3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018

THE PATTERNS THAT DON'T EXIST: Study on the effects of psychological human biases in data analysis and decision making

Dr. Parkavi A
Department of Computer Science and Engineering
Ramaiah Institute of Technology, Bangalore, India
parkavi.a@msrit.edu

Anam Jawaid
Department of Computer Science and Engineering
Ramaiah Institute of Technology, Bangalore, India
anamjawaid786.aj@gmail.com

Saima Dev
Department of Computer Science and Engineering
Ramaiah Institute of Technology, Bangalore, India
saimaa.dev@gmail.com

Vinutha M S
Department of Computer Science and Engineering
Ramaiah Institute of Technology, Bangalore, India
vinuthamsv@gmail.com

***Abstract*—Data analytics and decision making are among the most profound human-developed methods to examine data and draw various conclusions about the information they contain. To ensure accuracy in prediction, mathematical formulae have been developed. Error correction and uncertainty analysis have been used to obtain results closer to the real time expected outcome however; they still fall back on many cases.**

This paper analyses background human-behavioural processes that affect data analysis and consequent decision making. A major loophole that plays a role in the inaccuracy of predictions are psychological human biases, which are not taken into account while formulating, analysing and utilising data analysis techniques. Provided in this paper is the examination of lesser known behavioural traits their often neglected effects on data analytics and decision making, followed by the proposal of prejudice detection algorithm to detect the presence of one of the most commonly faced bias known as the confirmation bias.

Keywords---*Data Analysis, Decision Making, Cognitive human bias, Perceptual bias, Social Bias, Motivational Bias, Psychological behaviour, Senses.*

I. INTRODUCTION

Decision making is perceived to be a simple yes or no task. While the community of engineers know its complexity, due to automation of this technique using data analysis. From the study on effects of data analytics in decision making [1] we can say that data analytics helps organisation analyse their own internal and external data to enhance their knowledge, and then these results are used for the decision making process within the organisations. The major challenger to make this process extremely efficient and accurate, it is important to implement a successful decision making process based on data analysis. We can now say that both data analytics and decision making are interdependent on each other. Hence, any factor affecting one of these processes has a direct impact on result which further increases the importance of analysing background factors which have been affecting these procedures but are not more commonly studied or identified. What is lesser known, is how this process is influenced by underground human behavioural tendencies. Cognitive biases refer to deviation from rationality of judgement. These unlike other psychological tendencies, do not vary from person to person and follow a systematic pattern of deviation. While making any decision, the human brain tends to think it has taken into account all dimensions. This is not necessarily bad, but sometimes it leads to spurious results in analysis and bad decisions.

During information processing, every step from choosing the dataset, data modelling and obtaining results, human behaviour has a major influence. Some tendencies are known to have a direct impact. The ignorance of psychological tendencies and cognitive human behaviour usually leads to unaccounted uncertainty and results where analysts do not understand the cause of deviation of the actual result from the expected result.

According to studies, apart from cognitive biases, social and perceptual bias also account for an impact in data analysis. These biases are influenced by external environment, society and perception. These biases are less complex than cognitive biases. Although they seem superficial in phenomena and function, they have a complex relationship and impact in data analysis. Another form of a well-known bias called motivational bias plays a major role in decision making but does not affect data analysis. These are quite similar to cognitive biases but form their subset. Due to their interrelation, our study included analysis of these biases as well and their role in the process of data analysis and decision making

II. LITERATURE SURVEY

A review of Big Data Analysis [2] revealed the ways in which we can acquire supplementary value even from enormous unrelated information obtained from different institutions. If the data analysis methods used are incorrect or inaccurate, data is of no use to the analysts. To overcome this, Big Data Analytics was introduced. This form of science is involved when analysis is performed on humongous amount of data in actual or real time. This can be done as this technique can

ISBN: 978-1-5386-6078-2 © 2018 IEEE

predict use cases in a variety of fields that are using big data, even in extremity.

Data quality plays a major role in all research fields. As proposed in the paper [3] are affairs that are involved while investigating quality of the data used by analysts and also provides details related to favorable circumstances of related study. It also provides an analysis of how the quality of data can affect data analysis as a process on the whole, how effective it is and other impacts of it on business related activities involving intelligence. It also puts forth the progress of studies and research related to data quality that maybe used to create a real time application. Certain important quality attributes have been emphasized which are used by institutions. These are the most commonly used attributes such the accuracy of data; it's consistent and accomplished nature. Events that cause negative impact on the organization's environment could be detected using business intelligence process, resulting in faster response to the changes that are to be made. Decision makers need to consider effective business intelligence to satisfy the customers, by anticipating their needs.

While studying the impact of biases in Big Data [4] we discovered the most commonly encountered biases in machine learning. These biases are introduced due to the self-learning nature of this technique. One of them proposed is the class imbalance. These biases do not fall in our spectrum of study. Additionally, the debiasing technique for the identified biases is also put forward. This has a relevance to psychological biases as it is done by analyzing the specifications and making the analysts aware of their existence. This is the same technique used to eradicate psychological biases from most projects. It is also stated how the eradication of these biases helps obtaining results that are more reliable.

The paper based on visualization research framework [5] provides a lightweight framework that provides a reference to decide the method of research to be adopted when an analysts is trying to detect a particular bias during their study of reference for selecting research methods, when trying to identify a bias in visualization research. The framework categorizes biases into three discrete levels. Out of the ones put forward, our study will be based on social as well as perceptual biases due to their impact on data analysis and decision making. The paper further explains the ways in which different biases maybe analyzed using the three level differentiation. Further discussions are made on how to proceed on detecting the biases that are encountered during the research involving visualization, hoping that the framework provides a way for discovering new ideas.

Further study of biases lead to examination of confirmation bias [6] in analysis of information. Proposed is a method that maybe used to reduce this bias that merely involves considering the causes, different from the ones already taken into account for data analysis.

A compilation was obtained after the detailed study of various papers. This helped us shortlist most commonly known biases, their roles in various intelligence methods and ways to mitigate these. These observations were used to figure out the possible impact of biases in the field of data analysis and decision making.

III. BIASES IN THE LIGHT OF DATA ANALYSIS AND DECISION MAKING

A. Cognitive Bias

A study was conducted to test all biases that have either directly or indirectly had some effect on data analysis results. The most important bias; Cognitive bias is a methodical difference between the "correct" answer acquired according to rules and the analyst's actual answer. Confirmation bias is the most commonly observed cognitive bias that impacts analysis of data. In this psychological behaviour, a person begins to concentrate only on a singular idea proposed i.e. hypothesis. Since there is compulsion to prove the proposed idea, the person takes into account only the data favouring his conclusion. As proposed by Paul and Leonard, there is a natural tendency to outweigh the pros and cons regarding confirming and disconfirming the evidence [6]. It has been proved by researchers that this bias manifests itself in complex analysis tasks [8] .As a result of this bias, data selection and collection tends to fall towards the positive spectrum and hence away from the expected or real outcome. One popular example is the predictions of Clinton's win in 2016 US presidential elections. Similar to confirmation bias, behaviour to study is the Overconfidence bias. This occurs when the analysts provide assessments for a given set of criteria that are way above or too limited in extent from the actual performance. This directly leads to overestimation and over-precision in data analysis again leading to only positive conclusions. The major sectors impacted are legal, financial, defence and engineering due to quantitative estimates and data.

Myopic problem among cognitive bias directly affects data collection and modelling. This occurs when a problem is oversimplified and then its representation is used to model the system. This leads to an incomplete mental model of the problem statement as a whole. In simpler terms, the focus is always on the smaller alternatives based on objectives.

B. Social Bias

A larger category of biases known as Social biases refers to biases that play an important role in the judgment based on social levels and is affected by various cultural phenomenon. These biases occur on the highest level, due to systematic biases during socialization. A Famous bias in this category is the outgroup homogeneity bias which is the phenomenon in which some people observe various people that are excluding the people who belong to their own group in order to be more homogeneous than their peer in-group [4]. It is often caused by imperfect and stereotypical memory. Labelled data belonging in a scatter plot visualization results in better linear separability only if one of the labels corresponds to an out-group and the other to an in-group provided by the user. This trait can be seen as having a positive impact on data analysis when used for classifier techniques such as k-nearest neighbour classification when data is taken from both in-group and out-group yielding accurate results.

ISBN: 978-1-5386-6078-2 © 2018 IEEE

C. *Perceptual Bias*

Perceptual biases in contrast to other biases occur only at the perceptual level. This means it is not a result of complex psychological behaviour but mere inputs from senses. Significant in this category is the clustering illusion. This [9] [10] bias explains why some people observe patterns in minute sets of random information. The result of variance and existence of clustered data in any random data set is most commonly miscalculated. A typical example can be if you toss a coin, and get heads thrice in a row, people will directly conclude that the coin is biased and might feel confident about it. However tails - heads - tails is probably the same sequence as heads - heads - heads since the tosses are statistically independent. In case of visualization research, this can manifest itself as users obtaining scatter-plots that have low density, consequently over-interpret them and coming to judgements that don't occur so often. A remedy for this bias is developing null plots, which is basically visualization presenting simulated data accumulated from the null hypothesis. Another effect observed is the Weber-Fechner Law in which the differences between stimuli are observed with the help of a logarithmic scale. It takes more additional millimetres of measurement to determine two larger areas rather than two areas having lesser magnitude [5]. By understanding this phenomenon one can imply the need of additional data to analyse when the magnitude of data is large.

D. *Motivational Bias*

Motivational biases can be defined as the ones where the impact on behavior is mainly influenced by the importance of outcomes, results, events or consequences, which lead to the wrong choices. An illustration of this bias can be confirmation bias, which occurs due to the urge to prove the proposed hypothesis. These when studied need not [6] always be conscious. With respect to decision-making, it will straight forward lead to neglected the right result or using an overestimated data model to yield the required results.

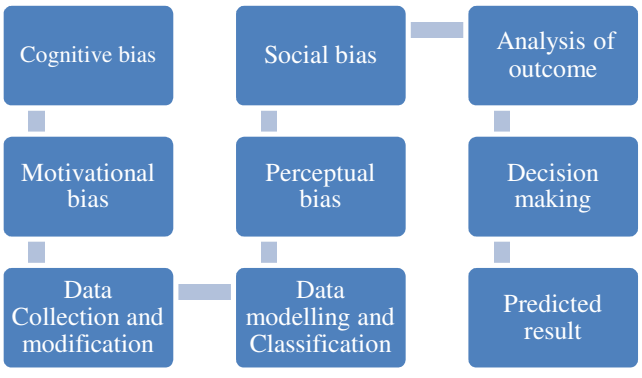

Figure 1: Effect of psychological traits on steps in data analysis and decision making

IV. IMPACTUAL ANALYSIS

A majority of psychological biases are seen to have impacted the data collection and modification phase along. These include motivation and cognitive biases such as confirmation bias. With a wrong data set having incorrect, inaccurate or incomplete data or bad quality data, as cited in the study of its impact on data analysis [3] it has been pointed out that improvement of data quality is extremely hard as poor quality data generates errors and increases workload. In the data modelling phase, analysts being unsatisfied by the results go back to the preparation phase and generate different attributes. For the model to work correctly, this has to be repeated several times leading to enormous amount of wastage. Economically, poor quality data has cost business more than 611 billion dollars.

Social and perceptual biases directly impact analysis models with some having a positive impact such as outgroup homogeneity trait. Other results of these are over-estimation, over-interpretation, under-estimation and low density plots as mentioned. Weber-Fechner-Law however, affects data collection and quality. These have a direct impact on business, research and technological sectors due to wrong results costing them billions of dollars every year for a minor overlook of human behaviour.

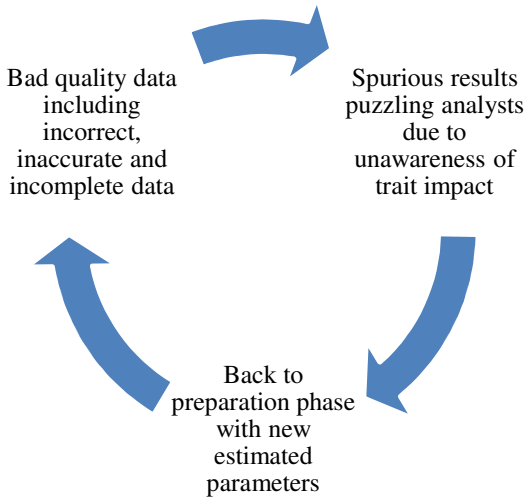

Figure 2: Recursive cycle produced due to bad data quality

V. METHODOLOGY

Leaving aside the various ways of removal of biases that can be done by re examination and psychiatric care, we propose a new methodology that can help in the removal of cognitive biases. Proposed here is the Prejudice Detection algorithm that should be followed before analysts go ahead with the project. This has been proposed for confirmation bias. Data collection begins with generating a questionnaire. The type of questions asked depends on the domain of the project and the biases to be investigated. let's consider confirmation bias with

ISBN: 978-1-5386-6078-2 © 2018 IEEE

the domain of prediction of possibility of road accidents. In this case the questions generated maybe

- Which of the following scenarios are best fitted with respect to your data?
- Which of the following parameters has been taken into consideration?

As stated before and confirmation bias analysts try to collect only the data that will yield positive results. In this case various options are given to check whether the data scientist has also considered the options where in road accidents do not occur or has only considered the scenarios that are prone to road accidents. Various options are provided such as with respect to question 1 the options provided are "over speeding bikes", "same lane for buses and cars" and "following Lane discipline". As you can observe over speeding bikes has the highest probability off road accidents. If the data analyst has chosen only scenarios like this or similar to this it indicates the possibility of confirmation bias. Similarly for the second question options are given such as "distractions such as use of mobile", "intoxication" and "none" .If the scientist choses "none" it indicates that the scientist has mostly accumulated diverse data and not only the data suited to his result.

The responses received from Data scientist are then accumulated. The collected data is cleaned to remove spurious tuples. Since the data is categorical, numeric values are assigned to each answer of every question. The parameters are assigned values based on the inclination of the data scientist to prove the result. In the given example "over speeding bike" will be assigned the highest value and "following Lane discipline" will we assign the lowest value with "same Lane drives" being assigned an intermediate value. Once numeric values have been obtained, cosine similarity is used to find the similarity index between the given answers of all the questions and the answers that guarantee to give positive results. If the similarity index is found to be sufficiently higher (threshold may be decided arbitrarily) the result will indicate the presence of confirmation bias. The same procedure of Data Collection and assignment of numerical value may be performed for other biases as well. The data model however will change according to the bias to be detected.

VI. RESULTS AND DISCUSSION

A detailed study was made on the underlying behavioural factors that influence data analysis and decision making process. These account for the uncertainty in results and the errors which analysts couldn't recognise the source or cause of. Cognitive, social, perceptual and motivational biases were introduced which revealed the impact of these traits in the analysis process. As argued by data scientists, these do have a direct effect on data analysis which can be seen in real time descripensies. A prejudice detection algorithm was put forward based on the study conducted on various biases that promises to predict the presence of confirmation bias if present in data analysis and decision making.

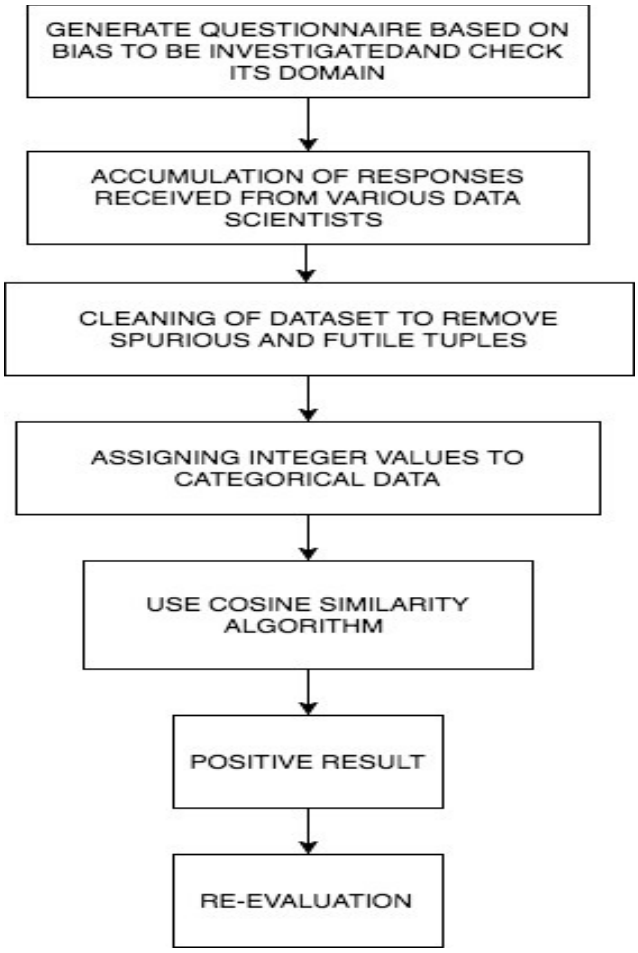

Figure 3: Prejudice Detection Algorithm

VII. CONCLUSION AND FUTURE SCOPE

Analysis of real time data and research helps us conclude the presence of impacts of behavioral traits in data analysis and decision making. They sometimes have been helpful but most commonly have led to downfalls and will further continue to do so if not taken care in the future. To eradicate the negative impacts, pattern recognition techniques can be developed for cognitive biases. An algorithm has been proposed that uses the cosine similarity algorithm to predict the presence of confirmation bias. This can be extended to other forms of biases by changing data analysis using cosine similarity phase as required for the particular bias under detection. Other types of biases require monitoring and can only be minimized with expert supervision and awareness of these tendencies. If resolved, this will lead to a drastic change in data analysis and decision making sector by reducing the percentage of wrong results by a considerable amount. Analysts also have a fallback on these traits to examine and easily point out the

ISBN: 978-1-5386-6078-2 © 2018 IEEE

inaccuracy or incorrectness which is often unaccounted for. Economically, it will save institutions a wholesome amount of money spent every year in the correctness of data analysis and decision making by reducing the impact of these traits.

REFERENCES

[1] "Evaluation of an aspect of Data Management", Stavros Mouroutis.

[2] "Big Data Analytics: A Literature Review Paper", Nada Elgendy and Ahmed Elragal, Springer International Publishing Switzerland 2014.

[3] "Impact of Biases in Big Data", Patrick Glauner, Petko Valtchev and Radu State ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

[4] "A Framework for Studying Biases in Visualization Research", Andre Calero Valdez ,Martina Ziefle,Michael Sedlmair .

[5] " Confirmation Bias in the Analysis of Remote Sensing Data", Paul E. Lehner, Leonard Adelman, Robert J. DiStasio Jr., Marie C. Erie, Janet S. Mittel, Sherry L. Olson.

[6] "Examining the effect of causal focus on the option generation process: An experiment using protocol analysis Organizational Behavior and Human Decision Processes", Adelman, L., Gualtieri, J., & Stanford, S. (1995)., 61, 54-66.

[7] "Cognitive and Motivational Biases in Decision and Risk Analysis", Gilberto Montibeller and Detlof von Winterfeldt.

[8] "Confirmation bias in Complex Analysis", Adelman Brant A. Cheikes, Mark J. Brown ,Paul E. Lehner and Leonard, IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans 38(3):584 - 592 · June 2008.

[9] "Biases and sensitivities in geometrical illusions. Vision research", M. Morgan, G. J. Hole, and A. Glennerster. 30(11):1793–1810, 1990.

[10] "Biases in judgments of separation and orientation of elements belonging to different cluster" , T. Seizova-Cajic and B. Gillam.. Vision research, 46(16):2525–2534, 2006.

ISBN: 978-1-5386-6078-2 © 2018 IEEE