https://doi.org/10.17576/apjitm-2022-1102-10
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 11 No. 2 December 2022: 123 - 132
e-ISSN: 2289-2192

Received: 6 April 2022
Accepted: 26 July 2022
Published: 1 December 2022

PRIVACY BASED CLASSIFICATION MODEL OF PUBLIC DATA BY UTILIZING TWO-STEPS VALIDATION APPROACH

MASNIDA HUSSIN
RAJA AZLINA RAJA MAHMOOD
NUR RAIDAH SALIM

ABSTRACT

Digital information has become a trend and is integral to modernizing and leveraging various resources in Information Technology (IT). Vast data and information can be obtained anytime and anywhere at our fingertips through ICT facilities. This is considered public data due to its being shared publicly, such as on social media. Public data can be arranged according to various criteria and formats. Users have a right to understand which data can be publicly shared and which data is supposed to be in a private state. However, people always misunderstand and mislead which data needs to be secured and which can be shared. It is further critical when this public data is already exposed to data breaches and data theft. In this work, we propose a data privacy classification approach for public data where this data resides on digital platforms. It aims to inform the public about the level of data privacy before they reveal it on open and free digital platforms. We employed three different privacy classes: low, medium, and high. In response to that, we identified entities of public data that refer to digital information platforms such as websites, mobile apps, and online systems. We then dug further into the data attributes of each entity. The public data attributes are sorted and passed to respondents to obtain their input regarding their decisions on which privacy class is suitable for the respective attribute. Based on the input from respondents, we then used a Naive Bayesian classifier to generate probability weightage for re-assigning the data attributes into the most suitable privacy class. This two-level data classification stage brings better perspectives on data privacy. This modified version of the public data privacy class is then verified by the respondents to analyze their preferences while measuring users' satisfaction. According to the results, our public data privacy classification model meets public expectations. Optimistically, well-organized data classification contributes to better data practices.

Keywords**:** public data; data classification model; data privacy; Naïve Bayesian classifier.

MODEL KLASIFIKASI BERASASKAN PRIVASI DATA AWAM DENGAN MENGGUNAKAN PENDEKATAN PENGESAHAN DUA PERINGKAT

ABSTRAK

Maklumat digital telah menjadi trend dan penting untuk memodenkan dan memanfaatkan pelbagai sumber dalam Teknologi Maklumat (IT). Data dan maklumat yang luas boleh diperolehi pada bila-bila masa dan di mana sahaja di hujung jari kami melalui kemudahan ICT. Ini dianggap sebagai data awam kerana ia dikongsi secara terbuka, seperti di media sosial. Data awam boleh diatur mengikut pelbagai kriteria dan format. Pengguna mempunyai hak untuk memahami data mana yang boleh dikongsi secara terbuka dan data mana yang sepatutnya berada dalam keadaan peribadi. Walau bagaimanapun, orang sentiasa salah faham dan mengelirukan data mana yang perlu dijamin dan yang boleh dikongsi. Ia lebih kritikal apabila data awam ini sudah terdedah kepada pelanggaran data dan kecurian data. Dalam kerja ini, kami mencadangkan pendekatan klasifikasi privasi data untuk data awam di mana data ini berada di platform digital. Ia bertujuan untuk memaklumkan kepada orang ramai tentang tahap privasi data sebelum mereka mendedahkannya di platform digital terbuka dan percuma. Kami menggunakan tiga kelas privasi yang berbeza; rendah, sederhana, dan tinggi. Sebagai tindak balas kepada itu, kami mengenal pasti entiti data awam yang merujuk kepada platform maklumat digital seperti laman web, aplikasi mudah alih dan sistem dalam talian. Kami kemudian menggali lebih jauh ke dalam atribut data setiap entiti. Atribut data awam disusun dan diserahkan kepada responden untuk mendapatkan input mereka berkenaan dengan keputusan mereka mengenai kelas privasi yang sesuai untuk atribut masing-masing. Berdasarkan input daripada responden, kami kemudian menggunakan pengelas Naive Bayesian untuk menjana pemberat kebarangkalian untuk memperuntukkan semula atribut data ke dalam kelas privasi yang paling sesuai. Peringkat klasifikasi data dua peringkat ini membawa perspektif yang lebih baik mengenai privasi data. Versi kelas privasi data awam yang diubah suai ini kemudiannya disahkan oleh responden untuk menganalisis pilihan mereka sambil mengukur kepuasan pengguna. Mengikut keputusan, model klasifikasi privasi data awam kami memenuhi jangkaan orang ramai. Secara optimis, klasifikasi data yang teratur menyumbang kepada amalan data yang lebih baik.

Kata kunci**:** data awam; model klasifikasi data; privasi data; Pengelas Bayesian Naïve

INTRODUCTION

The digital era significantly changed the way people lived by making daily life much easier. The emerging use of digital information leads to a wealth of information and meaningful resources that bridge communication gaps across the globe. The growth of digital information platforms, for example, online shopping, social media, and big data apps, offers diverse information sharing opportunities that lead to the realization of the culture of knowledge (Sanderson, et al., 2019; Wieringa, et al., 2021; Di Minin, et al., 2021). In this era, electronic data and records have been widely used in the collaboration of information among various entities, i.e., organizations and businesses. In response to the Public Sector Open Data (DTSA) (www.malaysia.gov.my available online) that aims to support the implementation of open data, the public should cultivate the basic skills to evaluate information and determine its importance. As described in the Public Sector Open Data Implementation Guidelines (www.malaysia.gov.my available online), every piece of data needed for clear definition and features is provided. As a result, it aids in justifying data/information privacy and measuring it when it is required to be publicly shared on social networking sites.

Note that even though the data can be publicly shared, some level of privacy needs to be implemented to ensure unintentional and unlawful manipulations of such data are prevented. The public also always misunderstands and misleads which data needs to be safeguarded and which can be shared (Mohamed, et al., 2012; Ahmad, et al., 2019). Some information might be unimportant to others but might be vital to some people. In these past studies (Gosain, et al., 2014; Wang, et al., 2019; Lim, Hyun-il 2020), there are methods of enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protection. The public data that can openly be discovered from the digital information platforms, including personal details (e.g., name, phone number, electronic mail, home/office address), job and affiliation, as well as social life (e.g., hobbies, family and friend contacts), requires privacy identification. In any case required for public awareness. It is for the public's benefit because it might happen that irresponsible parties have already exploited and illegally used the users' data without the owner's knowledge.

The public must be well-informed about the privacy of their data that is being shared on open digital information platforms. It motivates us to form a data classification model to provide public awareness of their data privacy. Our model consists of two main procedures: identification and verification. The identification procedure aims to accurately elucidate public data that has different privacy requirements. For example, some information may be unimportant to others but critical to others. Hence, the public's perspectives on the level of data privacy are important to study. Meanwhile, the verification procedure involves justifying the privacy level using a probability function. It is due to its diverse perspectives in terms of data privacy levels. Then, by using the weightage probability function from the Nave Bayesian scheme, it helps with accurately classifying the public data into the corresponding privacy class. These two-stage procedures can contribute to a better understanding of the privacy implications of public data and better usability for creating digital media content.

LITERATURE REVIEW

Public data in digital information platforms refer to data or information that is available and can be read through an electronic or digital platform. Such information is assembled and made available to the public by either an organization or for several personal purposes. Part of such data includes personal data, that leads to other related information i.e., health and social information. There are studies on information extraction of big consumer opinion data sets that are explained from various perspectives, including data acquisition, opinion target recognition, feature identification and sentiment analysis, opinion summarization and sampling (Jin, et al., 2019; Wang, et al., 2019; Lim, Hyun-il 2020; Di Minin, et al., 2021). The authors (Jin, et al., 2019) provide some checklists for understanding the consumer data, e.g., critical customer needs and customer opinions, for making effective comparisons among the available products. Their survey helps researchers and practitioners understand how consumer opinion data can be processed, analyzed, and exploited. Despite the fact that their studies are not in our domain, the public's opinions help the researchers gain a better understanding in order to produce a better product design. Meanwhile, in Di Minin, et al. (2021) the data was analyzed using a structural equation modeling technique that concerns the importance of severity, self-efficacy, and perceived vulnerability in data sharing through social networking sites.

It is necessary to have knowledge of the privacy level of public data to ensure people/the public get more awareness when they want to share any data on digital information platforms. Note that the processing of sensitive personal data, including personal details and health information, is protected under the Malaysia Personal Data Protection Act 2010. If the data is of high privacy, then they should check the data policy offered by the platforms. Note that every piece of data has its own confidential values (Enaizan, 2020; Song, et al., 2022). Prior to that, the public should understand that there are private elements to each piece of data, even though the data is reasonable to be publicly shared. For example, a car plate number might be considered less private information, but it becomes highly significant to the police officer, insurance company, etc. As a result, the public must be aware that they cannot simply share such information with the authorities. However, due to the rapid growth in data-collecting technologies (Wang, et al., 2019; Wieringa, et al., 2021), it has led to the emergence of new forms of data. A new direction in data definition and classification is required, particularly for the privacy levels of those that are visible on social media. For example, any person is able to know the other person's daily activities by reading the posting, and it's even more dangerous if the status is left for public view. Such scenarios might lead to dangerous conditions. Even though the public has a right to share their information on digital platforms, due to the increase in the number of cyber-attacks, they must be aware of what they can or cannot share publicly.

DATA CLASSIFICATION METHOD

We propose a data privacy model for classifying the confidential levels of an individual’s information that is shared publicly. "Public data," in our perspective, is the personal information that exists and can be freely read by anyone on digital platforms. In other words, we are forming a data privacy catalog so the public can refer to it while improving their awareness of their own data privacy status. Importantly, our privacy model was developed by involving public perspectives and a technical classification approach. This two-stage classification approach aims to ensure the data privacy status is accurate and relevant to the public.

In our preliminary study, there is a data observation made (Figure 1) to understand the criteria and features of each attribute in public data. In our data privacy model (Figure 2), the collected attributes of data from digital platforms will be verified for their privacy level by our informant (respondent). The feedback (output) from them will be the input for our classification approach. The attributes of the data are classified by using the Naïve Bayesian classifier. This approach has also been proposed in (Xue, et al., 2019) for controlling privacy by adding restrictions in accessing useful information from the databases. In our work, the classifier generated a weightage value to indicate the privacy level of each attribute in response to each privacy class. The attributes of data will be assigned to the privacy class that gives the highest weighted value. In developing the data privacy classification model (Figure 2) there are two stages comprised, data collection in stage 1 then, data evaluation and data verification in stage 2.

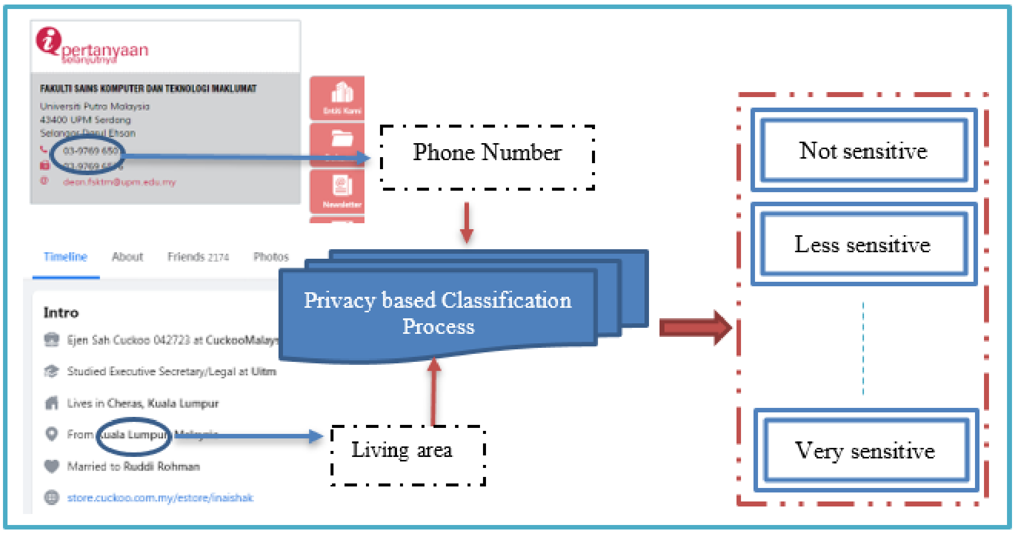


FIGURE 1. Data Observation Process

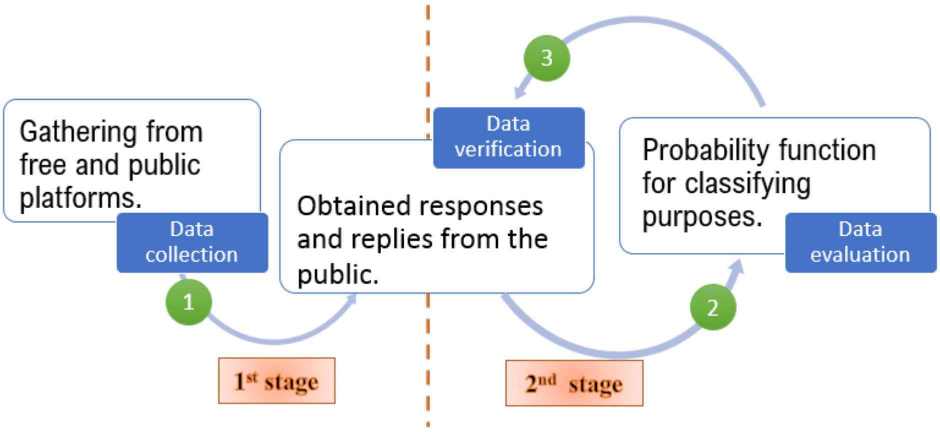

Gathering from free and public platforms.
Data collection
1
Obtained responses and replies from the public.
Data verification
3
Probability function for classifying purposes.
Data evaluation
2
1st stage
2nd stage

FIGURE 2. Data Classification Model

IDENTIFYING

This procedure consists of several activities as follows.

1. Collecting data from open and free public digital platforms
   At this stage, we identified open and free public digital platforms. In response to our public data definition, the chosen digital platforms in this work are social media, organization websites, a registration system, and a survey system. These platforms are represented as entities. The theme of data collection is focused on personal information, work-related information, and lifestyle information. Each entity might have more than one theme of data. For example, organization websites contain personal and work-related information. The data for each entity is listed and recorded, then refers as attributes.

2. Filtering the attributes
   The attributes are then classified according to their privacy. We divided data privacy into three categories based on the level of confidentiality: low, medium, and high. Low means the data does not need to be secretive; it can be revealed without any privacy concern. Meanwhile, high refers to the data that requires privacy in terms of its usability and discovery. For example, it is compulsory to ask the data owner's permission before the data can be used for other purposes. If that happens, then the owner has a right to call for their justice. The medium type of data means the data requires some level of privacy, but it is still negotiable to be used and exposed for other purposes.

3. Identifying respondents
   At this stage, the number of respondents is determined before contacting and communicating with them to obtain their permission to contribute to this work. It is very important due to the fact that the respondents are required to take part twice in the verification process. Specifically, their involvement in the early parts, when the data is required to be assigned into privacy groups, and later, after the data is assembled by our probability method. In this work, we identified three major groups of respondents according to age. By selecting age as the sample parameter, it means revealing the public data privacy perspectives of different maturities and generations. The respondents' profiles are selected randomly but guided. The sample respondents are classified into four groups: 13 to 20-years old, 21 to 30-years old, 31 to 50-years old, and more than 50-years old. For each group, we set the saturation level at 30. It means we only process 30 respondents' feedback from each age group. Even though there is more than 30 feedback in some groups, a balance in input for next processing is important in this work for fairness comparison.

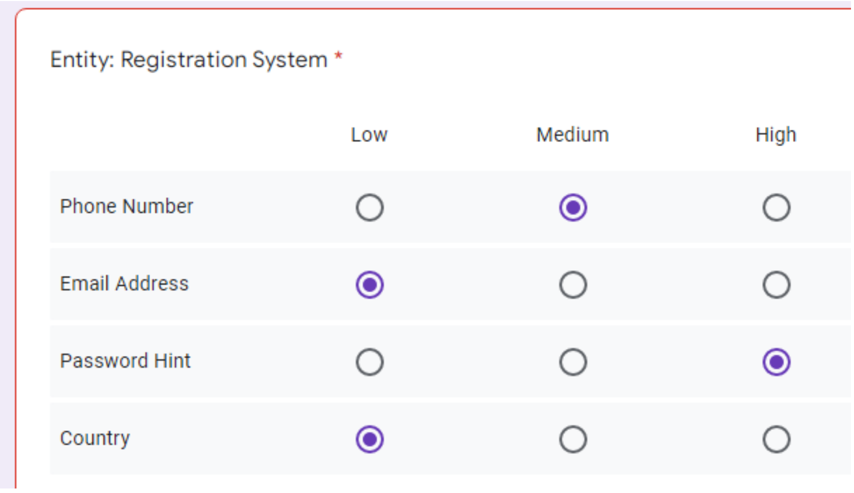
VERIFICATION

This procedure consists of several activities as follows.

1. Analyzing public perception in data attributes
   This activity means obtaining replies from the respondents. Previously, the instrument was created using an online form, as shown in Figure 3. Our low, medium, and high privacy levels are thoroughly elucidated in the instrument. The

respondent's roles are to choose a privacy group (i.e., low, medium, or high) by clicking on the level of privacy that is thought to be appropriate for the public attributes provided. The responses are then analyzed.

$$P(Entity|Att1) * P\ (Entity|\ Att2) * \ldots P\ (Entity|\ Attn)$$



Entity: Registration System *

	Low	Medium	High
Phone Number	○	●	○
Email Address	●	○	○
Password Hint	○	○	●
Country	●	○	○

FIGURE 3. The instrument using online form (Entity: Registration System)

2. Amending the data using a classifier algorithm.
This is one of the major activities in our work. The response from the public towards data privacy is technically insufficient to claim accuracy on public data privacy. Open opinion is required to be supported by further investigation. In this study, the probability function for efficient cataloging is a Naive Bayesian classifier. The classifier algorithm is designed by counting on how frequent the data attributes have been chosen for each privacy level. The idea is to compute the three probabilities, that is the probability of the data attributes being low, medium, or high. Initially, we compute the 'prior' probabilities for each of the classes of privacy. That is, by referring to responden answers where a proportion of each privacy class out of all the privacy from a population. The population used is 4 x 30 dimensional where 4 represents four groups of respondents and 30 is the number of respondents per group. By referring to the Bayes rule, the constant value, attributes in our case, A and privacy levels B are reflecting to each other. From this indication, the probability P(A|B) refers to a comparison made among the classification events where a factor of low, medium and high has been identified. Details on the formula (Equation 1 until 3) and its operation for each privacy classes are given as below.

$$P\ (Low\ |\ (Entity\ |\ Att_1 \ldots Att_n) = \frac{P\ (Entity,\ Att_1\ |\ Low) * P\ (Entity,\ Att_2\ |\ Low) * \ldots P\ (Entity,\ Att_n\ |\ Low)}{P(Entity|Att_1) * P\ (Entity|\ Att_2) * \ldots P(Entity|\ Att_n)} \quad [1]$$

$$P\ (Medium\ |\ (Entity\ |\ Att1..Attn) =$$
$$P\ (Entity,\ Att1\ |\ Medium) * P\ (Entity,\ Att2\ |\ Medium) * \ldots P\ (Entity,\ Attn\ |\ Medium)$$

$$P(Entity|Att1) * P\ (Entity|\ Att2) * \ldots P\ (Entity|\ Attn)$$

[2]

$$P\ (High\ |\ (Entity\ |\ Att1 \ldots Attn) =$$
$$P\ (Entity,\ Att1\ |\ High) * P\ (Entity,\ Att2\ |\ High) * \ldots P\ (Entity,\ Attn\ |\ High)$$

[3]

The Naïve Bayesian classifier generated a weightage for each privacy group according to input from the respondents' feedback. The data attribute will be allocated to the highest probability weightage of the privacy level where it can be low, medium, or high. We then rearranging the public data attributes into their respective privacy groups in accordance to its class, and refers it as modified version. Given example in Figure 4 that the *Country* indicates has Medium privacy meanwhile *Password Hint* is discovered has High privacy.

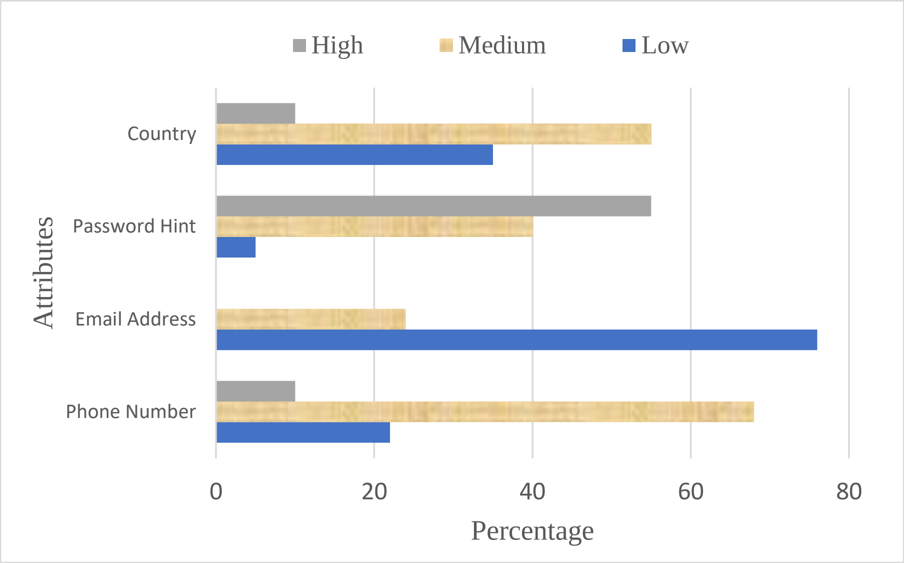


FIGURE 4. Result from Privacy Class (Entity: Registration System)

3. Analyzing public feedback on data privacy model.
   By using our modified version of public data with their respective privacy level, we then one more time asked the respondents to review the public data privacy. This time they are given a set of questionnaires developed by using a 5-point Likert scale that asks for their satisfaction towards the modified version of public data privacy. The respondents are required to review each of the data attributes whether the attributes are assigned to a suitable privacy class or not. Their responses have then been analyzed.

RESULTS AND DISCUSSION

The sample respondents who are sampling based on age are 13 to 20-years old, 21 to 30-years old, 31 to 50-years old and more than 50-years old. This selection aims to analyse various people's perspectives on privacy and personal data on digital platforms. Due to the numbers at each sampling group being not balanced, we then accumulated normalization value from the total respondents for merely considering the top 30 respondents' answers from each group. In this work we successfully identified 30 data attributes from various sources (Figure 5). We eliminate the redundancy of data by filtering the attributes' features similarity:

Based on Figure 6, it shows that most respondents convey positive feedback. The result also shows that, on average, approximately 80% of respondents are satisfied with our approach for the privacy data model. It is taken by considering both fully satisfied and satisfied feedback. Even though our number of data attributes is relatively small, nearly all of them are satisfied with the data attributes listed. We also collected their comments, and some of the comments mentioned that the data attributes are close and are used in their daily activities. There are also some suggestions to form a similar methodology in designing other data privacy models, i.e., critical, and confidential data.

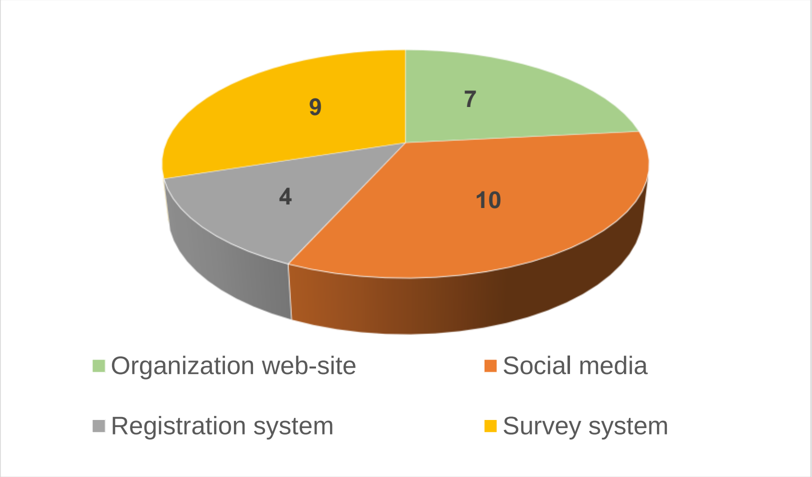


FIGURE 5. Data Collection Distribution based on Sources

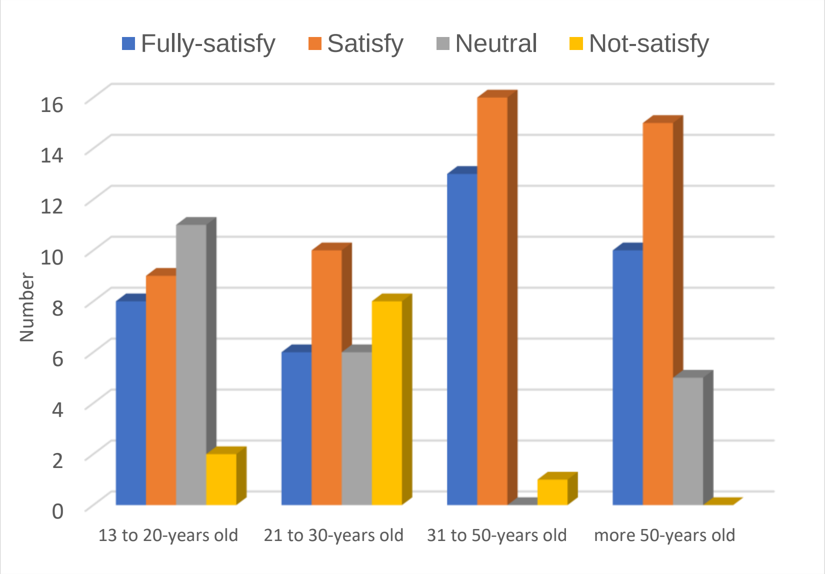


FIGURE 6. End-user Satisfaction Survey towards the Privacy Data Model

CONCLUSION

Development of digital procedures for data sharing and analysis is important in the digital era, especially with the advancement of mobile technologies and social media platforms. By ensuring the usefulness of digital platforms, effective information sharing can be achieved in many sectors for better business operations. Besides the advantages of such platforms, digital content (data/information) presents a challenge in system truthfulness. It attracts many cyber-attackers who steal and manipulate the data. Therefore, the public is supposed to be aware of which data should be shared and which should not. We believe the data model is more accurate when the real users (i.e., the public) get involved in the data modeling process. For this reason, our data privacy model involves the public's perspective in defining the data privacy level. Later, the Naïve Bayesian classifier further makes the classification realistic. Even though the data can be publicly shared, awareness of its privacy also needs to be considered to avoid the public being a victim of cyber-attacks. A better data catalog in terms of privacy level helps to regulate and refine digital procedures for data sharing and privacy policy. Optimistically, better data privacy awareness among digital platform users leads to cyber resilience.

REFERENCES

Ahmad, N. A., and Othman, N. 2019. Information Privacy Awareness Among Young Generation in Malaysia. Journal of Science, Technology, and Innovation Policy, 5(2), pp. 1-10.

Di Minin, E., Fink, C., Hausmann, A., Kremer, J., and Kulkarni, R. 2021. How to address data privacy concerns when using social media data in conservation science. Conservation Biology, 35(2), pp. 437-446.

Enaizan, O., Zaidan, A. A., Alwi, N. M., Zaidan, B. B., Alsalem, M. A., Albahri, O. S., and Albahri, A. S. 2020. Electronic medical record systems: Decision support examination framework for individual, security and privacy concerns using multi-perspective analysis. Health and Technology, 10(3), pp. 795-822.

Gosain, A., and Chugh, N. 2014. Privacy preservation in big data. International Journal of Computer Applications, 100(17).

Jin, J., Liu, Y., Ji, P., and Kwong, C. K. 2019. Review on recent advances in information mining from big consumer opinion data for product design. Journal of Computing and Information Science in Engineering, 19(1).

Lim, Hyun-il. 2020. Design of Security Level Management System for Enhancing Security of Data Access. Indian Journal of Computer Science and Engineering. 11. pp. 375-382. 10.21817/indjcse/2020/v11i4/201104234.

Mohamed, N., and Ahmad, I. H. 2012. Information privacy concerns, antecedents and privacy measure use in social networking sites: Evidence from Malaysia. Computers in Human Behavior, 28(6), pp. 2366-2375.

Open Government Data. Retrieved from https://www.malaysia.gov.my/portal/content/30024 (Available online).

Sanderson, T., Reeson, A., and Box, P. 2019, April. Optimizing Open Government: an economic perspective on data sharing. In Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance, pp. 140-143.

Song, J., Han, Z., Wang, W., Chen, J., and Liu, Y. 2022. A new secure arrangement for privacy-preserving data collection. Computer Standards & Interfaces, 80, pp. 103582.

Wieringa, J., Kannan, P. K., Ma, X., Reutterer, T., Risselada, H., and Skiera, B. 2021. Data analytics in a privacy-concerned world. Journal of Business Research, 122, pp. 915-925.

Wang, G., Lu, R., and Guan, Y. L. 2019. Achieve privacy-preserving priority classification on patient health data in remote eHealthcare system. IEEE Access, 7, pp. 33565-33576.

Xue, Q., Zhu, Y., anad Wang, J. 2019. Joint distribution estimation and naïve bayes classification under local differential privacy. IEEE transactions on emerging topics in computing, 9(4), pp. 2053-2063.

*Masnida Hussin[1,2], Raja Azlina Raja Mahmood[1], Nur Raidah Salim[2]

[1] Faculty of Computer Science and Information Technology, University Putra Malaysia (UPM), Serdang, Selangor.

[2]Institute for Mathematical Research (INSPEM), University Putra Malaysia (UPM), Serdang, Selangor.

{*corresponding author: masnida@upm.edu.my}