IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 September 2023, accepted 21 December 2023, date of publication 25 December 2023,
date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347194

RESEARCH ARTICLE

Runtime and Design Time Completeness Checking of Dangerous Android App Permissions Against GDPR

RYAN MCCONKEY AND OLUWAFEMI OLUKOYA
School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, BT7 1NN Belfast, U.K.
Corresponding author: Oluwafemi Olukoya (o.olukoya@qub.ac.uk)

ABSTRACT Data and privacy laws, such as the GDPR, require mobile apps that collect and process the personal data of their citizens to have a legally-compliant policy. Since these mobile apps are hosted on app distribution platforms such as Google Play Store and App Store, the app publishers also require the app developers who wish to submit a new app or make changes to an existing app to be transparent about their app privacy practices regarding handling sensitive user data that requires sensitive permissions such as calendar, camera, microphone. To verify compliance with privacy regulators and app distribution platforms, the app privacy policies and permissions are investigated for consistency. However, little has been done to investigate GDPR completeness checking within the Android permission ecosystem. In this paper, we investigate the design and runtime approaches towards completeness checking of sensitive ('dangerous') Android permission policy declarations against GDPR. In this paper, we investigate the design and runtime approaches towards completeness checking of dangerous Android permission policy declarations against GDPR. Leveraging the MPP-270 annotated corpus that describes permission declarations in application privacy policies, six natural language processing and language modelling algorithms are developed to measure permission completeness during runtime while a proof of concept Class Unified Modeling Language Diagram (UML) tool is developed to generate GDPR-compliant permission policy declarations using UML diagrams during design time. This paper makes a significant contribution to the identification of appropriate permission policy declaration methodologies that a developer can use to target particular GDPR laws, increasing GDPR compliance by 12% in cases during runtime using BERT word embedding, measuring GDPR compliance in permission policy sentences, and a UML-driven tool to generate compliant permission declarations.

INDEX TERMS Security and privacy protection, requirement engineering, regulatory compliance, GDPR, android permission, unified modelling language, privacy policy, NLP, data privacy, mobile applications.

I. INTRODUCTION

The EU's General Data Protection Regulation (GDPR) came into effect in 2018 and contains 99 articles and 173 recitals that apply to any company that processes or stores personal data for EU citizens even if the application is not EU-based [1]. The penalties for breaking GDPR laws in the most serious cases can be as high as €20 million or 4% of the annual turnover rate. In lesser scenarios, penalties and fines can still lead to reprimands and restrictions on obtaining and processing personal data which can become detrimental for a company or organization that needs to store personal information [2]. To protect access to sensitive information and actions, Android utilises app permissions to support user privacy [3]. While there are different base permission types in the Android ecosystem, they are characterized by a protection level that describes the risk implied in the permission. Dangerous permissions (aka runtime permissions) are one

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

of the select range of permissions types that the user has to accept and acknowledge. In the official description on Google developer documentation [4], dangerous permissions are, "*a higher-risk permission that would give a requesting application access to private user data or control over the device that can negatively impact the user*". Dangerous permissions carry the risk of revealing personal information and the identity of the user. The use of dangerous permission requires a privacy policy by law [5]. The need to access sensitive areas of a device to gain personal information is a decision taken by the application developer and must be defined in the application manifest file.[1] Developers are susceptible to errors when writing privacy policies that declare the collection, usage, processing and transfer of personal information in a meaningful and transparent way [6]. Such mistakes could lead to the developer inadvertently breaking GDPR laws and receiving a heavy fine, jeopardizing the company or organization they work for and tarnishing consumer transparency concerning the handling of personal data. Several studies [6], [7], [8], [9], [10], [11] have shown that developers struggled to embed privacy into software systems. These studies suggest that software developers who design systems that collect and process sensitive user data have difficulties with incorporating privacy requirements and protocols from regulatory authorities into software applications. The lack of decision support tools for applying data protection principles, privacy reasoning, and user privacy verification in software design is cited by developers as the main deterrent to incorporating GDPR principles into software development practises [7], [8], [12].

In evaluating Android permission completeness, a large-scale evaluation of 164156 Android apps was explored in [13] and [14] to investigate whether the privacy policy matches its dangerous permission request. The investigations have shown that app privacy policies and sensitive (or "dangerous") permission requests are not always transparent. Prior literatures [15], [16], [17], [18], [19], [20], [21], and [22] have demonstrated the discrepancy that exists in the Android and iOS ecosystem by evaluating sensitive data access through dangerous permissions, app's code, third party library, data dissemination practices, Android API usage, app privacy policies, library inclusion and other relevant metadata. The common denominator amongst these works of literature is the investigation of the trustworthiness of the app's privacy policies from a privacy and regulatory point of view. The conclusion of the privacy compliance analysis of mobile apps investigated in the literature is the prevalence of questionable privacy policies, inconsistencies, lack of transparency and non-compliance with regulatory requirements. A challenge that developers face is that developers must comply with privacy laws and there is no real methodology that exists to assist in the development of a privacy policy thus developers are trying to comply with regulations without the necessary knowledge of what language and explicit terms of language are needed to implement dangerous android permission-policy declarations (DAPD) [23]. This has resulted in many mobile application developers seeking guidance on Stack Overflow for the creation of compliant privacy policies [24], [25], [26]. The challenge of creating GDPR-compliant privacy policies becomes more evident as developers due to either confusion, ease of development, misuse or disregard requests for multiple permissions for the same information [27].

One way to mitigate the challenges developers face is by creating automated tools to assist small to medium-sized teams in the generation of permission policy snippets that are compliant with privacy laws. To create developer-centric solutions, this study investigates GDPR compliance of the dangerous Android permission-policy declarations used for each permission group in 270 mobile applications during runtime. The three-pronged approach investigates (i) the completeness of dangerous android permissions in fulfilling GDPR obligations, (ii) the feasibility of generating GDPR-compliant policies for sensitive permission requirements extracted from UML diagrams at design time, and (iii) evaluates if the GDPR is fit for purpose in describing android permission categories, the sensitive data requested, sensitive APIs, actions permissions represent and the semantic meaning. Since the GDPR contains articles and recitals that describe the data protection regulation an individual or organisation must comply with, while Android permission policy declarations are a developer's attempt to convey transparently information about apps accessing dangerous permission to collect sensitive data, it is therefore, necessary to investigate whether such permission-policy snippets are coherent, explicit, accurate, concise and transparent complies with GDPR as the benchmark. As a result of this, dangerous Android permission policy statements are a verified approach for completeness checking of privacy policies and applications. The contribution of this research is highlighted below:

- **Completeness Checking of Sensitive Android Permissions and GDPR:** To the best of our knowledge, this is the first work that evaluates completeness checking of articles of the GDPR and sentences declaring the request and usage of sensitive Android permissions. Most works of the literature [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39] evaluate completeness checking of applications and privacy policies against GDPR requirements. We investigated how well the permissions policy and categories adhered to GDPR. This was further backed by a thorough examination of the GDPR's suitability for verifying the accuracy of Android permissions.
- **Empirical Analysis of the Suitability of Diverse Natural Language Processing Techniques for Text Similarity:** We evaluate six NLP algorithms to measure GDPR compliance in the language and declarations used in different dangerous android permission declaration

[1] https://developer.android.com/training/permissions/requesting

methodologies at multiple textual dimensions. The algorithms investigated are Universal Sentence Encoder (USE) [40], Sentence Bert (SBERT) [41], Glove [42], Bi-Directional Encoder Representations (BERT) word embedding [43], N-Grams, Vector Space Modelling (VSM) [44], [45] and Fuzzy String Matching (FSM).

- **Requirements Engineering**: While other techniques have operationalized requirements from texts using statistical NLP [46], semantic frames [47], semantic parsing [48], domain-specific language [49], graphical modelling language [50], privacy-enhanced business process model and notation [51],information-flow labels [21], we used statistical NLP and UML mapping to identify permission-related requirement.
- **Privacy Policy Generation at Design Time Using UML Diagrams:** Using modelling languages for visualising a system at design time, we implement a solution that helps developers to generate compliant sensitive permission declarations using UML diagrams (class diagrams, activity diagrams etc) during design time. First, it scans the UML diagrams and checks which permission is required based on the classes, attributes, operations and relationships between objects and generates a privacy policy declaration for the specific sensitive permission based on a specified threshold.

In this paper, the term *DAPD* is used frequently. By *DAPD*, we mean statements in the app privacy policy explicitly or implicitly describing access to *dangerous* or *sensitive* Android permission declared in the app's manifest file. These statements are required to provide information about the sensitive data the application is collecting through dangerous permissions and how it will be processed. If the application is accessing multiple sensitive areas of a user's device, then, it is expected to find multiple *DAPDs* in the app policy since the app requires permissions for each area. By *DAPD methodologies*, we mean the different methods, application developers are using to provide these permission-policy snippets in their app privacy policy. We use *completeness* and *compliance* interchangeably. We are also aware of the debate around the use of terminologies, *privacy policies* and *privacy notices*, which are two distinct documents. The argument has been that privacy policy is internal, while privacy notices are external and customer-facing. As a result, privacy notices are statements that explain to visitors (users) how their data will be used and their privacy rights, but privacy policies are the company's guidelines for how employees should protect customer data.[2] For the sake of this study, an external customer-facing statement prepared by app developers that outlines how the app collects uses, and shares user data is referred to as the *app privacy policy*.

The rest of the paper is structured as follows. Section II reviews the literature on eliciting privacy and security requirements from GDPR for system compliance, completeness checking of privacy policies and applications, and natural language processing techniques for textual similarity in GDPR. The methodology is presented in Section III including the key components of the proposed framework for runtime and design time GDPR compliance checking using Android app permissions, the datasets used and the preprocessing steps, textual similarity algorithms implemented and the similarity metric used. Section IV demonstrates the results obtained from experiments designed to evaluate the proposed compliance-checking methodology. We also discuss the practical implications of the results from a developer and platform perspective. Section V discusses the limitations of the proposed approach and future directions, while Section VI concludes the work with a summary of the key findings and future work.

[2]https://termly.io/resources/articles/privacy-notice-vs-privacy-policy/

II. LITERATURE REVIEW

While there are works in literature [50], [52], [53], [54], [55], [56], [57] that have focused on extracting privacy-related and software requirements from GDPR, our work is focused on assisting developers with the compliance requirements associated with Android permissions declarations and UML design based on articles from the GDPR law. We provide a literature review of two key areas that relate to our work: (i) completeness checking of privacy policies, and (ii) completeness checking of software (applications) against data protection regulations.

A. COMPLETENESS CHECKING OF PRIVACY POLICIES

Completeness checking of privacy policies against GDPR was examined in [28] and [29] using a two-pronged approach that identifies privacy-related requirements in GDPR with privacy policies using a conceptual model of metadata traceable to GDPR articles. Abualhaija et al. [30] proposed an automated question-answering approach useful for discovering legal text passages related to compliance requirements to help requirements engineers embed privacy in the design of software systems. Lippi et al. [31] proposed *CLAUDETTE*, a web server that automates the detection of potentially unfair clauses in online contracts using machine learning and natural language processing on a corpus of 50 contracts, to accomplish AI-enabled consumer protection. Tesfay et al. [38] proposed *PrivacyGuide*, an end-user support tool for reading and understanding privacy policies using GDPR as the guide. Sanchez et al. [32] investigated the automation of privacy policy compliance as a multilabel text classification task using SVM. Each statement in a given policy is assessed and classified against each data protection goal listed in GDPR.

Using a dataset of 115 privacy policies, Mousavi et al. [39] used word embeddings, CNN and BERT for the multilabel classification of privacy policy paragraphs into predefined categories to produce a standard benchmark for privacy policy classification. Through the representation of data practice descriptions in privacy statements as semantic frames, Bhatia et al. [33] proposed an approach for identifying incompleteness in data action instances such as

collection, retention, usage and transfer. By modelling data-intensive applications (DIAS) as a dataflow, Guerriero et al. [34] proposed a framework for defining, enforcing and checking privacy policies in large-scale DIAs. Elwany et al. [58] produced an Optical Character Recognition (OCR) mechanism to analyze legal documentation by leveraging a fine-turned BERT model to understand and extract text from legal corpora. Elluri et al. [59] measured the semantic similarity of different GDPR laws with cloud privacy policies. Hegel et al. [60] used NLP and OCR in legal documents to extract visual features such as layout, style and text placement to extract important pieces of information through enhanced contextual understanding. Other approaches have used crowdsourcing techniques to investigate whether data practises and privacy goals can be reliably extracted from privacy policies through crowdsourcing for the completeness of privacy policy checking [35], [36], [37].

The major advantage of approaches in this area is that they provide an automated way of verifying whether the content of a privacy policy is complete according to the provisions of relevant data protection regulations such as GDPR. By designing completeness criteria based on data privacy goals or privacy-related provisions in the GDPR, these approaches can investigate violations in privacy policies. This approach has some limitations. Firstly, they do not investigate the problem at a personal data or sensitive user actions level in the privacy policy. Solutions are developed by extracting metadata from the GDPR for completeness checking. A violation has taken place, for instance, if a controller is not named in a privacy policy. Such information is vague about which sensitive or personally identifiable information was compromised. Second, a subjective interpretation and comprehension of GDPR articles are used in the construction of the criterion. Thirdly, the GDPR identifies personal data and special categories of data in its definition, which calls for various processing requirements. However, the problem is only broadly examined by present methodologies for completeness methods. Finally, the approaches are not generalizable as the multi-domain evaluation of the metadata identification and completeness approaches have not been verified. To replicate the methodologies of completeness checking based on metadata for other data protection regulations such as the California Consumer Privacy Act (CCPA), a new conceptual model of privacy-policy metadata through systematic qualitative and completeness checking criteria for privacy policies for CCPA would be developed that feeds into developing an automated solution. This required effort hinders the replication of the proposed methodologies.

B. COMPLETENESS CHECKING OF APPLICATIONS

Users are concerned about the privacy of applications they use, especially if sensitive user data is involved, as evidenced by user reviews of COVID-19 contact tracing apps [61]. Fan et al. [62] investigated GDPR compliance violations at the app privacy and code level in mobile health applications by verifying the completeness of privacy policy, the consistency of data collection and the security of data transmission. In an exploratory study, Kununka et al. [63] examined the data handling practices and privacy policy compliance of Android and iOS apps for discrepancies. Hatamian et al. [64] studied the extent to which COVID-19 contact tracing Android apps comply with the legal requirements of GDPR. Rahman et al. [13] proposed an automated machine learning solution to evaluate completeness checking in Android applications dangerous permissions against privacy policies and highlighted the non-transparent state of permission-policy declarations of dangerous Android permissions. Shezan et al. [48] developed an NLP-driven approach, *NLP2GDPR*, to automatically extract text from Android applications and generate a GDPR-compliant feature. Slavin et al. [19] created an approach that identifies privacy promises in mobile application privacy policies and checks against the code using information flow analysis to see if data is extracted outside of an application thus infringing on privacy policy declarations.

The approaches in this domain have made significant progress in compliance checking of applications. This is done by investigating the compliance level of different kinds of mobile applications with legal requirements in GDPR and investigating discrepancies in applications for violations. These approaches also go beyond the app privacy policies by checking for violations in the app code and permissions. One of the major limitations of these approaches is that they have not considered the three-pronged approach of completeness checking using the permissions, app privacy policy and GDPR for a robust view. Apps require privacy policies, and those policies must be GDPR-compliant and disclose sensitive data access that requires dangerous permissions. It is this limitation that influenced the proposed research. By examining related works on completeness checking of privacy policies against GDPR and completeness checking of Applications, identifying privacy-related requirements and NLP for semantic similarity in legal documents, it was observed that little or no empirical analysis has been conducted to measure compliance of permission policy statements for dangerous android permissions with GDPR and the generation of GDPR-compliance permission policy statements at design time using UML diagrams.

III. METHODOLOGY

The methodology investigated in this study is an NLP-based automated compliance checking of Android permissions-policy declaration against GDPR. We discuss the proposed methodology by describing the framework, dataset collection and pre-processing, language understanding algorithms for similarity matching and the similarity metric for measuring the distance between the vector representation of the permission policy and the GDPR corpus.

A. OVERVIEW OF FRAMEWORK

The methodology investigated in this study is an NLP-based automated compliance checking of Android permissions-policy declaration against GDPR. The proposed approach for checking the runtime and design time GDPR compliance using Android app permissions spans four different tasks. In the first task, we extract and process the text in GDPR using natural language processing algorithms. In the second task, we process the text from the annotated corpus that matches each dangerous android permission to declarations used by over 270 mobile applications. In the third task, we perform the completeness checking of the Android permission declarations against GDPR articles and recitals. In the final task, we extract permission requirements from UML diagrams for GDPR-compliant permission policy generation at design time. In general, our approach enables an implicit compliance checking of the software using the dangerous android permissions declaration and the class diagram against the articles and recitals in the GDPR. Our work concentrates on providing automation for all the tasks. Figure 1 shows the framework for measuring the completeness of dangerous Android permissions declarations in privacy policies against GDPR laws.

We propose a novel framework that leverages the MPP-270 annotated policy corpus that maps permission and privacy policy snippets of all the 10 dangerous permission categories and every GDPR article and are compared using six NLP algorithms at five textual dimensions to calculate a cosine similarity (CS) results as shown in Figure 1. The five different textual dimensions are represented at the sentence level using SBERT and USE, BERT at the word embedding level, FSM at a pure string level, VSM at a vectorization level and BERT and GloVe vectorizations are applied at the N-Gram level. The DAPD identified with the highest cosine similarity result is extracted for each GDPR article on every algorithm for all the permission categories. As a benchmark dataset for permission completeness, the open-source MPP-270 annotated corpus was developed in [13]. We used the annotated corpus to investigate the GDPR compliance of permission-policy statements. Details regarding the dataset's development and the human annotation process are available on the project website [14].

B. DATASET

The framework requires two input datasets - the GDPR [1] and MPP-270 [14] corpus. To measure DAPD compliance with GDPR laws, a GDPR corpus with suitable recitals was designed that contained every GDPR `article number`, `title` and `text` in a structured format. This corpus was self-created to be data analytic and includes suitable recitals. The MPP-270 Corpus which is an annotated corpus containing the methodologies that have been used to declare DAPD in 10 permission categories from 270 Android applications was used as a ground truth to match the semantic similarity against the text found in every GDPR article.

TABLE 1. 30 dangerous permission APIs categorized into 10 permission groups [65].

Permission Group	Dangerous Permissions
CALENDAR	READ_CALENDAR
	WRITE_CALENDAR
CAMERA	CAMERA
CONTACTS	READ_CONTACTS
	WRITE_CONTACTS
	GET_CONTACTS
LOCATION	ACCESS_FINE_LOCATION
	ACCESS_COARSE_LOCATION
	ACCESS_WIFI_STATE
	ACCESS_MEDIA_LOCATION
MICROPHONE	RECORD_AUDIO
PHONE_CALL	READ_PHONE_NUMBERS
	CALL_PHONE
	ANSWER_PHONE_CALLS
	ADD_VOICEMAIL
	USE_SIP
	READ_CALL_LOG
	WRITE_CALL_LOG
	PROCESS_OUTGOING_CALLS
PERSISTENTID	READ_PHONE_STATE
	ACCESS_NETWORK_STATE
SENSOR	BODY_SENSORS
	ACTIVITY_RECOGNITION
SMS	SEND_SMS
	RECEIVE_SMS
	READ_SMS
	RECEIVE_WAP_PUSH
	RECEIVE_MMS
STORAGE	READ_EXTERNAL_STORAGE
	WRITE_EXTERNAL_STORAGE

The annotated policy corpus describes three key pieces of information about the app, namely: i) the app identifier, in this case, the package name ii) its declared dangerous permissions extracted from the app manifest file, and iii) the permission-policy snippets extracted from the app privacy policy for each dangerous android permissions declared in the app manifest file. If the app did not declare the dangerous permission, then the value `'0'` is used in place of a policy text, while `'NOT_FOUND'` means that the 8annotators were unable to locate any permission-policy snippets for the declared dangerous permission [14]. These permission categories include CAMERA, MICROPHONE, PHONE_CALL, SENSOR, SMS, CALENDAR, CONTACTS, LOCATION, STORAGE and PERSISTENTID (cf Table 1). The list of permissions considered is consistent with 30 dangerous permission APIs categorized in 10 permission groups in MPP-270 [13], [14], an annotated policy corpus for mapping between dangerous android permissions and privacy.

C. DATASET PREPROCESSING

The GDPR and the MPP-270 corpus dataset were pre-processed for the N-grams, VSM, FSM and

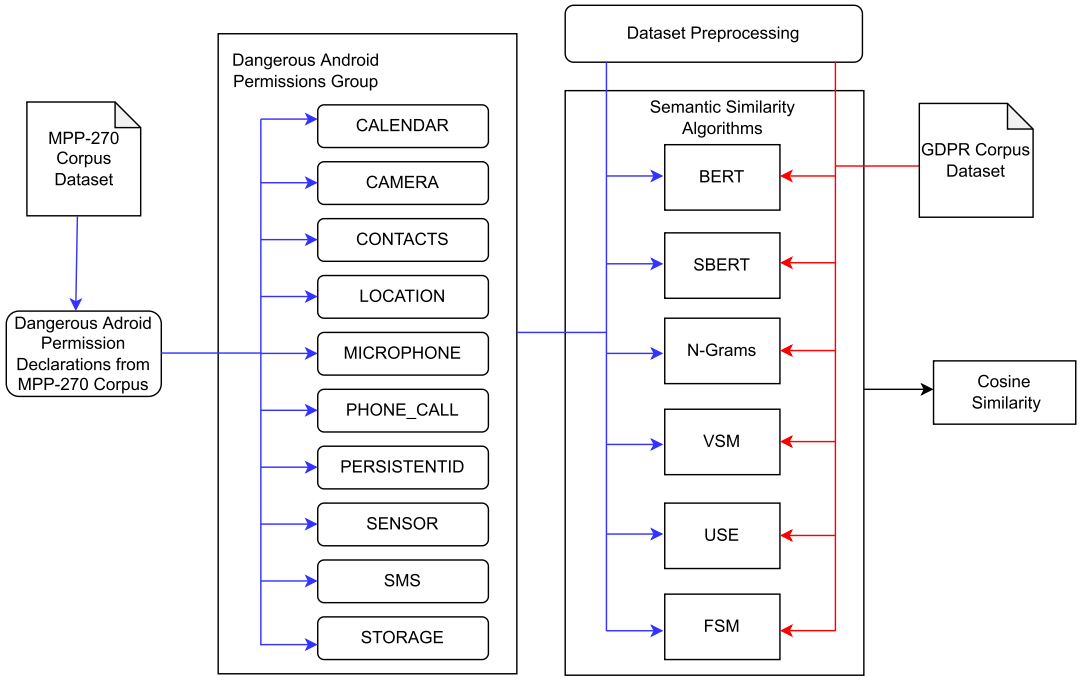


FIGURE 1. Completeness checking of dangerous android permissions declarations in app privacy policies against GDPR.

implementation of the BERT word embedding algorithms. The preprocessing steps include removing all stop words and punctuation and applying lemmatization. Lemmatization was applied over stemming for the reason that lemmatization stores more semantic context. Context is important while applying semantic similarity thus applying a form of stemming could cause the reduced words to become ambiguous or incorrect. Numbers were not removed as some articles are included with certain references to laws and directives which is considered an important aspect. For example, if a DAPD directly references a law or directive then the compliance should increase. For the implementation of SBERT and USE algorithms, removing stop words and lemmatization was not applied to maximize effectiveness and improve accuracy. This was because SBERT reads and takes into consideration the words left to right of each scanned word for each sentence to understand the sentence context. The MPP-270 dataset also had additional measures implemented to extract accurate information. For example, any value encountered in a column that was 0 or did not exist was not extracted for analysis and handled accordingly.

D. SEMANTIC SIMILARITY ALGORITHMS

The goal of the semantic similarity algorithms is to extract textual entities at different textual dimensions from the GDPR and MPP-270 annotated policy corpus. The output would take one of these forms - sentences, word embeddings, strings, vectors, and N-grams depending on the encoding methods of the algorithm. We describe the choice and methodology of the six algorithms implemented in the research below.

1) SENTENCE EMBEDDING

SBERT was implemented by encoding the meaning of the specified sentence with the rest of the index for both DAPD and GDPR laws. The SBERT algorithm implemented the pre-trained model `'all-mpnet-base-v2'` which is a model trained on 1 billion training pairs of data. SBERT was used as it takes into consideration the semantic context of every word in a sentence [41]. Both USE and SBERT used the cosine similarity function found in the `Scipy` toolkit for the result. Other sentence embedding techniques such as *InferSent* and *SentEval* were considered. However, the results in [41], highlight that an SBERT implementation outperforms *InferSent* and *SentEval*. USE on the other hand was implemented to gauge the embedding interpretation derived from using a question-and-answer pre-trained model. Implementing embedding at a sentence dimension allows the identification of areas that fail to conform to aspects of GDPR laws but also identifies the most compliant areas.

The sentence embedding techniques SBERT and USE enable the embedding encoding methodology. The outputs are completely different to each other with SBERT producing a vector embedding representation while USE outputs the results as a tensor object for each sentence. Universal Sentence Encoders have been used in [66] for encoding texts of the GDPR articles and privacy by design principles for automated text similarity tasks. Sentence embedding models have been utilised to detect dangerous Android permissions

in-app privacy policies in [18]. The use of USE and SBERT have yielded highly precise annotation in [67] for semantic matching between text associated with privacy controls and user queries.

2) WORD EMBEDDING

The BERT word embedding algorithm implements the pre-trained model `'bert-base-uncased'` which was trained on 110 million parameters of uncased text in the English language [43]. In this implementation, the context of each word is considered for the entire index for both text corpora using a tensor-based approach. BERT takes into consideration the words surrounding each word and contextualizes each word. As a result, two BERT implementations were investigated. The first BERT implementation was created with no major preprocessing techniques such as the removal of punctuation. Stop words were retained to investigate whether the use of stop words increased compliance because of the increased context the algorithm may derive from the overall sequence of tokens. The second implementation of BERT is a preprocessed implementation that uses N-grams. BERT has been used in [58] for understanding and analyzing legal documents and in [30] for extracting compliance requirements from GDPR. The word embedding technique uses BERT as the methodology and uses tokenized encoding. The output is a tensor object created using the sequence of input tokens from the sentence with individually tokenized words.

The transformer architecture [68], which makes use of bidirectional self-attention, is the foundation of BERT. The BERT's attention mechanism operates on a collection of queries (Q), keys (K), and values (V), each of which is a scaling dot product matrix. The dimension of Q and K is d_k, while the dimension of V is d_v. The weights on the values are obtained using a softmax function, and the matrix of the result is calculated as follows:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Each $Head_i$ trains its attention map using a group of random parameter matrices on the queries, keys, and values since Multi-Head attention comprises several attention layers operating concurrently [68] as shown below:

$$\begin{aligned}\text{Multi-Head}(Q, K, V) &= \text{Concat}\,(\text{Head}_1, \ldots,\ \text{Head}_h)\,W^O \\ \text{where } Head_i &= \text{Attention}\left(QW_i^Q, K\ W_i^K, V\ W_i^V\right)\end{aligned} \tag{2}$$

where the projections $W_i^Q, K\ W_i^K, V\ W_i^V, W^O$ are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v} W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$

The BERT base model (uncased) adopts Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) as training objectives to learn bidirectional representations. Given a sentence s = (s_1, s_2, s_3,...s_n), MLM randomly masks 15% tokens and replaces them with a special symbol [MASK]. Let us define T as the set of masked positions, s_T as the set of masked tokens, and $s_{\setminus T}$ as the sentence after masking. MLM pre-trains the model θ by maximizing the following objective:

$$\log P\left(s_T \mid s_{\setminus T}; \theta\right) \approx \sum_{t \in T} \log P\left(s_t \mid s_{\setminus T}; \theta\right). \tag{3}$$

3) N-GRAMS

The N-gram algorithm analyses and extracts the most common N-grams of two in every GDPR law and DAPD. The algorithm then identifies the DAPD N-gram with the highest cosine similarity against each GDPR law. The most common and highest cosine similarity N-grams are then embedded using GloVe. The GloVe implementation is trained using the `'glove.6B.300d.word2vec'` corpus and then semantically compared. The most common N-grams are also compared semantically to both corpora. N-grams were analyzed only at a bi-gram level as the Android permission category `SENSOR` which is also interpreted occasionally as `BODY_SENSORS` is the only category represented at a bi-gram level. A calculation to measure the semantic meaning of bi-grams was established in [69] and the benchmark synonymy value of two words was proposed to be 0.8025. Thus, 0.8025 will be the threshold value to conclude if the bi-grams between each corpus are compliant at the N-gram level dimension. GloVe and BERT were used for embedding and obtaining a cosine similarity measurement of the most common N-grams over other word embedding techniques. It was important to implement BERT to compare how a contextualized N-gram implementation compares to a fixed vector interpretation. The BERT results have been converted from a tensor flow numerical representation to a floating point. Implementing N-grams and GloVe embeddings is consistent with state-of-the-art techniques in tasks that involve mapping privacy policies with GDPR laws. N-grams were used in learning the GDPR data protection goals for completeness checking of privacy policies under GDPR in [32]. Pre-trained Glove Word embeddings were used in [28] and [29] for vector-space representations of text in the completeness checking of privacy policies against GDPR. Word embedding models like *GloVe, word2vec* and *fastText* were implemented in [70] for measuring the semantic correlation between sensitive Android permissions and app textual descriptions. N-Gram was implemented with three different encoding methods. The standard N-Gram implementation tokenizes the most common N-Grams and outputs the N-Gram as a tuple. The N-Gram BERT and GloVe implementations output the result as an array of vectors and a tensor object respectively.

4) VSM TFIDF

The vector space dimension statistical approach was utilised to describe the semantic similarity between the GDPR and sensitive Android permission-policy snippets using a TFIDF implementation utilising VSM. Shahmirzadi et al. [44] used Vector Space Modelling (VSM) to extract

metrics relating to patent-to-patent similarity to evaluate the performance of VSM on a variety of TFIDF variations and text similarity methodologies. According to these findings, the baseline Term Frequency-Inverse Document Frequency (TFIDF) implementation for VSM is an appropriate choice for determining text similarity, while other TFIDF versions were not beneficial. We implemented the algorithm since the findings from [44] demonstrated that TFIDF VSM is suitable for determining textual similarity at the vector level. For the cosine similarity result, the `Scikit-Learn` pairwise similarity function was used. The VSM technique enables TFIDF and uses tokenized vectors, the output is an array consisting of the term frequency of each tokenized word.

5) FUZZY STRING MATCHING

The FSM algorithm was implemented using the `TheFuzz`[3] toolkit to interpret GDPR and DAPD at a pure string dimension. This algorithm can recognise smaller changes to both text similarity, such an algorithm could assist in interpreting semantic similarity. Two FSM variants were chosen to measure text similarity, these include the *FSM Set Ratio* which finds the ratio of common tokens and calculates a similarity score and the *FSM Partial Ratio* which is a Levensthtein distance approach in which each word is tokenised with the accumulated common words in both strings for comparison. *Partial Ratio* was chosen due to its suitability in comparing strings that are not the same length while *Set Ratio* was chosen due to its flexible detection ability regarding the interpretation of out-of-order words and textual homographs. FSM calculates the score between the two corpora and also compares the score between the most common N-Grams extracted for every GDPR law and the most common N-Gram from the most similar DAPD. Match similarity produced by the Fuzzy String matching technique based on the Levenshtein distance was used in [71] for verifying GDPR compliance based on informed consent and in [72] for analysing the impact of GDPR on website privacy policies.

Other methods for representing textual dimensions in the domain of topic distribution and clustering algorithms were investigated for their suitability. Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Hierarchical Dirichlet Process (HDP) were all experimented with. Jenson-Shannon distance, Wasserstein Distance (WD) and Euclidean distance were used as distance metrics to attempt to find similarities between textual entities in the corpora. However, for these algorithms to produce findings that are dependable, stable, and consistent, a big corpus is required. Since the GDPR corpora are extensive and the MPP-270 corpus is short, this strategy was quickly determined to be inadequate. Table 2 shows the NLP algorithm techniques, methodologies, encoding method and output result.

[3]https://github.com/seatgeek/thefuzz

E. COSINE SIMILARITY

To measure the results of the USE, SBERT, BERT word embedding, N-gram and VSM algorithms, cosine similarity was adopted to interpret a measurement of similarity between indexes of the two corpora. The choice of cosine similarity measure for computing the statistical similarity between two textual entities is consistent with their effective use in document similarity tasks [73], [74], [75], [76]. Furthermore, the effectiveness of the cosine similarity measure has been validated in completeness checking tasks of mapping privacy policies against GDPR [29], [62], [77]. One of the aims of this study is to find permission-policy snippets that maximize integrability with GDPR compliance through semantic similarity, and cosine similarity seems to be the most suited for the task. Other metrics such as Wasserstein Distance (WD) were considered but this metric assumes the inputs are probability distributions while the algorithm implemented was represented using embedding and vectorization.

The cosine similarity for comparing two vectors is defined as follows:

$$\cos(X, Y) = \frac{X \cdot Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2}\sqrt{\sum_{i=1}^{n} (Y_i)^2}} \quad (4)$$

where X and Y are vector representations from the permission-policy declaration and GDPR corpus. A high cosine value indicates that permission declaration in the app privacy policy is closely related to an article in the GDPR and thus a completeness and compliance judgement can be made about the permission.

IV. EVALUATION

This study aims to answer three key questions and sub-questions that inform the experimental design.

- **RQ1**: Does the declaration for sensitive Android permission in the App permission policy contain meaningful and relevant information in line with GDPR articles and recitals about collecting and processing sensitive data?
  - Is the range of sensitive or dangerous Android app permissions supported by the Android ecosystem adequate and sufficient to fulfil GDPR obligations?
  - What is the level of GDPR compliance of DAPD used by developers?
  - Does GDPR use meaningful and relevant language to enhance the completeness checking of Android app permissions?
  - Are the declarations used in DAPD detailed enough?
- **RQ2**: How can we generate privacy policies for sensitive dangerous permissions requested by an app from the UML diagram in such a way to clearly and specifically inform users about sensitive data being requested, actions the permissions represent or their semantic meaning in line with data protection laws?
  - Is it feasible to assist mobile app developers in the automated generation of GDPR-compliant

TABLE 2. Comparing the different algorithms applied for semantic textual similarity between DAPD and GDPR compliance.

Example Dangerous Android Permissions Declaration: "...mobile network information such as the public land mobile network plmn provider id and internet protocol ip address"			
Technique	**Method**	**Encoding Method**	**Output Result**
Sentence Embedding	SBERT	Embedding	[-2.49856617e-02 -1.16642667e-02 -3.38957533e-02.....
Sentence Embedding	USE	Embedding	tf.Tensor([-1.74765370e-03 -4.58935387e-02 6.51189908e-02.....
Word Embedding	BERT	Tokenised Encoding	[tensor([[101, 4684, 2897, 2592, 2107, 2004, 1996, 2270, 2455, 4684....
Vector Space Modelling	VSM TFIDF	Tokenised Vectors	[[1 1 1 1 1 1 2 2 1 1 1 1]]
N-Gram	N-Gram	Tokenised	('mobile', 'network')
N-Gram	GloVe N-Gram	GloVe Vectors	('mobile', 'network') = [array([-1.64296463e-01, -1.09365014e-02.....
N-Gram	BERT N-Gram	BERT Embedding	('mobile', 'network') = tensor([[-4.4379e-01, 1.3519e-01, 1.1006e-01.....
Fuzzy String Matching	Partial Ratio	Tokenized	['mobile', 'network', 'information', 'public', 'land', 'mobile', 'network', 'plmn'...
Fuzzy String Matching	Set Ratio	Tokenized	['mobile', 'network', 'information', 'public', 'land', 'mobile', 'network', 'plmn'...

permission-policy snippets by extracting permission requirements from UML diagrams at design time?

- **RQ3**: Can we adequately conduct compliance by matching GDRP laws with Android permissions categories, APIs and permission-policy declarations?
  - To what extent is it possible to accurately classify dangerous Android permissions with GDPR?

To answer **RQ1**, experiments were conducted by mapping the permission-policy snippets of dangerous Android permissions for measuring completeness and compliance. To answer **RQ2**, UML diagrams in the form of XML data or raw PNG files are taken as input for requirements engineering and privacy policy generation for sensitive Android permissions for design time compliance. To answer **RQ3**, we analyse the results from the runtime analysis using permission-policy snippets in **RQ1** and design time analysis using UML diagrams to measure the effectiveness of GDPR compliance at design and runtime using Android permissions.

In presenting the results, we use some terms such as *average declaration, cosine similarity average,* and *highest average identified*. The *average declaration* is a metric calculated for each permission category in which every GDPR article is matched with every DAPD methodology with an average calculated from the resulting cosine similarity, a *cosine similarity average* is then derived from the resulting *cosine similarity average* for each GDPR and DAPD comparison. This can be described as an *average of averages*. An equivalent FSM score is calculated in Table 3. The *highest average identified declaration* metric on the other hand takes the cosine similarity result for the highest identified DAPD methodologies for each GDPR article in each permission category.

A. RQ1: COMPLETENESS CHECKING OF SENSTIVE ANDROID PERMISSIONS AND GDPR

To answer RQ1, the permission-policy snippets for the dangerous android permission with the highest cosine similarity are extracted for each GDPR law using all the textual similarity algorithms (cf Section III-D) for all permission categories (cf Table 1). The result of the experiments shows the most compliant dangerous Android permission policy declarations to use for each GDPR law. Table 3 shows the results from this experiment, the highest DAPD cosine similarity result for every GDPR law is compared to the average cosine similarity result for every GDPR law to visualize the compliance to GDPR increase when using the correct DAPD methodologies. Table 4 shows the FSM results for the average FSM DAPD compliance to GDPR compared to the highest FSM DAPD methodologies for every GDPR law. It is important to note the difference in scale and sensitivity of the textual similarity score of each algorithm for measuring compliance. For example, a cosine similarity of 0.50 for GloVe might be considered very high based on the nature of vectorization. While a BERT word embedding cosine similarity of 0.60 would be described as low and 0.80 as high based on the contextual nature of BERT and its ability to find similarities of long-distance words.

Table 3 shows low compliance when the cosine similarity results are derived from using VSM TFIDIF. This shows at a vector level the methodology to declare DAPD with GDPR laws does not comply. VSM TFIDF is the equivalent of searching for a word-to-word similarity and uses a term frequency to derive a result on how important certain words are. The compliance is low as certain words contained in GDPR laws are not being used in the DAPD methodologies. The level of compliance is expected to be low at this textual dimension considering the method VSM TFIDF functions, the level of compliance in some increases significantly to the point that in the case of the STORAGE category, the compliance increased by 325% to the highest cosine similarity average of 0.34. The compliance level may be low but this means 34% of the word and the associated term frequency comply between GDPR and using the highest resulting cosine similarity identified DAPD. Such results could indicate that to raise DAPD compliance, a developer could use contextually similar words. Table 4 shows the results from both the FSM algorithms. Using the highest DAPD increases GDPR compliance substantially for both algorithms. Although more reliable results were derived from the pre-processed word embedding technique. The USE results in Table 3 do not give good results although this was expected considering the pre-trained model was trained on a question-and-answer set.

TABLE 3. VSM, USE, SBERT, GloVe N-gram and BERT N-gram results of DAPD compliance against GDPR laws showing the difference between the highest and the average cosine similarity.

Dangerous Android Permission **CAMERA** Declarations (Cosine Similarity)								
	VSM TFIDF	USE	SBERT	Glove N_gram	BERT N_gram	BERT Word Embedding	Pre-Processed BERT Word Embedding	Overall Average Excluding N-gram
Highest Average Identified Declaration	0.38	0.170	0.31	0.40	0.61	0.72	0.80	0.48
Average Declaration	0.15	0.049	0.16			0.63	0.70	0.34
Percentage Increase	153.30%	246.94%	177.75%			14.29%	14.29%	41.18%
Dangerous Android Permission **MICROPHONE** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.42	0.17	0.33	0.15	0.46	0.76	0.81	0.50
Average Declaration	0.17	0.05	0.17			0.64	0.71	0.35
Percentage Increase	147.06%	240%	94.18%			18.75%	14.05%	42.86%
Dangerous Android Permission **PHONE_CALL** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.30	0.14	0.25	0.28	0.53	0.70	0.77	0.43
Average Declaration	0.11	0.052	0.15			0.66	0.72	0.34
Percentage Increase	172.72%	169.23%	66.60%			6.06%	6.94%	26.47%
Dangerous Android Permission **SENSOR** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.27	0.096	0.17	0.26	0.48	0.68	0.73	0.39
Average Declaration	0.17	0.05	0.11			0.64	0.70	0.33
Percentage Increase	58.82%	92%	54.54%			6.25%	4.29%	18.18%
Dangerous Android Permission **SMS** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.24	0.09	0.18	0.32	0.53	0.68	0.75	0.39
Average Declaration	0.14	0.052	0.14			0.66	0.73	0.34
Percentage Increase	71.43%	73.08%	28.57%			3.03%	2.74%	14.71%
Dangerous Android Permission **CALENDAR** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.27	0.09	0.19	0.26	0.48	0.64	0.72	0.38
Average Declaration	0.17	0.04	0.12			0.59	0.67	0.32
Percentage Increase	58.82%	125%	58.30%			8.47%	7.46%	18.75%
Dangerous Android Permission **CONTACTS** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.37	0.19	0.32	0.25	0.55	0.74	0.81	0.49
Average Declaration	0.13	0.04	0.15			0.62	0.70	0.33
Percentage Increase	184.62%	375%	113.30%			19.35%	15.71%	48.48%
Dangerous Android Permission **LOCATION** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.44	0.20	0.40	0.41	0.60	0.79	0.85	0.54
Average Declaration	0.20	0.05	0.18			0.66	0.73	0.36
Percentage Increase	120%	375%	122.2%			19.69%	16.44%	50%
Dangerous Android Permission **STORAGE** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.34	0.19	0.34	0.30	0.55	0.74	0.81	0.48
Average Declaration	0.08	0.03	0.15			0.62	0.69	0.31
Percentage Increase	325%	533.30%	126.60%			19.35%	17.39%	54.84%
Dangerous Android Permission **PERSISTENTID** Declarations (Cosine Similarity)								
Highest Average Identified Declaration	0.39	0.19	0.33	0.41	0.62	0.74	0.83	0.50
Average Declaration	0.13	0.04	0.17			0.64	0.73	0.34
Percentage Increase	200%	375%	94.18%			15.63%	16.70%	47.06%

In Table 5, the *highest* identified cosine similarity relates to the highest similarity value identified between a GDPR law and permission-policy declarations for each dangerous android permission category. The *average* highest cosine similarity relates to the derived highest cosine similarity result of every DAPD vs GDPR. In contrast, the *overall average* cosine similarity relates to the average derived result of each DAPD vs a corresponding GDPR law. The main issue identified relates to the average result of methodologies developers use to declare DAPD which are not compliant with GDPR. The contribution of this research is the identification of the most compliant DAPD for a developer to use for each corresponding GDPR law in each dangerous permission category. Analyzing the average cosine similarity results from Table 5, it is found that the `CAMERA` dangerous permission category has an overall average of 0.70 for DAPD compliance with GDPR, while the highest identified DAPD for every GDPR law in the category averages a cosine similarity score of 0.80. For the `MICROPHONE` permission category, using the highest identified DAPD methodologies for each GDPR law increases GDPR compliance to 0.81 which is a substantial increase from the overall compliance average of 0.71. This trend for increasing compliance continues for the `PHONE_CALL` category in which compliance rises from

TABLE 4. FSM partial and set ratio results of DAPD compliance against GDPR laws showing the difference between the highest and the average cosine similarity.

	FSM (Partial Ratio)	FSM (Set Ratio)	Overall Average
CAMERA - FSM Similarity			
Highest Average Identified Declaration	53.61	37.86	45.74
Average	16.75	16.4	16.57
Percentage Increase	218.75%	130.90%	175.90%
MICROPHONE - FSM Similarity)			
Highest Average Identified Declaration	50.84	38.56	44.70
Average	16.24	16.72	16.48
Percentage Increase	212.98%	130.58%	171.19%
PHONE_CALL - FSM Similarity)			
Highest Average Identified Declaration	48.96	32.40	40.68
Average	16.41	16.61	16.51
Percentage Increase	198.33%	95.02%	146.80%
SENSOR - FSM Similarity			
Highest Average Identified Declaration	18.75	27.13	22.94
Average	11.5	16.64	14.09
Percentage Increase	62.49%	62.96%	38.93%
SMS - FSM Similarity)			
Highest Average Identified Declaration	19.17	27.37	23.27
Average	13.13	20.87	17.00
Percentage Increase	45.96%	31.10%	36.84%
CALENDAR - FSM Similarity)			
Highest Average Identified Declaration	51.35	37.30	44.32
Average	23.09	21.90	22.50
Percentage Increase	122.32%	70.26%	96.99%
CONTACTS - FSM Similarity)			
Highest Average Identified Declaration	54.09	44.48	49.28
Average	17.64	17.54	17.59
Percentage Increase	206.56%	153.61%	180.17%
LOCATION - FSM Similarity)			
Highest Average Identified Declaration	54.17	43.93	49.05
Average	11.41	16.65	14.03
Percentage Increase	374.54%	163.79%	249.49%
STORAGE - FSM Similarity			
Highest Average Identified Declaration	59.60	51.92	55.76
Average	31.01	20.18	25.59
Percentage Increase	92.21%	157.25%	117.86%
PERSISTENTID - FSM Similarity			
Highest Average Identified Declaration	54.09	44.48	49.28
Average	17.64	17.54	17.59
Percentage Increase	206.56%	153.61%	180.17%

an overall average of 0.72 to 0.77. The SENSOR and SMS categories reveal the smallest increases in DAPD GDPR compliance in which the SENSOR permission category increases from an overall compliance result of 0.70 to 0.73 and the SMS category increases from an overall average result of 0.72 to the highest identified average of 0.75. The compliance for the CALENDAR permission category had an identified increase from 0.67 to 0.72. The SMS, SENSOR and CALENDAR permission categories have the lowest compliance increase, thus suggesting the overall quality of declaring DAPD is not good enough compared to the other dangerous permission categories. The CONTACTS dangerous permission category increases from 0.70 to 0.81 while LOCATION has a substantial increase in compliance from an overall average of 0.73 to the most compliant permissions increasing to 0.85. STORAGE increases in compliance from an overall average of 0.69 to 0.80 with the final dangerous permission category PERSISTENTID increasing from an overall average of 0.73 to 0.83. In some categories, using the identified highest complying DAPD can increase the average compliance for every GDPR law by nearly 20%.

With the SBERT results in Table 6, it is found that DAPD complies differently with different sentences in GDPR laws. Another contribution of this research is the identification of sections of GDPR sentences that are not covered or reduce DAPD compliance. Not only are the sections that reduce compliance identified, but the best DAPD methodology to comply best to that sentence is identified. Though these results express compliance issues, even the best methodologies that are used do not adequately cover certain sentences in parts of GDPR laws. This could reveal that more in-depth methodologies may be needed to comply with all sentences of GDPR laws. Table 6 represents an example in which two similar sentences in the same GDPR law use the same highest identified DAPD. The first aspect to note is that identifying the highest complying DAPD for each sentence significantly increases compliance with the GDPR law. The average cosine similarity compliance result for the first sentence in Table 6 is 0.29 while using the highest identified complying DAPD increases the compliance with GDPR to a cosine similarity value of 0.62. The second sentence in Table 6 is very similar to the first sentence but has a different context. The DAPD used is the same as the first sentence with the higher level of compliance, this shows that more in-depth details in declaring DAPD are needed for every sentence in the GDPR law to comply. The low compliance value for the second sentence shows that one specified declaration is not adequate to cover the entirety of GDPR laws. DAPD could in theory be mapped to each sentence of a GDPR law to derive the best level of GDPR compliance. The DAPD was not split into sentences to enhance investigation into the parts of GDPR that are lacking compliance when compared to the dangerous declaration permission methodology used. Some declaration methodologies are also too small to compare at the sentence level, for example, 'name and photo' are used as a declaration methodology to declare DAPD in the CAMERA category.

Also, Table 6 shows two different sentences from the GDPR law Right to Object are shown. Each DAPD methodology has the identified highest cosine similarity score declaration methodology. Each sentence from the same GDPR law has different methodologies to declare the DAPD with each sentence. The most important observation to note is the levels of compliance between the two methodologies and

TABLE 5. Comparison of the cosine similarity results between an identified DAPD and the corresponding GDPR article.

BERT Pre-processed Word Embedding Approach (Cosine Similarity = CS)				
Dangerous Android Permission Declaration With The Highest Cosine Similarity	Identified GDPR Text With Highest Cosine Similarity	Highest	Average	Overall Average
Dangerous Android Permission **CAMERA** Category				
such content may include your name social media username image likeness voice other identifiable information available in such user generated content you also have the right to withdraw consent where you have previously given your consent to the processing of your personal data for example by turning off camera access in your mobile device settings similarly we may ask for access to your camera in case you want to use certain features of our services	**Article 16 GDPR - Right to rectification** The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her. Taking into account the purposes of the processing, the data subject shall have the right to have incomplete personal data completed, including by means of providing a supplementary statement **Suitable Recitals: (65) Right of Rectification and Erasure**	0.89	0.80	0.70
Dangerous Android Permission **MICROPHONE** Category				
the following categories of your personal information may be shared within the scope of the disclosures described in the section disclosure above identifiers protected characteristics commercial information geolocation data internet or other electronic network activity information and audio electronic visual or similar information and inferences drawn from any of the above information categories the categories of information that we may automatically collect from you through the use of cookies and similar technologies include identifiers geolocation data commercial information audio electronic visual or similar information and internet or other electronic network activity information	**Article 44 GDPR - General principle for transfers** Any transfer of personal data which are undergoing processing or are intended for processing after transfer to a third country or to an international organisation shall take place only if, subject to the other provisions of this Regulation, the conditions laid down in this Chapter are complied with by the controller and processor, including for onward transfers of personal data from the third country or an international organisation to another third country or to another international organisation. All provisions in this Chapter shall be applied in order to ensure that the level of protection of natural persons guaranteed by this Regulation is not undermined. **Suitable Recitals: (101) General Principles for International Data Transfers, (102) International Agreements for an Appropriate Level of Data Protection**	0.86	0.81	0.71
Dangerous Android Permission **PHONE_CALL** Category				
mobile network information such as the public land mobile network plmn provider id and internet protocol ip address	**Article 4 GDPR - Definitions** **Suitable Recitals: (15) Technology Neutrality (24) Applicable to Controllers/Processors Not Established in the Union if Data Subjects Within the Union are Profiled (26) Not Applicable to Anonymous Data (28) Introduction of Pseudonymisation (29) Pseudonymisation at the Same Controller (30) Online Identifiers for Profiling and Identification (31) Not Applicable to Public Authorities in Connection with Their Official Tasks (34) Genetic Data (35) Health Data (36) Determination of the Main Establishment (37) Group of undertakings**	0.83	0.77	0.72
Dangerous Android Permission **SENSOR** Category				
this is the information regarding your location when you use a location-enabled service including the location of your device when you use parallel space collected from the gps wifi compass accelerometer or other sensors in your mobile device and the ip address of the device or internet service you use to access parallel space	Same as PHONE_CALL	0.80	0.73	0.70
Dangerous Android Permission **SMS** Category				
we also collect contact information if you choose to upload sync or import it from a device such as an address book or call log or sms log history which we use for things such as helping you and others find people you may know and for the other purposes listed below	Same as CAMERA	0.83	0.75	0.72
Dangerous Android Permission **CALENDAR** Category				
may use information from your calendar	**Article 31 GDPR - Cooperation with the supervisory authority** The controller and the processor and, where applicable, their representatives shall cooperate, on request, with the supervisory authority in the performance of its tasks. **Suitable Recitals: (82) Record of Processing Activities**	0.81	0.72	0.67
Dangerous Android Permission **CONTACTS** Category				
except in the circumstances of this paragraph or under the statutory requirement of law or authorized by user we will not publish or disclose users nonpublic phonebook or contact information	Same as SMS and CAMERA	0.90	0.81	0.70
Dangerous Android Permission **LOCATION** Category				
we will only collect such information if a location services for the mobile application is enabled and b the permissions in the mobile device allow communication of this information physical location or movements if you have previously allowed us access to your geolocation data you can opt out of making this information available to us by visiting your mobile devices settings for the relevant application or the settings page for the relevant game geolocation information when you use a smartphone or other mobile device to access our services with your permission we may collect your geolocation information to optimize user experience such as for localization accuracy language display or the provision of relevant advertising the type of information that we may collect includes your first and last name email address geolocation password or other identifying information you choose to provide the terms personal information or personal data as used in this privacy policy shall mean any information that enables us to identify you directly or indirectly by reference to an identifier such as your name identification number location data online identifier or one or more factors specific to you.	Same as SENSOR and PHONE_CALL	0.91	0.85	0.73
Dangerous Android Permission **STORAGE** Category				
any personally identifiable information provided by you will not be considered as sensitive if it is freely available and or accessible in the public domain like any comments messages blogs scribbles available on social platforms like facebook twitter etc any posted, uploaded, conveyed, gcommunicated by users on the public sections of the sites becomes published content and is not considered personally identifiable information subject to this policy	**Article 48 GDPR - Transfers or disclosures not authorised by Union law** Any judgment of a court or tribunal and any decision of an administrative authority of a third country requiring a controller or processor to transfer or disclose personal data may only be recognised or enforceable in any manner if based on an international agreement, such as a mutual legal assistance treaty, in force between the requesting third country and the Union or a Member State, without prejudice to other grounds for transfer pursuant to this Chapter. **Suitable Recitals: (115) Rules in Third Countries Contrary to the Regulation**	0.87	0.80	0.69
Dangerous Android Permission **PERSISTENTID** Category				
we also collect other kinds of information from you or other sources which we refer to as “other information” in this policy which may include but is not limited to device identification “id” which is a distinctive number associated with a smartphone or similar handheld device but is different than a hardware serial number internet connection means such as internet service provider “isp” mobile operator wifi connection service set identifier “ssid” international mobile subscriber identity “imsi” and international mobile equipment identity "imei" other information including but not limited to the IP address the isp college or organization that operates the network you test and network hardware and device identifiers such as your ssid or imei if the test is conducted on a smartphone	Same as LOCATION, SENSOR and PHONE_CALL	0.89	0.83	0.73

TABLE 6. Examples comparing how different sentences in the same GDPR law have dissimilar compliance for similar parts of the law.

Sentence BERT (Cosine Similarity = CS)				
GDPR Law Title	Sentence in GDPR Law	Highest Identified Dangerous Declaration	Highest CS	Average CS
Dangerous Android Permission STORAGE Category				
Conditions for Consent	The data subject shall have the right to withdraw his or her consent at any time.	ii accessing information stored on your device relating to your use of and engagement with websites and apps eg adobe connect meetings and crash reports and iii analyzing your content you can withdraw your consent to such activities at any time accessing information stored on your device relating to your use of and engagement with websites and apps eg adobe connect meetings and crash reports accessing information stored on your device which you allow us to receive through devicebased settings eg photos location and camera in order to provide certain functionality within our apps and websites and analyzing your content using techniques such as machine learning in order to improve our services and the user experience on other occasions where we ask you for consent we will use the information for the purposes which we explain at that time	0.62	0.29
Conditions for Consent	The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal.	Same DAPD as above.	0.41	0.08
Dangerous Android Permission MICROPHONE Category				
Right to object	Where the data subject objects to processing for direct marketing purposes, the personal data shall no longer be processed for such purposes.	personal information we sold or disclosed for a business purpose in the preceding number months we have sold or disclosed to one or more third parties the following categories of personal information that identifies relates to describes is capable of being associated with or could reasonably be linked directly or indirectly with a particular consumer identifiers such as a real name alias postal address unique personal identifier such as a device identifier cookies beacons pixel tags mobile ad identifiers and similar technology customer number user alias other forms of persistent or probabilistic identifiers online identifier internet protocol address email address account name and other similar identifiers characteristics of protected classifications under california or federal law such as age and gender online activity internet and other electronic network activity information including but not limited to browsing history search history and information regarding your interaction with websites applications or advertisements geolocation data sensory information audio electronic visual and similar information your preferences characteristics predispositions behavior and attitudes	0.66	0.31
Right to object	we will also request access to your photos media and files your devices camera and microphone your wifi connection information and your device id and call information	The controller shall no longer process the personal data unless the controller demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims.	0.48	0.27

the associated sentences. The first sentence in the 'Right to Object' GDPR law has an associated detailed and in-depth DAPD methodology that derives a cosine similarity score of 0.66. The average derived DAPD methodology for this sentence is only 0.31. On the other hand, the second sentence in the 'Right to Object' GDPR law in Table 6 has a less in-depth highest scoring DAPD methodology which derives a cosine similarity score of 0.48 while the average derived cosine similarity score is 0.27. This example shows that not all methodologies are equal in quality and depth. This also supports the fact that methodologies to declare DAPD may not have enough variation to comply with all aspects of GDPR laws. Again the results reinforce the increase in GDPR compliance by using the identified highest-scoring cosine similarity DAPD methodologies.

Table 7 shows the identification of the most frequent N-Grams found in the highest complying DAPD and corresponding GDPR law for each permission category. These N-grams are compared to derive a GloVe vectorization and BERT word embedding result to determine the contextual and global vectorization similarity between the N-grams. The dangerous permission categories CAMERA, PHONE_CALL, SMS and LOCATION give exact or close to similar results for both the BERT and GloVe results. For the other categories, BERT's contextualized and sensitive approach is more clear. For MICROPHONE, GloVe gives a result of -0.07 meaning the results are dissimilar while BERT gives a value of 0.39. This expresses the sensitivity of BERT to at least find some type of connection or context. A result of 0.39 is nearly the equivalent of a dissimilar GloVe result but as the pre-trained data is so large and BERT is far more sensitive than GloVe then the results tend to be inflated. Interestingly, no correlation can be found between the DAPD and GDPR laws found in Table 5 and the results derived from the N-Grams using the same DAPD and GDPR laws in Table 7. Considering the synonymy threshold of 0.8025 which was proposed in [69] as a threshold value to correlate a semantic meaning between bi-grams, only the dangerous permission category LOCATION reaches this threshold between the bi-grams identified in the DAPD and the associated GDPR law. All the other categories fail to meet this threshold, some of these categories have observable similarities such as the dangerous permission category CAMERA with the Android N-gram ('personal', 'data') and the GDPR N-Gram ('data', 'subject') with the respective cosine similarity results of 0.81 for GloVe and 0.75 for BERT. This could indicate that the entire context and syntactic structure of the DAPD may

TABLE 7. The most common N-Grams found in the GDPR and DAPD identified in table 5.

N-Gram Comparison With Contextualised Words						
Dangerous Permission Category	GDPR Law Title	Most Frequent Android Permission Policy N-Gram	Most Frequent GDPR Text N-Gram	GloVe	BERT	Table 5 BERT
CAMERA	Right to rectification	('personal', 'data')	('data', 'subject')	0.81	0.75	0.89
MICROPHONE	General principle for transfers	('audio', 'andor')	('third', 'country')	-0.07	0.39	0.86
PHONE_CALL	Definitions	('phone', 'number')	('personal', 'data')	0.52	0.47	0.83
SENSOR	Definitions	('background', 'activity')	('personal', 'data')	0.47	0.45	0.80
SMS	Right to Rectification	('also', 'collect')	('data', 'subject')	0.53	0.53	0.83
CALENDAR	Cooperation with the supervisory authority	('mobile', 'phone')	('controller', 'processor')	0.24	0.59	0.81
CONTACTS	Right to Rectification	('social', 'medium')	('data', 'subject')	0.44	0.66	0.90
LOCATION	Definitions	('personal', 'data')	('personal', 'data')	1	1	0.91
STORAGE	Transfers or disclosures not authorised by Union law	('collect', 'personal')	('third', 'country')	0.34	0.47	0.87
PERSISTENTID	Definitions	('device', 'identifier')	('personal', 'data')	0.37	0.61	0.89

be the reason for increasing compliance rather than using similar words and that the context is more important than the similarity of the text.

Based on the analysis using several algorithms, the textual similarity dimension with the highest similarity results was found to be the BERT word embedding implementation with the most accurate variant being the pre-processed implementation. Thus, a more in-depth analysis to compare GDPR and the associated text in the identified highest DAPD was conducted. The use of SBERT directly identifies where compliance is failing between each GDPR law and DAPD.

1) DEVELOPER PERSPECTIVE

Do the sentences used for declaring DAPD in Android created by developers map with relevant and meaningful information in the GDPR articles? Using the highest identified policy methodologies in Table 5, each DAPD and its corresponding GDPR law will be analyzed to investigate whether the mappings are meaningful. In some dangerous permissions, the mappings between the permission policy declaration and corresponding GDPR articles are meaningful. Permission categories with meaningful mappings are `PHONE_CALL`, `SENSOR`, `LOCATION` and `PERSISTENT_ID`. The similarity amongst these permissions is that they are all mapped with **Article 4 GDPR - Definitions**. This suggests that these section of the GDPR law implicitly or explicitly refers to the permission categories, sensitive Android APIs, sensitive data requested, actions the permissions represent and the semantic meaning of the permissions. For the `SENSOR` permission category, there are sections in **Article 4** that focus on *genetic, biometric data* and *data concerning health* (**Article 4(13), Article 4(14), Article 4(15)**). For `LOCATION` permission, *location data* is mentioned as part of *personal* and *profiling data* in sections of **Article 4 (1)** and **Article 4(4). Article 4(1)** uses *online identifier* as an example of personal data, which matches well with the permission-policy declaration for `PERSISTENT_ID`. Further, **Article 4** is linked with **Recital 30 - Online Identifiers for Profiling and Identification** which explicitly mentions *"online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags."*,[4] which matches with the sensitive data `PERSISTENT_ID` provides and the `PHONE_CALL` permission policy description.

There are other cases where contextually, both pieces of data from the permission policy description and the GDPR match but they do not in any way represent the sensitive data requested by the permissions or the actions they represented. Permissions in this category are `CAMERA`, `MICROPHONE` and `STORAGE`. Similarly, `CALENDAR`, `CONTACTS` and `SMS` permission policy descriptions and the corresponding match GDPR article lack contextual similarity and do not represent the actions of the permissions. *These findings imply that the use of SBERT for matching permission policy declaration with GDPR articles using cosine similarity shows that completeness checking of dangerous Android permission policy declaration against GDPR is achievable and can be automated.*

Why do a large number of DAPD lack GDPR Compliance? While the context among GDPR laws and the highest identified DAPD are similar, some GDPR laws have different aims. From a developer perspective, the issue regarding better compliance among other methodologies for declaring dangerous android permission policy may relate to a lack of context and difficulty targeting certain GDPR laws because there are no appropriate recitals and articles that accurately capture the permission category. As demonstrated in [13], another possible reason for the lack of compliance could be that these permissions, such as `SMS`, `CALENDAR` and `CONTACTS` are difficult to explicitly or implicitly declare in privacy policies even though they have been declared in the app manifest file.

[4]https://gdpr-info.eu/recitals/no-30/

2) PLATFORM PERSPECTIVE

Is the range of permission categories used in the Android ecosystem sufficient? It is difficult for a developer to comply with every GDPR law based on the limited range of dangerous permission categories. For example, it might be difficult for a developer to comply with articles on **'Territorial Scope'** while declaring the usage of the dangerous permission category which is more focused on complying with articles such as **'Conditions for Consent'**? Compliance with specific GDPR laws is more crucial than others. For instance, **Article 4 GDPR - Definitions** is very important as it defines different kinds of personal sensitive data protected under GDPR. This suggests that the Android ecosystem can develop permissions around these sensitive data that apply to mobile applications, and ensure that the permission-policy description aligns with the provisions in GDPR for the collection, transparency and processing requirements. On the other hand, increasing the number of dangerous permission categories may complicate and confuse the process of declaring compliant DAPD permissions for developers. However, the advantage of expanding the number of permission categories for compliance is that certain categories can be used to target crucial GDPR laws. As proven by the results in Table 5, carefully constructed DAPD can contextually comply with GDPR laws. One solution is that Google creates more dangerous permission categories based upon selections of GDPR laws thus allowing developers to target sections of GDPR. Since permissions on Android aim to support user's privacy by also protecting access to restricted actions and not just restricted data [3], the definition of the restricted actions can be influenced by **Chapter 2 (Art.5-11) Processing** and **Chapter 3 (Art.12-23) Rights of Data Subject** to create the required dangerous permissions. Determining which articles and recitals should be targeted could lead to other compliance issues and misinterpretation. As they neatly map to important categories of personal data in the GDPR, as shown in Table 8 where Y stands for Yes and N for No, we believe that the app permission categories supported by the Android ecosystem are sufficient. The metadata from `Storage` can be obtained to elicit `Location` information. Similarly, since `Storage` is also intended for storing any kind of media including images, then `Biometric data` can also map with storage. However, the recommendations on focusing on particular articles and recitals relevant to the app ecosystem could be implemented to make it easier for developers to comply with GDPR.

Is it that the language used for GDPR laws is not explicit enough? From Table 5 the larger articles that cover more scope tend to have higher compliance results. **Article 4 GDPR Definitions** is of one the articles with the most depth and scope. The average cosine similarity compliance result among every permission category for the highest identified corresponding DAPD is 0.84. This may indicate that GDPR laws that are less explicit and have a larger scope may make it easier for a developer to comply with the GDPR law.

Do the declarations that are used for DAPD need to be longer and more detailed? The results from Table 5 which derives the highest cosine similarity compliance results with the corresponding GDPR laws and Table 7 which details the cosine similarity between N-Grams indicate that the best method to create DAPD methodologies is to structure the methodologies in a similar syntactic and contextual structure rather than using the same words. As per the results from Table 6, the methodologies to declare DAPD do not comply well with all sections of the associated GDPR law. This may indicate that to comply with a high standard, the highest identified DAPD for each sentence may need to be used and conjoined into a longer more detailed declaration. The contribution from the results of the BERT sentence embedding techniques enables this to happen thus each sentence for each GDPR law can have a DAPD with the highest identified compliance. An argument can also be made about mismatches in explicit declarations used in GDPR and the terms used in DAPD policies. For example, Table 5 shows that although contextually the DAPD and associated GDPR law are consistent, the actual aim of the GDPR law is usually completely different. For example, the DAPD in the `CAMERA` dangerous permission category complies best with the GDPR law **'Right to rectification'**. Contextually both the permission and the GDPR law draw similarities but the aim of the GDPR law is different.

B. RQ2: PERMISSION-POLICY GENERATION AT DESIGN TIME WITH UML DIAGRAMS

To answer, RQ2, a proof of concept for class relationship UML design time compliance tool for automatic dangerous android permission policy generation was developed. This approach is focused on developers that use UML and is implemented using the results derived from the BERT word embedding since it generated the best results for completeness checking (cf IV-A). The sample class relationship UML image used for the Tesseract OCR Engine[5] text extraction component of the design time tool is sourced from a section of a large UML diagram that was used for an actual mobile application. For the XML data input, the entire XML data of the UML diagram used for the image snippet was used as data input. The rationale for using UML diagrams relates to a design time-oriented approach in which GDPR-compliant DAPD can be generated using information during the development process rather than at the end of a development life cycle. Such a method creates a new approach to developing GDPR-compliant applications. This approach saves a developer time, reduces 8error from a developer who isn't knowledgeable about data protection laws, reduces the likelihood of GDPR DAPD methodologies that are not compliant from being created for the privacy policy, removes costly legal fees associated with privacy policy creation, equips a developer with a tool to streamline DAPD and reveals increased transparency in the compliant

[5]https://github.com/tesseract-ocr/tesseract

TABLE 8. Mapping between article 4 GDPR - definitions and permissions category.

	Permission Category					
Article 4 GDPR Definitions	**CAMERA**	**LOCATION**	**MICROPHONE**	**PERSISTENTID**	**SENSOR**	**STORAGE**
Biometric Data	Y	N	Y	N	Y	Y
Location	N	Y	N	N	Y	Y
Genetic Data	Y	N	N	N	Y	N
Online Identifier	N	N	N	Y	N	N
Processing	Y	N	Y	N	N	Y

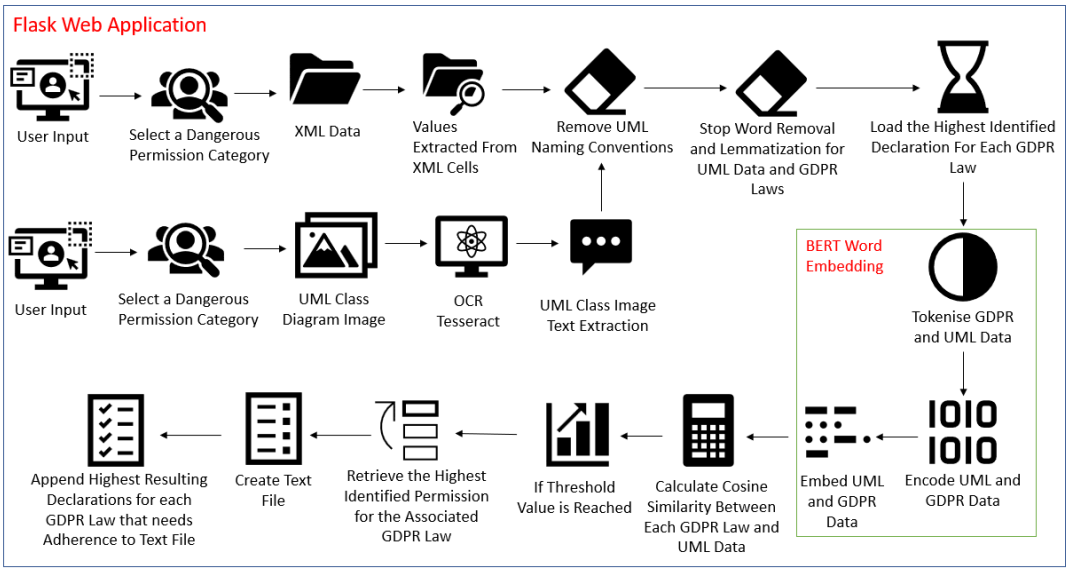


FIGURE 2. Proof of concept design time class relationship DAPD generator.

methodologies a developer should use to comply to each GDPR law.

Figure 2 describes the framework for the tool where a user (developer) can either upload the UML diagram as an image or XML file and the user is prompted to select a permission category. If the input is an image, the Tesseract OCR engine is used to detect all the words in the image, while all the values associated with the text in the XML would be extracted if the input was XML data. Regardless of the type of input, the extracted text would have developer naming conventions such as camel casing and underscores removed. The result of removing naming conventions leads to the separation of individual words from the original words. Preprocessing for both the input and every GDPR law takes place in which stop word removal and lemmatization is applied and the highest identified DAPD for each GDPR law is loaded. BERT word embedding using the `'bert-base-uncased'` pre-trained model is then used in which every GDPR law and the UML extracted data is tokenized, encoded, and embedded. Cosine similarity is then used to calculate the syntactic, semantic, and contextual similarity between the UML data and every GDPR law with a defined threshold value indicating the need to retrieve the associated DAPD derived from the highest identified DAPD found for every permission category. The retrieved DAPD is then saved to a text file which a developer can use in a privacy policy. With the use of the Tesseract OCR engine implementation, the solution is not limited to class relationship diagrams but other types of data falling into structural and behavioural diagrams could be used individually or as a collection to generate GDPR-compliant DAPD.

Table 9 demonstrates the extraction of UML data from an image is transformed into generated permissions when compared to laws that reach a user-defined contextual similarity threshold. All three laws identified generate the permission most contextually similar based on the inputted UML data. The threshold value used in such an example was 0.09 meaning each law would need an approximate contextual similarity of around 10% to automatically generate the DAPD. Another use of the tool is that once developers upload their UML, the tool automatically scans the UML against the dangerous android permission categories, and produces as an output, the dangerous android permissions that the application requires based on the UML and also generates an optimal permission policy description based on the MPP-270 corpus that complies with GDPR. Based on these results, developers can target specific articles of GDPR of interest for compliance, and also specify specific

TABLE 9. Demonstration of how UML data can be identified with GDPR laws using BERT to generate permission declarations (threshold value = 0.09).

Sample UML Data Input from Image [23]	GDPRD Laws that Reach Threshold When Matched	Generated Permission Policy
LoginViewActivity, LoginInController, FirebaseController, LoginAccount, ForgottenPasswordViewActivity, loginAccountEmailInput, getCurrentUserID,...	Monitoring of approved codes of conduct Competence of the lead supervisory authority Mutual assistance... Art.41 GDPR, Art.56, Art.61	profile information that you provide for your user profile eg id instagram id profile image information we collect we may collect and use the following personal information that identifies relates to describes is reasonably capable of being associated with...

TABLE 10. Dangerous permission occurrence in matched GDPR articles.

Dangerous Permission Category	Permission Category	Use of Sensitive APIs	Sensitive Data Requested	Actions Permissions Represent	Semantic Meaning
CAMERA	NM	NM	IM	IM	EM
MICROPHONE	NM	NM	EM	EM	EM
PHONE_CALL	NM	IM	IM	IM	IM
SENSOR	EM	IM	EM	EM	EM
SMS	NM	NM	NM	IM	IM
CALENDAR	NM	NM	NM	NM	NM
CONTACTS	NM	NM	IM	IM	IM
LOCATION	EM	IM	EM	EM	EM
STORAGE	IM	IM	IM	IM	IM
PERSISTENTID	IM	IM	EM	IM	EM

TABLE 11. Meaning of personal data.

Google	GDPR
- Personally identifiable information - Financial and payment information - Authentication information - Phonebook, Contacts, Device Location - SMS and call-related data - Health data, Health Connect data - Inventory of other apps on the device - Microphone, Camera - Other sensitive device or usage data	- Personally Identifiable Information - Identifiers (name, identification number, location data, online identifier) - Special categories (racial or ethnic origin, religious or philosophical beliefs, trade union membership, sex life & sexual orientation) - Pseudononymic data, Biometric data - Health data, Genetic Data - Person-related factors (physical, physiological, genetic, mental, economic, cultural or social identity

thresholds based on business needs of their requirements engineering. With the proof of concept in Figure 2 and the results in Table 9, it is possible to automate privacy policy generation for dangerous android permissions from UML diagrams. Originally a classification model was planned but the NLP approach was more suitable due to the small MPP-270 corpus dataset.

C. RQ3: SUITABILITY OF GDPR FOR COMPLETENESS CHECKING OF ANDROID APPLICATION PERMISSION POLICIES

We have demonstrated the possibility of inferring some dangerous android permissions from GDPR articles such as `PHONE_CALL`, `PERSISTENT_ID`, `SENSOR`, `LOCATION` by measuring compliance derived from textual similarity algorithms. With the results from runtime analysis (cf Section IV-A), the highest compliant DAPD was found to increase DAPD compliance to GDPR by 12% using BERT. With design time (cf Section IV-B), GDPR laws can be matched with text extracted from images or XML input of UML diagrams. We also combined the framework for the runtime and design time analysis to design an automated tool that generates GDPR-compliant permission-policy snippets for permission requirements inferred from the UML information.

To further corroborate the results from Table 5, we investigated whether the permission categories, sensitive Android API usage, the sensitive data they request, the actions these permissions represent, or their semantic meaning are implicitly or explicitly declared in the matched GDPR. As shown in Table 10, we denoted the result as *NM - Not Mentioned, IM - Implicitly mentioned* and *EM - Explicitly Mentioned*, which shows that some of the permissions in the Android ecosystem can be inferred and categorised from GDPR articles and recitals.

We argue that the GDPR is adequate for sensitive Android permission declaration completeness, as it includes Android permission policy or implicitly describe sensitive user data collection and processing. In some permission categories, the permission policy snippets can be matched explicitly with GDPR articles and recitals, and in some scenarios, the permission category is only implicitly covered in the GDPR. There are some reasons for the implicit matching in some permission categories. Firstly, some permission-relevant information from the MPP-270 could have explicitly matched GDPR articles and recitals if they contained relevant information about sensitive data collection. For example, **Article 4 GDPR - Definitions** provides permission policy information for biometric data which includes facial images and dactyloscopic data, which should have been directly mapped to the `CAMERA` permission category. However, due to the quality of the information provided in the MPP-270 for permission category, the policy information matched with **Article 16 GDPR - Right to Rectification**, which does not describe the permission or the actions it represents (cf Table 5).

Another reason for some of the *Not Mentioned* or *Implicitly Mentioned* cases in Table 10 is the language used by GDPR and Google in defining personal and sensitive user data. Table 11 shows the definition of personal and sensitive user data by Google[6] and GDPR (cf **Art. 4 GDPR - Definitions, Art.9 GDPR - Processing of special categories of personal data**). Voice is considered as personal data under GDPR, because it is information relating to an identified or identifiable natural person, and in some cases, voice recordings may constitute biometric information under GDPR. While Google uses clear and direct data for voice data defining the microphone as sensitive user data, it is lumped under PII or biometric data under GDPR, which is

[6]https://support.google.com/googleplay/android-developer/answer/13316080?sjid=18484364633345142 24-EU

ambiguous and generic. As a result, developers might find it easier to write permission policy snippets using languages that comply with Google Play Developer Programme Policies than GDPR. Another case in point is the CALENDAR permission category which allows an app to read, share and save a user's calendar data. This also falls under personal data because it is personal information stored on the user's contact card and it could contain PII, however, this permission and the action it represents is not explicitly covered in the GDPR. Table 11 and Table 10 further shows that all the categories of personal and sensitive user data are covered in the GDPR, hence, we can adequately conduct compliance by matching GDRP articles with Android permissions categories, APIs and permission-policy declarations.

To further argue that GDPR is suitable for completeness and compliance checking using Android permissions, we align Google Privacy and Terms of Service (ToS) with GDPR to investigate the similarities between sections in Google Privacy & Terms matches with articles in the GDPR. Since the Android operating system which supports app permissions investigated in this study is a platform owned by Google, investigating the completeness of Google Privacy & Terms against GDPR will provide additional insights into the suitability of GDPR. There are 16 and 11 sections respectively in the Google Privacy Policy and ToS respectively. To achieve this goal, BERT embeddings were used to match the different sections in the Google terms of service and privacy policies to GDPR. The closest matching GDPR articles are then identified using the articles with the highest cosine similarity. Table 12 shows the results of the analysis of Google TOS and privacy policies against GDPR. An interesting insight is that contextually, the different sections of Google TOS and privacy policies all match with articles in the GDPR with an average cosine similarity value of *0.83*, except the section on *Updates* in the ToS that matches with *Repeal of Directive 95/46/EC*. This may also prove how contextually structuring a permission-policy declaration would yield higher results as Google has contextually structured the majority of their TOS and privacy policies towards the GDPR article 4 Definitions, which is a key section of the GDPR that discusses the general provision of the regulation. We can therefore conclude that the GDPR is suitable for performing completeness for Google Privacy Policy and Terms of Service, which can be cascaded down to the Google platform such as the Android operating system that supports app permissions.

V. LIMITATION AND FUTURE WORK

One of the limitations of the research is the examined annotated policy corpus. The systematic mapping between the app privacy policy and Android permissions was done by manually annotating 270 Android application privacy policies. The apps were selected based on popularity measured by number of downloads and user ratings. Firstly, there are currently over 2.65 million apps and games in the Google Play Store[7] and 270 apps are not a representative of the app distribution. Additionally, apps (including games) on the Google Play store fall into 49 categories,[8] however, the top 270 apps used in the corpus only covered 13 app categories. The annotated policy corpus for mapping between permission and privacy considered 30 dangerous permission APIs, however, there are 42 dangerous permission APIs on the official Android API documentation.[9] This means that the coverage of the permission-policy snippet analysis for compliance was not investigated for some permission groups that are not part of the 10 considered dangerous permission categories or permissions added in newer versions of API releases. Finally, paid apps were not part of the selected apps for policy annotation. The implication of these selection biases is that permission policy behaviours might vary between apps and games, popular and non-popular apps, paid and free apps, and evaluated app categories vs non-evaluated app categories. The transparency of app privacy policies used in creating the gold standard dataset could be biased towards the selection criteria which are not a true representation of the app market. However, we argue that this limitation does not affect the findings in this research we focused on investigating the suitability of GDPR for completeness checking of permission-policy declaration. Since the corpus depends on human annotators to find permission-policy snippets in the app privacy policy for declared permission in the permission manifest file, this means that the corpus is highly subjective in interpreting privacy policies for permission transparency. This is due to the nature of privacy policy being ambiguous and subject to multiple interpretations, even among privacy and legal experts [78], [79], [80].

The semantic relationship of textual description bi-grams using GloVe, word2vec and Fasttext were investigated in [70], and the results revealed inaccuracies in the way each algorithm matches semantic and context-driven disambiguation between entities. Other findings suggested that word embedding techniques struggled in cases to produce an accurate result for words depicting similar meanings, which reveals the limitation of the technique to understand the same context that a human would interpret in certain bi-grams. Such an issue may have affected the performance of the GDPR completeness checking approach for dangerous android permission policy declaration as the word embedding techniques used may have at some point misinterpreted the semantic relationship with other words in the sentence transformer techniques or the N-Gram driven experiments. As the UML design time compliance tool is a proof of concept idea, the tool focuses solely on developers that use class relationship diagrams with UML in the software development cycle. This could alienate a proportion of developers who do not use UML during development or

[7]https://www.businessofapps.com/data/app-stores/
[8]https://support.google.com/googleplay/android-developer/answer/9859673?sjid=2389976692120545916-EU
[9]https://developer.android.com/reference/android/Manifest.permission

TABLE 12. Completeness checking of google privacy policy and terms of service against GDPR.

Google Privacy Policy - https://policies.google.com/privacy?hl=en-UK			
Sections	**GDPR Article**	**GDPR Category**	**Score**
Introduction	Article 4 - Definitions	General Provisions	0.78
Information that Google Collects	Article 4 - Definitions	General Provisions	0.76
Why Google collects data	Article 4 - Definitions	General Provisions	0.79
Your privacy controls	Article 4 - Definitions	General Provisions	0.75
Sharing your information	Article 4 - Definitions	General Provisions	0.87
Keeping your information secure	Article 4 - Definitions	General Provisions	0.84
Exporting & deleting your information	Article 4 - Definitions	General Provisions	0.78
Retaining your information	Article 33 - Notification of a personal data breach to the supervisory authority	Security of Personal Data	0.79
Compliance & cooperation with regulators	Article 44 - General principle for transfers	Transfers of personal data to third countries of international organisations	0.89
European requirements	Article 9 - Processing of special categories of personal data	Principles	0.86
About this policy	Article 7 - Conditions for consent	Principles	0.87
Related privacy practices	Article 4 - Definitions	General Provisions	0.82
Data transfer frameworks	Article 4 - Definitions	General Provisions	0.84
Key terms	Article 42 - Certification	Controller and Processor	0.80
Partners	Article 4 - Definitions	General Provisions	0.82
Updates	Article 89 - Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes	Provisions relating to specific processing situations	0.70
Google Terms of Service - https://policies.google.com/terms?hl=en-UK			
Introduction	Article 4 - Definitions	General Provisions	0.88
Your relationship with Google	Article 4 - Definitions	General Provisions	0.85
Using Google services	Article 4 - Definitions	General Provisions	0.88
Content in Google services	Article 4 - Definitions	General Provisions	0.88
Software in Google services	Article 4 - Definitions	General Provisions	0.84
In case of problems or disagreements	Article 4 - Definitions	General Provisions	0.90
About these terms	Article 4 - Definitions	General Provisions	0.85
Updates	Article 94 - Repeal of Directive 95/46/EC	Final Provisions	0.83
Definitions	Article 9 - Processing of special categories of personal data	Principles	0.89
List of services and service-specific additional terms	Article 7 - Conditions for consent	Principles	0.75
How Google handles government requests for user information	Article 4 - Definitions	General Provisions	0.84

have a UML class relationship diagram. For permissions requirement engineering, different sources of design time elements beyond UML, such as UI textual descriptions can be leveraged. UI textual descriptions have been employed in [70] for the semantical resolution of permission request patterns in Android apps. The texts may also describe access to restricted data or sensitive action. For example, a UI text field can have a description like *"Upload supporting files", "Take a photo" "Start recording an audio message"*, which are all accessing private user data or sensitive actions such as *STORAGE, CAMERA* and *MICROPHONE* protected by permissions. Regardless of the source of the design time information, whether they are UML diagrams or UI textual descriptions, we have demonstrated the relevance of our approach in automated permission policy generation. We have shown the utility of our method in automated permission policy generation, regardless of the design time element, whether they are UML diagrams or UI textual descriptions.

The solution could also be extended to other permission-declaring files such as iOS applications, browser extensions etc. A similar analysis for permission compliance could be investigated for other GDPR-like laws such as the California Consumer Privacy Act (CCPA), and Payment Card Industry Data Security Standard (PCI DSS). Concerning GDPR, an expanded empirical analysis could be conducted by implementing more textual representations thus expanding the scope past textual dimensions. The measurement of textual similarity was mapped in [81] with textual distance and representation highlighting the many combinations that can be used both textually and numerically to derive results for an enhanced conclusion regarding DAPD-GDPR compliance. The development of an application generating compliant and contextualized DAPD using machine learning based on the information from the UML could be investigated as this approach would require large amounts of data of DAPD to generate a compliant level of contextualized declarations which are unique for each UML application. The UML design time tool concept could be extended through the incorporation of a browser extension plug-in in which the tool scans the DAPD in a privacy policy and detects inadequate DAPD. Developers of applications could then be alerted if such declarations fall below a compliance threshold. This idea could be deployed by the Google Play Store as part of the approval process for users uploading applications in which the privacy policy has to have a compliant DAPD. This would require a substantial amount of training data which is not yet available. The UML design time tool concept could be expanded to include other structural and

behavioural UML diagram components such as flowcharts, entity relationship databases, and sequence and activity diagrams. Collecting information in these diagrams could assist in creating a more targeted and compliant DAPD. Another future direction is investigating other pre-trained models for language understanding such as MPNet [82], which combines masked and permuted language modelling.

VI. CONCLUSION

This paper investigates runtime and design time GDPR completeness checking using dangerous Android permissions. For runtime analysis, completeness checking was done by representing the permission policy declaration for each permission category requested in the app privacy policy. For design time analysis, UML class diagrams were utilized to extract permission requirements from the class elements and generate a permission-policy declaration that is GDPR-compliant. Through the results, we demonstrate the most compliant permission policy declarations for each permission category. As previously highlighted, developers lack the legal knowledge to develop compliant permission policy declarations. This paper contributes to the state-of-the-art by developing a tool to equip developers with apparatus to automatically generate compliant DAPD methodologies to GDPR and avoid non-compliant DAPD, this uses design time requirements without developer legal knowledge. We also demonstrated that the completeness of permission policy with GDPR articles could be substantially improved by applying a similar contextual structure to a targeted GDPR law rather than allocating the exact words in the DAPD. Other state-of-the-art solutions focus on generating requirements or taking already created privacy policies for textual analysis. This project combines NLP with semantic similarity to automatically generate compliant DAPD based on requirements using UML class diagrams.

One area of future work we are keen on exploring is the usability analysis of the proposed UML tool. Since the goal is to help actual developers with privacy policy generation and requirements elicitation with GDPR-compliant permission declaration using UML diagrams, a usability evaluation would help in measuring the extent to which learning and using that tool to achieve compliance goals, especially with their permission declaring systems such as browser extensions, mobile apps etc. The user's satisfaction with the usability evaluation process will serve as feedback into the tool development process to improve its effectiveness, efficiency, flexibility and robustness. This usable study contributes to building compliance tools that are developer-friendly and developer-centric. Another area of future work involves creating a larger benchmark annotated policy corpus for permission completeness. In this study, we leveraged MPP-270 which creates a mapping between permission requested (declared in the app manifest file) and permission-relevant information in the app privacy policies, created by manually annotating 270 Android application policies. With a large annotated corpus, a classification model built on machine learning algorithms could be integrated into our solution.

REFERENCES

[1] GDPR. (2018). *General Data Protection Regulation (GDPR)*. Accessed: May 2, 2023. [Online]. Available: https://gdpr-info.eu/

[2] Itgovernance. (2018). *GDPR Penalties and Fines—What's the Maximum Fine in 2023?* Accessed: May 2, 2023. [Online]. Available: https://www.itgovernance.co.uk/dpa-and-gdpr-penalties

[3] *Permissions on Android*, Android Developers, 2023. Accessed: May 15, 2023. [Online]. Available: https://developer.android.com/guide/topics/permissions/overview

[4] *Android Developers*, Permission Element, 2023. Accessed: May 20, 2023. [Online]. Available: https://developer.android.com/guide/topics/manifest/permission-element

[5] Sara Pegarella. (2023). *Android Permissions that Need a Privacy Policy—Termsfeed*. Accessed: May 23, 2023. [Online]. Available: https://www.termsfeed.com/blog/android-permissions-privacy-policy/

[6] A. Senarath and N. A. G. Arachchilage, "Why developers cannot embed privacy into software systems: An empirical investigation," in *Proc. 22nd Int. Conf. Eval. Assessment Softw. Eng.*, Jun. 2018, pp. 211–216.

[7] A. Alhazmi and N. A. G. Arachchilage, "Why are developers struggling to put GDPR into practice when developing privacy-preserving software systems?" in *Proc. Symp. Usable Privacy Secur. (SOUPS)*, Aug. 2020, p. 1.

[8] A. Alhazmi and N. A. G. Arachchilage, "I'm all ears! Listening to software developers on putting GDPR principles into software development practice," *Pers. Ubiquitous Comput.*, vol. 25, no. 5, pp. 879–892, Oct. 2021.

[9] A. Senarath, M. Grobler, and N. A. G. Arachchilage, "Will they use it or not? Investigating software developers' intention to follow privacy engineering methodologies," *ACM Trans. Privacy Secur.*, vol. 22, no. 4, pp. 1–30, Nov. 2019.

[10] M. Tahaei, A. Frik, and K. Vaniea, "Privacy champions in software teams: Understanding their motivations, strategies, and challenges," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2021, pp. 1–15.

[11] A. Alhazmi and N. A. G. Arachchilage, "A serious game design framework for software developers to put GDPR into practice," in *Proc. 16th Int. Conf. Availability, Rel. Secur.*, Aug. 2021, pp. 1–6.

[12] I. Omoronyia, U. Etuk, and P. Inglis, "A privacy awareness system for software design," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 29, no. 10, pp. 1557–1604, Oct. 2019.

[13] M. S. Rahman, P. Naghavi, B. Kojusner, S. Afroz, B. Williams, S. Rampazzi, and V. Bindschaedler, "PermPress: Machine learning-based pipeline to evaluate permissions in app privacy policies," *IEEE Access*, vol. 10, pp. 89248–89269, 2022.

[14] MPP-270. (2022). *Is the App Policy Permission-Complete? MPP-270: Annotated Policy Corpus for Mapping Between Permission and Privacy*. Accessed: May 2, 2023. [Online]. Available: https://sites.google.com/view/permpress/download-mpp-270-corpus?pli=1

[15] S. Zimmeck, Z. Wang, L. Zou, R. Iyengar, B. Liu, F. Schaub, S. Wilson, N. Sadeh, S. Bellovin, and J. Reidenberg, "Automated analysis of privacy requirements for mobile apps," in *Proc. AAAI Fall Symp. Ser.*, 2016, pp. 286–296.

[16] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. Reidenberg, N. C. Russell, and N. Sadeh, "MAPS: Scaling privacy compliance analysis to a million apps," *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 3, pp. 66–86, Jul. 2019.

[17] L. Yu, X. Luo, J. Chen, H. Zhou, T. Zhang, H. Chang, and H. K. N. Leung, "PPChecker: Towards accessing the trustworthiness of Android Apps' privacy policies," *IEEE Trans. Softw. Eng.*, vol. 47, no. 2, pp. 221–242, Feb. 2021.

[18] R. Baalous and R. Poet, "Utilizing sentence embedding for dangerous permissions detection in Android Apps' privacy policies," *Int. J. Inf. Secur. Privacy*, vol. 15, no. 1, pp. 173–189, Jan. 2021.

[19] R. Slavin, X. Wang, M. B. Hosseini, J. Hester, R. Krishnan, J. Bhatia, T. D. Breaux, and J. Niu, "Toward a framework for detecting privacy policy violations in Android application code," in *Proc. IEEE/ACM 38th Int. Conf. Softw. Eng. (ICSE)*, May 2016, pp. 25–36.

[20] A. Sunyaev, T. Dehling, P. L. Taylor, and K. D. Mandl, "Availability and quality of mobile health app privacy policies," *J. Amer. Med. Inform. Assoc.*, vol. 22, no. e1, pp. e28–e33, Apr. 2015.

[21] N. V. N. Kumar and R. K. Shyamasundar, "Realizing purpose-based privacy policies succinctly via information-flow labels," in *Proc. IEEE 4th Int. Conf. Big Data Cloud Comput.*, Dec. 2014, pp. 753–760.

[22] Á. Feal, P. Calciati, N. Vallina-Rodriguez, C. Troncoso, and A. Gorla, "Angel or devil? A privacy study of mobile parental control apps," *Proc. Privacy Enhancing Technol.*, vol. 2020, no. 2, pp. 314–335, Apr. 2020.

[23] M. Tahaei, K. Vaniea, and A. Rashid, "Embedding privacy into design through software developers: Challenges and solutions," *IEEE Secur. Privacy*, vol. 21, no. 1, pp. 49–57, Jan. 2023.

[24] M. Tahaei, J. Bernd, and A. Rashid, "Privacy, permissions, and the health app ecosystem: A stack overflow exploration," in *Proc. Eur. Symp. Usable Secur.*, Sep. 2022, pp. 117–130.

[25] M. Tahaei, T. Li, and K. Vaniea, "Understanding privacy-related advice on stack overflow," *Proc. Privacy Enhancing Technol.*, vol. 2022, no. 2, pp. 114–131, Apr. 2022.

[26] M. Tahaei, K. Vaniea, and N. Saphra, "Understanding privacy-related questions on stack overflow," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2020, pp. 1–14.

[27] M. Tahaei, R. Abu-Salma, and A. Rashid, "Stuck in the permissions with you: Developer & end-user perspectives on app permissions & their privacy ramifications," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2023, pp. 1–24.

[28] O. Amaral, S. Abualhaija, D. Torre, M. Sabetzadeh, and L. C. Briand, "AI-enabled automation for completeness checking of privacy policies," *IEEE Trans. Softw. Eng.*, vol. 48, no. 11, pp. 4647–4674, Nov. 2022.

[29] D. Torre, S. Abualhaija, M. Sabetzadeh, L. Briand, K. Baetens, P. Goes, and S. Forastier, "An AI-assisted approach for checking the completeness of privacy policies against GDPR," in *Proc. IEEE 28th Int. Requirements Eng. Conf. (RE)*, Aug. 2020, pp. 136–146.

[30] S. Abualhaija, C. Arora, A. Sleimi, and L. C. Briand, "Automated question answering for improved understanding of compliance requirements: A multi-document study," in *Proc. IEEE 30th Int. Requirements Eng. Conf. (RE)*, Aug. 2022, pp. 39–50.

[31] M. Lippi, P. Pałka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor, and P. Torroni, "CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service," *Artif. Intell. Law*, vol. 27, no. 2, pp. 117–139, Jun. 2019.

[32] D. Sánchez, A. Viejo, and M. Batet, "Automatic assessment of privacy policies under the GDPR," *Appl. Sci.*, vol. 11, no. 4, p. 1762, Feb. 2021.

[33] J. Bhatia, M. C. Evans, and T. D. Breaux, "Identifying incompleteness in privacy policy goals using semantic frames," *Requirements Eng.*, vol. 24, no. 3, pp. 291–313, Sep. 2019.

[34] M. Guerriero, D. A. Tamburri, and E. Di Nitto, "Defining, enforcing and checking privacy policies in data-intensive applications," in *Proc. IEEE/ACM 13th Int. Symp. Softw. Eng. Adapt. Self-Manag. Syst. (SEAMS)*, May 2018, pp. 172–182.

[35] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, N. A. Smith, and F. Liu, "Crowdsourcing annotations for websites' privacy policies: Can it really work?" in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 133–143.

[36] F. Liu, R. Ramanath, N. Sadeh, and N. A. Smith, "A step towards usable privacy policy: Automatic alignment of privacy statements," in *Proc. 25th Int. Conf. Comput. Linguistics, Tech. Papers (COLING)*, 2014, pp. 884–894.

[37] J. Bhatia, T. D. Breaux, and F. Schaub, "Mining privacy goals from privacy policies using hybridized task recomposition," *ACM Trans. Softw. Eng. Methodol.*, vol. 25, no. 3, pp. 1–24, Aug. 2016.

[38] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna, "PrivacyGuide: Towards an implementation of the EU GDPR on Internet privacy policy evaluation," in *Proc. 4th ACM Int. Workshop Secur. Privacy Analytics*, Mar. 2018, pp. 15–21.

[39] N. M. Nejad, P. Jabat, R. Nedelchev, S. Scerri, and D. Graux, "Establishing a strong baseline for privacy policy classification," in *Proc. 35th IFIP TC 11 Int. Conf. ICT Syst. Secur. Privacy Protection (SEC)*, Maribor, Slovenia. New York, NY, USA: Springer, Sep. 2020, pp. 370–383.

[40] D. Cer, Y. Yang, S.-Y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for English," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*, 2018, pp. 169–174.

[41] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Jan. 2019, pp. 3973–3983.

[42] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[44] O. Shahmirzadi, A. Lugowski, and K. Younge, "Text similarity in vector space models: A comparative study," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 659–666.

[45] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[46] J. D. Young, "Commitment analysis to operationalize software requirements from privacy policies," *Requirements Eng.*, vol. 16, no. 1, pp. 33–46, Mar. 2011.

[47] A. Arellano, E. Carney, and M. A. Austin, "Natural language processing of textual requirements," in *Proc. 10th Int. Conf. Syst. (ICONS)*, Barcelona, Spain, 2015, pp. 93–97.

[48] F. H. Shezan, Y. Lao, M. Peng, X. Wang, M. Sun, and P. Li, "NL2GDPR: Automatically develop GDPR compliant Android application features from natural language," 2022, *arXiv:2208.13361*.

[49] J. Caramujo, A. Rodrigues da Silva, S. Monfared, A. Ribeiro, P. Calado, and T. Breaux, "RSL-IL4Privacy: A domain-specific language for the rigorous specification of privacy policies," *Requirements Eng.*, vol. 24, no. 1, pp. 1–26, Mar. 2019.

[50] E. Vanezi, G. M. Kapitsaki, D. Kouzapas, and A. Philippou, "A formal modeling scheme for analyzing a software system design against the GDPR," in *Proc. ENASE*, 2019, pp. 68–79.

[51] P. Pullonen, J. Tom, R. Matulevičius, and A. Toots, "Privacy-enhanced BPMN: Enabling data privacy analysis in business processes models," *Softw. Syst. Model.*, vol. 18, no. 6, pp. 3235–3264, Dec. 2019.

[52] G. M. Riva, "Privacy architecting of GDPR-compliant high-tech systems: The PAGHS methodology," M.S. thesis, Dept. Elect. Eng., Math. Comput. Sci., Univ. Twente, Enschede, The Netherlands, 2019.

[53] M. Robol, M. Salnitri, and P. Giorgini, "Toward GDPR-compliant socio-technical systems: Modeling language and reasoning framework," in *Proc. IFIP Work. Conf. Pract. Enterprise Model.* New York, NY, USA: Springer, 2017, pp. 236–250.

[54] T. Antignac, R. Scandariato, and G. Schneider, "Privacy compliance via model transformations," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS&PW)*, Apr. 2018, pp. 120–126.

[55] H. Gjermundrød, I. Dionysiou, and K. Costa, "Privacytracker: A privacy-by-design GDPR-compliant framework with verifiable data traceability controls," in *Proc. Int. Conf. Web Eng.* New York, NY, USA: Springer, 2016, pp. 3–15.

[56] S. D. Ringmann, H. Langweg, and M. Waldvogel, "Requirements for legally compliant software based on the GDPR," in *Proc. OTM Confederated Int. Conf. Move Meaningful Internet Syst.* New York, NY, USA: Springer, 2018, pp. 258–276.

[57] O. Olukoya, "Assessing frameworks for eliciting privacy & security requirements from laws and regulations," *Comput. Secur.*, vol. 117, Jun. 2022, Art. no. 102697.

[58] E. Elwany, D. Moore, and G. Oberoi, "BERT Goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding," in *Proc. Workshop Document Intell. NeurIPS*, 2019, pp. 1–4.

[59] L. Elluri, K. Pande Joshi, and A. Kotal, "Measuring semantic similarity across EU GDPR regulation and cloud privacy policies," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2020, pp. 3963–3978.

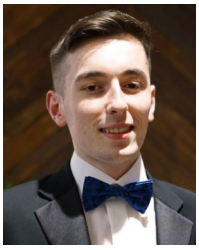
[60] A. Hegel, M. Shah, G. Peaslee, B. Roof, and E. Elwany, "The law of large documents: Understanding the structure of legal contracts using visual cues," 2021, *arXiv:2107.08128*.

[61] M. Bano, D. Zowghi, and C. Arora, "Requirements, politics, or individualism: What drives the success of COVID-19 contact-tracing apps?" *IEEE Softw.*, vol. 38, no. 1, pp. 7–12, Dec. 2020.

[62] M. Fan, L. Yu, S. Chen, H. Zhou, X. Luo, S. Li, Y. Liu, J. Liu, and T. Liu, "An empirical evaluation of GDPR compliance violations in Android mHealth apps," in *Proc. IEEE 31st Int. Symp. Softw. Rel. Eng. (ISSRE)*, Oct. 2020, pp. 253–264.

[63] S. Kununka, N. Mehandjiev, and P. Sampaio, "A comparative study of Android and iOS mobile applications' data handling practices versus compliance to privacy policy," in *Proc. 12th IFIP WG 9.2, 9.5, 9.6/11.7, 11.6/SIG 9.2. 2 Int. Summer School Privacy Identity Manag. Smart Revolution*, Ispra, Italy, Sep. 2017, pp. 301–313.

[64] M. Hatamian, S. Wairimu, N. Momen, and L. Fritsch, "A privacy and security analysis of early-deployed COVID-19 contact tracing Android apps," *Empirical Softw. Eng.*, vol. 26, no. 3, pp. 1–51, May 2021.
[65] *Manifest Permission Reference*, Android Developers, 2023. Accessed: May 22, 2023. [Online]. Available: https://developer.android.com/reference/android/Manifest.permission
[66] B. Brumen, "Automated text similarities approach: GDPR and privacy by design principles," in *Information Modelling and Knowledge Bases XXXII*, vol. 333. Amsterdam, The Netherlands: IOS Press, 2021, p. 213.
[67] R. Khandelwal, T. Linden, H. Harkous, and K. Fawaz, "PriSEC: A privacy settings enforcement controller," in *Proc. USENIX Secur. Symp.*, 2021, pp. 465–482.
[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[69] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Commun. ACM*, vol. 8, no. 10, pp. 627–633, Oct. 1965.
[70] O. Olukoya, L. Mackenzie, and I. Omoronyia, "Security-oriented view of app behaviour using textual descriptions and user-granted permission requests," *Comput. Secur.*, vol. 89, Feb. 2020, Art. no. 101685.
[71] T. R. Chhetri, A. Kurteva, R. J. DeLong, R. Hilscher, K. Korte, and A. Fensel, "Data protection by design tool for automated GDPR compliance verification based on semantically modeled informed consent," *Sensors*, vol. 22, no. 7, p. 2763, Apr. 2022.
[72] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The privacy policy landscape after the GDPR," *Proc. Privacy Enhancing Technol.*, vol. 2020, no. 1, pp. 47–64, Jan. 2020.
[73] S. I. Hajeer, "Comparison on the effectiveness of different statistical similarity measures," *Int. J. Comput. Appl.*, vol. 53, no. 8, pp. 14–19, Sep. 2012.
[74] P. Sitikhu, K. Pahi, P. Thapa, and S. Shakya, "A comparison of semantic similarity methods for maximum human interpretability," in *Proc. Artif. Intell. Transforming Bus. Soc. (AITB)*, vol. 1, Nov. 2019, pp. 1–4.
[75] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0144059.
[76] R. Xu and D. C. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, Nov. 2005.
[77] N. M. Nejad, S. Scerri, and J. Lehmann, "KnIGHT: Mapping privacy policies to GDPR," in *Proc. 21st Int. Conf. Knowl. Eng. Knowl. Manag. (EKAW)*, Nancy, France. New York, NY, USA: Springer, Nov. 2018, pp. 258–272.
[78] J. R. Reidenberg, T. Breaux, L. F. Cranor, B. French, A. Grannis, J. T. Graves, F. Liu, A. McDonald, T. B. Norton, and R. Ramanath, "Disagreeable privacy policies: Mismatches between meaning and users' understanding," *Berkeley Tech. LJ*, vol. 30, no. 1, pp. 39–88, 2015.
[79] S. Zimmeck and S. M. Bellovin, "Privee: An architecture for automatically analyzing web privacy policies," in *Proc. 23rd USENIX Secur. Symp. (USENIX Secur.)*, 2014, pp. 1–16.
[80] J. R. Reidenberg, J. Bhatia, T. D. Breaux, and T. B. Norton, "Ambiguity in privacy policies and the impact of regulation," *J. Legal Stud.*, vol. 45, no. S2, pp. S163–S190, Jun. 2016.
[81] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information*, vol. 11, no. 9, p. 421, Aug. 2020.
[82] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 16857–16867.



RYAN MCCONKEY received the integrated M.Eng. degree in software engineering from the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast. He is currently a Software Engineer in the defence industry. His research interests include efficient software system design, natural language processing, and artificial intelligence.

OLUWAFEMI OLUKOYA received the Ph.D. degree in computing science from the University of Glasgow, U.K. He is currently a Lecturer (Assistant Professor) with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, an EPSRC-NCSC Academic Centre of Excellence in Cyber Security Research (ACE-CSR), where he leads research in malware and attack technologies, systems security, privacy, and cybercrime.