Factors Affecting the Usability of Unstructured Big Data

Joshua Devadason¹, Rehan Akbar²

Abstract—Big data is a valuable asset for organisations as it analyses and help to understand the customers, changes within their business environment, market analysis and future trends. The big data is multifaceted (different data types and versatile), and mostly exists in unstructured formats. The extraction of value from this data is challenging. The usability and productivity of this multifaceted unstructured data is greatly compromised. A number of factors and associated reasons affect the usability of unstructured big data. The present research work investigates these factors and associated reasons behind the usability issues of multifaceted unstructured big data. The identification of these factors contribute to develop solutions to reduce the lack of usability of highly unstructured big data. A detailed study of existing literature followed by survey questionnaire has been conducted to identify the factors and their reasons. Descriptive statistics has been used to analyse and interpret the data and results.

Keywords—Big data, Data types, Factors, Multifaceted, Unstructured, Usability, Limitations

I. INTRODUCTION

Big data has value that is highly desired by many organisations and companies. The ability to fully exploit this value would ensure better business operations, better customer services and even greater profit margins generated. Big data, in last two years alone has generated close to 2.5 Quintillion bytes of data, and is never defined by its volume [1] but by what an organisation or company is capable of doing with it [2]. Extracting this value is not as simple as it looks. Sometimes, data come in a nice structured and tabulated form which can be easily converted into information that can be understood and utilised. However, data mostly is not in this nice form rather in an unstructured format and many different forms or multifaceted. Big data information usually originates from both humans and machines. Humans tend to create text-based information which reside in documents, emails, social media activities, surveys, location data, website content, and even through text messages. Machine generated information includes satellite images, scientific data, photographs and video, radar or sonar data and many others. Unstructured big data do not have a predefined schema in its arrangement [3], while multifaceted big data exist in many forms such as text, images, video, audio, and several other formats. Multifaceted unstructured big data has the combination of both of these characteristics. Today, analysing this type of data manually is considered a better option because most of the technologies cannot support this type of data extraction.

It is not impossible to process and utilize multifaceted unstructured data forms to derive value (information & knowledge) from it, but the process is difficult and challenging. However, not all the companies and organisations have the ability or the technical skills to do this due to which this data ends up being discarded. This type of data also need to be stored which is usually a problem for organisations as most of the organisations rely on relational database systems which do not support data in an unstructured or multifaceted form [4]. The rate of growth of big data is increasing every year with improvements in technology and newer capabilities. With this trend, the rate of deriving valuable information being discarded is also increasing every year. If this data is made usable, the value of the derived information can be highly beneficial. It can lead to business operations being able to run in less time and at lower cost along with far better intelligent decisions being made [2]. Furthermore, organisations and businesses would be able to detect problems and fix them before they occur, better understand their customers and their buying habits to serve them better, to determine the risk portfolios in short period of time, and finally, the ability to find defects, defects, and issues in near real-time. With such a high value, it is required to fully utilise the big data, both structured and unstructured, regardless of its multifaceted form.

Big data, a proven valuable asset for organisations, is underutilized to its full value and potential, simply because of the problems faced during the extraction of information process from multifaceted unstructured data. Even worse is the situation when unusable data is discarded permanently. Big data is much easier to handle and process when it is in a structured and ordered form [5]. It can easily be stored in a relational database, and translating into decisions and ideas for a company. An information that does not confine itself into an organisation’s database is usually discarded [6]. Most of this type of data is unstructured and the total is expected to increase at least 62% every year. This trend needs to be stopped, and solutions are required to solve the problem. As the volume of big data is increasing every year, so to further the advancement of information extraction, the factors behind the usability issues of multifaceted unstructured big need to be determined to find solutions. To solve the problem of usability issues of multifaceted unstructured big data, it must be considered as a business problem rather than just a complex problem that can be solved by technology [1]. The identification of these factors would help in new technologies and advancements in data analysis and information derivation.

This research work serves three main objectives. First objective is to determine these factors. A thorough literature in the domain of data analysis, big data, and information retrieval has been explored as preliminary study. Based on the literature study, a questionnaire survey has been prepared and conducted to determine the factors, which, initially are four, and as second objective, the reasons behind the factors, as some issues are related to external or unrelated reasons. The respondents of the survey are professionals working on big data and analytics in real life conditions. The final objective is

Joshua Devadason (email: jotube2904@1utar.my).
Rehan Akbar (email: rehan@utar.edu.my).
Faculty of Information and Communication Technology
Universiti Tunku Abdul Rahman
Kampar, Perak Malaysia.
to propose a solution. However, the proposed solutions may not resolve or alleviate all the problems is usability of unstructured data but would be a significant contribution of this research work in addressing the usability problems.

II. LITERATURE REVIEW

The literature review in the area of big data is very complex and time consuming because of broad spectrum data analysis and information extraction methodologies, multiple data types, diversified approaches and work in this area. After thorough literature study, four factors that generally affect the usability of unstructured big data have been identified and discussed here.

A. Storage Issues

When performing information extraction on multifaceted unstructured big data, it is important for the information to be stored so it can be retrieved used at a later time. Big data is not defined by its volume, according to [2], as its importance and value comes from what organisations are capable of doing with the information. Despite this, big data can be massive and large in its volume, according to [7], not to mention its characteristic of high velocity. An article titled “What is Big Data” by the computer company IBM according to [1] showed that in the span of just two years, about 2.5 Quintillion bytes of data has been created or generated by mankind. At the high rate of data creation and its large volume demands that database systems must be able to ingest it, process it, store it, and allow for later retrieval within a short period of time. It sounds simple, but there are challenges to implementing such a database system.

Relational database management systems (RDBMS) or SQL–based database systems are the considered to be the most popular in deployment [4]. According to [4], RDBMS is highly unsuited for it because the characteristics of multifaceted unstructured big data which does not follow a predefined schema in its arrangement. With the high velocity of incoming data, the RDBMS is unable to cope with. The volume of the data also creates problems when it comes to the storage of the information. According to a study [8], examples of how companies like Facebook can generate over 500 terabytes of data in their data mining operations of its users. The study also reveals that Walmart gathers close to 2.5 petabytes of data of customer transactions every hour. This not only shows how large big data can be, but also the rate of growth of the information over time. The companies and organisations will be required to upgrade their storage solutions very frequently in order to keep up with this demand. The cost of these storage devices are also a big challenge [9] that varies depending on the memory size. This is still cheaper, when considering the alternative, solid state drives though they provide faster performance compared to hard drives [10].

NoSQL, according to [11], is considered to be the better alternative to RDBMS or SQL-based database systems when it comes to the storage of multifaceted unstructured big data. NoSQL database systems offer some advantages over their RDBMS counterparts, according to [12]. Running NoSQL as the database system is also considered much cheaper than SQL-based database systems. The best feature of NoSQL is that it is completely free to use, as the system is open source [13]. In fact, if these limitations can be overcome, then NoSQL database systems are the prime candidate for the storage of multifaceted unstructured big data. Therefore, volume and management of big data makes storage as an important factor that affect the processing of the data and information extraction.

B. User Interaction and Involvement

Multifaceted unstructured big data generally has the characteristics of data with low degree of organisation and turning this mess into something that can be understood and utilised is no easy challenge, according to [14]. To make matters worse according to [15], data with these characteristics would continue to show growth at a rate of 62% a year, leading the logical method of approach towards automation for the extraction of information from big data. There is however always the risk of important information being discarded simply because automated machines are unable to process some portions of big data. So, according to [16], user interaction is a requirement in ensuring the quality of extracted information in Information Extraction (IE) applications. This is because it is an inherently imprecise process, hence human interaction is required when processing and managing multifaceted unstructured big data. However, with most users more likely to experience far better ease of use, better control and efficient multitasking abilities using graphical user interface (GUI) applications, the quality can be limited to the confines of the functionalities provided the developer, according to [17]. This in turn, would drive most IE applications to be developed using command line tools or CLI, to retain most functionalities being available to the user, though most users may be unable to adjust to this type of application, due to its lack of user friendliness.

As information extraction is considered an inherent process [16], the user interaction is important to improve its quality. The users referred here are both developers and application users, whereby to develop an application that can explain query result provenance, incorporating user feedback, developing novel user interfaces, and developing novel interaction modes. User interaction should be kept to a minimum when performing information extraction from multifaceted unstructured big data. In the long run, more functionalities would have to be automated rather than aided. Most industries today are already moving towards automated solutions as it can save time and cost in business operations [18]. However, the real improvement would be that human error would be heavily reduced or even eliminated, depending on the industry. So, while user interaction would greatly improve the quality of information extracted from big data, most companies and organisations would avoid it in favour of solutions that would offer automation in its process instead.

C. Visualisation Tools

Big data, whether unstructured or multifaceted must be extracted and processed so to transform large amounts of data into something that can be understood and can be used to further improve a company or an organisation in its business operations. The best way to achieve this result is to convert this large amount of data into graphs and pictorials, which would allow the information to be easily understood by most people. The process is simple and easy to accomplish when the data is in a structured and tabular form, as the data can easily be processed. There are multifaceted forms such as textual, non-textual, human generated or machine generated. Human generated unstructured data can be made up of text-based files that can be sourced from websites and other internet sources like emails, social media data, communication data, business application data, mobile data,
or media while machine generated data can be made up of satellite imagery, scientific data, digital surveillance or sensor data. The situation becomes far more challenging for data analyst, when the data is both multifaceted and unstructured. As of today, only about 20% of the industries worldwide actually have the technical skill and ability to be able to process this type of data [15], indicating that data that can’t be processed are either ignored or discarded.

The ordinary data analysing applications are unable to process this information. However, it has become possible for computers to be able to handle and process large amounts of data at high velocity, as the rapid evolution of the blend of art and science sets the corporate landscape [19]. This allows tools that can process large volumes and high velocity multifaceted unstructured big data into information of value to users, available out there in the market. These tools include FCS Analytics, Cluvio, QlikSense, SPSS by IBM, MongoDB, Phocas Software, Cyfe, MicroStrategy Enterprise Analytics and many others, according to [20], some more popular than others. For example, eBay, an e-commerce organisation utilises a tool called Tableau, which has the capability to translating large datasets in large and complex forms into spontaneous visualisations. This results give eBay the ability to better serve their customers and conduct sentiment analysis. However, these tools do have their limitations.

For one, these off the shelf tools for processing multifaceted unstructured big data, show poor performance in functionalities, scalability, and response time. However, the biggest limitation, according to [8], will have to be the poor performance in conducting data visualisation. When conducting visualisation on big data in any form, the aim is to represent key information and knowledge more instinctively and effectively. If these tools can’t manage this well, then data is very likely discarded or ignored. The reason for these limitations is not just due to the large sizes and high dimensions of big data, but its multifaceted and unstructured form as well. These challenges would have to be overcome before multifaceted unstructured big data can be fully utilised.

D. Limitations in Existing Programming Languages

When it comes to developing programs for any kind of use, there are many types of popular programming languages available for development. According to [21], SQL, C++, Java, and Perl are among top used languages in software development. Even modern-day gaming applications are written in C++, according to [22]. These programming languages are considered quite powerful for most types of software development, but have certain limitations when developing application for information extraction. This is because popular programming languages follow procedural based rules and instructions, performing its tasks set by the developer by obeying a set of rules and instructions, creating a problem when there is a need to develop, debug, understand, modify and optimise the application [23]. When it comes to developing Information extraction programs for processing big data, popular programming languages like Java, C++, Perl, or SQL can be considered highly unsuited for the task. It becomes even more challenging when the data to be processed and extracted is in multifaceted and unstructured forms as well. A study conducted at the University of Wisconsin-Madison [16] explored the possibility of solving the limitations of popular programming languages by developing a custom programming language. This programming language would be highly suited for developing programs and software that perform information extraction processes on big data. This custom language is called xlog, a declarative language that works using multiple Datalog-like rules, allowing it to be used to design and develop programs capable of handling both static and dynamic types of data. The research claims that one of the key advantages of this custom programming language is its ability to be far more amenable to query optimisation techniques. The study emphasizes that newer programming languages are required to solve the limitations of popular programming languages of today and that this solution could be achieved.

Developing a custom programming language can present itself with its own set of challenges. According to an article [24], the process is expected to be difficult and long before a stable and working language is developed, because the scope of the work itself is expected to expand. Also, the matter of design and syntax must be taken into consideration. Support for this new programming language must also be developed, along with all the necessary learning materials, the runtime libraries with the necessary features such as I/O performance, memory allocation, and transcendental functions. The scope of the work is large and the development can take years, with odds of success against the developer. With so much effort and no guarantee of success, it can be understandable why most developers of information extraction programs would still rather work with existing programming languages and put up with their limitations.

III. RESEARCH METHODOLOGY

The process of any research or study always begins with the identification of the research problem, then performing literature reviews on relevant topics, ensuring that the research purpose is specified, performing data collection based upon this determined purpose, analyse the data and interpret it, and then finally, report and evaluate the study [25]. The qualitative and quantitative approaches are two main types of research methods. The qualitative research works by investigating a phenomenon and their relationships by using something that is measurable and done in a systematic way, usually with numbers [26]. This type of methodology is not suitable for this research, as time is limited and performing a qualitative study require long hours of observations, along with appointments with interviewees, which would have to be made in advance and finally, even longer hours analysing the responses collected from the study’s participants. Quantitative research works by conducting a study to better understand the social world through its wide array of dimensions, to get a better view of texture, situations and life stories that happen every day, imaginings, understandings, and experiences of the respondents in the study [27]. This type of study serve as the methodology of this research, as the responses provided by the participants would be structured, allowing easy and unbiased analysis of the collected responses. This is important since the time frame for this study is limited.

The quantitative method using questionnaire survey has been used in this research work. The questionnaire comprises of two parts. The first part comprises of the demographic data of the participants such as age, gender, company name, type of company, designation in company, number of years of experience within the industry, and finally, state, and country of origin of the participants. The demographic questions ensure that the participants are expert in their fields, and are very familiar with big data, to prevent inaccurate answers in the data collection process of this study. The respondents are
experts with at least 5 years of working experience as data analyst, big data researcher, big data practitioner, data scientist, or related fields. In the second part of the questionnaire, the questions have been set based on the factors determined through the literature review, as 5-6 questions per factor, using the Likert scale of (1) Strongly Disagree to (5) Strongly Agree. The selection of the number of participants is done as described by [28], while the confidence level is set at 95% and the margin of error at 8% [29]. The response distribution is estimated to be 20%, and the sample size is [40].

The descriptive statistical analysis has been performed [30], [31] to analyse the collected data and its interpretations. The questions are subject to statistical parameters, where the mean and standard deviation is calculated to provide the average of the squared differences and to determine how spread out the responses are respectively. The small value of standard deviation indicates similar responses from the respondents and vice versa. However, ± 1 standard deviation is considered good. The order in terms of high impact to low impact w.r.t the mean and standard deviation is 1) user interaction requirement, 2) storage issues, 3) limitations of visualisation tools, and 4) insufficient features in popular programming languages. Interestingly, decreasing mean and increasing standard deviation is the clear indication of the order of the impact of these factors and how experts value them. However, for the insufficient features in popular programming languages, the too large spread i.e. 3.02 shows the tendency of disagreement of the experts over this factor.

The detailed analysis and findings for each of the factors have been presented in the following sections.

A. Insufficient Features in Popular Programming Languages

With highest mean and lowest standard deviation, the user interaction requirement is believed to have the highest effect on the usability of unstructured data. 1.70 standard deviation which is less than 2 shows general agreement of the experts on this factor. The standard deviation describes the distribution of the responses from the mean [34]. The smaller value of standard deviation indicates similar responses from the respondents and vice versa. However, ± 1 standard deviation is considered good. The order in terms of high impact to low impact w.r.t the mean and standard deviation is 1) user interaction requirement, 2) storage issues, 3) limitations of visualisation tools, and 4) insufficient features in popular programming languages. Interestingly, decreasing mean and increasing standard deviation is the clear indication of the order of the impact of these factors and how experts value them. However, for the insufficient features in popular programming languages, the too large spread i.e. 3.02 shows the tendency of disagreement of the experts over this factor.

The results of the survey have been summarized in Table 4.1.

| Factors Explored                                      | N  | Mean | Standard Deviation |
|-------------------------------------------------------|----|------|--------------------|
| 1. Insufficient Features in Popular Programming Languages | 96 | 3.06 | 3.02               |
| 2. User Interaction Requirement                        | 96 | 4.47 | 1.70               |
| 3. Storage Issues                                      | 96 | 4.02 | 2.05               |
| 4. Limitations of Visualisation Tools                  | 96 | 3.97 | 2.36               |

The detailed analysis and findings for each of the factors have been presented in the following sections.

IV. RESULTS AND DISCUSSION

The survey has been conducted among 96 participants, 25 to 58 years old, with 6 to 36 years of experience. Figure 4.1 and 4.2 show the gender and designation of the respondents.

![Gender](image1.png)

**Figure 1. Gender**

![Designations](image2.png)

**Figure 2. Designations**

Figure 3 shows the sectors of the economy, data is collected from and respondents belong to.

![Economy Sectors](image3.png)

**Figure 3. Economy Sectors**
developed to overcome these challenges and limitations of the existing popular programming languages.

Table 4.2. Insufficiency in Programming Languages

| Query                                                      | Mean | Strongly Agree (%) | Agree (%) |
|------------------------------------------------------------|------|---------------------|-----------|
| 1. Developing an information extraction program to extract and process incoming data from different sources is difficult and challenging. | 3.94 | 10.42               | 78.13     |
| 2. Some of the challenges on this type of program development include the insufficient functionalities of popular programming languages. | 2.95 | 3.13                | 6.25      |
| 3. These insufficient functionalities in program development are one of the reasons for extracted information being discarded instead of being utilised. | 3.04 | 1.04                | 15.63     |
| 4. Popular programming languages are not well suited for developing a program for information extraction. | 1.50 | 0                   | 3.13      |
| 5. I believe a custom programming language is needed to overcome the limitations of existing programming languages. | 4.64 | 71.86               | 19.79     |
| 6. Developing a custom programming language would cost a high amount of time and money. | 3.20 | 0                   | 21.88     |

Table 4.2 shows that the majority of the participants strongly disagreed that complete automation of information extraction is difficult, in order to reduce these mistakes caused by human error. It can be concluded that the participants of this study do strongly believe that the requirement of user interaction during information extraction is a factor that affects the usability of multifacetated unstructured data. A reason for this factor could be that human error still exists, even with the involvement of human interaction, and tools utilised in the process of information extraction must be simple, in order to reduce these mistakes caused by human error. However, the proposal of automation of the entire process of information extraction to eliminate human error, cannot be considered as a solution to this factor because most of the respondents neither agreed nor disagreed with it. This may be due to the fact that while automation may resolve human error, there is no guarantee that the information extracted from multifacetated unstructured data is done with quality and accuracy.

Table 4.3. User Interaction Requirement

| Query                                                      | Mean | Strongly Agree (%) | Agree (%) |
|------------------------------------------------------------|------|---------------------|-----------|
| 7. User interaction is required to extract and process incoming data to ensure quality of information extracted from it. | 4.88 | 86.46               | 11.46     |
| 8. User interaction is required to extract and process incoming data to ensure accuracy of information extracted from it. | 4.86 | 84.38               | 13.54     |
| 9. There is still a chance of human error despite the usage of user interaction in the information extraction process. | 4.84 | 82.29               | 15.63     |
| 10. Programs with simple user interfaces are required to reduce the rate of human error in data extraction. | 4.84 | 83.33               | 14.58     |
| 11. Complete automation of information extraction and processing of incoming data would help to eliminate human error. | 2.91 | 0                   | 0         |

Table 4.3 shows that majority of the participants strongly agreed that user interaction is important to ensure quality and accuracy of the information extracted from multifacetated unstructured data. They also believe in the occurrence of human error, despite that user interaction exists and if it is a requirement then programs should have simple interface (GUI) to reduce the rate of human error during information extraction. However, 89% of the participants neither agreed nor disagreed that complete automation of information extraction and processing of incoming data would help to eliminate human error. It can be concluded that the participants of this study do strongly believe that the requirement of user interaction during information extraction is a factor that affect the usability of multifacetated unstructured data. A reason for this factor could be that human error still exists, even with the involvement of human interaction, and tools utilised in the process of information extraction must be simple, in order to reduce these mistakes caused by human error. However, the proposal of automation of the entire process of information extraction to eliminate human error, cannot be considered as a solution to this factor because most of the respondents neither agreed nor disagreed with it. This may be due to the fact that while automation may resolve human error, there is no guarantee that the information extracted from multifacetated unstructured data is done with quality and accuracy.
C. Storage Issues

Table 4.4. Storage Issues

| Query                                                                 | Mean  | Strongly Agree (%) | Agree (%) |
|---------------------------------------------------------------------|-------|--------------------|-----------|
| 12. Information extracted from collected data usually does not follow a defined schema. | 3.72  | 0                  | 71.86     |
| 13. It is difficult to easily process and store this type of data in some SQL-based database systems. | 3.17  | 0                  | 16.67     |
| 14. Storage space upgrades are frequently required to keep supporting and managing incoming data into the database system. | 3.86  | 0                  | 87.5      |
| 15. Frequent storage upgrades help to discard unimportant or low-priority data to save cost. | 4.51  | 51.04              | 46.86     |
| 16. A better solution would be a database system that does not utilise SQL-based systems. | 4.81  | 79.17              | 18.75     |

As given in Table 4.4, most of the respondents neither agreed nor disagreed that it is difficult to easily process and store this type of data in some SQL-based database systems. Most participants agree that information extracted from collected data usually does not follow a defined schema and that storage space upgrades are frequently required to keep supporting and managing incoming data into the database system. However, 51% strongly agree and 46.9% agree that the reasons for data being discarded frequently is due to the cost requirements of frequent storage upgrades. Most participants do strongly agree that when it comes to storage options, a database system that does utilise SQL would be a better solution for the storage of multifaceted unstructured big data. From here, it can be inferred that most participants feel that storage is a factor, when it comes to the usability issues of multifaceted unstructured big data. Most do agree that storage space upgrades can be a contributing reason to the storage issue as a factor, but do really feel that multifaceted unstructured big data would be too challenging to have it stored in SQL-based database systems. To conclude, storage issues are a contributing factor affecting the usability of multifaceted unstructured big data, and that cost being the reason for its existence and a solution to this would be to utilise non-SQL-based database systems. However, in contrast to the literature conducted earlier, most participants neither agree nor disagree that the difficulty of easily process and store this type of data in some SQL-based database systems can be considered as one of the reasons for this factor’s existence.

D. Limitations of Visualisation Tools

Table 4.5. Visualisation Tools

| Query                                                                 | Mean  | Strongly Agree (%) | Agree (%) |
|---------------------------------------------------------------------|-------|--------------------|-----------|
| 17. Usage of visualisation tools to process incoming multifaceted unstructured big data is helpful most of the time. | 4.14  | 12.5              | 85.42     |
| 18. These tools tend to have poor performance in functionalities, scalability, and response time with varying size and complexity of the data sets. | 3.08  | 0                  | 8.33      |
| 19. These tools tend to be limited when it comes to the quality of conducting data visualisation on incoming multifaceted unstructured big data. | 3.16  | 0                  | 15.63     |
| 20. I usually discard or ignore data that is not processed by these tools. | 4.75  | 82.29             | 9.38      |
| 21. Visualisation tools available in the market do not always meet the requirements to process large volumes of data. | 4.73  | 78.13             | 15.63     |

Table 4.5 shows that most of the participants agree that the usage of visualisation tools to process incoming multifaceted unstructured data is helpful most of the time, but 78% to 82% of participants strongly agree that data is discarded or ignored when it cannot be processed by these tools, and the market today does not have visualisation tools that meets all the requirements for processes large amounts of data. 81% and 89% of the participants however, chose neither agree nor disagree on these tools having limitations and poor performance when conducting data visualisation on multifaceted unstructured big data. From here, it can be inferred that visualisation tools available in the market today can contribute to the usability faced in multifaceted unstructured big data. From here, it can be inferred that most participants agree that visualisation tools that process multifaceted unstructured big data is considered helpful most of the time, but strongly agree that data that can’t be processed is usually discarded and that visualisation tools available in the market today are insufficient for process multifaceted unstructured big data. This shows that the limitations of visualisation tools contribute as an affecting factor in the usability of multifaceted unstructured big data. However, the participants mostly disagree that the reason for this factor’s existence is due to poor performance in functionalities, scalability and response time, and poor quality of conducting data visualisation on incoming multifaceted unstructured big data. So, while this study can conclude that better software and programs must be developed to overcome the limitations of
the visualisation tools in the market today, it cannot determine the limitations themselves.

E. Reliability Analysis

To conduct the reliability analysis of the collected data, the Cronbach’s Alpha (α) has been calculated to check and verify for internal consistency. The results of the analysis are given in Table 4.6.

| Factor                          | Number of Questions (K) | Cronbach’s alpha (α) |
|---------------------------------|-------------------------|----------------------|
| 1. Insufficient Features in Popular Languages | 6                       | 0.709                |
| 2. User Interaction Requirement | 5                       | 0.706                |
| 3. Storage Issues               | 5                       | 0.634                |
| 4. Limitations of Visualisation Tools | 5                       | 0.712                |

The factors, insufficient features in popular programming languages and user interaction, scored 0.709 and 0.706 respectively, and the factor, limitations of visualisation tools, scored 0.712, all within the acceptable range, whereby the responses from the participants are considered acceptable for a study. However, when it comes to storage issues as a factor, the score is 0.634, which is lower than the acceptable range of 0.7. The reason for this low score is due to the collected responses being inconsistent, hence making storage issues as a factor affecting the usability of multifaceted unstructured big data. Scores between 0.6 and 0.7 are considered questionable in any study, but as the study is conducted among a small sample size of 96 participants, the responses can be generally acceptable, according to [35].

V. Conclusion

Big data, when processed and converted into information, is considered highly valuable to most organisations and companies. It is challenging to extract information from multifaceted unstructured data in large volumes. The four factors identified in this research work are believed to contribute to address this issue. To develop solutions it is necessary to identify the factors and reasons behind the problem, which this research work has done successfully with its initial findings. The four factors as identified in the research are subjective and fundamental level mainly related to the tools and management of data. With limited scope of the research, it is hard to judge the impact of these factors on the usability of unstructured data and their correlation with each other and other unexplored factors. The contribution of the research is significant and maintains the novelty to addressing the challenges of non-usability and unproductivity of big data.

A. Future Work

Some limitations in the present research could be improved in future studies. First, findings are very generalised while it needs more precise data, analysis and explanations for which more qualitative studies are required. Besides, the limitations of data visualisation tools needs to be explored further as a factor affecting the usability of multifaceted unstructured big data, which could not be determined in this study. As storage issues scored a value lower than 0.7, it would be wise to consider exploring this factor as well with a large sample size. Further, more affecting and critical factors should be explored for effectively addressing the issues of usability and unproductivity of big data problems.

REFERENCES

[1] BDW Editor, “The Value In Unstructured Data,” 2016. [Online]. Available: http://blog.bigdataweek.com/2016/08/01/value-unstructured-data/. [Accessed 2 March 2018].
[2] SAS Institute Inc., “Big Data; What it is and why it matters,” 2017. [Online]. Available: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html. [Accessed 11 February 2018].
[3] A. N. F. M. K. Judith Hurwitz, Big Data For Dummies, Paperback, 2013.
[4] F. Stroud, “Top 10 Enterprise Database Systems of 2017,” 2017. [Online]. Available: https://www.serverwatch.com/server-trends/slideshows/top-10-enterprise-database-systems-to-consider-2015.html. [Accessed 7 March 2018].
[5] C. Taylor, “Structured vs. Unstructured Data,” 2017. [Online]. Available: https://www.datamation.com/big-data/structured-vs-unstructured-data.html. [Accessed 3 March 2018].
[6] People Data Labs, “The Problem with Searching through Unstructured Data,” 2018. [Online]. Available: https://www.peopledatalabs.com/blog/searching-unstructured-data. [Accessed 11 February 2018].
[7] D. Gewirtz, “Volume, velocity, and variety: Understanding the three Vs of big data,” 2016. [Online]. Available: http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/. [Accessed 26 March 2018].
[8] S. Uthayasankar, M. K. Muhammad, I. Zahir and W. Vishanth, “Critical analysis of Big Data challenges and analytical methods,” Journal of Business Research, vol. 70, no. 1, pp. 263-286, 2017.
[9] A. Klein, “Hard Drive Cost Per Gigabyte,” 2017. [Online]. Available: https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/. [Accessed 7 March 2018].
[10] J. Martin, “The best SSD 2018,” 2018. [Online]. Available: https://www.techadvisor.co.uk/test-centre/storage/best-ssd-2018-3235200/. [Accessed 7 March 2018].
[11] A. Salehinia, “Comparisons of Relational Databases with Big Data: a Teaching Approach,” South Dakota State University, Brookings, 2017.
[12] mongoDB, “What Is A Non Relational Database,” mongoDB, 2018. [Online]. Available: https://www.mongodb.com/scale/what-is-a-non-relational-database. [Accessed 26 July 2018].
[13] NoSQL, “NoSQL,” 2018. [Online]. Available: http://nosql-database.org/. [Accessed 7 March 2018].
[14] BrightPlanet, “Structured vs. Unstructured Data,” 2012. [Online]. Available: https://brightplanet.com/2012/06/structured-vs-unstructured-data/. [Accessed 7 March 2018].
[15] Ciklum, “BIG DATA AND THE CHALLENGE OF UNSTRUCTURED DATA,” 2017. [Online]. Available: https://www.ciklum.com/blog/big-data-and-the-challenge-of-unstructured-data/. [Accessed 4 March 2018].
[16] D. AnHai, N. Jeffrey F., R. Raghu, B. Akanksha, C. Xiaoyong, C. Fei, C. Ting, C. Eric, D. Pedro, G. Byron, G. Chaitanya, H. Jiannsheng, S. Warren and V. Ba-Quy, “Information Extraction Challenges in Managing Unstructured Data,” SIGMOD Record, vol. 37, no. 4, pp. 14-20, 2008.
[17] Computer Hope, “Command line vs. GUI,” 2017. [Online]. Available: https://www.computerhope.com/issues/ch000619.htm. [Accessed 7 March 2017].
[18] C. Trobridge, “What comes after Automation for Software Development?,” 2016. [Online]. Available: https://www.ibm.com/developerworks/community/blogs/invisibl
read/entry/What_comes_after_Automation_for_Software_Development?lang=en. [Accessed 7 March 2018].

[19] S. I. Inc., “Data Visualization What it is and why it matters,” 2018. [Online]. Available: https://www.sas.com/en_my/insights/big-data/data-visualization.html. [Accessed 7 March 2018].

[20] Big Data Software, “Big Data Software,” 2018. [Online]. Available: https://www.capterra.com/big-data-software/. [Accessed 7 March 2018].

[21] K. Bouwkamp, “The 9 Most In-Demand Programming Languages of 2016,” 2016. [Online]. Available: http://www.codingdojo.com/blog/9-most-in-demand-programming-languages-of-2016/. [Accessed 6 March 2018].

[22] R. Bhargava, “What Is the Best Programming Language For Games,” 2017. [Online]. Available: https://www.freelancer.com/community/articles/what-is-the-best-programming-language-for-games. [Accessed 6 March 2018].

[23] study.com, “Object-Oriented Programming vs. Procedural Programming,” 2018. [Online]. Available: https://study.com/academy/lesson/object-oriented-programming-vs-procedural-programming.html. [Accessed 6 March 2018].

[24] W. Bright, “So You Want To Write Your Own Language?,” 2014. [Online]. Available: http://www.drdobbs.com/architecture-and-design/so-you-want-to-write-your-own-language/240165488. [Accessed 6 March 2018].

[25] A. S. Fischler, “Quantitative Research Methods,” 2010. [Online]. Available: https://education.nova.edu/Resources/uploads/app/35/files/arc_doc/quantitative_research_methods.pdf. [Accessed 7 March 2018].

[26] P. D. Leedy, Practical research: planning and design, 1st Edition ed., New Jersey: Prentice-Hall, 1993.

[27] J. Mason, Qualitative Researching, 2nd Edition ed., London: SAGE Publications Ltd, 2002.

[28] Gert Van Dessel, “How to determine population and survey sample size?,” 2013. [Online]. Available: https://www.checkmarket.com/blog/how-to-estimate-your-population-and-survey-sample-size/. [Accessed 8 March 2018].

[29] DataStar, “What Every Researcher Should Know About Statistical Significance (October 2008),” 2008. [Online]. Available: http://www.surveystar.com/startips/oc2008.pdf. [Accessed 28 July 2018].

[30] W. M. Trochim, “Descriptive Statistics,” 20 October 2006. [Online]. Available: https://www.socialresearchmethods.net/kb/statdesc.php.

[31] Fundamental of Statistics, “Fundamental of Statistics,” 8 October 2012. [Online]. Available: http://www.statisticshowto.com/cronbachs-alpha-spss/. [Accessed 29 July 2018].

[32] Institute for Digital Research and Education, “WHAT DOES CRONBACH’S ALPHA MEAN?,” 2018. [Online]. Available: https://stats.idre.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/. [Accessed 29 July 2018].

[33] Stephanie, “Cronbach’s Alpha: Simple Definition, Use and Interpretation,” 2014. [Online]. Available: http://www.statisticshowto.com/cronbachs-alpha-spss/. [Accessed 29 July 2018].

[34] DataStar, Inc., “How to Interpret Standard Deviation and Standard Error in Survey Research,” 2013. [Online]. Available: https://www.greenbook.org/marketing-research/how-to-interpret-standard-deviation-and-standard-error-in-survey-research-03377. [Accessed 8 August 2018].

[35] P. H. C. H. Moss SC, “Reliability and validity of the PAS-ADD Checklist for detecting psychiatric disorders in adults with intellectual disability,” Journal of Intellectual Disability Research 1998, vol. 42, no. 42, pp. 173-183, 1998.