Quality Assessment of Big Data

[1]Dikshit Kotian, [2]Mr.Manjunath H R, [3]Dhanush Kumar, [4]Ashwitha Salian, [5]Deeksha Hegde

[1,3,4,5]Student, [2]Associate Professor
Alva's Institute of Engineering and Technology

***Abstract*: Due to the tremendous volume of generated and processed data from various application domains, Big Data has gained enormous momentum over the past few years. It is estimated that 80% of all data generated is unstructured nowadays. It has been identified that evaluating the quality of Big Data is essential to guarantee dimensions of data quality including, for example, completeness and accuracy. Current initiatives are still under investigation for unstructured data quality assessment. In this paper, we propose a model for quality assessment to handle Unstructured Big Data (UBD) quality. A repository of data quality manages relationships between dimensions of data quality, quality metrics, features methods of extraction, methods of mining, types of data and domains of data. A sample analysis provides an UBD data profile. UBD Quality Assessment Model that handles all processes from UBD to Quality Report profiling exploration. The model provides an initial blueprint for unstructured Big data quality estimation.**

***Keywords*: Big data, quality Assessment, survey**

Introduction

Big data is commonly defined as the way we collect, store, manipulate, analyze, and gain insight from heterogeneous data that is rapidly increasing. Most of the new generated data is unstructured due to the increase of unlimited mobile and human generated data from social media that in an unstructured way combines text, pictures, audio, video. Unstructured data is a fast-growing phenomenon, say industry analysts, compared to all other data types. According to a survey carried out by [1], it will increase by as much as 800 percent over the next five years.

Assessment of Big Data Quality is an important phase integrated into pre-processing data. It is a phase in which data is prepared according to the requirements of the user or application. When the data is well defined with a schema or a tabular format, it becomes easier to evaluate its quality as the data description helps to map the attributes to quality dimensions.

Big Data Overview

Big data values

McKinsey & Company observed how big data created values following in-depth research on U.S. healthcare, EU public sector management, U.S. retail, global manufacturing, and global personal location data. Through research on the five core industries representing the global economy, the McKinsey report highlighted that big data could give full play to the economic function, enhance the productivity and competitiveness of businesses and public sectors, and create huge benefits for consumers. McKinsey summarized the values big data could create: if big data could be used creatively and effectively to improve efficiency and quality, the potential value of the U.S. medical industry gained through data could exceed USD 300 billion, thereby reducing U.S. spending.

Challenges of Big Data

In the big data era, the sharply increasing data deluge poses enormous challenges in data acquisition, storage, management, and analysis. Traditional systems of data management and analysis are based on the system of relational database management (RDBMS). Such RDBMSs, however, apply only to structured data other than semi-structured or unstructured data. Furthermore, more and more expensive hardware is being used by RDBMSs.

Some literature[4-6] discuss obstacles in the development of big data applications. The key challenges are listed as follows:
Data representation: many datasets have heterogeneity levels in type, structure, semiconductor, organization, granularity, and accessibility. The purpose of data representation is to make data more relevant to computer analysis and user interpretation. An improper representation of data, however, will reduce the value of the original data and may even hinder effective data analysis.

Redundancy reduction and compression of data: in datasets there is generally a high level of redundancy. Redundancy reduction and data compression are effective in reducing the entire system's indirect costs on the assumption that the data's potential values are not affected.

Data life cycle management: Data is generated at unprecedented rates and scales compared to the relatively slow progress of storage systems, pervasive sensing and computing. We face many pressing challenges, one of which is that such massive data could not be supported by the current storage system.

Analytical mechanism: masses of heterogeneous data are processed within a limited time by the analytical system of big data. Traditional RDBMSs, however, are designed strictly with a lack of scalability and expandability that could not meet the performance requirements.

Data confidentiality: Because of their limited capacity, most large data service providers or owners currently can not effectively maintain and analyze such huge data sets. To analyze such data, they need to rely on professionals or tools to increase potential safety risks.

Energy management: Mainframe computing systems ' energy consumption has drawn much attention from both economic and environmental perspectives. With increased data volume and analytical demands, big data processing, storage and transmission will inevitably consume more and more electricity.

Expendability and scalability: Big data analytics system must support current and future datasets. The analytical algorithm must be able to process ever more complicated and expanding datasets.

Cooperation: Big data analysis is an interdisciplinary research that requires experts working together in different fields to harvest big data's potential. A comprehensive architecture of the big data network must be established to help scientists and engineers in different fields access different types of data and make full use of their expertise to work together to achieve the analytical goals.

Big Data Lifecycle

PHASE I: DATA GENERATION

BUSINESS DATA: IT and digital data use has been instrumental in boosting the business sector's pro-accessibility for decades. All companies estimate the volume of business data worldwide to double every 1.2 years. Internet business transactions, including business-to-business transactions and business-to-consumer transactions, will reach 450 billion a day. The ever-increasing volume of business data requires more effective real-time analysis in order to gain additional benefits. Amazon handles millions of back-end operations and queries from over half a million third-party sellers every day, for example. More than 1 million customer transactions are handled by Walmart every hour.

NETWORKING DATA: Networking has penetrated human lives in every possible way, including the Internet, the mobile network and the Internet of Things. Typical network applications, considered big data sources on the network, include, but are not limited to, searching, SNS, websites, and clicking streams. At record speeds, these sources generate data, demanding advanced technologies. For instance, in 2008, Google, a representative search engine, handled 20 PBs a day. Facebook stored, accessed, and analyzed more than 30 PBs of user-generated data for social network applications.

SCIENTIFIC DATA: Very large datasets are being generated by more and more science-based applications, and the development of several disciplines relies heavily on the analysis of massive data

DATA ATTRIBUTES: Heterogeneous data with unprecedented complexity are generated by pervasive sensing and computing across the natural, business, Internet and government sectors and social environments.

NIST introduces Big Data classification attributes that are listed below.

Volume is the sheer data sets volume. ·

Variety refers to the structured, semi-structured and unstructured data form.

Velocity of data generation and requirement in real time.

The ability to join multiple datasets is Horizontal Scalability.

Two categories are included in the Relational Limitation.

PHASE II: DATA ACQUISITION

Information gathering alludes to the way toward recovering crude information from true items. The procedure should be well structured. Something else, incorrect information accumulation would affect the consequent information investigation method and eventually lead to invalid outcomes. In the meantime, information accumulation techniques not just rely upon the material science qualities of information sources, yet in addition the targets of information investigation. Accordingly, there are numerous sorts of information accumulation techniques. In the subsection, we will first center around three basic strategies for enormous information gathering, and afterward address a couple of other related strategies.

Sensors are utilized normally to gauge a physical amount what's more, convert it into a discernible advanced sign for preparing (also, perhaps putting away). Sensor types incorporate acoustic, sound, vibration, car, compound, electric flow, climate, weight, warm, and nearness. Through wired or remote systems, this data can be exchanged to an information gathering point. Wired sensor

systems influence wired systems to associate an accumulation of sensors and transmit the gathered data. This situation is reasonable for applications in which sensors can effectively be sent and oversaw. For instance, numerous video reconnaissance frameworks in industry are presently constructed utilizing a single Ethernet unshielded turned pair per advanced camera wired to a focal area (certain frameworks may give both wired and remote interfaces) [9]. These frameworks can be conveyed in open spaces to screen human conduct, such as robbery and other criminal practices. On the other hand, remote sensor systems (WSNs) use a remote system as the substrate of data transmission. This arrangement is best when the precise area of a specific marvel is obscure, especially when the condition to be observed does not have a framework for either vitality or correspondence. As of late, WSNs have been broadly examined and connected in numerous applications, such as in condition research [10], [11], water checking [12], structural designing [13], [14], and natural life living space observing [15]. The WSN normally comprises of countless spatially appropriated sensor hubs, which are battery-fueled minor gadgets. Sensors are first sent at the areas specified by the application necessity to gather detecting information. After sensor arrangement is finished, the base station will spread the system setup/the board as well as accumulation order messages to all sensor hubs. In light of this demonstrated data, detected information are accumulated at various sensor hubs and sent to the base station for further handling. [16] offer an itemized dialog of the previous. A sensor based information accumulation framework can be considered as a digital physical framework. As a matter of fact, in the scientic try space, numerous strength instruments, for example, attractive spectrometer, radio telescope, are utilized to gather analyze information. They can be viewed as an uncommon kind of sensor. In this sense, explore information gathering frameworks additionally have a place with the class of digital physical framework.

Log: a standout amongst the most generally conveyed information gathering strategies, are produced by information source frameworks to record exercises in a specied le position for consequent investigation. Log files are helpful in practically every one of the applications running on advanced gadgets. For instance, a web server regularly records every one of the snaps, hits, get to and different traits made by any site client in an entrance log le. There are three primary sorts of web server log le groups accessible to catch the exercises of clients on a site: Common Log File Format (NCSA), Extended Log Format (W3C), and IIS Log Format (Microsoft). Every one of the three log le positions are in the ASCII content group. Then again, databases can be used rather than content les to store log data to improve the questioning efficiency of gigantic log stores . Different models of log file-based information accumulation incorporate stock ticks in financial applications, execution estimation in system observing, and traffic the board. Rather than a physical sensor, a log le can be seen as "programming as-a-sensor". Much client executed information accumulation programming has a place with this classification.

Web Crawler: A crawler is a program that downloads and stores website pages for an internet searcher. Around, a crawler begins with an underlying arrangement of URLs to visit in a line. Every one of the URLs to be recovered are kept and organized. From this line, the crawler gets a URL that has a specific need, downloads the page, identities every one of the URLs in the downloaded page, and adds the new URLs to the line. This procedure is rehashed until the crawler chooses to stop. Web crawlers are general information accumulation applications for site based applications, such as web crawlers and web reserves. The creeping procedure is controlled by a few approaches, including the choice strategy, return to approach, consideration arrangement, and parallelization strategy. The choice arrangement imparts which pages to download; the return to approach chooses when to check for changes to the pages; the amenability strategy averts over-burdening the sites; the parallelization strategy organizes appropriated web crawlers.

DATA TRANSMISSION

IP BACKBONE

The IP spine, at either the locale or Internet scale, gives a high-limit trunk line to exchange huge information from its inception to a server farm. The transmission rate and limit are controlled by the physical media and the connection the executives strategies.

Physical Media are commonly made out of numerous fiber optic links packaged together to expand limit. When all is said in done, physical media should ensure way decent variety to reroute traffic if there should be an occurrence of disappointment. Link Management concerns how the signal is transmitted over the physical media. IP over Wavelength-Division

Multiplexing (WDM) has been created over the past two decades. WDM is innovation that multiplexes numerous optical transporter flag on a solitary optical fiber utilizing various wavelengths of laser light to convey various sign. To address the electrical data transfer capacity bottleneck confinement, Orthogonal Frequency-Division Multiplexing (OFDM) has been considered as a promising possibility for future rapid optical transmission innovation. OFDM permits the range of person subcarriers to cover, which prompts an additional information rate flexible, light-footed, and asset efficient optical system. So far, optical transmission frameworks with up to limits of 40 Gb/s per direct have been sent in spine systems, though 100 Gb/s interfaces are currently monetarily accessible and 100 Gb/s arrangement is normal before long. Indeed, even Tb/s-level transmission is predicted in the close future. Due to the difficulty of sending improved system conventions in the Internet spine, we should pursue standard Web conventions to transmit enormous information. Notwithstanding, for a territorial or on the other hand private IP spine, certain choices may accomplish better execution for specific applications.

Data Center Transmission: At the point when enormous information is transmitted into the server farm, it will be traveled inside the server farm for situation change, handling, etc. This procedure is alluded to as information focus transmission. It generally connects with server farm arrange engineering and transportation convention:

Data Center Network Architecture: A server farm comprises of different racks facilitating an accumulation of servers associated through the server farm's inward association arrange. Most present server farm interior association systems depend on item switches that configure an authoritative fat-tree 2-level or 3-level engineering. Some different topologies that expect to make more efficient server farm systems can be found in .Due to the innate lack of electronic bundle switches, expanding correspondence transmission capacity while at the same time lessening vitality utilization is difcult.

DATA PRE-PROCESSING

As a result of their various sources, the gathered informational collections may have various dimensions of value regarding commotion, repetition, consistency, and so forth. Exchanging and putting away crude information would have important expenses. On the interest side, certain information examination techniques and applications may have exacting necessities on information quality. Thusly, information preprocessing methods that are intended to improve information quality ought to be set up in huge information frameworks. In this subsection, we briey study current inquire about endeavors for three ordinary information pre-handling systems.

Integration: Information combination systems intend to join information dwelling in various sources and furnish clients with a perspective on the information. Information incorporation is an experienced field in customary database look into. Already, two methodologies won, the information distribution center technique and the information organization technique. The information distribution center technique, otherwise called ETL, comprises of the accompanying three stages: extraction, change furthermore, stacking. The extraction step includes associating with the source frameworks and choosing and gathering the vital information for investigation preparing. The change step includes the utilization of a arrangement of principles to the removed information to change over it into a standard configuration. The heap step includes bringing in removed and changed information into an objective stockpiling foundation. Second, the information organization technique makes a virtual database to question and total information from divergent sources. The virtual database does not contain information itself but rather contains data or metadata about the real information and its area. Be that as it may, the "store-and-draw" nature of these two approaches is unacceptable for the elite needs of spilling or pursuit applications, where information are much more unique than the inquiries and must be handled on the y.

Cleansing: The information purifying system alludes to the procedure to decide wrong, inadequate, or nonsensical information and afterward to correct or evacuate these information to improve information quality. A general structure for information purging comprises of five corresponding steps: Search and distinguish mistake occurrences ;Correct the mistakes; Document mistake cases and blunder types; and Modify information section strategies to decrease future blunders. Additionally, group checks, culmination checks, sensibility checks, and breaking point checks are ordinarily considered in the purifying procedure. Information purifying is commonly considered to be fundamental to keeping information reliable and refreshed what's more, is in this way generally utilized in numerous fields, for example, banking, protection, retailing, broadcast communications, and transportation. Current information cleaning procedures are spread crosswise over various spaces. In the internet business space, albeit a large portion of the information are gathered electronically, there can be not kidding information quality issues.

Redundancy Elimination: Information excess is the reiteration or superuity of information, which is a typical issue for different datasets. Information repetition pointlessly expands information transmission overhead what's more, causes hindrances for capacity frameworks, including squandered extra room, information irregularity, decreased dependability what's more, information debasement. In this way, numerous scientists have proposed different repetition decrease strategies, for example, repetition recognition and information pressure.These techniques can be utilized for various datasets or application conditions and can make significant benefits, likewise to gambling presentation to a few negative elements. For example, the information pressure strategy represents an additional computational load in the information pressure and decompression forms. We ought to evaluate the tradeoff between the benefits of excess decrease and the going with weights. Information excess is exemplified by the developing sum of picture and video information, gathered from broadly sent cameras. In the video observation space, tremendous amounts of excess data, including fleeting excess, spatial repetition, factual excess and perceptual excess, is hidden in the crude picture and video files . Video pressure procedures are generally used to diminish repetition in video information. Numerous significant gauges (e.g., MPEG-2, MPEG-4, H.263, H.264/AVC) have been fabricated and connected to reduce the weight on transmission and capacity. For summed up information transmission or capacity, the information duplication method is a particular information pressure method for dispensing with copy duplicates of rehashing information. In a capacity deduplication process, a one of a kind lump or fragment of information will be dispensed an identification (e.g., hashing) and put away, and the identification will be added to an identification list. As the deduplication examination proceeds, another lump partner with the identification, which as of now exists in the identification list, is viewed as a repetitive piece. This piece is supplanted with a reference that focuses to the put away piece. Thusly, just a single case of any bit of given information is held. Deduplication can incredibly lessen the sum of extra room and is especially significant for huge information capacity frameworks. Notwithstanding the information pre-handling strategies depicted above, different tasks are fundamental for specific information objects. One model is highlight extraction, which plays a basic job in regions, for example, mixed media look and DNA investigation. Regularly, these information objects are depicted by high-dimensional element vectors (or focuses), which are composed away frameworks for recovery. Another precedent is information change, which is ordinarily used to deal with circulated information sources with heterogeneous composition and is especially valuable for business datasets. We should consider together the attributes of the datasets, the issue to be tackled, execution necessities and different components to pick an appropriate information pre-handling plan.

PHASE III: DATA STORAGE

Capacity Infrastructure Equipment foundation is in charge of physically putting away the gathered data. The capacity foundation can be comprehended from alternate points of view. To start with, capacity gadgets can be classified dependent on the special innovation. Common stockpiling advancements incorporate, yet are most certainly not constrained to, the accompanying.

Random Access Memory (RAM): RAM is a type of PC information stockpiling related with unstable kinds of memory, which loses its data when controlled off. Current RAM incorporates static RAM (SRAM), dynamic RAM (DRAM), and stage change memory (PRAM). Measure is the prevalent type of PC memory.

Magnetic Disks and Disk Arrays: Magnetic circles, such as hard circle drive (HDD), are the essential segment in present day stockpiling frameworks. A HDD comprises of one or progressively inflexible quickly turning circles with attractive heads orchestrated on a moving actuator arm to peruse and compose information to the surfaces. In contrast to RAM, a HDD holds its information notwithstanding when controlled off with much lower percapacity cost, however the read and compose tasks are much slower. On account of the high use of a solitary huge limit circle, plate clusters amass various plates to accomplish enormous limit, high access throughput, and high accessibility at much lower costs.

Storage Class Memory: Storage class memory alludes to non-mechanical capacity media, for example, ash memory. As a rule, ash memory is utilized to develop strong state drives (SSDs). Not at all like HDDs, SSDs have no mechanical parts, run all the more unobtrusively, and have lower get to times and less dormancy than HDDs. In any case, SSDs stay more costly per unit of capacity than HDDs.

These gadgets have diverse execution measurements, which can be utilized to manufacture a versatile and elite enormous information stockpiling subsystem. More insights regarding capacity gadgets advancement can be found in [17]. Of late, half and half approaches [18], [19] have been proposed to construct a various leveled capacity framework that consolidates the highlights of SSDs also, HDDs in a similar unit, containing a huge hard circle drive and a SSD reserve to improve execution of much of the time gotten to information. A run of the mill design of multi-level SSD-based capacity framework is appeared in Fig. 8, which comprises of three parts, i.e., I/O demand line, virtualization layer, and cluster. Virtualization layer acknowledges I/O demands and dispatches them to volumes that are comprised of degrees put away in varieties of various gadget types. Current business SSD-based multi-level frameworks from IBM, EMC, 3PAR and Compllent as of now gain satisfied execution. Nonetheless, the major difficulty of these frameworks is to figure out what blend of gadgets will perform well at least expense. Second, stockpiling framework can be comprehended from a organizing design viewpoint. In this classification, the capacity subsystem can be composed in various ways, counting, yet not constrained to the accompanying.

Direct Attached Storage (DAS): DAS is a capacity framework that comprises of a gathering of information stockpiling gadgets (for precedent, various hard circle drives). These gadgets are associated legitimately to a PC through a host transport connector (HBS) with no capacity arrange between them also, the PC. DAS is a basic stockpiling expansion to a current server.

Network Attached Storage (NAS): NAS is le-level stockpiling that contains numerous hard drives orchestrated into sensible, repetitive capacity holders. Contrasted and SAN, NAS gives both capacity and a le framework, and can be considered as a le server, though SAN is volume the board utilities, through which a PC can get plate extra room.

Storage Area Network (SAN): SANs are devoted systems that give square dimension stockpiling to a gathering of PCs. SANs can merge a few stockpiling gadgets, for example, circles and plate clusters, and make them available to PCs with the end goal that the capacity gadgets seem, by all accounts, to be privately joined gadgets.

Conclusion

In this paper, firstly, we introduce big data and the current challenges of it. Then we focus on the three phases of the value chain of big data, i.e., data generation, data acquisition, data storage. For each phase, we introduce the general background, discuss the technical challenges, and review the latest advances. Although big data is a hot research area with great potential in both academia and industry. There is a compelling need for a rigorous and holistic definition of big data, a structural model of big data, a formal description of big data, and a theoretical system of data science. Big data transfer involves big data generation, acquisition, transmission, storage, and other data transformations in the spatial domain. As discussed, big data transfer usually incurs high costs, which is the bottleneck for big data computing. However, data transfer is inevitable in big data applications. It is difficult for current and mature batch processing paradigms to adapt to the rapidly growing data volume and the substantial real-time requirements. Two potential solutions are to design a new real-time processing model or a data analysis mechanism. In the traditional batch-processing paradigm, data should be stored first, and, then, the entire dataset should be scanned to produce the analysis result. Much time is obviously wasted during data transmission, storage, and repeated scanning. There are great opportunities for the new real-time processing paradigm to reduce this type of overhead cost. Due to the value-sparse feature of big data, a new data analysis mechanism can adopt dimensionality reduction or sampling-based data analysis to reduce the amount of data to be analyzed.

References

[1]. Ikbal Taleb, Mohamed Adel Serhani, Rachida Dssouli," Big Data Quality Assessment Model for Unstructured Data", 2018.
[2]. Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, , Aida Mustapha," Data Quality: A Survey of Data Quality Dimensions",2012.
[3]. M. Chen, S. Mao, Y. Liu, "Data Quality: A Survey of Data Quality Dimensions", 2014.
[4]. Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. Proc VLDB Endowment 5(12):2032–2033
[5]. Chaudhuri S, Dayal U, Narasayya V (2011) An overview of business intelligence technology. Commun ACM 54(8): 88–98
[6]. Agrawal D, Bernstein P, Bertino E, Davidson S, Dayal U, FranklinM, Gehrke J, Haas L, Halevy A, Han J et al (2012) Challenges and opportunities with big data. A community white paper developed by leading researches across the United States
[7] J. Manyika *et al.*, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Glob. Inst.*, pp. 1–137, 2011.
[8] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," *IEEE Access*, vol.2, pp.652–687, 2014.
[9] V. Chandramohan and K. Christensen, ``A _rst look at wired sensor networks for video surveillance systems," in *Proc. 27th Annu. IEEE Conf.Local Comput. Netw. (LCN)*, Nov. 2002, pp. 728_729.
[10] L. Selavo *et al.*, ``Luster: Wireless sensor network for environmental research," in *Proc. 5th Int. Conf. Embedded Netw. Sensor Syst.*, Nov. 2007,pp. 103_116.
[11] G. Barrenetxea, F. Ingelrest, G. Schaefer, M. Vetterli, O. Couach, and M. Parlange, ``Sensorscope: Out-of-the-box environmental monitoring," in *Proc. IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, 2008, pp. 332_343.
[12] Y. Kim, T. Schmid, Z. M. Charbiwala, J. Friedman, and M. B. Srivastava, ``Nawms: Nonintrusive autonomous water monitoring system," in *Proc. 6th ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2008, pp. 309_322.
[13] S. Kim *et al.*, ``Health monitoring of civil infrastructures using wireless sensor networks," in *Proc. 6th Int. Conf. Inform. Process. Sensor Netw.* Apr. 2007, pp. 254_263.
[14] M. Ceriotti *et al.*, ``Monitoring heritage buildings with wireless sensor networks: The Torre Aquila deployment," in *Proc. Int. Conf. Inform.*
Process. Sensor Netw., Apr. 2009, pp. 277_288.
[15] G. Tolle *et al.*, ``A macroscope in the redwoods," in *Proc. 3rd Int. Conf. Embedded Netw. Sensor Syst.*, Nov. 2005, pp. 51_63.
[16] F. Wang and J. Liu, ``Networked wireless sensor data collection: Issues, challenges, and approaches," *IEEE Commun. Surv. Tuts.*, vol.13, no. 4, pp. 673_687, Dec. 2011.
[17] K. Goda and M. Kitsuregawa, ``The history of storage systems," *Proc. IEEE*, vol. 100, no. 13, pp. 1433_1440, May 2012.
[18] J. D. Strunk, ``Hybrid aggregates: Combining SSDS and HDDS in a single storage pool," *ACM SIGOPS Oper. Syst. Rev.*, vol. 46, no. 3, pp. 50_56, 2012.
[19] G. Soundararajan, V. Prabhakaran, M. Balakrishnan, and T. Wobber, ``Extending SSD lifetimes with disk-based write caches," in *Proc.8th USENIX Conf. File Storage Technol.*, 2010, p. 8.
[20]. Chandrasekhar Rangu, Shuvojit Chatterjee, Srinivasa Rao Valluru," Text Mining Approach for Product Quality Enhancement", 2017.