PERTANIKA
JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

An Attribute-based Data Privacy Classification Through the Bayesian Theorem to Raise Awareness in Public Data Sharing Activity

Nur Aziana Azwani Abdul Aziz[1], Masnida Hussin[1,2]* and Nur Raidah Salim[2]

[1]Department of Communication Technology and Networking, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia
[2]Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia

ABSTRACT

The growth of the digital era with diverse existing electronic platforms offers information sharing and leads to the realization of a culture of knowledge. Vast amounts of data and information can be reached anywhere at any time, fingertips away. These data are public because people are willing to share them on digital platforms like social media. It should be noted that not all information is supposed to be made public; some is supposed to be kept private or confidential. However, people always misunderstand and are misled about which data needs to be secured and which can be shared. We proposed an attribute-based data privacy classification model using a Naïve Bayesian classifier in this work. It aims to identify and classify metadata (attributes) commonly accessible on digital platforms. We classified the attributes that had been collected into three privacy classes. Each class represents a level of data privacy in terms of its risk of breach. The public (respondent) is determined according to different ages to gather their perspective on the unclassified attribute data. The input from the survey is then used in the Naïve Bayesian classifier to formulate data weights. Then, the sorted privacy data in the class is sent back to the respondent to get their agreement on the class of attributes. We compare our approach with another classifier approach. The result shows fewer conflicting reactions from the respondents to our approach. This study could make the public aware of the importance of disclosing their information on open digital platforms.

Keywords: Naïve Bayes, privacy classification, public data attribute

ARTICLE INFO

Article history:
Received: 03 February 2023
Accepted: 14 June 2023
Published: 24 November 2023

DOI: https://doi.org/10.47836/pjst.32.1.14

E-mail addresses:
azianaazwani00@gmail.com (Nur Aziana Azwani Abdul Aziz)
masnida@upm.edu.my (Masnida Hussin)
nurraidah@upm.edu.my (Nur Raidah Salim)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

The digital platforms transformed people's lives by making daily tasks much easier. In this era, data and records have been widely used to improve storing, accessibility, and sharing, such as the collaboration of information among various entities, i.e., organizations and businesses. Moreover, the fact that the data is stored in the Cloud computing environment and can reside anywhere beyond the geographical boundary might cause the user to lose control over their data. Data communication via digital platforms makes news, rumors, feedback, comments, reports, and other information available for retrieval and response by anyone. Such information might consist of confidential data that has been implicitly shared for several purposes. Some information might not be important to users but might be vital to others. The publicly accessible information includes personal details (e.g., name, phone number, electronic mail, home, or office address), job and affiliation, and health information (e.g., medication lists, diagnostic tests, physical assessments, and history observations). The high demand for such data input collected on the digital platform has confused in determining the sensitivity levels of the data. This unnoticed data-sharing activity results in data leaks and exploitation by irresponsible parties.

An open digital platform is an online space where people can interact with it because technological advancements have drastically altered people's daily lives. Public data is mostly shared in an open digital space. There are lots of public data attributes that can be found on open digital platforms. Some examples of public data attributes are gender, age, education status, hometown, nationality, and email address. In our project, the information refers to "public data," which anyone can access from digital platforms such as social media and company websites. Noticeably, the open digital platform creates great potential for the economy and society; however, it may invade someone's privacy as public data is shared freely. The Australian Cyber Security Centre advises people to remain cautious when disclosing too much personal information online (Analysis & Policy Observatory, 2020). Someone who discloses their personal information risks becoming a victim of identity theft, stalking, and harassment. For instance, protecting personal information can help avoid phishing scams. The term "phishing" refers to a scam when fraudsters send emails or pop-up messages that seem to have been sent by a bank, a government agency, or an online retailer. The message directs the victims to a website or phone number that they can contact to update their account information or to receive a reward. It can imply that something negative will occur if the victims ignore it. It implies that someone's privacy has been violated.

Individuals' perceptions of privacy vary, which may result in a lack of knowledge. For instance, some individuals could believe that their workplace is unrelated to privacy, while others would hold the opposite opinion. In our work, we focus on developing a data privacy classification using the Bayesian theorem to identify and classify the public data attributes into different levels of privacy (i.e., low, medium, and high). Then, a data catalog containing our data attributes and their privacy classes is created and meant to be shared with the public to increase their data privacy awareness. The remainder of this paper follows the study background, public data classification process, results, discussion, and concluding remarks.

LITERATURE REVIEW

Public data refers to available data and information accessed through an electronic or digital platform. Organizations or individuals gather and make this information available for various purposes. According to the MyGoverment (2019), each piece of data must have a precise definition and set of characteristics. This document guideline aims to support the implementation of open data, where the public should cultivate the basic skills to evaluate information and determine its importance. It assists in justifying data and information privacy and measuring before the information is publicly shared on any digital platform. The data can be explained from various perspectives, including data acquisition, opinion target recognition, feature identification, sentiment analysis, opinion summarization, and sampling (Sanderson et al., 2019; Reza et al., 2020).

The ability of an individual to disclose just certain personal information about themselves is known as privacy. In the social media era, people became concerned about their data being visible online. However, the privacy policy on social networking sites is unclear and not well-defined. Salim et al. (2022) and Cain and Imre (2022) identified the different privacy options the social media site offers users to protect their information. These security options strengthen users' ability to divulge information while allowing the information privacy settings to those requiring it. Another researcher stated that sharing information on social media demands a lot of security and privacy settings (Rehman et al., 2022). They also stated that it is difficult to enforce data privacy on social media when people are willing to reveal their personal information publicly. Ravn et al. (2019) specified that users disclose much personal information on Instagram without realizing it is putting them in danger. From a Facebook security perspective, Rashid and Zaaba (2020) mentioned that Facebook users frequently store and exchange different types of personal data, which raises the risk of privacy breaches.

Researchers and practitioners conduct surveys to understand how users' opinions can be processed, analyzed, and used to raise public data privacy awareness. Their survey findings are classified into several key levels for classifying the users' responses in various variables. Bibhu et al. (2021) conducted a survey analysis that revealed that around 56% of people are very concerned about their privacy being publicly shared. Algarni (2019) highlighted that those four sensitivity classifications—high, medium, low, and unclassified are mostly used by researchers to leverage the data's sensitivity. They also mentioned that information that the public can view falls under the "unclassified" category. The information on the medium level is intended for a specific group of people, and the information on the high level is the information that cannot be revealed, is extremely sensitive, and can be accessed by certain people only. Wu et al. (2021) highlighted that the privacy concept is based on 4 types, person, preference, event, and trait, for protecting privacy leakage. They employed a deep learning model and an ontology-based classification approach for grouping those privacy features.

The Bayesian network as a classifier has a transparent probability function and considers the prior information of samples, making Naive Bayes a simple and powerful classification model (Liu et al., 2013; Vu et al., 2022). The Bayes theorem is used by Abraham et al. (2019) to find highly confidential and less confidential data. They proposed a method of classifying emails according to their security levels, which are categorized as highly confidential and lowly confidential emails, using a Naïve Bayesian classifier. Vu (2022) proposed a privacy-preserving Naive Bayes classification based on secure multi-party computation. While achieving a high level of privacy, their model has higher accuracy compared to another classifier. Those works, however, do not investigate public data attributes. Meanwhile, Wibawa et al. (2019) showed the ability of the Naive Bayes classifier to classify the quality of a journal that ranks in Q1, Q2, Q3, Q4, and NQ. Their classifier approach classified the quality of journals and achieved 71.60% accuracy. Zanella-Béguelin et al. (2022) employed the Bayesian method for interpreting differential privacy to obtain a posterior for the confidence interval of the false positive and false negative rates of membership inference attacks. The result is promising but has not been analyzed from the public usability point of view. Besides utilizing the Naïve Bayes classifier, Shallal et al. (2020) proposed a *k-nearest neighbors* algorithm (*k-NN*) to divide Internet of Things (IoT) data into three security levels. However, the computation process of *k-NN* is very difficult due to the procedure of classification, which will utilize whole training samples. However, their three security levels significantly make the classification process more effective and less complex. Our work employs the Naïve Bayes classifier by classifying data privacy into three levels. The ranking of data attribute privacy is initially established from the viewpoint of the public. The Bayes theorem will then be used to quantify each data attribute. We solicit the public's opinion once more to confirm that they agree with the classified data attribute.

METHOD OF CLASSIFICATION PROCESS

This discussion details how the Naive Bayes method retrieves each attribute's privacy class. Additionally, we described each stage of the privacy classification process.

Naïve Bayes Classifier for Privacy Classification

A classifier is a machine learning model using certain attributes to distinguish between objects. The Bayes theorem can be used to calculate the likelihood of A occurring given the presence of B. A is the hypothesis in this case, while B is the proof or evidence (Vu, 2022). The features are thought to be independent. That is, one feature's existence does not change another's behavior. As a result, it is known as naïve. The fundamental idea of Bayes' theorem is shown in Equation 1. P (A | B) is the posterior probability, where the conditional probability distribution represents what parameters are likely after observing the data object (Liu et al., 2013). P (B | A) is the likelihood function, expressing the probability of falling under a specific category or class. P (A) is the prior probability distribution representing prior knowledge or uncertainty about a data object.

$$P\ (A\ |B) = \frac{P\ (B\ |\ A)\ \times\ P\ (A)}{P\ (B)} \qquad [1]$$

By adding new variables that matched our public data attributes privacy setting, we modified Equation 1—given that Equation 2 modifies the Bayes theorem used in this work.

$$P(weightage\ |\ privacy\ level) = \frac{P\ (privacy\ level\ |weightage)\ \times\ P\ (weightage)}{P\ (privacy\ level)} \qquad [2]$$

It is described using Equation 2, where the posterior probability is P (weightage|privacy level) or P(W|PL). It is the probability that the weight is True, given the evidence from the chosen privacy level based on the public count. The public count is the number of votes from survey respondents on the chosen privacy level. P (privacy level|weightage) or P(PL|W) defines the probability of the chosen privacy level based on the public count if the weight is true for the likelihood function. In other words, it uses the sum of the two preceding values. The prior involved in this study is P (weightage), which is the probability of assumed weightage, and P (privacy level), which is the evidence and is the probability of the chosen privacy level based on the public count.

Methodology for Privacy Classification

This work takes several stages for the classifying process (Figure 1). The stages are inter-operable with each other to ensure the consistency of input and output between stages.

Data Observation and Collection. We observed the public data attributes and collected 15 from various sources based on Table 1. We believe there are 15 data attributes that the public normally shares

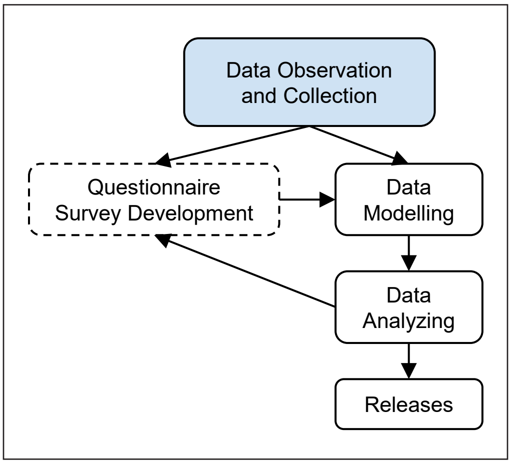

Data Observation and Collection
Questionnaire Survey Development
Data Modelling
Data Analyzing
Releases

Figure 1. Privacy classification process on an open digital platform. We then designed the respondent to represent the public. We identified 40 people as our respondents (Budiu & Moran, 2021), representing a range of ages. The focus of 40 respondents as hard-core Internet users is sufficient for obtaining comprehensive insights on data attribute privacy levels. The respondents come from a variety of social backgrounds, including friends, relatives, and individuals selected at random, and are not part of any specific group. The respondent is engaged in our study through emails based on their willingness to be part of this data collection. The age factor chosen is significant in analyzing their digital data-sharing behaviors. University students and young adults between 20 and 29 are the first age group. Another age group is between 30 and 50 since they have more experience dealing with privacy concerns in real life. Respondents over 50, who are older individuals or may be retirees and are potentially exposed to data privacy exposure, make up the last age group. As a result, there are 3 age groups: 13 respondents between the ages of 20 and 29, 14 respondents between the ages of 30 and 50, and 13 respondents above 50. Figure 2 displays the pie chart of the respondents' age group.

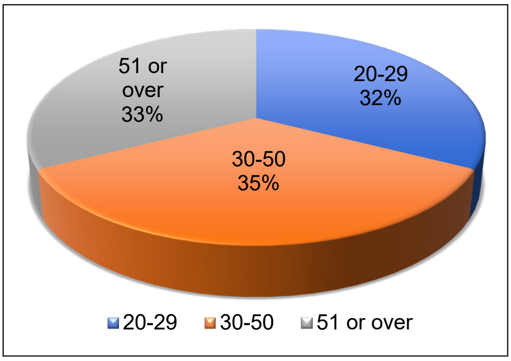


Figure 2. Respondents' age group

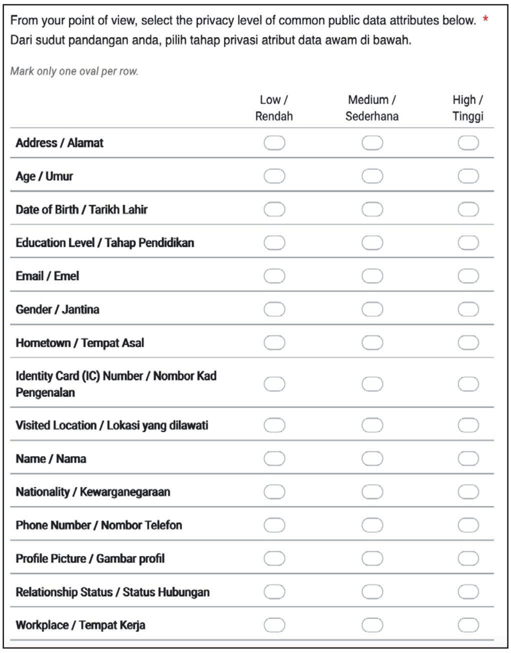
Table 1
Public data attributes and their sources

Public data attributes	Sources
Address	Dokuchaev et al. (2020)
Age	Reza et al. (2020)
Date of Birth	Reza et al. (2020)
Education Level	Wu et al. (2021)
Email	Reza et al. (2020)
Gender	Reza et al. (2020)
Hometown	Reza et al. (2020)
Identity Card Number	Wu et al. (2021)
Visited Location	Dokuchaev et al. (2020)
Name	Dokuchaev et al. (2020)
Nationality	Dokuchaev et al. (2020)
Phone Number	Indeed (2021)
Profile Picture	Wu et al. (2021)
Relationship Status	Salim et al. (2022)
Workplace	Salim et al. (2022)

Data Modeling. We created a survey through Google Forms to determine the respondents' view on the privacy level of the selected public data attributes. Figure 3 shows the survey question that requires the respondents to select the privacy level of 15 public data attributes based on their point of view. The public data attributes are classified into three privacy classes: low, medium, and high. It aims to observe initial respondents' awareness of public data privacy. Their response will then become the foundation element in our Bayes Theorem model. The privacy classes—high, medium, or low are represented as follows:

- **High privacy class**: The attribute that significantly affects individuals' privacy and may cause security risks. For example, when data classified in this privacy class gets into the wrong hands, it may result in people having their privacy compromised, identities stolen, or fraud committed in their names.

- **Medium privacy class**: The attribute that has little effect on the privacy of individuals and may slightly cause security risks. The real damage may lessen in this privacy class compared to the high privacy class. However, it could occur to any individual as the data in this privacy class might be a turning point for the security risks.
- **Low privacy class**: The attribute that does not significantly affect individuals' privacy and may not cause security risks. For example, if the individual decides to expose data that belongs in this privacy class, the invasion of privacy for this person may not occur. It somehow will not impact the individual's daily life as it is not really involved with real damage.



From your point of view, select the privacy level of common public data attributes below. *
Dari sudut pandangan anda, pilih tahap privasi atribut data awam di bawah.

Mark only one oval per row.

	Low / Rendah	Medium / Sederhana	High / Tinggi
Address / Alamat	○	○	○
Age / Umur	○	○	○
Date of Birth / Tarikh Lahir	○	○	○
Education Level / Tahap Pendidikan	○	○	○
Email / Emel	○	○	○
Gender / Jantina	○	○	○
Hometown / Tempat Asal	○	○	○
Identity Card (IC) Number / Nombor Kad Pengenalan	○	○	○
Visited Location / Lokasi yang dilawati	○	○	○
Name / Nama	○	○	○
Nationality / Kewarganegaraan	○	○	○
Phone Number / Nombor Telefon	○	○	○
Profile Picture / Gambar profil	○	○	○
Relationship Status / Status Hubungan	○	○	○
Workplace / Tempat Kerja	○	○	○

Figure 3. Survey questions in the first-time verification process

The Bayesian value is calculated for each public data attribute using the designated Bayes theorem from Equation 2. It includes the probability of assuming weight and the probability of choosing a privacy level based on a public count. The public count is the number of votes from the respondents' feedback on the chosen privacy level. The weighting is designed to prioritize the privacy of public data attributes. It is determined by 15 public data attributes, divided into 3 classification classes. Hence, every rank increment of a public data attribute is assumed to be 0.33. Each attribute's weight is considered a multiple of 0.33, with the highest privacy having the most weight. The highest weight is 4.95, given to the highest privacy public data attribute, while the lowest weight is 0.33, given to the lowest privacy public data attribute. This data modeling is repeated for the second time after the respondents' responses are analyzed and formulated through the Bayesian theorem. The two-time public verification process ensures that our data privacy model is reliable and accurately reflects our users' perspectives.

Data Analysis. This stage analyzes the result obtained from two ranking steps: (1) the respondents' responses and (2) the Bayesian value. Public data attributes are initially ranked based on the survey analysis according to the public count. For instance, one attribute will be classified into the "high privacy" class if most respondents select that privacy level for that attribute. Then, the respondents' perspectives were further analyzed using the Bayesian data model. The public data attributes are further ranked and compared between the survey's public count and the Bayesian value. The later discussion details the comparison of public data attributes' rankings and their privacy classes.

Releases and Feedback. The public data attributes catalog is developed based on the latest privacy ranking generated from a two-step classifying process. This catalog aims to raise public awareness about the privacy of their information and their right to keep it private.

RESULT AND DISCUSSION

We presented the results of the classification process according to mixed-method research, which employs quantitative research (by utilizing the Bayesian theorem) and qualitative research (by conducting surveys to get feedback from the respondents). Then, the final ranking of public data attributes will be given.

The Ranking of the Privacy Level of each Public Data Attribute Based on the Number of Public Counts

According to the first survey output, the public data attributes classified into a high privacy class are the identity card number, phone number, visited location, and address. Table 2 shows the number of counts from the public for each privacy class of the public data attribute.

Based on Table 2, we can see that the first row of public data attributes has the highest privacy as the public count is the highest, which is 25 votes, and the subsequent row follows in descending order of public count, as the address has 20 votes, the phone number has 18 votes, and visited location has 16 votes. Most of the common public data attributes fall into the medium privacy class that is ranked accordingly, which includes workplace, education level,

Table 2
Public count for each attribute based on chosen privacy class

Public data attributes	Privacy class		
	Low	Medium	High
Identity Card Number	0	15	25
Address	7	13	20
Phone Number	9	13	18
Visited Location	10	14	16
Workplace	8	16	16
Education Level	9	27	4
Age	12	20	8
Date of Birth	10	19	11
Email	11	19	10
Relationship Status	10	18	12
Profile Picture	11	16	13
Hometown	11	16	13
Name	14	14	12
Nationality	21	12	7
Gender	21	13	6

age, date of birth, email, relationship status, profile picture, hometown, and name because the workplace has 16 votes, education level has 27 votes, age has 20 votes, date of birth has 19 votes, email has 19 votes, relationship status has 18 votes, profile picture has 16 votes, hometown has 16 votes, and name has 14 votes. While nationality has 21 votes and gender has 21 votes, both are classified as having low privacy. However, public data attributes with similar public counts on certain privacy classes are now classified into medium privacy classes. It is evident when the workplace falls within the medium to high privacy class, and the name falls within the low to medium privacy class since both have the same public count.

The Ranking of the Privacy Level of each Public Data Attribute Based on the Number of Public Counts Versus Bayes Value

The Bayes value is assessed using Equation 2 for each public data attribute. Table 3 is an example of calculation in producing Bayes value. The value is calculated and assigned to each public data attribute initially ranked based on the public count (Table 4).

We can see that the Bayes value is disorganized due to the different weights of each data attribute and that it contradicts its public count. We then sorted the ranking of public data attributes according to Bayes value and produced the new ranking list shown in Table 5. The difference only affects the high-priority class, where the address has replaced the identity card number rank, the phone number has replaced the address rank, and the identity card number has replaced the phone number rank. Meanwhile, other public data attributes do not require re-ranking or exchanging the privacy classes.

The output of classifying the public data attributes in Table 5 was then distributed again to the former respondents. Here, we assert a two-step verification process that ensures the dependability of our public data attributes catalog. This second survey required the respondents to state whether they “agree” or “disagree” with the current privacy level for each attribute, as depicted in Figure 4. Table 6 shows the result of the second survey analysis, which is the number of respondents that agree and disagree with the stated privacy class of each public data attribute.

Table 3
Example data calculation (High privacy class)

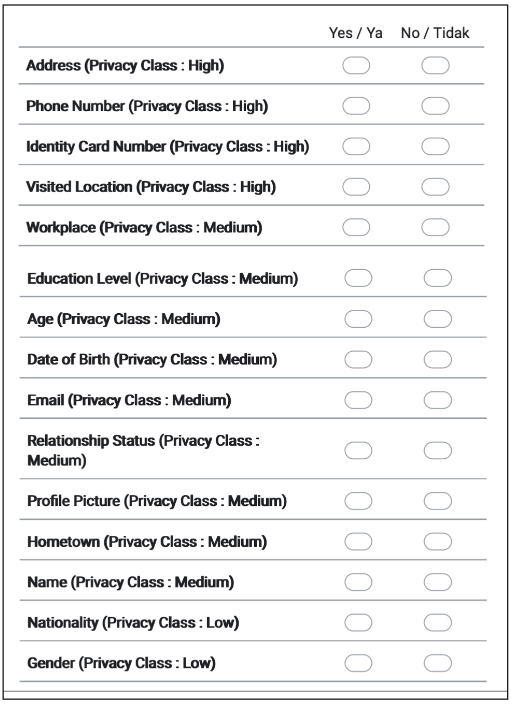
#	Public data attributes	P (PL| W)	P (W)	P (PL)	Bayes value
1	Identity Card Number	5.575	4.95	25/40 = 0.625	44.154
2	Address	5.12	4.62	20/40 = 0.5	47.308
3	Phone Number	4.74	4.29	18/40 = 0.45	45.188
4	Visited Location	4.36	3.96	16/40 = 0.4	43.164

Table 4
Initial ranking of public data attributes based on the public count

Privacy class	Public data attribute	Bayes value
High	Identity Card Number	44.154
	Address	47.308
	Phone Number	45.188
	Visited Location	43.164
Medium	Workplace	36.572
	Education Level	19.433
	Age	20.611
	Date of Birth	17.312
	Email	13.543
	Relationship Status	10.692
	Profile Picture	8.456
	Hometown	5.676
	Name	3.790
Low	Nationality	1.489
	Age	0.537

Table 5
Final ranking of public data attributes after arranging Bayes value

Privacy class	Public data attribute	Bayes value
High	Address	47.308
	Phone Number	45.188
	Identity Card Number	44.154
	Visited Location	43.164
Medium	Workplace	36.572
	Age	20.611
	Education Level	19.433
	Date of Birth	17.312
	Email	13.543
	Relationship Status	10.692
	Profile Picture	8.456
	Hometown	5.676
	Name	3.790
Low	Nationality	1.489
	Gender	0.537



	Yes / Ya	No / Tidak
Address (Privacy Class : High)	○	○
Phone Number (Privacy Class : High)	○	○
Identity Card Number (Privacy Class : High)	○	○
Visited Location (Privacy Class : High)	○	○
Workplace (Privacy Class : Medium)	○	○
Education Level (Privacy Class : Medium)	○	○
Age (Privacy Class : Medium)	○	○
Date of Birth (Privacy Class : Medium)	○	○
Email (Privacy Class : Medium)	○	○
Relationship Status (Privacy Class : Medium)	○	○
Profile Picture (Privacy Class : Medium)	○	○
Hometown (Privacy Class : Medium)	○	○
Name (Privacy Class : Medium)	○	○
Nationality (Privacy Class : Low)	○	○
Gender (Privacy Class : Low)	○	○

Figure 4. Survey questions in the second-time verification process

Table 6
Number of respondents that agree and disagree

Public data attributes	Privacy class	Number of respondents	
		Agree	Disagree
Address	High	39	1
Phone Number		37	3
Identity Card Number		38	2
Visited Location		32	8
Workplace	Medium	34	6
Education Level		33	7
Age		32	8
Date of Birth		34	6
Email		32	8
Relationship Status		30	10
Profile Picture		30	10
Hometown		33	7
Name		36	4
Nationality	Low	34	6
Gender		33	7

Comparing with the *k-NN* Learning Machine Method

We also compare the Naïve Bayes classification with the *k-nearest neighbors* algorithm (*k-NN*) for an unsorted data set given in Table 7. In *k-NN* learning, the number of trained data will be kept as a sample to identify different classes = (high, medium, low). The training samples are taken from (Shallal et al., 2020), and the unsorted data set is given as UnS= {da_1, da_2, …, da_n} where (da_1, da_2, da_3,….., or da_n) is independent of each other. For the *k* neighbor, we set *k* = the ratio of da_n and 15, where 15 is the total of data sampling. The comparison result in Table 8 shows slight differences in the data sorted within the group. We also request that our respondents (the same group of 40 participants) give their opinion on the classification using the *k-NN* algorithm. It is shown in Table 9 that the five data received disagreeing responses from respondents regarding the group. It might be due to the *k-NN* algorithm's required training sample, where a few training cycles lead to misclassification. By comparing our Naïve Bayes approach, which received agreement for all data attributes, we conclude that our approach is suitable for identifying the respective privacy levels for the data attributes.

Table 7
Unsorted data set

Public data attributes
Address, Age, Date of Birth, Education Level, Email, Gender, Hometown, Identity Card Number, Visited Location, Name, Nationality, Phone Number, Profile Picture, Relationship Status, Workplace

Table 8
Comparison group of data classification

Privacy class	Public data attribute through Bayes value	Privacy class	Public data attribute through k-NN value
High	Address	High	Address
	Phone Number		Email
	Identity Card Number		Visited Location
	Visited Location	Medium	Age
Medium	Workplace		Gender
	Age		Name
	Education Level		Workplace
	Date of Birth		Education Level
	Email		Identity Card Number
	Relationship Status		Date of Birth
	Profile Picture		Hometown
	Hometown		Profile Picture
	Name		Phone Number
Low	Nationality	Low	Nationality
	Gender		Relationship Status

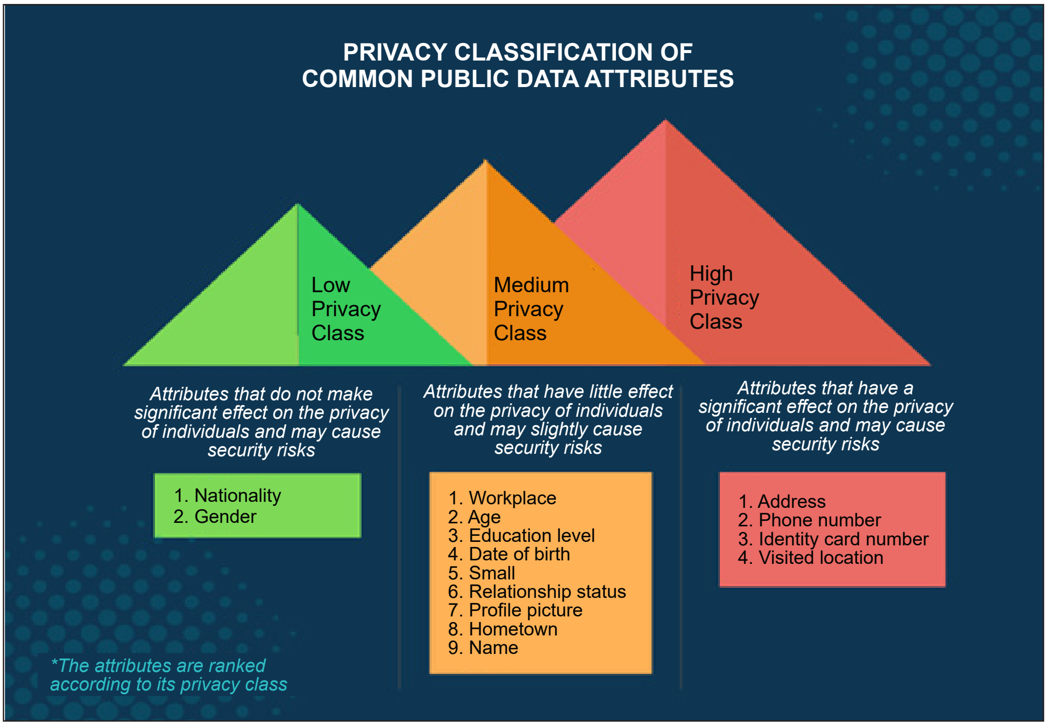
Finally, we formed a catalog of public data attributes for easier notifying the users on privacy levels of their common information, given in Figure 5.

CONCLUSION

The development of digital data sharing and analysis procedures is important in the digital era, especially with the advancement of mobile technologies and social media platforms. The data privacy policy should be revealed to ensure public rights and the usefulness of available electronic and digital media. Aside from a thorough understanding of public data privacy, awareness of information sharing via digital media improves socio-economic conditions. It implicitly attracts many investors to make

Table 9
Number of respondents that agree and disagree with k-NN approach

Public data attributes	Privacy class	Number of respondents	
		Agree	Disagree
Address	High	39	1
Email		10	30
Visited Location		32	8
Age	Medium	32	8
Gender		7	33
Name		36	4
Workplace		34	6
Education Level		33	7
Identity Card Number		3	37
Date of Birth		34	6
Hometown		33	7
Profile Picture		30	10
Phone Number	Low	2	38
Nationality		34	6
Relationship Status		14	26



Attributes that do not make significant effect on the privacy of individuals and may cause security risks

1. Nationality
2. Gender

Attributes that have little effect on the privacy of individuals and may slightly cause security risks

1. Workplace
2. Age
3. Education level
4. Date of birth
5. Small
6. Relationship status
7. Profile picture
8. Hometown
9. Name

Attributes that have a significant effect on the privacy of individuals and may cause security risks

1. Address
2. Phone number
3. Identity card number
4. Visited location

*The attributes are ranked according to its privacy class

Figure 5. The catalogue

investments in the country. This work may lead one to perceive that public data also has privacy. We used the Naïve Bayesian classifier to classify public data attributes into low, medium, and high privacy classes. Mixed-method research in verifying the classification process further makes our public data attributes privacy ranking list reliable. Furthermore, this acknowledges that public data privacy has its level that should be highlighted. It may encourage individuals to be more cautious when exposing their information, especially on open digital platforms.

ACKNOWLEDGEMENTS

This work has been supported by Cyber Security Malaysia (CSM) through a collaboration project of CSM-INSPEM, Universiti Putra Malaysia under Public Data Classification Program-Privacy and User Right parked under Tabung Amanah INSPEM (vot.no.: 6300410-10065).

REFERENCES

Abraham, A., Kanjamala, E. R., Thomas, E. M., & Akhila, G. P. (2019). Email security classification of imbalanced data using naive Bayes classifier. *International Journal of Wireless Communications and Network Technologies*, *8*(3), 16-20. https://doi.org/10.30534/ijwcnt/2019/04832019

Algarni, A. (2019). A survey and classification of security and privacy research in smart healthcare systems. *IEEE Access*, *7*, 101879-101894. https://doi.org/10.1109/ACCESS.2019.2930962

Analysis & Policy Observatory. (2020). *ACSC Annual Cyber Threat Report: July 2019 to June 2020*. Australian Cyber Security Centre. https://apo.org.au/node/308071 https://www.cyber.gov.au/acsc/view-all-content/advice/personal-information-and-privacy

Bibhu, V., Salagrama, S., Lohani, B. P., & Kushwaha, P. K. (2021). An analytical survey of user privacy on social media platform. In *2021 International Conference on Technological Advancements and Innovations (ICTAI)* (pp. 173-176). IEEE Publishing. https://doi.org/10.1109/ICTAI53825.2021.9673402

Budiu, R., & Moran, K. (2021). *How many participants for quantitative usability studies: A summary of sample-size recommendations*. Nielsen Normal Group. https://www.nngroup.com/articles/summary-quant-sample-sizes/

Cain, J. A., & Imre, I. (2022). Everybody wants some: Collection and control of personal information, privacy concerns, and social media use. *New Media & Society*, *24*(12), 2705-2724. https://doi.org/10.1177/14614448211000327

Dokuchaev, V. A., Maklachkova, V. V., & Statev, V. Y. (2020). Classification of personal data security threats in information systems. *T-Comm, 14*(1), 56-60. https://doi.org/10.36724/2072-8735-2020-14-1-56-60

Indeed. (2021). *A guide to data classification (with types and examples)*. Indeed. https://www.indeed.com/career-advice/career-development/data-classification

Liu, S., Zhu, M., & Yang, Y. (2013). A Bayesian classifier learning algorithm based on optimization model. *Mathematical Problems in Engineering*, *2013*, Article 975953. https://doi.org/10.1155/2013/975953

MyGoverment. (2019). *Mygov - The government of Malaysia's Official Portal*. MyGoverment. https://www.malaysia.gov.my/portal/content/30588

Rashid, A. F. A., & Zaaba, Z. F. (2020). Facebook, Twitter, and Instagram: The privacy challenges. In *2020 International Conference on Promising Electronic Technologies (ICPET)* (pp. 122-127). IEEE Publishing. https://doi.org/10.1109/ICPET51420.2020.00032

Ravn, S., Barnwell, A., & Neves, B. B. (2019). What is "publicly available data"? exploring blurred public-private boundaries and ethical practices through a case study on Instagram. *Journal of Empirical Research on Human Research Ethics, 15*(1-2), 40-45. https://doi.org/10.1177/1556264619850736

Rehman, S. U., Manickam, S., & Al-Charchafchi, A. (2022). Privacy calculus model for online social networks: A study of Facebook users in a Malaysian University. *Education and Information Technologies, 28*, 7205-7223. https://doi.org/10.1007/s10639-022-11459-w

Reza, K. J., Islam, M. Z., & Estivill-Castro, V. (2020). Protection of user-defined sensitive attributes on online social networks against attribute inference attack via adversarial data mining. In *Information Systems Security and Privacy: 5th International Conference, ICISSP 2019* (pp. 230-249). Springer International Publishing. https://doi.org/10.1007/978-3-030-49443-8_11

Salim, S., Turnbull, B., & Moustafa, N. (2022). Data analytics of social media 3.0: Privacy protection perspectives for integrating social media and Internet of Things (SM-IoT) systems. *Ad Hoc Networks, 128*, Article 102786. https://doi.org/10.1016/j.adhoc.2022.102786

Sanderson, T., Reeson, A., & Box, P. (2019). 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). ACM Publishing. https://doi.org/10.1145/3326365.3326383

Shallal, Q. M., Hussien, Z. A., & Abbood, A. A. (2020). Method to implement K-NN machine learning to classify data privacy in IOT environment. *Indonesian Journal of Electrical Engineering and Computer Science, 20*(2), 985-990. https://doi.org/10.11591/ijeecs.v20.i2.pp985-990

Vu, D. H. (2022). Privacy-preserving Naive Bayes classification in semi-fully distributed data model. *Computers & Security, 115*, Article 102630. https://doi.org/10.1016/j.cose.2022.102630

Vu, D. H., Vu, T. S., & Luong, T. D. (2022). An efficient and practical approach for privacy-preserving Naive Bayes classification. *Journal of Information Security and Applications, 68*, Article 103215. https://doi.org/10.1016/j.jisa.2022.103215

Wibawa, A. P., Kurniawan, A. C., Murti, D. M., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes classifier for journal quartile classification. *International Journal of Recent Contributions from Engineering, Science & IT (IJES), 7*(2), 91-99. https://doi.org/10.3991/ijes.v7i2.10659

Wu, J., Li, W., Bai, Q., Iko, T., & Moustafa, A. (2021). Privacy information classification: A hybrid approach. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2101.11574

Zanella-Béguelin, S., Wutschitz, L., & Tople, S. (2022). Bayesian estimation of differential privacy. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2206.05199