Creating and analysing privacy policies of Malaysia e-commerce using personal data protection act

Auwal Shehu Ali[1,2], Zarul Fitri Zaaba[1], Manmeet Mahinderjit Singh[1], Nor Badrul Anuar[3], Mohd Ridzuan M. Shariff[4]

[1]School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia
[2]Department of Computer Sciences, Bayero University Kano, Kano, Nigeria
[3]Centre of Research for Cyber Security and Network, Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia
[4]Strategic Research and Advisory, Cyber Security Malaysia, Menara Cyber Axis Cyberjaya, Cyberjaya, Malaysia

Article Info

Article history:

Received Jul 11, 2024
Revised Jan 7, 2025
Accepted Mar 9, 2025

Keywords:

Datasets
Internet
Machine learning
Personal data protection act
Privacy policy
Website

ABSTRACT

Despite legally binding agreements between users and website owners, users often overlook website privacy policies due to their length and complexity. Transparency in these policies is crucial, particularly in Malaysia, where regulatory agencies face challenges ensuring compliance with the personal data protection act (PDPA) of 2010 due to intricate language and complex legal clauses. Machine learning has been used to analyse privacy policies under various legal frameworks, but no dataset currently exists for the Malaysian PDPA. Thus, to bridge this gap, we introduce a pilot corpus of 50 privacy policies specifically tailored to the Malaysian PDPA. This dataset is analysed and made available for academic research, offering insights into privacy regulations and identifying trends in privacy policy transparency. Our findings pave the way for the development of tools to enhance compliance with PDPA standards and improve policy readability for users. The corpus also serves as a foundation for further research in privacy and data protection, encouraging the exploration of automated approaches for policy analysis and regulatory oversight.

Corresponding Author:

Zarul Fitri Zaaba
School of Computer Sciences, Universiti Sains Malaysia
11800 Pulau Pinang, Malaysia
Email: zarulfitri@usm.my

1. INTRODUCTION

Privacy policies are legally binding documents that inform users how websites collect, use, share, and manage their data [1]. Despite their ubiquity across the internet, research shows that privacy policies are often lengthy and challenging for users to fully comprehend [2]. As a result, many consumers do not take the time or effort to engage with these documents in detail. In Malaysia, privacy policies are regulated under the personal data protection act (PDPA) of 2010, which defines personal information as any data capable of identifying an individual, directly or indirectly, in commercial transactions [3]. These regulations have a broad scope, impacting many websites, applications, and online services [4].

Extracting and identifying key components within privacy policies enables organisations to assess and mitigate risks related to data protection and privacy breaches. However, ensuring compliance with the PDPA presents significant challenges [5]. Large-scale datasets of privacy policies are crucial for training machine learning models, as demonstrated by [6], [7], who examine online privacy policies on a wide scale comparable to the size of the internet.

Moreover, efforts to overcome the challenges posed by the complex readability of privacy policies have increased in the last decade. Tools and research methods that aim to tackle the length and difficulty in understanding these policies, including those utilising artificial intelligence techniques like machine learning and natural language processing for automated summarisation of privacy policy [6]–[9], would necessitate access to or greatly benefit from extensive collections of web privacy policies.

Several researchers have developed datasets of website privacy policies for private use, with some being made publicly available for the research community [9], [10]. However, a dataset addressing Malaysian data regulations and encompassing a broad collection of website privacy policies remains absent [5]. Acknowledging the significant variations in privacy policy content across different types of websites is essential. For example, cookies are commonly associated with sites that track browsing habits. At the same time, banking and healthcare websites often collect more detailed information, such as location histories and unique personal identifiers like fingerprints. These differences necessitate distinct privacy protection measures tailored to the nature of the data collected [11].

Instead of relying solely on traditional web crawling methods to gather privacy policies from websites, we employed the open-content directories available on Similarweb.com to identify the specific domains of interest. Similarweb.com estimates overall website traffic and offers insights into the main traffic sources of competitors, including referral websites, social media platforms, and frequently used search terms. In contrast, BuiltWith.com tracks over 2,500 e-commerce technologies across over 26 million e-commerce websites, offering detailed exportable data on factors such as expenditure, revenue, employee numbers, social media presence, industry classification, location, and rankings.

For this study, we curated and managed a collection of 50 website links, organised according to a hierarchical ontology. Utilising Similarweb.com presents a distinct advantage over simple web crawling methods, as it offers domain-specific information and insights into the market sector and categories associated with privacy policies. This comprehensive approach allows for more in-depth analysis and comparison of privacy policies across various e-commerce websites. The research presents several key findings, including: i) we present an initial dataset of privacy policies specifically sourced from Malaysian e-commerce websites; ii) we employ a manually curated hierarchy of 50 web links from Similarweb.com, enabling a targeted analysis of privacy policies within the Malaysian e-commerce sector; and iii) we publicly disclose details of our data collection to improve the ability to replicate our findings.

The rest of the article is organised in this order. Section 2 offers an overview of the existing research using openly accessible datasets of website privacy policies and highlights the specific area this paper aims to address. Section 3 further explains how the privacy policy dataset is constructed. Section 4 highlights results and findings, section 5 presents a discussion, and section 6 provides a conclusion and future work.

2. RELATED WORK

Machine learning and artificial intelligence methods to analyse privacy policies have been increasingly popular in the last ten years. The presence and existence of privacy policy datasets provide the basis of most of the previously stated strategies. This section focuses on the existing research and studies conducted on publicly available collections of privacy policy documents. Several scholars have focused on manually annotating collections of privacy policy documents. Scholars' manual annotation of these corpora hinders their ability to include more than a few hundred privacy policies. OPP-115 and APP-350 are the most often utilised corpora data privacy systems. OPP-115 is a dataset containing a collection of up to 115 diverse privacy policies, whilst APP-350 contains 350 datasets of diverse policies. Widely used for privacy research, the OPP-115 dataset [10] offers a collection of 115 privacy policies. Developed with the usable privacy policy (UPP) work [12]. The dataset comprises 115 website privacy policies gathered through the Amazon Alexa service. An unquestionable benefit of the corpora is the inclusion of annotations, and a labelling framework devised by the creators. The content encompasses a range of situations involving the utilisation of private information and provides details regarding the specialists responsible for annotating the texts. Multiple annotators assigned labels to each policy. It enabled the creation of more than 20,000 annotations that represent different elements of personal data utilisation. In their work, the authors of [13] established connections between the detailed annotation scheme and the principles outlined in data regulation like general data protection regulation (GDPR).

There are additional manually labelled collections of data that contain fewer than 1,000 policies. For example, there are 236 policies in one collection [14], 400 policies in another collection [8], [11], [15], 45 policies in another collection [16], and 64 policies in a fourth collection [17]. Although these corpora are helpful for machine learning classification in the E.U. region, they are inadequate for classification in other regions like Asia due to their regulation.

Some datasets collected from the internet lack annotation and are more significant in size but are not accessible to the general public. For example, there are data sets consisting of 9,295 policies [18] and 130,000 policies [6]. A limited number of web privacy policies, such as a corpus of 1,010 policies [15], are accessible to the public.

Several collections of mobile data usage agreements, specifically those found on the Android app store, are accessible. For instance, Kumar *et al.* [19] analysed 150,000 policies from Google Play. At the same time, Sunyaev *et al.* [20] examined a set of 183 privacy policies for health-related cross-platform mobile apps. A significant aspect of the MAPS framework [9] is its ability to assess the privacy policies of over a million Android applications. It offers a valuable resource of 441,626 app privacy policies, allowing analysis based on app categories. The privacy disclosures of mobile apps have garnered significant attention, primarily due to research examining the source code of a mobile app in conjunction with its privacy statement [21]–[23]. Privacy policies of websites, however, are also significant. Large discrepancies exist between what mobile app privacy policies and web privacy policies disclose. For example, cookies are mainly relevant to browsers' website policies. In general, applications installed on mobile devices can acquire more precise location data than websites. Therefore, mobile applications should explicitly state how they handle this location information in their privacy policies. However, there is an absence of substantial collections of privacy policy document datasets explicitly focusing on Malaysian PDPA websites, regardless of these variations.

A notable aspect of this study is the emphasis on gathering and examining privacy policies created by e-commerce websites in Malaysia. These websites are distinguished by their ability to process a greater variety of data types and to share data with other data processing companies. A new collection of specialised policies has been developed to address the challenge of analysing how certain elements are presented in privacy policies, along with a tailored methodology. The key characteristic of the crawler described in this research is its ability to gather data from e-commerce platforms based on specific parameters. The flexible search function of this web crawler guarantees that users only find highly relevant privacy policies that directly address their needs.

3. METHOD

3.1. Dataset structure and development

This section presents the methodology for selecting and organising privacy policies from Malaysian e-commerce websites, obtaining annotations, and curating the corpus. The goal is to develop a comprehensive collection of privacy policies from various Malaysian e-commerce platforms. This will be achieved by systematically crawling and extracting privacy policies from selected websites. The corpus will encompass a diverse set of e-commerce websites in Malaysia, capturing variations in the content and structure of privacy policies. In order to ensure the dataset's integrity, the collected policies will undergo manual review and annotation by domain experts, given the complexities of interpreting privacy policy language, as highlighted in prior research [12]. Therefore, this step is crucial for producing a high-quality training and evaluation dataset.

3.2. Similarweb.com

Similarweb.com is a digital intelligence platform providing tools for studying website traffic and market research. It serves firms that are looking to comprehend the digital environment. Their services cover website traffic statistics, which include data on visitor demographics and engagement indicators. In addition, Similarweb allows organisations to conduct competitor analysis, providing valuable information about their competitors' website traffic, marketing tactics, and audience demographics. This data enables organisations to enhance their website performance, discover new marketing prospects, and make informed decisions regarding their online presence. The privacy policy collection we possess comprises 50 URLs of Malaysian e-commerce websites. These websites are ranked based on the number of visitors they receive. In June 2024, we obtained the URLs from Similarweb.com. As anticipated, several sites on Similarweb.com were not functioning correctly, and we had to handle various exceptions issued by the URLs on Similarweb.com.

3.3. Identifying hyperlinks to privacy policies of the e-commerce corpus

At this stage, we verified the existence of hyperlinks that lead to possible privacy policies on every individual page within Similarweb.com. As previously noted by others [13], typical labels for hyperlinks to privacy policies often consist of terms such as "privacy policies," "term of service," "privacy notice," and "data privacy," and similar terms are typically mirrored in the URL. An established convention in collecting privacy policies has been to search for patterns of these phrases in the URLs [9], [13]. However, in prior studies [9], [13], selecting these phrases has been primarily random. To obtain a comprehensive collection of keywords, we manually evaluated a pre-existing dataset collection of 400 privacy policies, accompanied by their respective website addresses, which was employed to extract terms found explicitly within the URLs. We extract and utilise terms of service, privacy, statement, policy, terms, and help.

We analysed all links on Similarweb.com pages, focusing on those containing any specified keywords within their target URLs. It is important to note that a single URL on Similarweb.com may yield multiple potential privacy policy options, as we obtain all the links on the Similarweb.com page that meet this keyword criteria.

A noteworthy observation concerns the mean count of potential privacy policy links extracted from a valid Similarweb.com page. While analysing Similarweb pages, we found a high number of potential privacy policy links, often containing keywords like "legal" that might not be directly relevant. Despite removing duplicates and irrelevant pages, this approach resulted in an unpredictable decrease in potential privacy policies identified per webpage.

3.4. Eliminating repetitive links to privacy policies

We have noticed the presence of repetitive web addresses in the collection of potential policy content. We eliminate repetitive entries whilst retaining URLs that are present in many categories. Whilst a Similarweb.com page and its privacy policy can logically belong to two categories; we aim to avoid redundancy by eliminating duplicate URLs within a single category.

3.5. Privacy policies content

In the following stage, we will focus on retrieving the privacy policy content associated with each unique candidate URL. We retrieve the prospective websites and extract their primary text by eliminating boilerplate content. Initially, particular extensions were encountered when attempting to access the candidate URLs, which were eliminated. After removing such exceptions from the list of unique candidate policy URLs, we acquired the content of the proposed policy URL and eliminated any unnecessary or repetitive information.

3.6. Eliminating pages not in English

To identify the language of the extracted privacy policies, we utilised the Python module Lang detect, a language detection library derived from Google. Non-English documents were excluded from further analysis due to our inability to verify their content. Most of the policies in our dataset were written in English, accounting for 98% of the total. This finding aligns with previous research indicating that privacy policies across different languages exhibit similar patterns. While our dataset included some European languages, it lacked major Asian languages like Japanese, Russian, Chinese, and Korean. This suggests a potential gap in our data collection process, as Similarweb.com includes websites in these languages. As an illustration, [9], [10] conducted a web crawl to gather privacy regulations and discovered that the collection encompassed a multilingual aspect, featuring languages like Italian, Dutch, German, Spanish, French, and English. Whilst our dataset includes several European languages, it is notably short of major Asian languages like Japanese, Russian, Chinese, and Korean. This is surprising because Similarweb.com has webpages in these languages, suggesting a potential gap in our data collection process. Nevertheless, the Malaysian language is included in our top 10.

3.7. Annotating scheme and process

An annotation scheme was developed to categories the data practices outlined in privacy policies. A handful of industry experts selected many data practice categories and associated description qualities gathered from various privacy rules, refining the creation process to guarantee the scheme represented the actual policy text. Following talks between the experts, the annotation technique was subsequently applied to more policies and improved throughout several iterations [12]. Based on the existing approaches, the nine distinct types of data practices make up the final annotation scheme proposed:

- First party collection/use: methods and rationale behind the data collection practices of a service provider.
- Third-party sharing/collection: methods of sharing or collecting user data by 3rd parties.
- User choice/control: alternatives that consumers can choose from and manage.
- User access, edit, and deletion: what users may do with their data regarding access, editing, and deletion.
- Data retention: duration of data storage for users.
- Data security: safeguarding of user data.
- Policy change: when and how users might expect updates to their privacy policies.
- Global and targeted audiences: actions relevant solely to a subset of the population (e.g., children, Asians, or Malaysian residents).
- Other: more subsections for introductory or broader text, contact details, and procedures not addressed elsewhere.

Each data practice fits into one of the nine abovementioned categories, defined by a particular set of qualities associated with that category. An instance of the user preference/control data practice is linked to four essential attributes (choice type, choice scope, personal information type, and purpose) and one noncompulsory attribute (user type). The annotation scheme establishes a predetermined range of possible values for each property. To align the data practice with the policy text, each characteristic can also be linked to a specific section in the privacy policy.

3.8. Extracting legitimate policies

One of the most significant challenges encountered during dataset development was accurately determining which documents constituted privacy policies. We explored various machine-learning techniques to address this task, including regular expressions and machine-learning algorithms. However, we found that a more straightforward approach based on keyword analysis proved surprisingly effective. By focusing on the term "privacy" within URLs, we could curate a collection of privacy policies with exceptional accuracy and completeness. Our analysis revealed that websites containing the term "privacy" at least twice were highly likely to contain actual privacy policies, eliminating the potential for false positives caused by sites that merely include the term in their headers or footers. Furthermore, a pre-existing collection of documents containing specific URLs with relevant keywords significantly enhanced the effectiveness of our approach. This provided a valuable starting point for identifying potential privacy policies and refining our selection criteria.

Despite its simplicity, our classification technique achieves great precision and recall when applied to the completed privacy policy dataset of 50 policies. We conducted a manual labelling process on the English candidate privacy policy sample. We conducted a comprehensive analysis of the entire text, not just the title, of these policies to determine whether they are genuine privacy policies that include terms of service and information on privacy or if they are not privacy policies. Additionally, we allocated a binary vector of category-specific labels to each policy segment, where each component within the vector indicates the existence or lack of a data practice section in the portion. We utilised a vector consisting of nine components, all of which were derived from PDPA practice categories except other. We generated the highest quality data for this topic by employing a streamlined approach to consolidation. Whenever multiple annotators agreed that a particular element was present in a section, we assigned that section with the corresponding element label.

This study presents a methodology for constructing a corpus of Malaysian e-commerce privacy policies. By leveraging Similarweb.com, we identified potential privacy policy URLs based on keyword analysis and extracted their content. A rigorous annotation process was employed to label policy segments with relevant data practice categories. This methodology effectively identified privacy policies within the Malaysian e-commerce landscape, ensuring a high-quality dataset for further analysis.

4. RESULTS AND FINDINGS

The dataset used consisted of 1,542 segments drawn from 50 privacy policies. Utilising the Paragraph2Vec model [24] within the Gensim framework [25], each segment was encoded into a compact vector representation. This method captured the semantic relationships between terms in the privacy policy lexicon, acknowledging the specialised nature of the vocabulary while recognising the lack of full standardisation. Each policy clause was assigned a binary vector indicating category assignments, where each element represented a specific category, marked as 1 if the clause belonged to that category and 0 otherwise. This resulted in a vector with nine components, excluding the "other" category, which reduced the components to eight, derived from established privacy practice categories. A simple aggregation method was employed to generate gold-standard data: when multiple annotators agreed on a category for a given segment, the corresponding label was assigned to that section.

We selected two interpretable machine learning models, logistic regression (LR) and support vector machines (SVM), to predict category labels for segments within privacy policies, addressing the multi-class classification problem, as similarly explored by [10], [12]. LR and SVM are advantageous due to the transparency these models offer in explaining their classification decisions. This transparency is critical in privacy analysis, as it categorises policies as compliant or non-compliant and provides insights into the reasoning behind these classifications. By narrowing the label space to those in the training set, we ensured the focus remained on relevant categories. Additionally, we employed a sequence labelling technique, drawing from prior research that applied hidden Markov models (HMMs) to analyse privacy policy language [26]. Our approach differs from previous studies in that we use labels derived from the PDPA with an expert-designed annotation scheme rather than relying on themes generated through unsupervised methods.

Furthermore, based on our model, every concealed condition is associated with a distinct binary vector representing a specific combination of categories in the training data. The HMMs transition probabilities describe the inclination of privacy policy authors to arrange themes in comparable sequences. Due to the distinct real-valued vector representation of each segment in Paragraph2Vec, it was not feasible to directly derive an emission probability distribution from the training data. Consequently, we executed the K-Means++ algorithm utilising the scikit-learn toolkit on the segment vector representations and allocated each segment to a cluster. The emission probability distribution accurately depicted the characteristics of a specific class and produced the segment that is seen as a cluster. The two distributions are determined experimentally using the training data, and Viterbi decoding is employed to obtain the optimal labelling sequence during prediction. Table 1 shows the precision, recall, and F-1 scores. We choose precision, recall, and F-1 because they provide a more comprehensive picture of our privacy policy classification model's performance than accuracy alone, as did by [10], [12] since we are dealing with an imbalanced dataset of the privacy policy. This will help us understand our model on how well it identifies relevant cases (recall) and avoids irrelevant classifications (precision), allowing us to make informed decisions about its effectiveness.

Table 1. Precision/recall/F-1 for the two models

Category	SVM			LR		
	Precision	Recall	F-1	Precision	Recall	F-1
First-party collection/use	0.87	0.84	0.85	0.87	0.84	0.85
Third-party sharing/collection	0.68	0.71	0.64	0.78	0.85	0.83
User choice/control	0.56	0.87	0.78	0.76	06.4	0.72
User access, edit, and deletion	0.48	0.78	0.61	0.68	0.59	0.62
Data retention	0.74	0.74	0.71	0.74	0.74	0.71
Data security	0.68	0.64	0.66	0.68	0.64	0.66
Policy change	0.54	0.84	0.61	0.54	0.74	0.61
International and specific audiences	0.48	0.78	0.53	0.68	0.64	0.66
Others	0.84	0.94	0.78	0.74	0.74	0.71

5. DISCUSSION

The study presented a corpus of 50 privacy policies extracted from Malaysian e-commerce websites, offering a valuable resource for analysing the characteristics and readability of these policies in the specific context of the Malaysian online landscape. The average word length of privacy rules in this corpus is around 681, with a minimum of 231 words and a maximum of 5,959 words. The corpus comprises policies from over 50 distinct top-level domains (TLDs). The domains com.my and .com account for a significant portion of the corpus, with com.my spanning 52% and .com covering 48%. The corpus sources for country-level domains are limited to .my, which only covers one specific geographic region. A text's readability refers to the ease with which it may be understood or comprehended, influenced by the writing style [27]. Readability and length influence Internet users' choices to read or disregard a privacy policy [1]. Previous research has indicated the difficulty of understanding privacy policies, particularly in the European Union. This study offers an opportunity to investigate these challenges within the Malaysian context, potentially revealing unique insights and implications for improving privacy policy readability and user comprehension [28].

We assessed the readability of the policies in the corpus by employing the Flesh-Kincaid grade level (FKG) measure, commonly utilised in prior work and primarily recognised as the most prevalent metric for this purpose. The FKG measure expresses the readability score regarding a U.S grade level. The calculated FKG score produced a mean value of 15.34, corresponding to a standard deviation 5.3. The score can be regarded as an average equivalent to 15.34 years of education in the United States, which is about three years of college education needed to comprehend a privacy policy. However, Fabian *et al.* [28] discovered that the average FKG score is 13.6 in their study on the readability of privacy regulations, which involved analysing 50,000 documents. A reader is typically expected to have obtained a high school diploma or some college education, completing approximately 16 years of formal education.

For future research, we will priorities expanding our dataset to include a broader range of categories and sectors, ensuring greater representativeness and addressing the current limitations. A comparative analysis of privacy policies from websites that handle sensitive personal information versus those that do not will be a focal point. Furthermore, we intend to broaden our data collection by incorporating policies written in various languages, diversifying our dataset, and facilitating more comprehensive analyses.

6. CONCLUSION

This paper presents the development of a unique corpus consisting of 50 privacy policies from Malaysian e-commerce websites, sourced via Similarweb.com and aligned explicitly with the PDPA. To our knowledge, this is the first dataset explicitly focused on PDPA compliance. To verify quality and precision, we manually annotated ten candidate privacy regulations (20% of the dataset). Using average precision, we found that many corpuses have privacy regulations reflecting industry standards. We found repeated privacy policy URLs, possibly owing to parent company ownership and standard templates.

This dataset holds considerable potential for advancing research in privacy policy analysis, particularly in comparing policy compliance, summarisation, and applying machine learning algorithms requiring annotated datasets. However, it is essential to acknowledge the limitations of this study. The dataset, while valuable, consists of only 50 privacy policies, which may not adequately represent the diversity of Malaysia's e-commerce landscape. Furthermore, our focus on e-commerce policies restricts the applicability of our findings to other sectors, such as healthcare, banking, and communications. Additionally, manual annotation, although improving accuracy, may introduce subjective biases. Future research should aim to expand the dataset to cover a broader range of industries and increase the number of annotated privacy policies to enhance the dataset's representativeness and reliability. This would contribute to a more comprehensive understanding of privacy policy practices across different sectors and improve the generalizability of the findings.

ACKNOWLEDGMENTS

The authors would like to thank The Ministry of Higher Education Malaysia for supporting this study.

FUNDING INFORMATION

The Ministry of Higher Education Malaysia supported this work for the Fundamental Research Grant Scheme with Project Code: FRGS/1/2023/ICT07/USM/02/1.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Auwal Shehu Ali	✓	✓			✓	✓		✓	✓	✓				
Zarul Fitri Zaaba	✓	✓			✓	✓		✓		✓		✓		✓
Manmeet Mahinderjit Singh	✓	✓								✓		✓		
Nor Badrul Anuar	✓									✓				
Mohd Ridzuan M. Shariff	✓									✓				

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

The authors state that there is no conflict of interest.

DATA AVAILABILITY

The data that supports the findings of this study are available from the corresponding author, [ZFZ], upon reasonable request.

REFERENCES

[1] A. S. Ali, Z. F. Zaaba, M. M. Singh, and A. Hussain, "Readability of websites security privacy policies: a survey on text content and readers," *International Journal of Advanced Science and Technology*, vol. 29, no. 6 Special Issue, pp. 1661–1672, 2020.

[2] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies 2008 privacy year in review," *I/S: A Journal of Law and Policy for the Information Society*, vol. 0389, pp. 543–568, 2008.

[3] K. H. Hassan, "Personal data protection in employment: new legal challenges for Malaysia," *Computer Law & Security Review*, vol. 28, no. 6, pp. 696–703, Dec. 2012, doi: 10.1016/j.clsr.2012.07.006.

[4] A. S. Ali, Z. F. Zaaba, and M. M. Singh, "Privacy during epidemic of covid-19: a bibliometric analysis," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 1, pp. 587–596, Feb. 2023, doi: 10.11591/eei.v12i1.4460.

[5] H. Baskaran, S. Yussof, A. A. Bakar, and F. A. Rahim, "Data sharing using PDPA-compliant blockchain architecture in malaysia," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 147–157, 2023, doi: 10.14569/IJACSA.2023.0140515.

[6] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: automated analysis and presentation of privacy policies using deep learning," *Proceedings of the 27th USENIX Security Symposium*, pp. 531–548, Feb. 2018.

[7] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna, "I read but don't agree: privacy policy benchmarking using machine learning and the eu GDPR," in *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, New York, New York, USA: ACM Press, Apr. 2018, pp. 163–166, doi: 10.1145/3184558.3186969.

[8] R. N. Zaeem, R. L. German, and K. S. Barber, "PrivacyCheck: automatic summarization of privacy policies using data mining," *ACM Transactions on Internet Technology*, vol. 18, no. 4, p. 53, Apr. 2018, doi: 10.1145/3127519.

[9] S. Zimmeck *et al.*, "MAPS: scaling privacy compliance analysis to a million apps," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 3, pp. 66–86, Jul. 2019, doi: 10.2478/popets-2019-0037.

[10] S. Wilson *et al.*, "The creation and analysis of a website privacy policy corpus," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016, pp. 1330–1340, doi: 10.18653/v1/p16-1126.

[11] R. N. Zaeem and K. S. Barber, "A study of web privacy policies across industries," *Journal of Information Privacy and Security*, pp. 1–17, Nov. 2017, doi: 10.1080/15536548.2017.1394064.

[12] N. Sadeh *et al.*, "The usable privacy policy project: combining crowdsourcing, machine learning and natural language processing to semi-automatically answer those privacy questions users care about," C*arnegie Mellon University Technical Report CMU-ISR-13-119, Institute for Software Research, School of Computer Science,* Dec 2013.

[13] M. Srinath, S. Wilson, and C. L. Giles, "Privacy at scale: introducing the privaseer corpus of web privacy policies," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, Apr. 2021, pp. 6829–6839, doi: 10.18653/v1/2021.acl-long.532.

[14] V. B. Kumar *et al.*, "Finding a choice in a haystack: automatic extraction of opt-out statements from privacy policy text," in *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, Apr. 2020, pp. 1943–1954, doi: 10.1145/3366423.3380262.

[15] R. N. Zaeem and K. S. Barber, "The effect of the gdpr on privacy policies: recent progress and future promise," *ACM Transactions on Management Information Systems (TMIS)*, vol. 12, no. 1, pp. 1–20, Mar. 2018, doi: 10.1145/3389685.

[16] W. B. Tesfay, S. Kiyomoto, P. Hofmann, T. Nakamura, and J. Serna, "PrivacyGuide: towards an implementation of the eu gdpr on internet privacy policy evaluation," *ACM International Workshop on Security and Privacy*, vol. 18, pp. 15–21, Mar. 2018, doi: 10.1145/3180445.3180447.

[17] E. Costante, Y. Sun, M. Petković, and J. den Hartog, "A machine learning solution to assess privacy policy completeness," in *Proceedings of the 2012 ACM workshop on Privacy in the Electronic Society*, New York, NY, USA: ACM, Oct. 2012, pp. 91–96, doi: 10.1145/2381966.2381979.

[18] S. Zimmeck *et al.*, "Automated analysis of privacy requirements for mobile apps," in *Proceedings 2017 Network and Distributed System Security Symposium*, Reston, VA: Internet Society, 2017, pp. 286–296, doi: 10.14722/ndss.2017.23034.

[19] V. B. Kumar, A. Ravichander, P. Story, and N. Sadeh, "Quantifying the effect of in-domain distributed word representations: a study of privacy policies," in *CEUR Workshop Proceedings*, 2019, pp. 46–52.

[20] A. Sunyaev, T. Dehling, P. L. Taylor, and K. D. Mandl, "Availability and quality of mobile health app privacy policies," *Journal of the American Medical Informatics Association*, vol. 22, no. e1, pp. e28–e33, Apr. 2015, doi: 10.1136/amiajnl-2013-002605.

[21] B. Andow *et al.*, "Policylint: investigating internal privacy policy contradictions on google play," in *Proceedings of the 28th USENIX Security Symposium*, 2019, pp. 585–602.

[22] B. Andow *et al.*, "Actions speak louder than words: entity-sensitive privacy policy and data flow analysis with POLICHECK," in *SEC'20: Proceedings of the 29th USENIX Conference on Security Symposium*, 2020, pp. 985–1002, doi: 10.5555/3489212.3489268.

[23] W. Enck *et al.*, "TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones," *ACM Transactions on Computer Systems*, vol. 32, no. 2, 2014, doi: 10.1145/2619091.

[24] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *31st International Conference on Machine Learning, ICML 2014*, vol. 4, pp. 2931–2939, 2014.

[25] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, 2010.

[26] R. Ramanath, F. Liu, N. Sadeh, and N. A. Smith, "Unsupervised alignment of privacy policies using hidden markov models," in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, Association for Computational Linguistics, 2014, pp. 605–610, doi: 10.3115/v1/p14-2099.

[27] T. Ermakova, B. Fabian, and E. Babina, "Readability of privacy policies of healthcare websites," in *12. Internationale Tagung Wirtschaftsinformatik (Wirtschaftsinformatik 2015), Osnabrück, Germany*, 2015.

[28] B. Fabian, T. Ermakova, and T. Lentz, "Large-scale readability analysis of privacy policies," in *Proceedings - 2017 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2017*, Association for Computing Machinery, Inc, Aug. 2017, pp. 18–25, doi: 10.1145/3106426.3106427.

BIOGRAPHIES OF AUTHORS

Auwal Shehu Ali holds a Bachelor's and Master's degree in Computer Science from Bayero University, Kano (BUK), Nigeria. Currently, he is pursuing a Ph.D. in Computer Science at Universiti Sains Malaysia (USM), where his research focuses on the intersection of privacy and machine learning. His areas of expertise include privacy-preserving machine learning (PPML), usable cybersecurity and privacy, information security, privacy policy analysis, and federated learning. He can be contacted at email: auwal@student.usm.my.

Dr. Zarul Fitri Zaaba is a Senior Lecturer born in Kuala Lumpur, Malaysia, in 1981. He received his B.Sc. in Information Technology from the Universiti Utara Malaysia in 2003, M.Sc. in Information Security from the Royal Holloway University of London in 2007, and PhD from the University of Plymouth, the United Kingdom, in 2014. He is the author of more than 60 articles in journals and conference proceedings. He is currently a Senior Lecturer at the School of Computer Sciences, Universiti Sains Malaysia (USM). Before joining USM, he worked with the government and semi-government sectors for a few years. He specializes in information, privacy, web, and human security. He is also actively involved in science, technology, engineering, and mathematics (STEM) education and has been appointed as a Master Trainer in computational thinking. He can be contacted at email: zarulfitri@usm.my.

Associate Professor Dr. Manmeet Mahinderjit Singh is a Cybersecurity Educator and Researcher at the School of Computer Sciences, University Sains Malaysia (USM). She graduated from The University of Queensland, Australia, in 2012 with a Ph.D. in Data Security. She obtained her M.Sc. and B.Sc. degrees in Computer Science, specializing in Security and Distributed Systems. Her expertise is in information security, threat intelligence, data security, digital risks, smart devices security (IoT and mobile security) and cybersecurity prevention and detection solutions. She has an interest in cyber-criminology and sensor networks. She can be contacted at email: manmeet@usm.my.

Nor Badrul Anuar received his Master of Computer Science from the University of Malaya in 2003 and a Ph.D. at the Centre for Information Security and Network Research, University of Plymouth, the UK, in 2012. He is a Professor at the Faculty of Computer Science and Information Technology, University of Malaya. He has authored over 128 research articles and many conference papers locally and internationally. His research interests include information security, intrusion detection systems, data sciences, highspeed networks, artificial intelligence, and library information systems. He can be contacted at email: badrul@um.edu.my.

Dr. Mohd Ridzuan M. Shariff has been a career officer in the Malaysian Armed Forces for 28 years. Among the highlights of his military career was his involvement with the United Nations Protection Force (UNPROFOR) in Bosnia Herzegovina at the height of the Balkan War in 1995. After retirement, he joined the Institute of Diplomacy and Foreign Relations (IDFR), Ministry of Foreign Affairs, Malaysia in 2010. At IDFR, he trained and conducted lectures for local and international diplomats on strategic management, international relations, diplomacy, crisis management, and cross-cultural communications. He left IDFR in July 2020 to join Cyber Security Malaysia until now. He can be contacted at email: Madduan64@gmail.com.