EXPLORING THE ETHICAL IMPLICATIONS OF BIG DATA AND DATA MINING

Abstract

Big data analytics is a rapidly developing phenomena that is influenced by interactions between people, businesses, and society. The ethical ramifications for these stakeholders are still little understood and experimentally underexplored. Predictive analytics helps with future projections while descriptive analytics provides past data. The heterogeneous mass of digital information generated by companies and people is known as "Big Data" because of its characteristics (large volume, variety of formats, and speed of processing), which necessitate specialised and cutting-edge computer storage and analysis tools. We outline the fundamental ideas behind the ethical concerns raised by big data analytics and data mining. Then, in order to support the ethical use of stakeholder theory and discourse ethics to analyse big data analytics, we offer recommendations for how to strike a balance in interactions between individuals, businesses, and society. This article aims to define the principles, issues, and applications of big data and data mining as well as the significance of big data and data mining analytics.

Key words: Big Data, Data Mining, Big data Analytic, Ethical Implication.

Authors

Shailendra Chourasia
Assistant Professor,
CSE, GGCT,
shailendrachourasia@ggits.org

Pankaj Kumar Jain
Assistant Professor,
CSE, GGCT,
pankajjain@ggct.co.in

I. INTRODUCTION

Big data analytics use computers to extract patterns, correlations, and other insights from vast and complex data sets (Martin, 2015). Big data analytics may be used in many ways to increase economic and social value, including healthcare, public safety, and service innovations. However, it has recently been under fire for having unsavoury effects on a number of stakeholders. Concerns about privacy violations, substantial individual profiling, or customer discrimination have been made public (Zwitter, 2014).These issues highlight a contradiction between stakeholder values where organisations' objectives and incentives do not line up with those of people and society. We are therefore unaware of how to equitably distribute the benefits and costs of big data among various stakeholder groups.

In order to address the inherent value conflict in big data analytics systems, we employ a stakeholder approach (Mitchell, Agle, & Wood, 1997) to explore the interrelationships among various stakeholders. For two reasons, we do this. First, they urged discourse ethics to be used by IS researchers to answer moral conundrums .According to discourse ethics, morality develops through fair discussions among parties involved in an ideal speech environment (i.e., where everyone is on an even playing field). Secondly, by emphasising We respond to Markus' (2015) request that IS researchers look into how big data analytics affect various stakeholders, as well as researcher suggestions that figuring out how businesses' "non-responsible" use of big data affects people and society is a top research priority. Despite the fact that recent research has raised ethical issues with big data analytics, it hasn't investigated stakeholder perspectives or relationships.

The industry with the fastest growth over the past four years has been information technology (IT), which during the next two years is anticipated to grow by between $340 million and $1 trillion. Millions of individuals work as IT professionals in fields related to computers. However, the IT sector has a negative name for unethical and unprofessional behaviour. The area of IT that is the most specialised is big data and data mining. Globally, digital data is rapidly expanding, from 150 exabytes in 2005 to 1200 exabytes in 2010. In the upcoming years, data growth is projected to be 40%. From 2007 through 2020, the amount of digital data is projected to grow 44 times, doubling roughly every 20 months (Chowdhury, 2014). These industries are crucial for establishing professionalism and moral behaviour because of this. With the aforementioned issue in mind, I conducted research for the following question –

Why are ethics and professionalism vital in the field of information technology, namely in the fields of big data analytics and data mining?

Individuals' personal information is used by big data and data mining analysts. Businesses gather customer data using a variety of technologies and use it for data analysis, particularly for prediction analysis. However, it's crucial to consider how they use customer data. The research question for this paper is in favour of the United Nations Economic Commission. It will contribute to further study on "Big Data for Development: Assessing the Fitness in Monitoring and Evaluation (M&E) System" for any work that raises the question, "How might Big Data help monitor more precisely and current social, economic, and environmental phenomena? In order to address this genuine demand, I provided scholarly and professional viewpoints on IT professionalism and ethical issues with a focus on the Big Data and Data Mining field.

II. RESEARCH BACKGROUND

Massive Data Analysis The three Vs of technology—volume, variety, and velocity—or its underlying algorithms' capacity to provide insights —have been used most frequently to conceive big data analytics up until this point. Big data analytics is viewed by us as a socio-technical phenomenon that has an impact on all parties. is, however, constrained by such a technology focus. We augment the technological view by identifying three social processes that aim at and have an impact on people based on an analysis of the developing literature: The first three are algorithmic decision-making, data sourcing, and data sharing. First, many big data applications use people as a means of collecting data. The "catch-all-you-can" strategy is used by businesses and government agencies to get as much data as possible from people. This method measures people's daily activities, especially for the organisation performing the analytics. Second, until its value is exhausted, data collected from individuals is transferred from one entity to another. This reasoning has produced a secondary market for businesses to exchange or sell customer data. People struggle to understand why data-driven services are necessary in order to harvest and exchange client data. Third, organisations use algorithms to profile people—sometimes unintentionally—based on their race, ethnicity, gender, and socioeconomic status and to limit their options.

In this article, we emphasise the use of big data analytics by businesses in their interactions with consumers, especially when those businesses employ big data analytics to provide consumers with services and goods. We contend that when organisations gather, examine, distribute, and/or sell people's data without those people's informed or sincere consent, ethical problems result.

Discourse Ethics: Ethics is the study of how people ought to behave and what defines truthful conduct. There are several conventional methods of thinking about ethics (such as utilitarian, kantian, and aristotelian), yet each has drawbacks. The utilitarian approach, for instance, makes it difficult to predict the effects of one's actions in the modern world, and in human societies, Selecting the majority's interests over minority interests could be discriminatory. Here, we examine the ethics of big data analytics using discourse ethics as a framework. introduced discourse ethics to the IS literature and argued in favour of its ability to address moral concerns that face IS practitioners and scholars. Discourse ethics, which has its roots in research, is a more contemporary theory that integrates older theories of ethics, particularly Kantian and utilitarian views. Discourse ethics encompasses the idea of universalism and mostly focuses on morality. According to universalism, moral standards apply to activities that are equally good for everyone, regardless of setting or community.

The discourse process, or the process of communication action, is at the heart of discourse ethics. According to Habermas, a deliberative process is the most effective way for stakeholders to arrive at pragmatist, ethical, and moral norms. Only those norms, he contends, can be said to be valid, in the eyes of everyone who is affected and who is a participant in a practical conversation. This strategy requires that ethical norms be fairly contested among relevant parties and cannot already exist or be imposed. The ethical discourse will be shaped by the stakeholders' actual dialogues, which they should iteratively revise over time. This demands the establishment of the ideal speech environment, which calls for the freedom to discuss, assert, and contest one's beliefs as well as the participation of all parties on an equal footing.

In our proposal, we employ discourse ethics as our overarching ethical framework and argue that ethical big data analytics will develop from stakeholders' participation in and formation of ethical discourse. However, the degree of equality and satisfaction with the ideal speech environment among stakeholders will determine their capacity to create such a discourse. Because discourse ethics does not explain how to identify stakeholders and their salience, stakeholder theory (Mitchell et al., 1997) is used to identify, categorise, and analyse the stakeholders and their interrelationships.

Some IT professionals work with specific Big Data management frameworks. Big Data is a subset of the 3Vs, which stand for volume, velocity, and variety. It involves the examination of real-time data, particularly to find trends over time or historical analysis explains that "Data mining is an interdisciplinary field that incorporates methods from database management, statistics, machine learning, and artificial intelligence. Furthermore, data mining is the process of retrieving, summarising, and summarising data from various databases, according to Anonymous (1998). Data originates from consumers or from persons in both scenarios. Laws governing privacy, copyright, patents, and trademarks can prevent unethical and unprofessional activity in the IT sector, particularly in the fields of big data and data mining.

III. DESCRIBE BIG DATA AND DATA MINING

What is Big Data?

In modern culture, there is a tremendous amount of data that has been growing exponentially. The term "Big Data" refers to the creation and usage of technologies that convey the appropriate information from this data to the suitable user at the appropriate time. The issue is dealing with the constantly increasing volumes of data, which is in addition to dealing with the complexity of managing more different forms and complicated and interrelated data.

It is a complex polymorphic item, and as such, its definition varies according to the communities that are interested in it as a user or provider of services. Big Data, a technology created by the web's leading lights, offers itself as a means of giving everyone real-time access to enormous databases. Due to the fact that different sectors have varying definitions of what constitutes a huge amount of data, it is extremely difficult to describe big data precisely. Instead than referring to a particular collection of technologies, it describes a class of approaches and technologies. This is a new field, and as we work to understand how to apply this new paradigm and reap its advantages, the definition is changing.

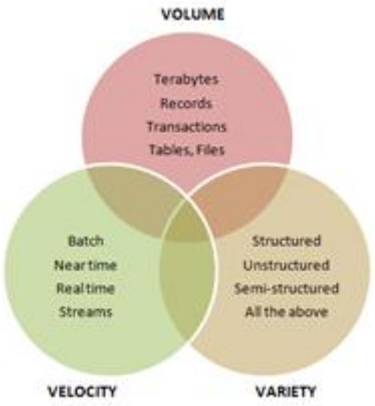
Characteristics of Big Data: Big Data describes vast, significantly larger datasets (volume), more diverse datasets (diversity), including organised, semi-structured, and unstructured data, and significantly faster datasets (velocity) than before. These are the 3V.

3V's

Volume: indicates how much data is produced, saved, and used by the system. The need to leverage the data as well as the expansion in the amount of data being collected and stored account for the volume increase.

Variety: shows how the range of data kinds that an information system can handle has expanded. The links and link types between these data are complicated as a result of this multiplication. The potential uses of a raw piece of data are similarly subject to diversity.

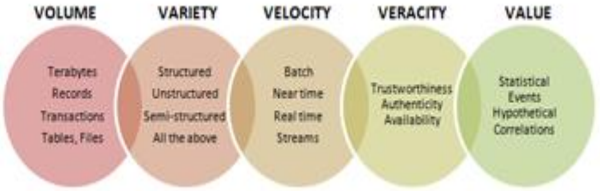
Velocity: shows how often data is created, collected, and shared. Due to the fact that the data are transmitted in a stream, real-time analysis is necessary.



5V's: Two more "V"s that is significant to this traditional classification is:

Veracity: degree of correctness, reliability, and uncertainty of the data and the data sources.

Value: the value and potential derived from data.



Big Data Analytics: Data that can be stored, processed, and computed more effectively than it can using conventional databases and data analysis techniques is referred to as "big data," which is a broad word. The use of tools and procedures that may be utilised to analyse and identify patterns in massive amounts of data is required when using big data as a resource. Structured data analysis advances due to the variety and speed of data processing. Because of the diversity of the data, it is no longer sufficient to merely assess the data and produce reports; the systems in place also need to be able to enable data analysis. Analysis involves automatically detecting the relationships between varieties of rapidly changing data in order to facilitate its use.

Big Data Analytics: Types

There are following types of Big Data Analytics

Descriptive Analytics: It entails posing the query: What's going on?

A set of historical data is produced during the initial step of data processing. Data organisation and pattern recognition are aided by data mining techniques. Future probabilities, patterns, and an understanding of likely events are provided by descriptive analytics.

Diagnostic Analytics: It entails posing the query: Why did it occur?

Analytical diagnostics looks for the root causes of a problem. It is used to determine an event's cause. This personality type seeks out and makes an effort to comprehend the causes of events and actions.

Predictive Analytics: It consists of asking the question: What is likely to happen?

It forecasts the future using historical data. It's all about making predictions. Predictive analytics analyses current data and creates scenarios of potential outcomes using a variety of techniques, including data mining and artificial intelligence.

Prescriptive Analytics: It entails posing the query: What ought to be done?

It is dedicated to choosing the appropriate path of action. Predictive analytics helps with future projections while descriptive analytics provides past data. Prescriptive analytics uses these variables to identify the best answer. Data mining is the process of searching through large data sets for pertinent or useful information. Businesses are supposed to gather enormous amounts of data, some of which may be homogeneous or amassed automatically. From those huge sets, decision-makers require access to smaller, more focused data fragments. Data mining is used to find the bits of knowledge that will guide corporate decisions and advise leadership.

Many software programmes, including analytics tools, may be used in data mining. It can be automated, in which case different employees submit different information requests to the database. In general, processes involving highly complex search techniques that provide tailored and focused results are referred to as data mining. To discover a certain column of expenses or accounts receivable for a particular working year, for instance, a data mining programme may search through dozens of years' worth of financial data. Big Data is important because of what you can do with it, not how much data we have. We can use data analysis to make wise decisions that save money and time.

Data Mining and Big Data: Data mining is the process of evaluating data from many angles and turning it into valuable information. It is sometimes referred to as data discovery or knowledge discovery. Businesses use this information to boost sales and lower operating costs. Among the many instruments used in data analysis are the software programmes used in data mining. Users of the software can classify and summarise the data patterns found after analysing it from various points of view. Strictly speaking, data mining is the process of identifying patterns or connections among vast quantities of related databases. The examination of massive datasets automatically or semi-automatically is the genuine data mining task. This is done to aid in the extraction of unexpected and previously unidentified data patterns. They consist of mining sequential patterns, cluster analysis of data sets, and finding anomalies in records. In these operations, database techniques like spatial indices are frequently used.

Following these steps, the patterns serve as a summary of the input data and can be applied to more in-depth analyses using machine learning or predictive analytics. For instance, data mining techniques can be used to find various categories of data. This is the procedure for evaluating bigger data sets in an effort to find relevant information. Examples of this data include consumer preferences, market trends, hidden patterns, and unidentified relationships. The results of the analytics typically result in increased operational effectiveness, greater marketing effectiveness, and new revenue potential.

Thanks to big data analytics, data scientists, predictive modellers, and other analytics specialists can now analyse enormous amounts of transaction data. Big data analytics can be used to study data that normal business programmes might not have found.

Challenges to Handle Big Data: Because there is such a wealth of complex and raw data available, programmers must make decisions. These huge datasets can be gathered, kept, and analysed in a variety of ways by an organisation. The Organization can even employ powerful big data tools to more quickly and effectively store, access, and manage the structured and unstructured data gathered from multiple sources. When working with large data sets, there aren't many obstacles to overcome. These are a few difficulties:

Handling a Large Amount of Data: Making decisions is made more difficult by the quantity of information. Over the past few years, there has been a significant increase in the amount of data that businesses can access. They know everything about a consumer, including their tastes, how they react to different smells, and the amazing new restaurant that just debuted in Italy last weekend. This data exceeds what can be computed, stored, and retrieved. The administration of this data rather than its accessibility is where the challenge resides. In addition to the expansion of unstructured data, data is now accessible in a variety of formats, including audio, video, social media, data from mobile devices, etc. One of the most recent methods developed to manage this data is the combination of relational and NoSQL databases. MongoDB, a built-in component of the MEAN stack, serves as an example of this. Distributed computing solutions like Hadoop are also available to manage the volume of Big Data.

Data Security: Data security is the main issue with data growth. Many companies claim to have problems with data security. In fact, this presents a bigger challenge than a lot of other data-related problems do. Data is obtained by businesses from many different sources, some of which cannot be trusted to be secure or compliant with organisational needs. They need to use a variety of data collection approaches to keep up with the demand for data. As a result, discordant data yield inconsistent analytical outcomes. There can be security risks because this data is available from so many different sources. We might never be able to determine which data channel is vulnerable, endangering the organization's data security and providing access to hackers. It is increasingly essential to implement data security best practises for safe data collection, storage, and retrieval.

Data Complexity: The enormous amounts of data that are updated every second must be managed by organisations. A retail company wanting to track consumer behaviour may find real-time data from previous customer transactions helpful. Two tools for data analysis that can be used for the same thing are Veracity and Velocity. Frameworks, compute engines, visualisation engines, ETL engines, and other necessary inputs are among them. In addition to the static data that is always available, it is critical for organisations to keep up with this data. As a result, better insights will be created and decision-making abilities will be enhanced.

Shortage of Skilled Resources: Big Data experts are in short supply at the moment. In their efforts to enhance their usage of Big Data and develop more effective Data Analysis platforms, several organisations have begun to raise this issue. The existing lack of skilled data scientists and analysts makes "number crunching" difficult and insight development time-consuming. Again, training new hires can be expensive for a company working with cutting-edge technology. Instead, a lot of people are concentrating on automation tactics that employ AI and ML to provide insights, but this also calls either highly skilled staff or the outsourcing of gifted programmers.

IV. CONCLUSION

A complex social phenomenon with a built-in dualism is big data analytics. Indeed, it offers opportunities for human society to advance, but it also creates ethical challenges for those participating. Information technology is the industry that is growing at the highest rate. There are many people that work in information technology (IT). We utilise IT to make life easier everywhere we go, but occasionally, it also creates issues. There are numerous examples of IT specialists abusing technology. It is challenging to prevent young practitioners from acting unethically and to improve their reputation. Government agencies, social networking platforms, and even organisations gather data on specific people. Of course, it matters how they use this information. On occasion, IT infrastructures are constructed without taking into account the needs of the entire population. Governments and political parties utilise big data unethically to control or manipulate the population. Big data analytics and data mining experts analyse data to make society and daily life better. This sector is vital for lowering crime and antisocial behaviour, and it could also aid in the fight against disease.

REFERENCES

[1] Sorell, Tom, Nasir Rajpoot, and Clare Verrill. "Ethical issues in computational pathology." Journal of Medical Ethics 48.4 (2022): 278-284.

[2] Stahl, Bernd Carsten. "Responsible innovation ecosystems: Ethical implications of the application of the ecosystem concept to artificial intelligence." International Journal of Information Management 62 (2022): 102441.

[3] Ida Someh, Michael Davern, Christoph F. Breidbach, Graeme Shanks (2019), Ethical Issues in Big Data Analytics: A Stakeholder Perspective, Communications of the Association for Information Systems, volume 44,article 34 ISSN: 1529-3181.

[4] Youssra Riahi, Sara Riahi, "Big Data and Big Data Analytics: Concepts, Types and Technologies" in International Journal of Research and Engineering ISSN: 2348-7860 (O) | 2348-7852 (P) | Vol. 5 No. 9 | September-October 2018 | PP. 524-528.

[5] Martin, K. E. (2015). Ethical issues in the big data industry. MIS Quarterly Executive, 14(2), 67-85.

[6] Markus, M. L., & Topi, H. (2015). Big data, big decisions for science, society, and business. Bentley University.

[7] Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decisionmaking: A call for action on the long-term societal effects of "datification". Journal of Strategic Information Systems, 24(1), 3-14.

[8] Chowdhury, Amin (2014). Dissertation Proposal on Big Data for Development: Assesing the fitness in M&E system. Module Assignment, Sheffield, Sheffield Hallam University.

[9] Zwitter, A. (2014). Big data ethics. Big Data & Society, 1(2).

[10] Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. Academy of Management Review, 22(4), 853-886.