

Design of a Decision Making Model for Integrating Dark Data from Hybrid Sectors

Sonam V Maju¹ and O.S Gnana Prakasi²

¹Doctoral Research Scholar, Computer Science and Engineering Department, Christ (Deemed to be University), Bangalore
Email: Sonam.maju@res.christuniversity.in

²Assistant Professor, Computer Science and Engineering Department, Christ (Deemed to be University), Bangalore
Email: gnana.prakasi@christuniversity.in

Abstract—The research on Dark data, from its definition to identification and utilization is a widely identified and encountered research problem since 2012 when Gartner defined Dark data as every possible information that an organization collects, process, analyze and store throughout regular business activities, but usually fails to make use of the stored information for other suitable purposes. The presence of dark data and its impact has been experienced by every sector, these data occupy large storage and remain unused. In this paper, we analyze Dark Data and proposed a design model to utilize dark data from multiple sectors and providing a solution to any critical situation a person might be in. For eg: Multiple cash transactions from an organizational bank account in a hospital successively over a period of 2-3 days may indicate a health emergency of any particular employee from that organization. Thus we are considering institutional data, medical data, and banking data in which machine learning algorithms can contribute huge changes in the current system and can help the decision-makers to make better decisions. The paper also proposes a few techniques and methods for the conversion of unstructured dark data to structured one and some extraction techniques for data using NLP and Machine Learning.

Index Terms— Dark data, Natural Language Processing, Entity based Recognition, Electronic Health Record.

I. INTRODUCTION

In July 2012, a researcher at Gartner mentioned the term Dark Data in a blog post. This was the first time the term 'Dark data' was used and hence Gartner coined the term dark data [1] They note that dark data is processed and stored "just in case it might come in handy" at a later point. In order to make maximum utilization of dark data, it's very important that the users of every business organization should ask the right questions at the right time. In subsequent literature, Gartner is credited with coining the term. An enormous amount of discussions happened on dark data after that and the research work under the same is in high demand even now. The initial challenge was to identify the presence of Dark data in different sectors and applications, then come the Identification of dark data from a pool of data and finally utilization of dark data for better decision making. Dark data is there everywhere around us from our smartwatches to small object detection sensors, from personal computers to institutional HPC, from hospitals to banks. This can justify what IBM claims [2] in the press conference that 80% of the data around us is dark data and not all the collected IoT sensor data is used for the

analysis. Stating Gartner's explanation on dark data "The information that every organization collect, process and store during regular business activities, but generally fail to use for other useful purposes".

George W Bush, ex-President of USA, in his 2004 State of the Union Address, broached the idea of computerizing the health records. The health sector has completely upgraded after that, everything changed inside out, the concept of Electronic Health Records not only made the clinical documentation a structured one but this introduction of EHR opened great and enormous paths to the researchers also. There are still more aspects that need to be explored in the EHR to make our life more convenient. Under the HIPPA regulations, the clinical data need to be stored and secured for years, so the amount of dark data in healthcare databases consumes a large amount of storage space. All the health monitoring sensor readings will contribute the collection of additional data from patients, doctors, clinics, healthcare centers and hospitals. Along with the monitor sensors, several devices are designed and used to give reminders to patients to for correcting their posture, taking their medicine, and even act as personal assistants. The information from the radiology images will not always be extracted and stored, but the images will be there in the storage and thus the data from these images will act as dark data. Hence, The amount of dark data which is present in the health care domain is very huge as there are multiple wearable smart devices to monitor each parameter of the patient, the smart pillbox, the sensors used for patient monitoring, and finally the structured but unused information in the EHR all these will contribute a vast amount of dark data making health sector as a prominent sector when it comes to the presence of dark data.

The finance sector is another protruding sector that contributes to the increasing amount of dark data by the number of emails, notification messages, advertisement messages, feedbacks collected by the users, and a lot more. There will be a lot of information in the customer feedbacks and the social media reviews, which can actually help the marketing department and the customer care section. Along with that tracking, the purchase details of the customer on a regular basis can actually be an advantage for both the bank as well as the customers. All these are collected, send, and saved hoping that these will come in handy in the future. Even all the transaction information which is highly confidential can be used for other purposes which at the end of the day can be used for the betterment of human life. Multiple usages of the same collected and processed information need to be fished out. Especially in the bank sector where almost all information of the customer is readily available, considering the privacy and confidentiality as prime factors majority of other associations can be done and better services can be provided, that too with the same collected and stored information.

Identification and utilization of dark data from the pool of data is a very challenging task. Not only challenging but is very important to treat Dark data as a top priority because of the escalating amount of data growth and in no time there will not be sufficient space for our high-priority data. In addition to that, the number of money companies and organizations are spending on the storage options is on a high scale and the interesting part is that majority of the storage spaces are occupied by redundant, trivial, and dark data.

The opportunity and possibilities of utilization of dark data in Hybrid Sector Applications are in trend among researchers. As the scope of interdisciplinary researches shoots up the scope of the Hybrid Sector Applications will always be one area that deserves to be explored more. In this paper, for the utilization of dark data, the health and Finance sector is considered and a model is also proposed by connecting the two sectors on the medical emergency financial backgrounds.

Hence, the introduction covered all the possibilities, derivations, and impacts of dark data. The second section, Literature Survey explains the importance of identification of dark data and utilization of the same by considering different sectors and applications through various research articles and research works. The third section explains the need for the integration of dark data from the different sectors for better decision-making. The fourth section introduces a new model for the integration of dark data from different sectors under an application perspective. The fifth section elaborates few techniques that can be used for the implementation and data analysis. The final section is the conclusion that wraps the overall motive of this research work.

II. LITERATURE SURVEY

David J. Hand contributes a glimpse of what dark data is and how it can be utilized, and the terrific results obtained while implementing on applications [3]. The article gives a clear explanation of dark data with genuine models like Coronavirus 19, Autism, and the 1986 Challenger Space Shuttle catastrophe. The article has referenced the presence of dark data in medical services and clarified the prospects of examination in obscurity information medical care. A few investigations are embraced without the adequate idea being given to the beginning and provenance of the information.

The article [4] characterizes the dark data within an organization as,

- i. Data exists but hard to recover or utilize.
- ii. Data is disregarded, never considered. These data may occupy huge storage but may not be used. [5] Quotes the valuable of dark in the EHR or the Electronic Medical Record (EMR) to monitor a patient's present condition, to record diagnostic evidence, procedures performed on the patient, and past treatment results. EHR is one of the valuable resources for largescale medical data analysis. The progress of technology and hospital information system (HIS) is widely promoting and depending on these EMR/HER. [6] Classifies the treasure for large-scale medical analysis of clinical/health data into three formats namely: Structured data in a particular format with easy accessibility, Semi-Structured data which includes file description, flowcharts, and unstructured data with no particular format.

[7] A myriad of data will be collected by the Healthcare organizations in the course of regular business activities and Hyper-personalization is implemented and developed in many industries, particularly consumer-centric sectors by utilizing these dark data. [8] shows dark data and hyper-personalization in healthcare with examples and few statistics and finally some solutions like, Centralization of Person/Patient Data, Centralization of Person/Patient Data, Data Security and information protection and data mining/Analytics are explained. [9] eSCRIBE templates: patient-specific reports (prescription, setup, follow-up, ...) together with dedicated VBA for Word code are used with a twofold purpose: i) to process and visualize the information stored as patient custom properties; and ii) to store the data in real-time generated during the RT process for each patient as structured JSON files, turning "dark-data" into "structured data".

Finance sector is having more dark data when compared to other possible sectors and in banking sector, dark data is identified in all forms like mail content, transaction reports, social media content, and the banking research data, etc. Dark data is mostly unstructured and is very difficult to analyze but when its analyzed by structuring the content, the benefits it provides for banking and also for retaining its customers are commendable. The data stored in banks can be categorized into two, structured and unstructured. The highly organized structured data like, the information entered in a prearranged layout such as account holder name, date, account number, etc. Unstructured data doesn't have any predefined formats. Like the customer feedback emails, social media reviews, etc. [10] The paper examines dark data, its significance, hazards included, the necessity and procedures to convert the unstructured dark data to structured section of useful information. It examines the uses of dark data in the bank sector by examining and converting unstructured into structured data. Also, the authors are clarifying various strategies of Natural Language Processing for the conversion of unstructured to structured data. The paper shows the execution of Natural Language Processing methods like Named Entity Recognition - to categorize components of input text natural language and Topic Modelling to recognize the theme around which the text revolves.

The paper concludes by stating that Named Entity Recognition will work perfectly for identifying the entities and in order to identify the intent and hidden patterns Topic modelling can be used. Researches on the Utilization of Dark data by considering two sectors are very rare and is the importance of this paper. The next section will explain the proposed model of the same at the application level.

III. NECESSITY OF INTEGRATING DARK DATA FROM DIFFERENT SECTORS

Employees regularly find it difficult to bear the entire cost during medical emergencies. Hence, companies do provide medical health policies. The health insurance policies providing by the employers can really help their staff in case of an accident, disease, or any similar health issues. Compensation is not just the wages you get through the pay check. It can also include health coverage, as well as other benefits or other taxes or services paid by your employer. [11] A survey has been conducted among the Indian urban youth asking about the life in lockdown, most of the employed respondents specified about the health insurance they obtained through their employer as of April 2020. Almost 14 percent of the part-time workers and interns will be having private health insurance. While, nearly 39% of employees working on full time mode had health insurance provided by the companies where they work. The above information is available in Statista Research Department, Oct 16, 2020. Also, [12] medical insurance for employees is compulsory in India to post the nationwide COVID-19 lockdown in 2020. As per the circular released on April 1, 2020 which is issued by the Insurance Regulatory and Development Authority of India (IRDAI), it is mandatory to have a medical insurance. Corporations were requested to monitor and follow a Standard Operating Procedure post lockdown in order to resume work, and offering insurance to the workforce was a part of it.

The banking industry is engendering a large amount of data on the day to day basis, where formerly this large amount of collected data was not properly used in banks. But presently, banks are able to utilize this data to guide the customers and thereby extenuating the important objective of marketing. [13] the article discussed on tracking the money movements and thus helping the banks to avert major disasters and frauds thereby helps to understand customer behaviour. The major adopters of big data technologies are the above mentioned banking and financial industry. International Banks have started to hitch the power of data to originate efficacy across different sections of their functioning. Big data can be considered as a tool which helps an organization to identify, utilize, manipulate, and manage large amount of data in the financial Big data analytics industry in a stipulated time and helps in the management of storage medium that chains such capacious data. Enhanced reporting can be done by getting access to a huge volume of data, which contains different needs of diverse customers. At that point, the Banks will offer their requirements in a evocative way. The Bank sector delivers the precise information required by the customer by using big data.

Apart from banking sector, there are several other sectors where the impacts of dark data can be studied. The health sector is another impact area of dark data where identification and utilization of dark data can contribute massively to the early prediction of diseases and also to identify the association between the diseases. For example: “Numerous cash transactions from an organizational bank account in a hospital consecutively over a period of 2-3 consecutive days may indicate a health emergency of any particular employee from that organization”. The example mentioned in the previous paragraph is just one example of many where dark data can be utilized for other beneficial purposes. In this, when an employee does a money transaction several times successively in consecutive days, the bank can trigger an organization with the need of financial assistance to that employee if needed. So in this design model, we integrated the dark data from the Banking sector and organization sector to help the employees.

Figure 1 represents the types of dark data prevailing in various organization sectors described in this research work. It can be observed that there is a lot of data which may presently be of no use but may prove beneficial in future.

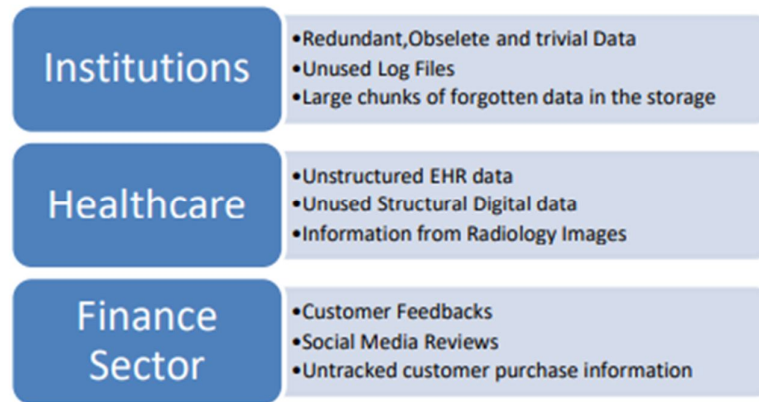


Figure 1. Description of dark data

IV. PROPOSED DESIGN MODEL

The following is the design of the proposed model at the application level. This design model is an example of amalgamating the health insurance perspective from an organizational side along with the Finance sector and Health sector. The institutional dark data, which contains the personal details of every employee as well as the log files of every employee and the salary credentials/bank statements. In the Bank sector, in which even though all data are processed and secured, analysis of transaction details can be considered as the dark data used in this application. When it comes to the health sector, dark data utilization is the data collected at the time of payment especially through debit cards and UPIs. Compared to cash transactions, the digital payments process consumes more storage space and at the end of the day, this data is not properly analyzed and used. By identifying and utilizing these dark data present in these sectors, the quality of the decision-making process can be upgraded to the next level.

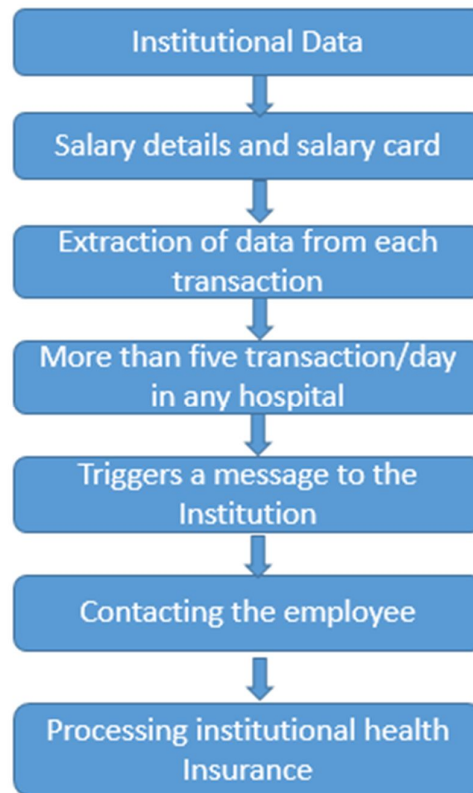


Figure 2. Flow of the design model

By integrating the dark data present in the institutional account, its transactions, and corresponding Bank along with the hospital Electronic Health records, pharmacy, and other bills, we can support an employee in medical aid physically, mentally, and more importantly financially. The institutional data comprises of both official and personal information including the confidential credentials of each employee for the formality purposes. These data can be used for helping and supporting an employee, by extracting the information and using it in different applications with the employee's permission. Of all such data. Here in this design model, the money transaction of each employee using his salary credit/debit card is tracked and is used for identifying medical emergencies. Once a medical emergency is identified, the necessary help and support can be given to the employee without any external intervention. The detailed description of the design model along with the technical explanation is provided in the following paragraphs.

Every time an employee does a money transaction, the bank can extract and categorize information using Natural language processing. The Entity Named Recognition can be used for labelling of data and classification of Information. Named Entity Recognition algorithm can be used to identify the relevant keywords/nouns from an input string. Different frameworks and algorithms are available to extract the relevant information using Machine Learning and Deep Learning. Various frameworks are available that supports deep learning NLP using Python, SpaCy, Keras, Tensor-Flow, and Theano, etc.. One can also use the Natural Language Toolkit to implement NER using traditional ML methods.

Once an employee does a money transaction several times successively in consecutive days, the corresponding bank employee can trigger an update message to the institution stating the status quo and the hospital transaction summary. The respective admins of the institution can check the transaction statements and can personally contact the employee and check with the where about. Suppose if the patient is in need of financial assistance, and normally a national insurance scheme will take days to sanction the amount. In this case the institution can process the institutional health insurance fund and transfer the same to the employee

The application of combining two independent sectors for the utilization of dark data is very much challenging. Institutional Data itself is a pool of dark data right from the log files to payroll. Identifying and Extracting Dark data from the institutional level can start by converting all unstructured data to structured format. The next section shows few techniques for extraction of the dark data from the pool of the data.

V. EXISTING TECHNIQUES HELPS TO EXTRACT AND UTILIZE DARK DATA

[14] In Deep Dive, the framework accepts unstructured information and creates an organized yield table, the entire cycle of change from unstructured to organized, applying a user-defined function to every document in order to get the candidate extraction and for the feature extraction, a User-defined feature function id is also applied to the candidate. Labelling of candidates for obtaining the output table and for that user-defined distant supervision rules are used. Finally, a graphical method is used to calculate the probability of each candidate. Once the data is structured then comes the extraction of information and inbuilt NLP and text mining are the two methods that can be used for the same.

The huge amount of data and its complexity of unstructured data to open up numerous new opportunities for the analyst. For the knowledge discovery and extraction of information from the textual context, there are two widely used methods are Text mining and natural language processing [15] Usually, text mining is the process of extracting valuable and useful patterns from text data, whereas NLP technologies are used as knowledge discovery. NLP techniques provides text classification, categorization, clustering of documents, identifying similar documents, information extraction, summarization, etc. Research area of text mining for identifying patterns in texts which comprises of search and recovery, text analysis, navigation and exploration of documents, extraction of keywords from a group of texts, knowledge management. Though information recovery and additional methods of text mining regularly using word stemming, and many more urbane techniques from Natural Language Processing which have been seldom used.

Text mining attempts to discovering designs in text-based unstructured documents dependent on the substance. NLP attempt to arrive at ideas of writings by means of explicit calculations. Text mining tries to identify forms in textual unstructured files based on the contents. NLP tries to reach concepts of texts through specific algorithms.

[16] The paper provides a complete evaluation of systems and analyses key data management encounters and techniques. The author motivates on three complementary lines of work: (1) integration of Machine Learning algorithms and languages along with the data management systems such as RDBMSs, (2) adapting data management techniques such as query optimization, and compression to new systems that target ML workloads, (3) combining data management and ML algorithms to design systems that improve ML lifecycle-related tasks. The identification and utilization of unstructured data using machine learning and natural language processing. The [17] authors are pointing out five areas for the easy implementation of the technology. The paper explains its objective by considering two case Studies, (i) Automated sales practice identification for increased compliance efficiency and regulatory coverage (ii) NLP/ML assisted workflow for marketing material review. In the first case study, by using the EY approach, the bank achieved a more than 90% capture rate of sales practice issues, reducing the risk of regulatory matters requiring (immediate) attention. In the second case study, the EY approach reduced the review time per document by 95%, and it did more than just minimize the resource requirements.

Conversion of unstructured data to structured data has main three states depicted by [18], a database containing unstructured data, data in its intermediary state, and effectively structured data. The user uploads the resumes and mechanically stored in a database. This database stores all the resumes in text format. This is fed into the system. The method uses Natural Language Tool Kit libraries and approaches to the reoccurrence of structured data to the database. Structuring of the data involves various methods like n-grams that split sentences into n-words, named entity recognition that tags entities like names, geographical location, etc., compiling the data using regular expressions for objects that have regularity like address, email ID, phone numbers, etc. Conversion of unstructured data into a structured form using Natural Language Processing needs to undergo various steps. Creating a web application for file upload, Conversion of file from the format .pdf to .txt, and finally, extracting information.

The paper [19] deliberates the identification of unstructured data, its extraction, and the classification of web pages, and uses Extensible Markup Language (XML) document for transforming these web pages into a structured format, and well ahead kept into a multimedia database. The influence of this research work is in the tactic of seizing and seizing the unstructured data and the effectiveness of a multimedia database to hold the data

of similar kind. The stored data can definitely provide aids to numerous communities like researchers, students, lecturers, and IT professionals because this information can contribute to all kinds of development, decision making, day-to-day operations, and other future prospects.

Once the extraction is also complete, then the flow of the work is more or less needs to be acquired manually. Like contacting the person and enquiring about the where about. The below table represents the abstract form of above technologies.

TABLE I: TECHNOLOGIES FOR UTILIZING DARK DATA

Technology	Method	Application
Deep Dive	Unstructured data is organized, by applying a user-defined function is used to obtain the candidate extraction and for the feature extraction, a User-defined feature function id is also applied to the candidate. Labelling of candidates for obtaining the output table and for that user-defined distant supervision rules are used. Finally, a graphical method is used to calculate the probability of each candidate.	Extraction of Dark Data
Text Mining	Text mining helps in finding patterns from the textual unstructured data based on contents.	Dark data mining
Natural Language Processing	NLP try to extract the concepts of text data via specific algorithms. NLP techniques provides text classification ,by identifying and labelling categories of related documents, information extraction, summarization etc.	Dark data mining
Machine Learning and DBMS	Proposed lines of work are: (1) integrating Machine Learning algorithms with existing data management systems like RDBMSs, (2)adapting different data management techniques like query optimization, and compression of data to new systems that target Machine Learning workloads. (3) Merging data management and ML ideas in order to build efficient systems that improve ML lifecycle-related tasks.	Data Management
Machine Learning and NLP	Classification and categorization through Machine learning and Relation Extraction methods from the bag of words through natural language processing.	Utilization of Dark data
Natural Language Toolkit	Natural Language Tool Kit libraries and approaches to the reoccurrence of structured data to the database. Structuring of the data involves various methods like n-grams that split sentences into n- words, named entity recognition that tags entities like names, geographical location, etc., compiling the data using regular expressions for objects that have regularity like address, email ID, phone numbers, etc	Unstructured to Structured Data

VI. CONCLUSION

Indeed, the data, information, technology and all the services are for the betterment of human living. Understanding the pros, cons and limitations of one sector and solving it using the pros of other sector is what we have used here. In this application it was necessary to focus on employee-employer analytics and the correlation between finance and health sector as it benefits both the sectors equally.

The highlight of the paper is the understanding of one major issue and its hidden solution that every employee has to face and its solution from the already existing resources. The paper shows the example model of utilizing dark data in the hybrid level by integrating the institutional data, medical data and the banking data. In addition, few Machine learning and Natural Language Processing techniques are discussed for the extraction of Dark data from the pool of data. The idea of identifying dark data and utilizing it in an application level is a leap towards utilizing dark data to an employee's improvement in preserving better employee-employer relationship and redefining the meaning of coexistence.

REFERENCES

- [1] <https://devopedia.org/dark-data>
- [2] <https://developer.ibm.com/technologies/analytics/articles/ba-databecomesknowledge-3/>
- [3] D. J. Hand : Dark data, Wiley Online library, The Royal Statistical Society, Significance, vol. 17, no. 3, pp. 42–44, June 2020.
- [4] Vicky Seehusen, Edgar Maldonado : Using a Roadmap in the Back Alleys of Dark Data , Journal of Technology Research , Volume 9, 2020
- [5] <https://blog.westerndigital.com/dark-data-in-healthcare>
iot/#:~:text=This%20type%20of%20data%20is%20what%20we%20call%20dark%20data.&text=As%20sensors%20become%20more%20compact,perspiration%2C%20movement%2C%20or%20diet.
- [6] Wencheng Su, ZhipingCai ,Yangyang Li, Fang Liu, Shengqun Fang, and Guoyan Wang : Data Processing and Text Mining Technologies on Electronic Medical Records: A Review, Journal of Health care Engineering, 2018.
- [7] Western Digital, Accenture : EMBRACING DARK DATA VALUE OF DATA A “Win-Win-Win” Through Dark Data And Hyper-personalization In Healthcare
- [8] Srinivas Iyengar Vice President at TietoEVRY, Technologist, Adjunct Professor at Dayanand Sagar University, Healthcare AI, Population health, Global healthtech : Dark Data is the Holy Grail of Healthcare, CIO Insider
- [9] A. Wals Zurita , H. Miras del Río , A. Bertolet: Walking towards our dark data in real time, XX Congreso, June2019
- [10] K. Munot, N. Mehta, S. Mishra, and B. Khanna: Importance of dark data and its applications, 2019 IEEE Int.Conf. Syst. Comput. Autom. Networking, ICSCAN 2019, pp. 1–6, 2019, doi: 10.1109/ICSCAN.2019.8878789.
- [11] <https://www.statista.com/statistics/1170400/india-employee-health-insurance-by-employment-type/>
- [12] <https://www.acko.com/group-health-insurance/is-gmc-policy-mandatory-for-employees-in-india/>
- [13] Pritee Chunarkar-Patil, Akshanda Bhosale: Big data analytics, Rajiv Gandhi Institute of IT and Biotechnology, Volume 2 Issue 5 – 2018
- [14] Ce Zhang, Jaeho Shin, Christopher Ré Stanford University Palo Alto: Extracting Databases from Dark Data with DeepDive, CA Michael Cafarella University of Michigan Ann Arbor, MI Feng Niu Lattice Data, Inc. Palo Alto,CA,
- [15] F. S. Gharehchopogh Z. A. Khalifelu: Analysis and Evaluation of Unstructured Data: Text Mining versus Natural Language Processing. IEEE,2011
- [16] Data Management in Machine Learning: Challenges, Techniques, and Systems, SIGMOD’17, May 14-19, 2017, Chicago, IL, USA c 2017 ACM. DOI: <http://dx.doi.org/10.1145/3035918.3054775>
- [17] Sameer Gupta, Darrin Williams, Sukumar Pant, Talia Visaggi : Harnessing insights in unstructured data through NLP and ML, Ernst & Young LLP
- [18] Sivaramakrishnan N, Vandana V, Vishali M, Dharshana S G, Subramaniaswamy V and Umamakeswari A : CONVERSION OF UNSTRUCTURED DATA TO STRUCTURED DATA WITH A PROFILE HANDLING APPLICATION
- [19] Siti Z. Z. Abidin, Noorazida Mohd Idris and Azizul H. Husain Faculty of Computer and Mathematical Sciences, Malaysia : Extraction and Classification of Unstructured Data in WebPages for Structured Multimedia Database via XML, IEEE,2010