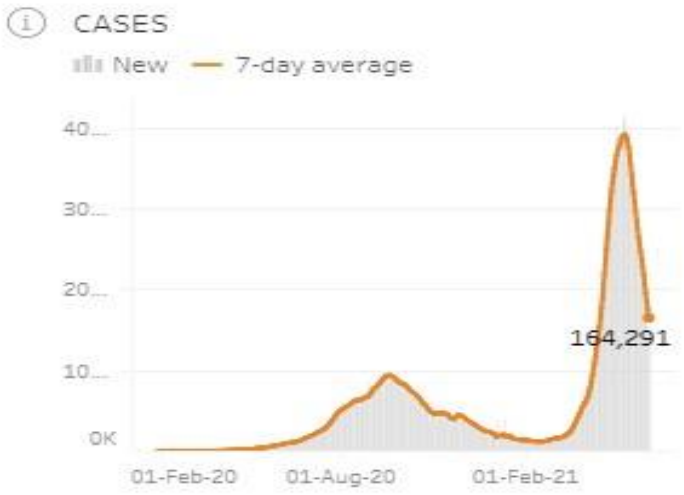


Figure 3. COVID cases

Figure 4. COVID deaths

XI. Power Bi

Power BI is a cloud based business intelligence provider suite by Microsoft. It is used to convert raw data into meaningful facts via the usage of intuitive visualizations and tables. One can effortlessly analyze facts and make crucial enterprise decision based totally on it. Power BI had sure capabilities for records visualization and evaluation through making sharable reports, dashboards, and apps.

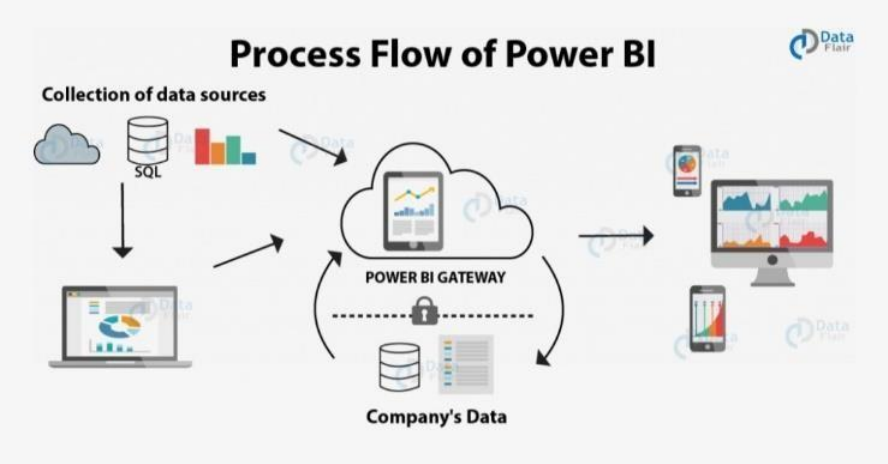


Figure 5. Process flow in Power BI

a. Pros

- Affordability
- Custom visualization
- Excel integration
- Data connectivity
- Data accessibility
- Interactive visualization

b. Cons

- Tables with complex relationships are difficult to handle.
- Less configuration of visuals.
- Rigid formulas
- Handling of large data volume is difficult
- Complex to understand and master.

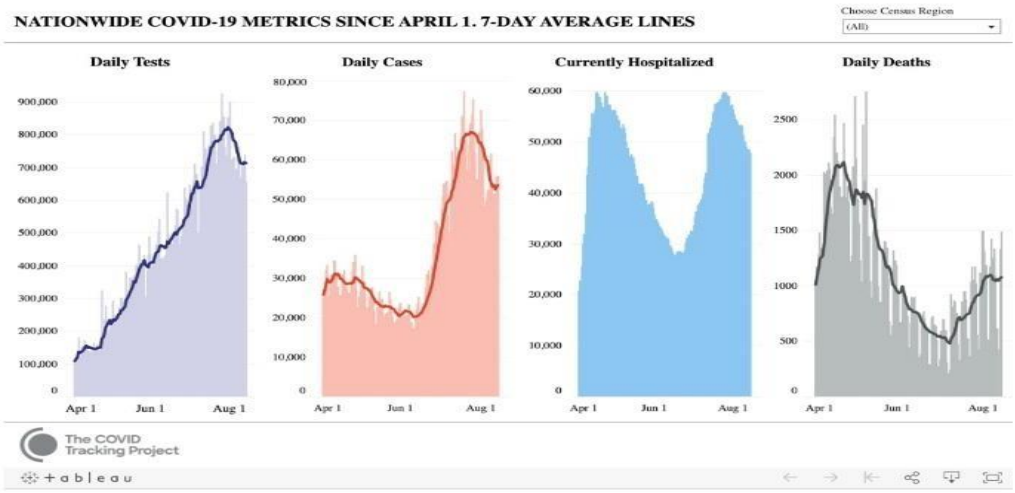


Figure 6. Graph of COVID Cases

The graph created using tableau shows the details of the comparison of daily confirmed covid cases, daily tests, currently hospitalized and death rate for 3 months. This visualization provides a better way to understand the data which in fact gives awareness to common people.

XII. What is Needed for a Good Tool ?

a. Every visualization tool is anticipated to transform raw records into eye beautiful charts and graphs and allow them to bring the hidden message in the records.

b. Tool should additionally have the functionality to examine and present the data in a digestible way.

c. The tool must be suitable in creating understandable data reports and dashboards by acquiring and aggregating massive and complex data from different sources.

d. Tool should be designed in a way that every user regardless of ability can learn to use it.

e. Filtering, processing etc of massive data need to be smooth.

f. The visualization have to be plenty and intuitive for any technical or non-technical person to apprehend it and draw significant insights from it.

g. Considering these elements Power BI and Tableau works hand in hand. The assessment of both these gives an idea approximately which one suits fine for covid data visualization.

XIII. Tableau or Power BI Suit Best for COVID-19

Tableau	Power BI
Tableau can handle a huge amount of data with better performance.	Power BI can handle a small amount of data.
24 different types of data visualizations are available for the users using Tableau.	Power BI provides huge data points to offer data visualization. It is offering quite 3500 data points for drilling down dataset.
Tableau suits the best for large amount of data found in the cloud.	Works better with a massive amount of data.
Used by analysts and experienced users.	Used by naïve and experienced users.
Can connect to numerous data sources.	Connects limited data sources
Has excellent customer support. For discussion tableau has a large community forum.	It provides less customer support to the one who uses it with a free power BI account.

Table 2. Comparison of Tableau and Power BI

- Tableau can cope with huge quantity of data with higher performance while power bi can cope with a limited volume of data.

- Tableau works pleasant when there is a huge data in the cloud but power bi doesn't work higher with large quantity of data.

- Since the data associated with COVID 19 is of massive extent it is able to be handled correctly using tableau. Tableau can create almost any sort of visualization with their platform, from a simple chart to innovative and interactive visualizations. Tableau is a high-quality choice for developing maps in addition to different sorts of charts. Covid data is collected specifically via path maps, the usage of GPS tracking etc. So for the evaluation and visualization of those sorts of data tableau will be the best choice.

Conclusion

The persisted spread of corona virus disease has affected the world terribly. A tremendous quantity of data iscollected regarding it to control the unfold of disease. These data may be treated with the assist of Big data tools. It appears that the present day human is far greater superior as to be stricken by a colourful eye-catching picture. Psychological research shows that 90% of all the information that people understand comes from their sense of sight. So for the right visualization and analysis of covid, Big data visualization tools can be used. Among the more than one Big data visualization tools to be had, there may be one answer that honestly sticks out first-rate for COVID data- “Tableau”, which enables to represent this sizeable data in so many visuals through which even common people get awareness of the disease.

Acknowledgment

I would like to thank my guide Ms. Remya Raveendran for her proper guidance and support for helping me to complete this research paper.

References

[1] C.K. Leung, "Uncertain frequent pattern mining", Frequent Pattern Mining, pp. 417-453, 2014.

[2] C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, Y. Wen and A. Cuzzocrea, "Big Data Visualization and Visual Analytics of COVID-19 Data," 2020 24th International Conference Information Visualisation (IV), 2020, pp. 415-420, doi: 10.1109/IV51561.2020.00073.

[3] Shikah J. Alsunaidi , Abdullah M. Almuhaideb, Nehad M. Ibrahim, Fatema S. Shaikh, Kawther S. Alqudaihi, Fahd A. Alhaidari, Irfan Ullah Khan, Nida Aslam & Mohammed S. Alshahrani, “Applications of Big Data Analytics to Control COVID-19 Pandemic” Sensor 2021,21,2282. https//doi.org/10.3390/s21072282.

[4] Nicola Luigi Bragazzi , Haijiang Dai , Giovanni Damiani, Masoud Behzadifar, Mariano Martini & Jianhong Wu (2020),” How Big Data & Artificial Intelligence Can Help Better Manage the COVID-19 Pandemic”, Int. J. Environ. Res. Public Health 2020,17,3176:doi.10.3390/ijerph17093176.

[5] Althaf Rahaman.S, Sai Rajesh, .Girija Rani“Challenging tools on Research Issues in Big DataAnalytics”(2018), Department of Computer Science, GITAM (Deemed to be University), Visakhapatnam, India.

[6] Nada Elgendy & Ahmed Elragal, “Big Data Analytics: A Literature Review Paper, Department of Business Informatics & Operations”, German University in Cairo (GUC), Cairo, Egypt, Springer International Publishig Switzerland 2014.

[7] Ritu Ratra, Preeti Gulia, “Big Data Tools & Techniques: A Roadmap for Predictive Analytic”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-2, December, 2019 4986 Published by: Blue Eyes Intelligence Engineering & Sciences Publication.

[8] Yubo Chen, Carson K. Leung, Siyuan Shang, Qi Wen, "Temporal Data Analytics on COVID-19 Data with Ubiquitous Computing", Parallel & Distributed Processing with Applications Big Data & Cloud Computing Sustainable Computing & Communications Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom) 2020 IEEE Intl Conf on, pp. 958-965, 2020.

[9] Brüssow, H. The Novel Coronavirus—A Snapshot of Current Knowledge. Microb. Biotechnol. 2020, 13,607–612. [CrossRef]

[10] Wang CJ, Ng CY, Brook RH. Response to COVID- 19 in Taiwan: big data analytics, new techology, & proactive testing. JAMA. 2020 doi:10.1001/ jama.2020.3151. [PubMed][Cross][Google Scholar].

[11] W. Jentner, D.A. Keim, “Visualization and visual analytic techniques forpatterns”, in High-Utility Pattern Mining, 2019, pp. 303-337.

Statement about the ownership and other particulars about newspapers
JOURNAL OF THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

(To be published in the first issue every year after the last day of February)

FORM IV

(See Rule 8)

1.	Place of the Publication	: The Maharaja Sayajirao University of Baroda The Maharaja Sayajirao University of Baroda, Vadodara - 390 002.
2.	Periodicity of its Publication	: Science & Technology two times in a year.
3.	Printer's Name	: Shri Jatin H. Somani
	Nationality	: Indian
	Address	: Manager, The Maharaja Sayajirao University of Baroda Press, (Sadhana Press), Vadodara - 390 001.
4.	Publisher's Name	: Prof. C. N. Murthy
	Nationality	: Indian
	Address	: The Maharaja Sayajirao University of Baroda, Vadodara - 390 002.
5.	Editor's Name	: Prof. C. N. Murthy
	Nationality	: Indian
	Address	: The Maharaja Sayajirao University of Baroda, Vadodara - 390 002.
6.	Names & Addresses of Individuals who own the newspaper and partners or shareholders holding more than one percent of the total capital	: The Maharaja Sayajirao University of Broda, Vadodara - 390 002.

I, C. N. Murthy declare that the particulars given above are true to the best of my knowledge and belief.

Prof. C. N. Murthy

JOURNAL OF THE MAHARAJA SAYAJIRAO UNIVERSITY OF BARODA

The Journal is mainly intended to publish original research papers contributed by the teachers and research scholars of The Maharaja Sayajirao University of Baroda after undergoing a peer review process. All submitted articles are subject to anti plagiarism check before being sent for review. Proceedings of Conferences held within The Maharaja Sayajirao University of Baroda are also accepted for publication subject to the individual articles being peer reviewed.

The Journal is issued every year in three parts. These parts are devoted respectively to topics relating to 1. Science & Technology, 2. Humanities and 3. Social Science. All manuscripts for review to the Science & Technology part should be addressed to :

The Editor (Science & Technology),
Journal of The M.S. University of Baroda,
A.I.C.S. Training Centre,
Vadodara - 390 002.

email: editor.msujst@gmail.com

Contributors are requested to submit manuscripts, by email attachment which should be typed with double spacing. Authors are advised to retain a copy of the paper that they send for publication as the Editors cannot accept responsibility for any loss of manuscripts not accepted for publication in the Journal.

Authors of papers printed in the Journal are entitled to one copy of the issue free of charge.

ADVERTISEMENT

All advertising matters should be sent so as to reach the office of the Journal a month before the publication of the Journal. Further information on rates and space can be had on writing to **editor.msujst@gmail.com.**

Printed at : The M.S. University of Baroda Press (Sadhana Press), Palace Road, Vadodara - 390 001.
MSUP-5027-200-22

"The full-blown lotus growing out of the lake symbolises the emergence of the mind and its triumph over matter. The flame rising from the center of the lotus is the flame of the human knowledge, spreading light and learning for the coming generations. The motto inscribed below the lotus defines the purpose and existence of life which is love of beauty, goodness and intellectual curiosity."

महाराजा सयाजीराव विश्वविद्यालय गीत

अमे वडोदराना विद्यापीठनां सपनां सारवनारा
अमे ज्योत जलावी सृष्टि नवली सहसा सर्जनहारा.

अमे गगनकुसुम कर धरनारा
अमे मगन मगन थई फरनारा
अगनबाथ अमे भरनारा
अमे दैन्यतिमिरने हरनारा.

श्री सयाजी विद्यापीठना ज्ञानदीपने धरनारा
सत्यं शिवं सुन्दरम् नो मंत्र अनंतर भणनारा.

Printed at: The Maharaja Sayajirao University of Baroda Press (Sadhana Press), Vadodara-390 001