IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 29, 2021, accepted October 14, 2021, date of publication October 28, 2021,
date of current version November 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3123950

A BERT Based Approach to Measure Web Services Policies Compliance With GDPR

LAVANYA ELLURI[1], (Member, IEEE),
SAI SREE LAYA CHUKKAPALLI[2], (Graduate Student Member, IEEE),
KARUNA PANDE JOSHI[1], (Senior Member, IEEE),
TIM FININ[2], (Member, IEEE), AND ANUPAM JOSHI[2], (Fellow, IEEE)
[1]Department of Information Systems, University of Maryland at Baltimore County (UMBC), Baltimore, MD 21250, USA
[2]Department of Computer Science, University of Maryland at Baltimore County (UMBC), Baltimore, MD 21250, USA

Corresponding authors: Lavanya Elluri (lelluri1@umbc.edu) and Karuna Pande Joshi (karuna.joshi@umbc.edu)

This work was supported in part by NSF Phase I Industry-University Cooperative Research Centers (IUCRC) University of Maryland at Baltimore County (UMBC), Center for Accelerated Real-time Analytics (CARTA) under NSF Award 1747724, and in part by an award of IBM Research.

ABSTRACT Data confidentiality is an issue of increasing importance. Several authorities and regulatory bodies are creating new laws that control how web services data is handled and shared. With the rapid increase of such regulations, web service providers face challenges in complying with these evolving regulations across jurisdictions. Providers must update their service policies regularly to address the new regulations. The challenge is that regulatory documents are large text documents and require substantial human effort to comprehend and enforce. On the other hand, web service provider privacy policies are relatively short compared to the regulatory texts, so it is hard to determine if an organization's policy document addresses the regulation's essential elements. We have developed a framework to automatically compare web service policies with regulatory policies to measure how closely the web service provider complies with a regulation. In this paper, we present our framework's details along with the results of analyzing a corpus of 3,000 privacy policies against GDPR. Our framework uses BiLSTM multi-class classification and a BERT extractive summarizer. We evaluate the framework's efficacy by checking the context similarity score between summarized GDPR and web service provider privacy policies.

INDEX TERMS Web service privacy policies, deep learning, context extraction, BERT summarization, knowledge discovery.

I. INTRODUCTION

Web service providers are increasingly storing their users' personal information. This could be information a user generates in interacting with the site, like browsing patterns and user transactions, to facilitate a better user experience. It could also be pictures, videos, and other data provided by a user. The providers often share a significant portion of this consumer data with third parties for additional analysis to improve their businesses. Hence, even though this shared data provides many beneficial services to users, its confidentiality and use remain a concern for many consumers. Due to the increase in sensitive information used by web services, regulatory authorities worldwide are formulating data protection regulations to protect their users' data. Examples are the European Union's General Data Protection Regulation (EU GDPR) [1]–[3], Australian Privacy Principles (APP) [4], Canada's Personal Information Protection and Electronic Data Act (PIPEDA) [5], and the California Online Privacy Protection Act (CalOPPA) [6]. Web service providers must adhere to these regulations if they are using data from users in these regions. This increase in complex data protection regulations has resulted in tremendous legal compliance challenges for web service providers. Whenever businesses update their privacy policies, they need to check compliance against these regulations and laws.

As an example of such regulations, we focus on GDPR. GDPR has become a key regulation as its penalties can be significant in a data breach when a service provider uses EU users' data. Web service providers include GDPR rules

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

in their policies to make their systems robust, operable in Europe, and acceptable to consumers. GDPR covers many rules for organizations utilizing European users' data [7], [8]. It has become one of the most impactful regulations, even outside Europe, as it has significant penalties. Organizations must pay large fines of up to ten million euros or 2% of their entire global revenue in the case of a data breach [9], [10]. Furthermore, according to GDPR, data subjects have the right to question web service providers and consumers.

As organizations change or update their privacy policies, they need to ensure that these policies continue to comply with GDPR. GDPR has complex rules in articles associated with various roles involved in using web services. This regulation document is massive, consisting of 99 articles where each article has a lengthy description. Currently, the process of ensuring compliance with a regulation is labor-intensive and time-consuming. Human intervention is required to extract the key summary rules from the text manually. Web service privacy policies often incorporate summaries from the regulation documents like rules for service providers and consumers. However, these privacy policy text documents are short when compared to the colossal regulatory documents. End users often must deal with the complexity of understanding the regulation rules and relating them to a service provider's stated privacy policies.

Web service providers exploit a user's data in a variety of ways. To perform any analysis on customer datasets and relate it to an individual user, Personally Identifiable Information (PII) is needed. As mentioned, the privacy of PII and other data managed by service providers is often of significant concern to consumers, and regulatory bodies worldwide are announcing their jurisdiction specific data protection laws. The increase in comprehensive regulations causes significant regulation compliance challenges to protect user's data. For example, GDPR identifies the rules that apply to organizations using any EU individual data like name, address, email address, medical records, social media posts. Service policy documents, currently accessible only in textual format, require substantial human intervention in terms of time and effort to check for compliance and prevent huge penalties in case of data breaches.

Our automated framework can help the web service providers to recognize missing areas in their privacy policy by using the classified GDPR class, and the overall compliance rate helps to make updates to their privacy policy to be more compliant to GDPR. It identifies the GDPR class by using BiLSTM multi-class classification and further summarizes the extracted context using a Bidirectional Encoder Representations from Transformers (BERT) based text summarizer to obtain the similarity score between the privacy policy and GDPR document. Besides saving organizational human resources for checking compliance, it also helps by generating a summary of the privacy policy, enabling potentially inconsistent policies in an organization to be more easily recognized and corrected. While a more detailed evaluation is currently underway, the preliminary assessment of our novel approach has been promising. We have carefully reviewed the relevant literature and have not found reliable work or baselines for comparing the privacy policies of web service providers with GDPR requirements. This is not surprising since GDPR is still relatively new, having been implemented only in 2018. Therefore, our research on automatically comparing other policies to GDPR is novel.

The structure of the remaining paper is as follows. Section II gives background and related work on short text classification using machine and deep learning methods. In Section III, we describe the technical approach of our framework, covering data collection, text classification, creating knowledge graphs from information extracted from the documents, text summarization, and an evaluation of our results. Finally, we conclude this paper in Section IV and identify ongoing and future work.

Our technical research relies on several AI and machine learning technologies, including text classification, word embeddings, and text summarization.

II. RELATED WORK

In this section, we review similar research conducted in this area. While researchers have used machine learning or deep learning methods applied to solve different text classification problems on complete documents, but they are not efficient in analyzing short or incomplete text. Classification, in the end, can certainly help in specific scenarios where the user wants to find documents on topics already developed by others. However, classification with a document in progress like a privacy policy is not supported.

A. MACHINE LEARNING FOR SHORT TEXT CLASSIFICATION

Automatic policy document categorization is essentially the classification challenge due to its unstructured textual format. Advanced machine learning approaches are applied to text categorization [11]–[13] problems such as fraud detection [14], recommendation systems [15], social media sentiment analysis [16] and cyber-attacks detection [17]. Li and Jain [18] evaluated naive Bayes, decision trees, nearest neighbor, and subspace classification algorithms to solve the document classification problem on Yahoo News items. Since the number of dimensions was high, they employed dimensionality reduction approaches to make the problem more tractable. The authors found the performance of naive Bayes and subspace classifiers to be best, and the performance of decision trees increased after using a boosting technique. Also, they have tried using a combination of classifiers and states that it may not constantly improve the accuracy. Another experimental result was observing a significant increase in classification performance due to including feature selection techniques. The performance of naive Bayes classifiers also improved as the number of features increased and was found to be ineffectively with small feature sets.

Much work has been done on classifying short text segments such as social media tweets and comments into a set

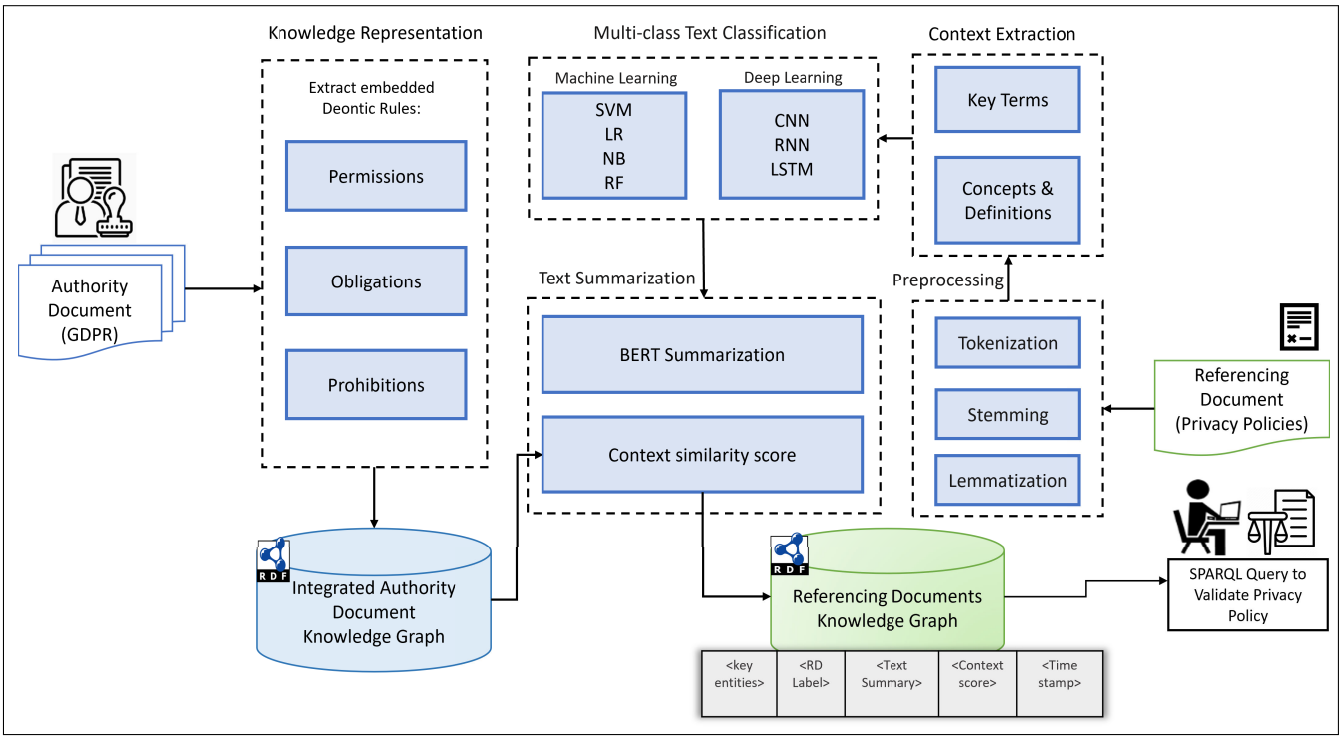


FIGURE 1. Architecture flow to extract knowledge from authority and referencing documents using NLP, DL, BERT and semantic web approaches.

of categories. Lee *et al.* [19] used tweets along with trending topics and definitions to label each category. Text-based and network-based classification techniques were applied to the data set and it was found that the network-based classifier performed better. Other research [20] used a simple bag-of-words representation to classify the tweets into news, opinions, deals, events, and private messages. Other research works [21]–[23] has explored sentiment analysis on short texts to identify the user opinion based on various machine learning approaches. They found that using fewer features made it harder to generalize the new set of documents but improved time and space complexities.

We have not identified research that uses short policy descriptions to classify the context incrementally in real-time. We provide a new approach for users who need to update an existing privacy policy or are writing new content based on current or proposed data protection regulations.

B. DEEP LEARNING APPROACHES

Deep learning is a comparatively advanced approach to machine learning, with improved automatic feature engineering from the textual data. Several deep learning approaches have made significant models in handwriting generation, text generation, image recognition, and image caption generation [24]–[26]. The effective procedures of deep learning are also observed in Natural Language Processing (NLP), like topic categorization [27], text classification [28], Part-of-Speech tagging (POS) [29], etc. To perform document classification using machine learning approaches, we need to do tokenization, stemming, lemmatization, and more steps to clean our text and to end up with relevant words. In deep learning, we do not have to go through these steps as the neural networks take care of all such activities and learn by themselves.

In the paper [30], author presents extreme multi-label text classification based on seven state-of-art techniques XML-CNN, FASTXML, FastText, SLEEC, CNN-Kim, Bow-CNN, and PD-Sparse. The author has developed a new architecture based on CNN-Kim like CNN; their model learns rich feature representations by passing the document to multiple convolutional filters. This model works on a dynamic max-pooling method to learn more features that need greater attention from different document sections. They have also included another hidden layer between the output layer and max-pooling to decrease the model size and increase the model's performance. Experimental results show that XML-CNN attained the best results on the datasets which they have used to evaluate.

All these approaches that we have discussed contribute to text classification based on supervised and unsupervised techniques. However, there is limited research on context matching of short text documents, especially if we want to compare context similarity between documents where one document is a short text, and another is a large textual dataset.

III. TECHNICAL APPROACH

This section describes our framework for automatic context extraction and comparison of short text documents, like

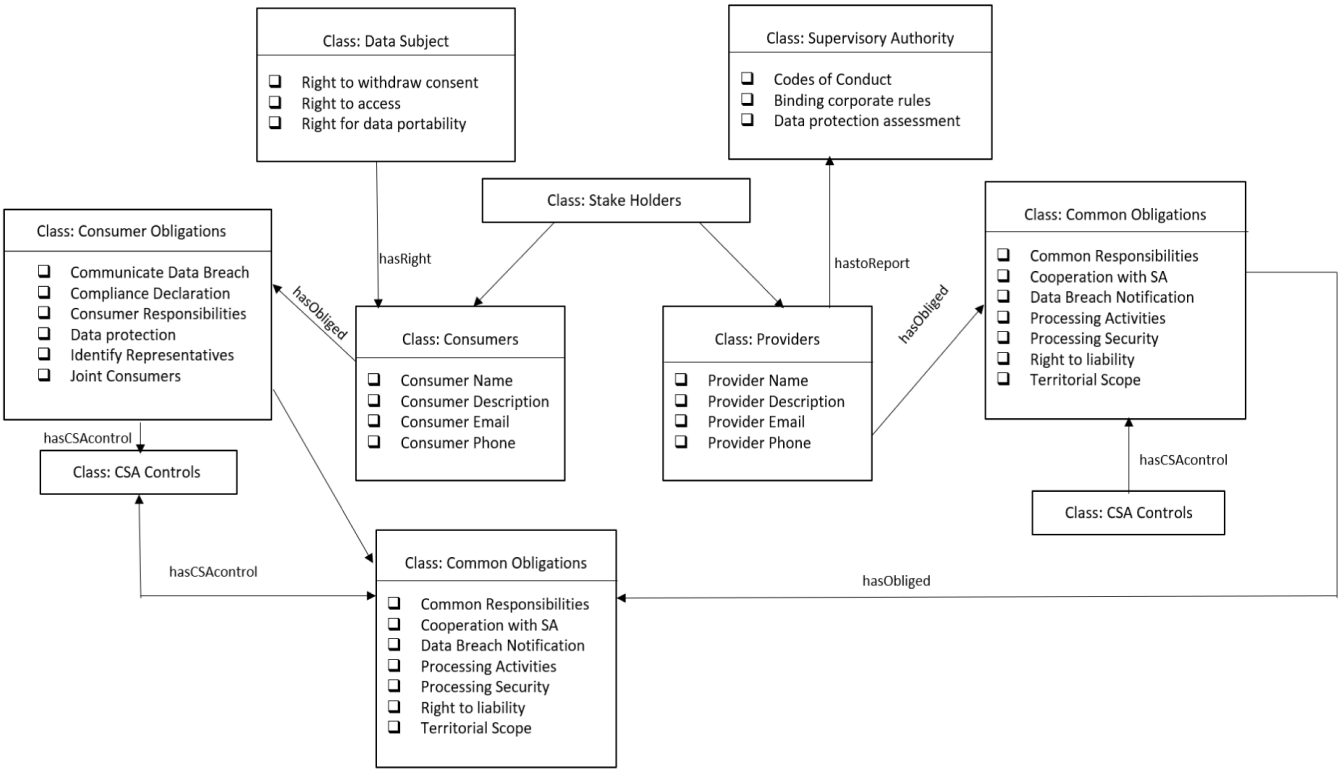


FIGURE 2. The authority document knowledge graph represents information extracted from the GDPR and is supported by an OWL ontology.

TABLE 1. Key entities from GDPR.

GDPR Entities	Occurrence
controller	1008
processor	528
data subject	375
supervisory authority	331

service provider policies. Figure 1, illustrates the overall system architecture of our framework. In this paper, we focus on applying this framework to the GDPR and comparing it with 3000 Web service privacy policies. Our system automatically predicts the GDPR entities present in a privacy policy and provides a context similarity score. We have used different Natural Language Processing (NLP) approaches, Machine Learning, and Deep Learning models for extracting key regulatory entities in a privacy policy. We then extracted the context based on the entities extracted to summarize the text using the BERT summarizer. We also construct knowledge graphs with the key information about the regulations and policies using the Semantic Web languages RDF [31]–[34] and OWL [35]–[38]. The knowledge graphs are stored in an Apache Jena server that can be queried to retrieve or visualize information. We developed tools of this framework using Python language.

Our framework is divided into six different phases after extracting the most frequent entities from the GDPR document. The six phases of our methodology are:

- **Authority Document Knowledge Graph:** In this phase, we mine relevant articles based on key entities from a regulatory (or authoritative) document and populate them as classes and data properties in the authority document knowledge graph. Section III-A has a detailed explanation of how GDPR elements were populated in the Authority Document knowledge graph.
- **Data Collection**: In this phase, we build a data corpus of the short text policies to compare with the authoritative document. For this study, we gathered two datasets of web service provider privacy policies, one to correspond to EU-based privacy policies and the second demonstrating policies worldwide. All the 3000 policies were downloaded with the latest version to make sure they were created after GDPR was published i.e., May 2018 and are available publicly. We constrain our assessment to web service policies in the English language.
- **Multi-class Text classification**: This phase consists of an integrated approach to extract context from referencing documents based on the classes of the authoritative document identified in phase 1. We used machine learning and deep learning techniques for this phase. Details on how this approach is applied to extract GDPR context from privacy policies are covered in section III-C.
- **Text Summarization**: In this phase, we obtain the context from the privacy policies based on the class extracted in phase 3. We then summarize the context of

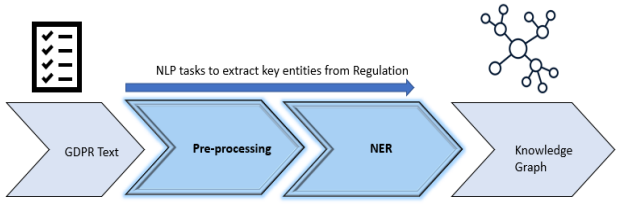


FIGURE 3. Approach for entities extraction to create knowledge graph.

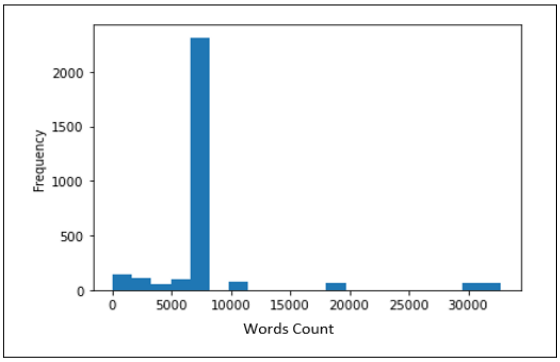


FIGURE 4. Words count distribution in web service privacy corpus.

the privacy policy and GDPR using BERT text summarization approach to obtain the context similarity score between the authority document and the referencing document.

- **Referencing Documents Knowledge Graph**: We created a referencing document knowledge graph that stores the instances of all web service privacy policies obtained by applying this framework to GDPR, as described in section III-E. We utilized Protege [39] to create this knowledge graph.
- **Validation**: We validated the results stored in the knowledge graph by utilizing a set of organization policies adhering to GDPR and another set of organizational policies that are not adhering to GDPR. Section III-F has detailed experimental investigation results.

A. AUTHORITY DOCUMENTS KNOWLEDGE GRAPH

In the first stage of our system, we analyzed the huge GDPR repository to extract relevant chapters and checklists. In the papers [40], [41], authors manually identified the specific key terms and developed a knowledge graph. In this work, we applied approaches like Named Entity extraction to determine the most frequently occurring entities in the GDPR corpus and populated our knowledge graph to store all the key entities. Figure 2 shows the authority knowledge graph for GDPR. As the privacy policies are short, they often would not include all the rules from the regulation document. Therefore, based on rules populated in our knowledge graph, we determined top four entities associated with GDPR. We noticed an alignment of GDPR rules in privacy policies are based on key entities rather than complete rules during the development. We elaborate on the definitions of these key entities we have identified as part of this work.

1) ENTITIES EXTRACTION

To identify the key entities of the web service policies we extracted the high frequency entities using Named Entity Recognition (NER). Figure 3 shows the process of extracting the knowledge from regulation and populating into the knowledge graph. These entities are helpful to further classify the privacy polices automatically into a set of predefined classes. The key entities identified by us were:

Controller: This entity is the legal agency that operates alone or jointly and determines the objectives of processing the personal data. Controllers can make judgments about all the processing actions.

Processor: Processor entity is a legal authority or agency that processes the personal data on behalf of the controller. They act on behalf of the controller under their permission.

Data Subject: People are identified as Data subject entity. They are individuals about whom the controller collects information in connection with their business operations.

Supervisory Authority: An authority established under the GDPR. This entity is responsible for supervising the application of the GDPR to safeguard the data subject rights and accelerate the free flow of Personal Identifiable (PII) EU user's data.

The count for each entity in the regulation is shown in Table 1. In this stage, we also checked for rules pertaining to entities like data subject and supervisory authority in the GDPR. Below are a few rules from the GDPR document.

Data Subject (GDPR): *"Effective protection of personal data throughout the Union requires the strengthening and setting out in detail of the rights of data subjects and the obligations of those who process and determine the processing of personal data, as well as equivalent powers for monitoring and ensuring compliance with the rules for the protection of personal data and equivalent sanctions for infringements in the Member States"* [2].

Supervisory Authority (GDPR): *"Aware that a personal data breach has occurred, the controller should notify the personal data breach to the supervisory authority without undue delay and, where feasible, not later than 72 hours after having become aware of it, unless the controller can demonstrate, following the accountability principle, that the a personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons"* [2].

2) DEONTIC RULES

Modal logic covers various other forms of reasoning such as deontic logic and temporal logic [42]–[44]. Deontic rules are statements containing permissions and obligations, and temporal logic characterizes time-based constraints. Deontic logic consists of four types of modalities i.e., permissions / rights, obligations, dispensations and prohibitions.

1. Permissions / Rights: Permissions are rules that define the rights or consents for an entity.

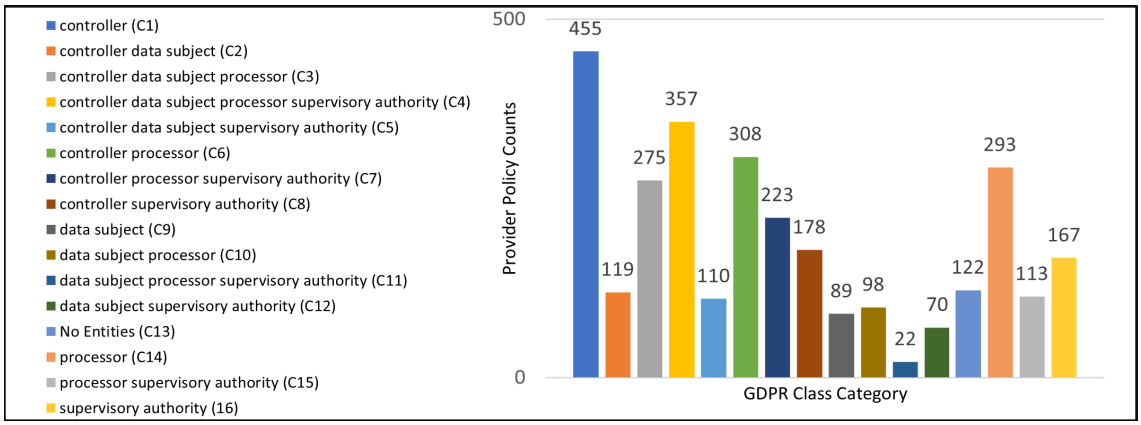


FIGURE 5. GDPR class categories in privacy policies.

2. Obligations: Obligations expressions are the enforced actions that an entity must accomplish.

3. Dispensations: Dispensations refer to optional expressions and describe non-mandatory conditions.

4. Prohibitions: Prohibitions are the phrases that indicate the acts which are prohibited.

In this research, we have used permissions and obligations to label sentences into any one of them. Sentences with verbs like 'could', 'may','can' were categorized as permissions, and sentences with verbs like 'must','shall','should' were classified as obligations. Below mentioned are some examples of our context:

Permissions (GDPR):

"Children merit specific protection with regard to their personal data, as they may be less aware of the risks, consequences and safeguards concerned and their rights in relation to the processing of personal data" [2].

"Regulation governing the lawfulness of personal data processing, establish specifications for determining the controller, the type of personal data which are subject to the processing, the data subjects concerned, the entities to which the personal data may be disclosed, the purpose limitations, the storage period and other measures to ensure lawful and fair processing" [2].

Obligations (GDPR):

"The right to the protection of personal data is not an absolute right; it must be considered in relation to its function in society and be balanced against other fundamental rights, in accordance with the principle of proportionality" [2].

"When the processing has multiple purposes, consent should be given for all of them. If the data subject's consent is to be given following a request by electronic means, the request must be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided" [2].

"The controller or processor should compensate any damage which a person may suffer as a result of processing that infringes this Regulation. The controller or processor should be exempt from liability if it proves that it is not in any way responsible for the damage. The concept of damage should be broadly interpreted in the light of the case-law of the Court of Justice in a manner which fully reflects the objectives of this Regulation [2].

B. DATA COLLECTION

Due to the wide acceptance and rapid engagement of EU GDPR in and outside EU, providers have started addressing GDPR rules in their service policies. We extracted the latest policy documents of more than 3000 organizations across the globe trying to adhere to GDPR. These organizations provide web-based businesses that use EU user's PII data in various domains like social media, health care, e-commerce, cloud infrastructure, and many more crucial areas. While downloading the organizations' privacy policies, we ensured the latest version of each organization is considered for this analysis. Most of the privacy policies have word ranges around 7000 to 8000. Figure 4 shows the total word count distributions among the privacy policies.

We considered only policies that were created after GDPR publish date, i.e., May 2018, as we wanted to consider the policies trying to comply with GDPR. Next, we extracted the entities from all these privacy policies identified in phase 1. The unique combinations for all the 3000 policies using these four entities are the 16 classes shown in Figure 5. We observed that one of the class out of these 16 classes doesn't have any privacy policies. All the policy descriptions and the associated entities were stored collectively for predicting the GDPR class using multi-class classification. Though we extracted all the policies that claimed to adhere to GDPR, we found some policies that didn't have any entities related to GDPR and they are classified under "No Entities".

C. MULTI CLASS TEXT CLASSIFICATION FROM PRIVACY POLICIES

We have chosen four machine learning and three deep learning models for this phase. At first, we predicted the classes using the traditional machine learning approaches such as Naive Bayes (NB), Random Forest (RF), Logistic Regression

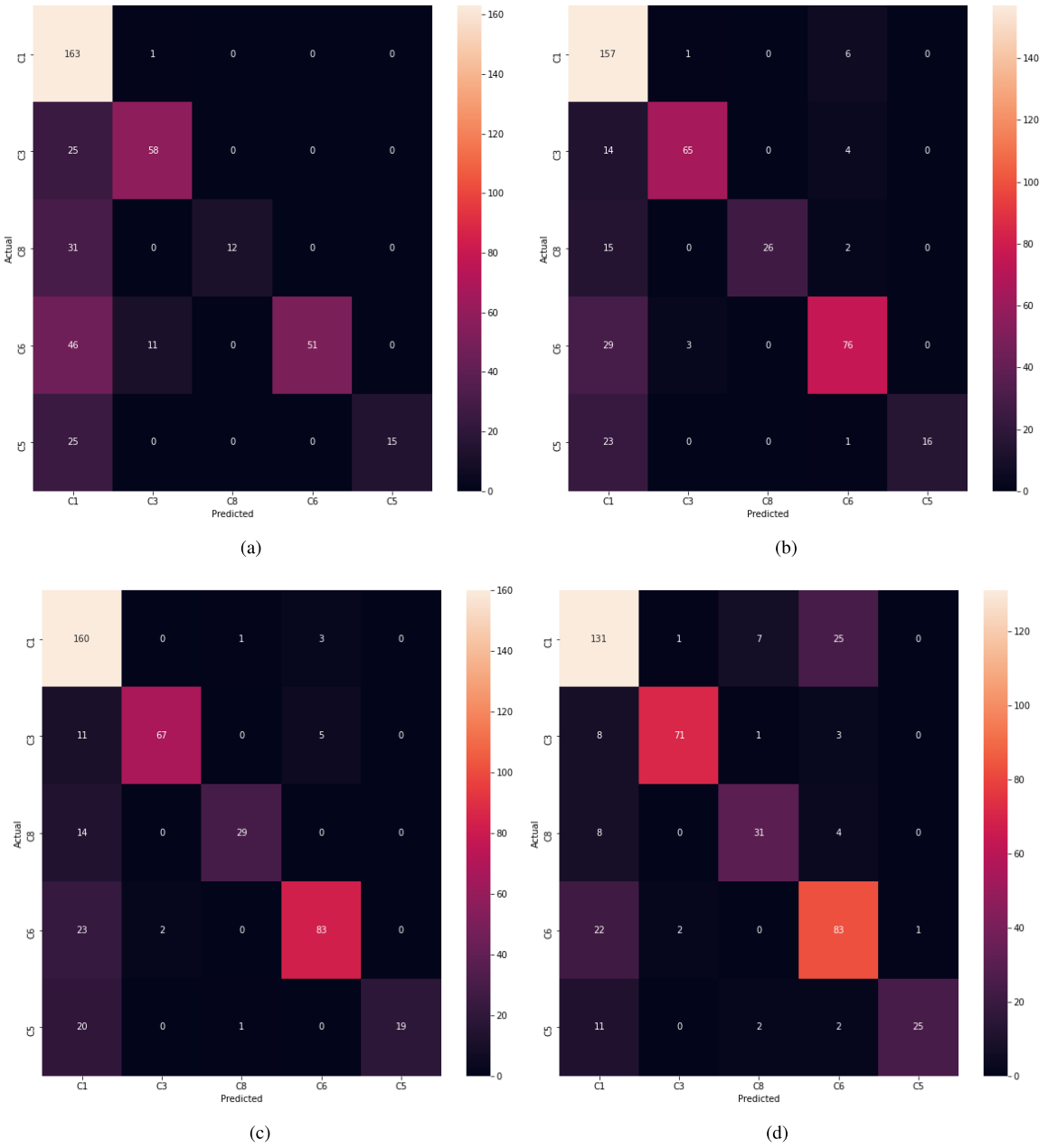


FIGURE 6. Confusion matrix of machine learning models on the provider policies corpus (a) NB; (b) LR; (c) RF; (d) SVM.

(LR), and Support Vector Machine (SVM) to check if the predictions are accurate before applying deep learning models for multi-class classification. We considered the 70% of the data as training set and the test size as 30%. The confusion matrix is shown for the five GDPR classes as shown in Figure 6 after applying all the machine learning models to the corpus. NB and LG models misclassified more policy documents when compared to RF and SVM models. As depicted in Figure 6 SVM model predicted most of the instances accurately. However, we wanted to check if there would be any further improvement in accuracy by applying the improvised deep learning models.

We next utilized convolutional neural network (CNN), long short-term memory (LSTM) and bidirectional LSTM (BiLSTM) deep learning approaches to predict the GDPR classes in privacy policies corpus of web services. The detailed explanation of the CNN model used in this framework is presented in the paper [28]. In the first layer of the architecture, the extremely useful n-gram features will be extracted, and then the embeddings of each word are stored. It will then pass through the pooling layer to create feature vectors and then transform the preceding convolution to a higher-level conceptual view. Finally, the dense layer summarizes the patterns of generated feature vectors and

TABLE 2. Evaluation metrics of text classification models applied in our framework.

Model	Precision	Recall	F-Measure	Accuracy
Naive Bayes	0.74	0.58	0.58	0.58
Logistic Regression	0.76	0.66	0.66	0.66
Random Forest	0.80	0.74	0.74	0.74
Support Vector	0.75	0.72	0.73	0.72
CNN	0.65	0.72	0.73	0.59
LSTM	0.70	0.75	0.74	0.68
BiLSTM	0.79	0.76	0.78	0.75

generates predictions for the resultant web service privacy policy.

LSTMs [45]–[47] and BiLSTMs [48]–[50] are enhanced versions of basic recurrent neural networks (RNNs). The underlying concept behind LSTMs is the memory units, which preserve historical knowledge over time. BiLSTMs contain two LSTMs that combine the extensive periods of background knowledge from both forward and backward directions during a specific time frame. This facilitates the hidden layer to store both the past and the future knowledge. Also, they learn long-term dependencies between the words in a series without collecting unnecessary knowledge. Therefore, we applied all these state-of-the-art semantic deep learning models such as CNN, LSTM and BiLSTM for more accurate predictions on the web service provider privacy policies corpus. Figure 7 shows the number of epochs and validation accuracy of the three deep learning models considered in this framework. The accuracy rates of CNN and LSTM models are 59% and 68%, similar to NB and LR models where the accuracy rates are 58% and 66%. However, the BiLSTM model has improved the accuracy rate to 75% compared to the RF and SVM model scores i.e., 74% and 72% as shown in Table 2. We studied both the simpler machine learning algorithms and the deep learning model to evaluate which works better. One of the key contributions of this paper is to empirically demonstrate which approach works better for comparing small text datasets. We observed an improvement in our model using the more expensive neural network approach. We expect that the difference will become more significant as the corpus size increases and plan to evaluate this in our future work.

Since the privacy descriptions are different in each privacy document, we initially thought it would be hard for the machine to understand them. However, we see that accuracy for this kind of problem is quite considerable. Also, we expect that by increasing the corpus size, there could be an improvement in the accuracy score. After we successfully predicted the GDPR class present in a privacy policy, we further wanted to evaluate the results by summarizing the text related to these entities using BERT text summarizer and then validating the context similarity scores using policies adhering to GDPR versus policies not adhering.

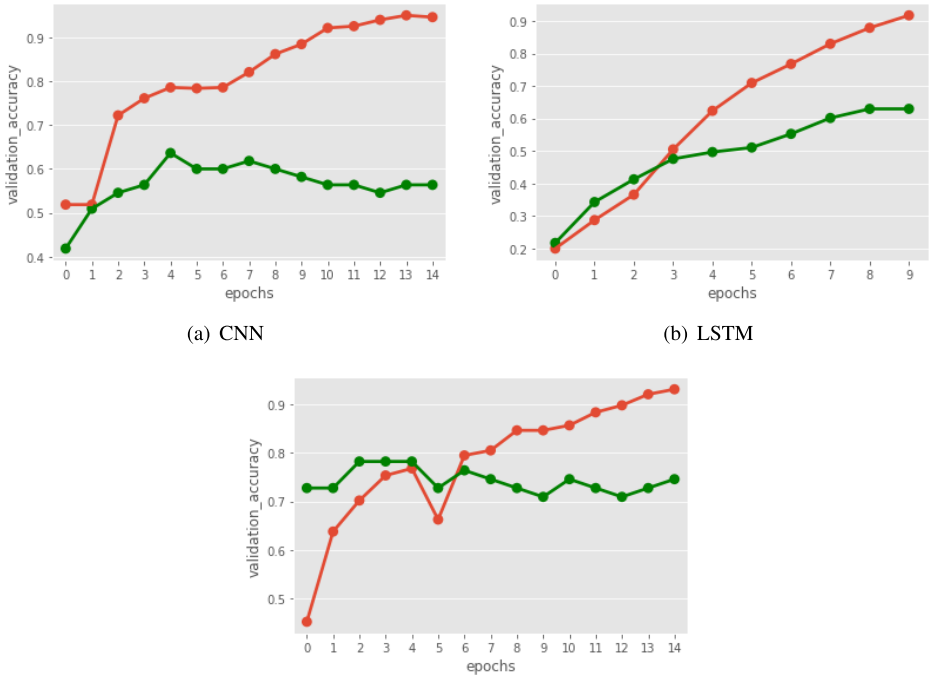
D. TEXT SUMMARIZATION

Text summarization is a technique in NLP where long texts are compressed to short texts without losing important information [51]. With the increase in textual datasets from various sources, the need for automatic text summarization tools has risen predominantly. The traditional manual summarization method is complicated and time-consuming for converting lengthy documents to short texts and so Automated text summarization methods preferred. There are two different approaches for automatic text summarization, viz. abstraction and extraction.

In the abstractive text summarization approach [52], the output summary obtained from the textual document contains text that may or may not be present in the original document as the sentences in summary are generated and not extracted. Therefore, the generated textual summary represents the important context of the textual document with less grammatical inaccuracies. In case of extractive text summarization approach [53], a scoring function is utilized to consider important sentences from the original document and concatenate them to form a summary. Unlike the abstractive text summarization approach, the sentences present in summary may not be grammatically accurate, but the vital points of the text are not modified.

In order to perform text summarization task, we have considered key entities from the predicted GDPR class described in section III-C and extracted the context from the privacy policy document related to these key entities. Further, we have incorporated extractive summarizer with BERT and K-Means [54] approach in our work. BERT is pre-trained on Wikipedia and BooksCorpus [53]. The pre-trained multilayer model is bidirectional as it considers both left and right contexts in all layers.

Our BERT based summarizer approach for policy documents is shown in Figure 8. We tokenize the sentences present in the policy document using BERT tokenizer and use them as an input to the pre-trained BERT model. Each word in the sentences is converted to a vector representation. Multiple layers can be chosen for creating embeddings with the help of default pre-trained BERT model. The final output layer of the BERT model generates a matrix in the form of $N \times W \times E$ where N represents number of sentences, W the tokenized words and E the embedding dimensions. However, the results do not provide us with best embeddings for sentences. Therefore, generated output embeddings are averaged out to generate $N \times E$ matrix in the N-2 layer for better representation. Further, the generated $N \times E$ matrix provided by the BERT is used for clustering. We used the K-Means algorithm [55] for clustering the embeddings obtained from



(a) CNN (b) LSTM

(c) BiLSTM

FIGURE 7. Multi-class identification accuracy scores on the polices using deep learning models CNN, LSTM and BiLSTM.

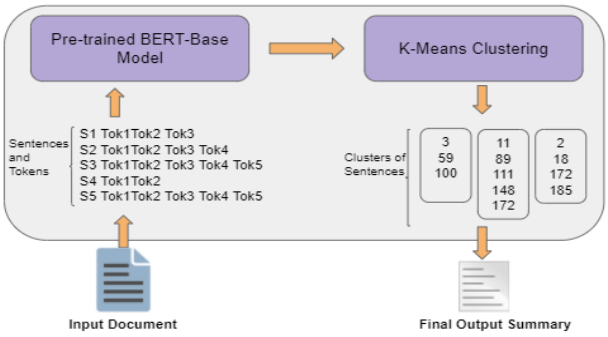


FIGURE 8. BERT based text summarization method for policy documents.

BERT model, where we chose the default value for K, which represents the number of clusters. Consequently, sentences closest to the centroid are considered for the final output summary.

E. REFERENCING DOCUMENTS KNOWLEDGE GRAPH

We used Protege [39] to create a document knowledge graph that contains all the extracted results of web service policy document from the previous sections. This ontology is essential to query if a policy document has mentioned any key entities that occur in GDPR. In addition, users can check the summary of the identified class from a privacy policy. This knowledge graph has a class ''Web_Service_Provider'' that holds several data properties shown below.

- **Predicted GDPR Class** property holds the identified GDPR class for a given web service provider policy document from the classification model used in our framework. This further helps to extract the summary of the key entities present in the class.
- **Extracted Text Summary** property represents the text summary extracted using BERT with the help of key entities found in the provider policy document. This helped us to compare the context of GDPR and the privacy policy texts.
- **Context Score** property holds the value obtained while comparing the text summary of key entities from the GDPR and privacy policy document. This score would give a sense to organization or the end user to know how much of the document is adhering to regulation.
- **Policy Date and Result Date** These two properties talk about the policy extracted date from the web and the date of the results identified using our framework.

These properties hold information crucial to query and retrieve a privacy policy's information. We used RDF [31]–[34] and OWL [35]–[38] languages to store the rules extracted from privacy policies used for this research. Our ontology hosted in the public domain can be quickly accessed and queried by organizations dealing with GDPR. Figure 9 shows the Azure instance of our knowledge graph. We can query the results using the SPARQL query language [56]–[59] as shown below:

```
SELECT ?dataproperty ?value WHERE {
    RD:Azure ?dataproperty ?value}
SELECT ?dataproperty ?value WHERE {
    RD:Amazon ?dataproperty ?value}
```

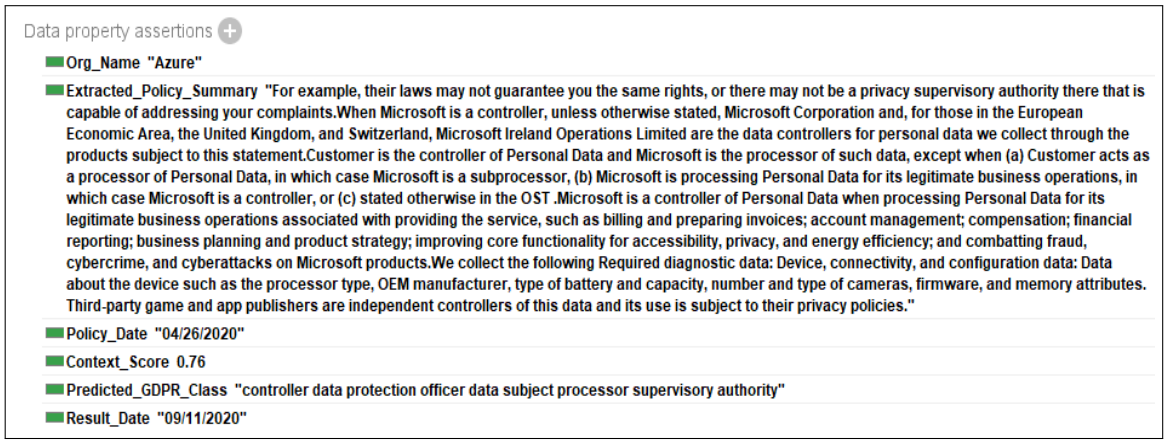


FIGURE 9. Azure policy knowledge extraction results.

F. VALIDATION

We validate our work by calculating the cosine similarity metric of the summarized GDPR and summary extracted from privacy policy documents by using the key entities present in the predicted GDPR class from section III-C.

We selected five privacy policy documents for our work that are compliant with GDPR. Further, we extracted sentences from each privacy policy document that are associated with any of the four entities such as controller, processor, data subject and supervisory authority. The entity extracted privacy policy documents are labeled as dataset A.

We selected another set of five privacy policy documents that were not compliant with GDPR and summarized each privacy policy document with the help of BERT extractive summarizer described above in section III D. Out of the five privacy policies, we extracted versions of Adobe and Azure privacy policies timestamped before the GDPR published date (i.e. May 2018) as they would not address any rules related to GDPR. The other three policies considered did not address GDPR rules as these organizations do not deal with European user's data. These documents are labeled as dataset B.

Similarly, we also summarized GDPR document by extracting the sentences based on four key entities and further summarized the obtained document with the help of BERT summarization model described earlier. We labeled the document as summarized GDPR.

We determine whether the privacy policy texts are compliant or not compliant with GDPR by measuring the cosine similarity using two of the document embedding algorithms such as TF-IDF and BERT.

1) COSINE SIMILARITY

This well known similarity measure determines the similarity between two documents (represented as vectors) irrespective of their size [60]. The similarity score is determined by the angle between the two vectors using the following formula [61].

$$Cos\ \alpha = \frac{A \times B}{|A| \times |B|} = \frac{\Sigma_{i=1}^{n} A_i \times B_i}{\sqrt{\Sigma_{i=1}^{n} (A_i)^2} \times \sqrt{\Sigma_{i=1}^{n} (B_i)^2}} \quad (1)$$

In our approach, the documents are represented as vectors using BERT or TF-IDF.

2) TF-IDF

This method, very commonly used in the IR community, is a combination of term frequency and inverse document frequency [62]. It measures how often a term occurs in a document (tf) in relation to how often it appears in all documents in the corpus (df). Then the inverse document frequency (idf) of the word is calculated as:

$$idf\,(word)_i = \log\left(\frac{Total\ number\ of\ documents}{df\,(word)_i}\right) \quad (2)$$

The vector of tf-idf scores shown in equation 3 are calculated by multiplying the term frequency (tf) of each document with the value of inverse document frequency (idf) obtained from equation 2.

$$tf\,(word)_{ij} \times idf\,(word)_i \quad (3)$$

The dot product of the obtained tf-idf vectors determines the cosine similarity score where the values usually lies in the range of 0 and 1.

As depicted in Table 3, tf-idf scores for dataset A are comparatively higher when compared to the scores in dataset B shown in Table 4 as the context similarity of the documents depends on the cosine similarity scores. The cosine similarity scores of the privacy policy documents in Table 3 are higher, meaning that these documents closely align with the summarized GDPR documents. In contrast, the policy documents in Table 4 are not in compliance with summarized GDPR due to their low similarity scores.

TABLE 3. Policy documents compliant with GDPR.

Policy	TF-IDF	BERT
Twilio	0.40	0.86
Atalassian Data Processing	0.48	0.84
Signaturit	0.44	0.81
Cloudinary	0.59	0.79
Azure	0.56	0.76

TABLE 4. Policy documents not compliant with GDPR.

Policy	TF-IDF	BERT
Pacer Healthcare	0.13	0.57
Adobe - Before GDPR 2018	0.14	0.56
Gimkit	0.10	0.54
Azure - Before GDPR 2018	0.04	0.48
Intellijoy	0.01	0.46

3) BERT

One of the newest approaches in language modeling utilizes a transformer architecture and the attention model to assign an embedding for each word. These embeddings are context-aware due to the bidirectional representation of the BERT. In our work, we have incorporated a pre-trained BERT model [63]–[65] for embedding the sentences of our documents in the corpus by applying a pooling operation to the output of BERT. The BERT base model consists of 12 layers of transformer blocks, 12 attention heads and 110 million parameters.

Table 3 and Table 4 shows the similarity scores of BERT embedded privacy policy documents and summarized GDPR. As shown, the context similarity scores with BERT have improved significantly when compared with tf-idf scores for the policy documents present in dataset A and dataset B.

Also, the cosine similarity scores for all the policies in dataset A scored higher than that of dataset B. This states that documents in dataset A are in compliant with GDPR than with dataset B policies. BERT score helps an organization know if their privacy policy is more likely or less likely adhering to GDPR. If the BERT score is low, it means that that organization needs to make significant changes to the privacy policy and vice versa. Our method helps an organization to perform the checks without any human intervention.

IV. CONCLUSION AND FUTURE WORK

Web service providers need to ensure that their policies comply with regulations like the EU's GDPR. These policies are short pieces of text when compared with regulation documents, and both sets of documents are available only in textual format. It currently requires significant human intervention to ensure that the policy document, e.g., a privacy policy, covers the regulations in GDPR. We have developed a novel framework using Deep Learning and Semantic Web approaches to capture the knowledge embedded in regulatory and web service policies. This knowledge is then stored in knowledge graphs corresponding to the authority and referencing documents which can be queried and reasoned over to determine if a policy is compliant with the authoritative or regulatory document. In this paper, we include the results of our study by comparing web service privacy policies with GDPR rules.

We began by identifying critical entities in the GDPR document and then used combinations of entities as GDPR classes. Next, we applied machine learning and deep learning to determine whether the GDPR class exists in a privacy policy document. After identifying the class using a multi-class text classification approach, we summarized the extracted text using BERT text summarization. Finally, we compared the summaries of authority and referencing documents to obtain a context similarity score. The extracted summary and the context similarity scores were stored in the referencing document knowledge graph. We validated our approach against five privacy policies adhering to GDPR and five that were not adhering to GDPR.

Since there is no existing reliable work or baselines for checking privacy policy compliance with GDPR, our framework will significantly benefit service providers by automatically identifying the uncovered GDPR areas in their privacy policy documents. This will help prevent organizations from running afoul of the regulations and having to pay penalties for not adhering to them. In this paper, we did stick to identifying missing GDPR areas and checking for compliance rate. As the regulation document is approximately 100 pages, whereas the privacy policies usually tend to be less than ten pages, so it would be hard to perform rule-by-rule analysis, so we used BERT summarizer to summarize the documents and check for compliance rate. However, we will consider extracting the rules as part of our future work and see their impact. Also, as part of our future work, we plan to create an automated tool that takes the results of our framework and populates the referencing knowledge graph. We also plan to work with a legal advisor to evaluate and score our results manually.

ACKNOWLEDGMENT

This work was supported in part by NSF Phase I Industry-University Cooperative Research Centers (IUCRC) UMBC: Center for Accelerated Real-time Analytics (CARTA) under NSF Award 1747724 and an award by IBM Research.

REFERENCES

[1] P. Voigt and A. von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. New York, NY, USA: Springer, 2017.
[2] (May 25, 2018). *2018 Reform of EU Data Protection Rules*. European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
[3] M. L. Jones and M. E. Kaminski, "An American's guide to the GDPR," *Denver Law Rev.*, vol. 98, p. 93, Jun. 2021.
[4] *A Breach of a Registered APP Code is an Interference With the Privacy of an Individual*, Privacy Act 1988 Pt III Div 3, Australia, 1998.
[5] *Personal Information Protection and Electronic Documents Act*, Wikipedia Contributors, Wikipedia, Web, San Francisco, CA, USA, Apr. 2021.
[6] *Online Privacy Protection Act*, Wikipedia Contributors, Wikipedia, Web, San Francisco, CA, USA, Apr. 2021.
[7] 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.
[8] D. A. Tamburri, "Design principles for the general data protection regulation (GDPR): A formal concept analysis and its evaluation," *Inf. Syst.*, vol. 91, Jul. 2020, Art. no. 101469.

[9] J. Ruohonen and K. Hjerppe, "The GDPR enforcement fines at glance," *Inf. Syst.*, 2021, Art. no. 101876.

[10] R. N. Zaeem and K. S. Barber, "The effect of the GDPR on privacy policies: Recent progress and future promise," *ACM Trans. Manage. Inf. Syst.*, vol. 12, no. 1, pp. 1–20, Mar. 2021.

[11] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surv.*, vol. 34, no. 1, pp. 1–47, 2002.

[12] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[13] V. B. Kobayashi, S. T. Mol, H. A. Berkers, G. Kismihók, and D. N. Den Hartog, "Text classification for organizational researchers: A tutorial," *Organizational Res. Methods*, vol. 21, no. 3, pp. 766–799, Jul. 2018.

[14] L. Elluri, V. Mandalapu, and N. Roy, "Developing machine learning based predictive models for smart policing," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2019, pp. 198–204.

[15] C. Basu, H. Hirsh, and W. Cohen, "Recommendation as classification: Using social and content-based information in recommendation," in *Proc. AAAI/IAAI*, 1998, pp. 714–720.

[16] F. Neri, "Sentiment analysis on social media," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2012, pp. 919–926.

[17] D. A. Simanjuntak, H. P. Ipung, C. Lim, and A. S. Nugroho, "Text classification techniques used to faciliate cyber terrorism investigation," in *Proc. 2nd Int. Conf. Adv. Comput., Control, Telecommun. Technol.*, Dec. 2010, pp. 198–200.

[18] Y. Li and A. K. Jain, "Classification of text documents," *Comput. J.*, vol. 41, no. 8, pp. 537–546, 1998.

[19] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary, "Twitter trending topic classification," in *Proc. IEEE 11th Int. Conf. Data Mining Workshops*, Dec. 2011, pp. 251–258.

[20] R. Batool, A. M. Khattak, J. Maqbool, and S. Lee, "Precise tweet classification and sentiment analysis," in *Proc. IEEE/ACIS 12th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2013, pp. 461–466.

[21] A. Mahendran, "Opinion mining for text classification," *Int. J. Sci. Eng. Technol.* vol. 2, no. 6, pp. 589–594, 2013.

[22] A. Sun, "Short text classification using very few words," in *Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2012, pp. 1145–1146.

[23] M. Chen, X. Jin, and D. hen, "Short text classification improved by learning multi-granularity topics," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1776–1781.

[24] D. J. Hemanth, J. Anitha, A. Naaji, O. Geman, D. E. Popescu, and L. H. Son, "A modified deep convolutional neural network for abnormal brain image classification," *IEEE Access*, vol. 7, pp. 4275–4283, 2018.

[25] D. J. Hemanth, J. Anitha, and L. H. Son, "Brain signal based human emotion analysis by circular back propagation and deep Kohonen neural networks," *Comput. Elect. Eng.*, vol. 68, pp. 170–180, May 2018.

[26] C. N. Giap, L. H. Son, and F. Chiclana, "Dynamic structural neural network," *J. Intell. Fuzzy Syst.*, vol. 34, no. 4, pp. 2479–2490, Apr. 2018.

[27] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 103–112.

[28] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–6.

[29] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.

[30] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 115–124.

[31] B. Quilitz and U. Leser, "Querying distributed RDF data sources with SPARQL," in *Proc. Eur. Semantic Web Conf.* Berlin, Germany: Springer, Jun. 2008, pp. 524–538.

[32] O. Lassila and R. Swick. (1999). *Resource Description Framework (RDF) Model and Syntax Specification*. [Online]. Available: wwwConsortium

[33] S. Decker, S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks, "The semantic web: The roles of XML and RDF," *IEEE Internet Comput.*, vol. 4, no. 5, pp. 63–73, Sep. 2000.

[34] J. Broekstra, A. Kampman, and F. Van Harmelen, "Sesame: A generic architecture for storing and querying RDF and RDF schema," In in *Proc. Int. Semantic Web Conf.* Berlin, Germany: Springer, Jun. 2002, pp. 54–68.

[35] D. McGuinness and F. Van Harmelen, *OWL Web Ontology Language Overview, W3C Recommendation*. Cambridge, MA, USA: World Wide Web Consortium, 2004.

[36] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, "SWRL: A semantic web rule language combining OWL and RuleML," *W3C Member Submission*, vol. 21, no. 79, pp. 1–31, 2004.

[37] S. Bechhofer, H. F. Van, J. Hendler, I. D. L. Horrocks McGuinness, P. F. Patel-Schneider, and L. A. Stein, "OWL web ontology language reference," *W3C Recommendation*, vol. 10, no. 2, pp. 1–53, 2004.

[38] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler, "OWL 2: The next step for OWL," *J. Web Semantics*, vol. 6, no. 4, pp. 309–322, Nov. 2008.

[39] M. A. Musen, "The Protege project: A look back and a look forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, Jun. 2015, doi: 10.1145/2757001.2757003.

[40] L. Elluri, A. Nagar, and K. P. Joshi, "An integrated knowledge graph to automate GDPR and PCI DSS compliance," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 1266–1271, doi: 10.1109/BigData.2018.8622236.

[41] L. Elluri, K. P. Joshi, and A. Kotal, *Measuring Semantic Similarity across EU GDPR Regulation and Cloud Privacy Policies*. Baltimore, MD, USA: UMBC Student Collection, 2020.

[42] K. I. Manktelow and D. E. Over, "Deontic reasoning," in *Perspectives on Thinking and Reasoning: Essays in Honour of Peter Wason*, 1995, pp. 91–114.

[43] S. Beller, "Deontic norms, deontic reasoning, and deontic conditionals," *Thinking Reasoning*, vol. 14, no. 4, pp. 305–341, Nov. 2008.

[44] L. T. McCarty, "Permissions and obligations," in *Proc. IJCAI*, vol. 83, Aug. 1993, pp. 287–294.

[45] A. Graves, "Generating sequences with recurrent neural networks," 2013, *arXiv:1308.0850*.

[46] D. S. Sachan, M. Zaheer, and R. Salakhutdinov, "Revisiting LSTM networks for semi-supervised text classification via mixed objective function," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6940–6948.

[47] X. Bai, "Text classification based on LSTM and attention," in *Proc. 13th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Sep. 2018, pp. 29–32.

[48] A. Graves, N. Jaitly, and A.-R. Mohamed, "Hybrid speech recognition with deep bidirectional LSTM," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 273–278.

[49] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Apr. 2019.

[50] G. Xu, Y. Meng, X. Qiu, Z. Yu, and X. Wu, "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522–51532, 2019.

[51] Q. Wang, P. Liu, Z. Zhu, H. Yin, Q. Zhang, and L. Zhang, "A text abstraction summary model based on BERT word embedding and reinforcement learning," *Appl. Sci.*, vol. 9, no. 21, p. 4701, Nov. 2019.

[52] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," 2015, *arXiv:1509.00685*.

[53] S. Narayan, S. B. Cohen, and M. Lapata, "Ranking sentences for extractive summarization with reinforcement learning," 2018, *arXiv:1802.08636*.

[54] D. Miller, "Leveraging BERT for extractive text summarization on lectures," 2019, *arXiv:1906.04165*.

[55] A. Agrawal and U. Gupta, "Extraction based approach for text summarization using k-means clustering," *Int. J. Sci. Res. Publications* vol. 4, no. 11, 2014, pp. 1–4, 2014.

[56] Wikipedia Contributors. (Jan. 20, 2021). *SPARQL*. Wikipedia, The Free Encyclopedia. Accessed: Apr. 4, 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=SPARQL

[57] B. DuCharme, *Learning SPARQL: Querying and Updating With SPARQL 1.1*. Sebastopol, CA, USA: O'Reilly Media, 2013.

[58] M. Schmidt, M. Meier, and G. Lausen, "Foundations of SPARQL query optimization," in *Proc. 13th Int. Conf. Database Theory (ICDT)*, 2010, pp. 4–33.

[59] F. Hogenboom, V. Milea, F. Frasincar, and U. Kaymak, "RDF-GL: A SPARQL-based graphical query language for RDF," in *Emergent Web Intelligence: Advanced Information Retrieval*. London, U.K.: Springer, 2010, pp. 87–116.

[60] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *Proc. 4th Int. Conf. Cyber IT Service Manage.*, Apr. 2016, pp. 1–6.

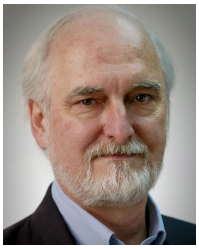
[61] J. V. R. Murthy, "Clustering based on cosine similarity measure," *Int. J. Eng. Sci. Adv. Technol.*, vol. 2, no. 3, pp. 508–512, 2012.

[62] M. Alodadi and V. P. Janeja, "Similarity in patient support forums using TF-IDF and cosine similarity metrics," in *Proc. Int. Conf. Healthcare Informat.*, Oct. 2015, pp. 521–522.
[63] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.
[64] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.
[65] X. Ma, Z. Wang, P. Ng, R. Nallapati, and B. Xiang, "Universal text representation from BERT: An empirical study," 2019, *arXiv:1910.07973*.

LAVANYA ELLURI (Member, IEEE) received the master's degree in management information systems from the University of Houston–Clear Lake. She is currently pursuing the Ph.D. degree in information systems (IS) with the University of Maryland, Baltimore County. She is also a Senior Database Engineer at REI Systems Inc., Sterling, VA, USA. Her research interests include data science, cloud computing, data security and privacy, semantic web, and text mining.

SAI SREE LAYA CHUKKAPALLI (Graduate Student Member, IEEE) received the B.Tech. degree in computer science from the Gandhi Institute of Technology and Management University, India. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Electrical Engineering, UMBC. Her research interest includes the intersection of artificial intelligence and cybersecurity for enhancing security and privacy in cyber-physical systems.

KARUNA PANDE JOSHI (Senior Member, IEEE) received the bachelor's degree in computer engineering from the University of Mumbai, India, and the M.S. and Ph.D. degrees in computer science from UMBC. She has extensive experience of working in the industry primarily as the IT Program/Project Manager at the International Monetary Fund. She is currently an Associate Professor of information systems and the UMBC Director of the Center for Accelerated Real-Time Analytics (CARTA). She also directs the Knowledge Analytics Cognitive and Cloud (KnACC) Laboratory. She teaches courses in big data, database systems design, and software engineering. She has published over 50 articles. Her research interests include data science, cloud computing, data security and privacy, and healthcare IT systems. Her research was supported by ONR, NSF, DoD, GE Research, and Cisco. She was twice awarded the IBM Ph.D. Fellowship from UMBC.



TIM FININ (Member, IEEE) received the degree from MIT and the University of Illinois at Urbana–Champaign. He has over 40 years of experience in applying AI to problems in information systems and language understanding. He has held positions at UMBC, Unisys, the University of Pennsylvania, and the MIT AI Laboratory. He is currently the Willard and Lillian Hackerman Chair of engineering and a Professor of computer science and electrical engineering with UMBC. His current research interests include knowledge graphs, analyzing and extracting information from text, and cybersecurity. He is a fellow of ACM and AAAI and was a recipient of the IEEE Technical Achievement Award.

ANUPAM JOSHI (Fellow, IEEE) received the B.Tech. degree from IIT Delhi, in 1989, and the M.S. and Ph.D. degrees from Purdue University, in 1991 and 1993, respectively. He is currently an Oros Family Professor and the Chair of the Department of Computer Science and Electrical Engineering, University of Maryland at Baltimore County (UMBC). He is also the Director of the Center for Cybersecurity, UMBC. He has published over 250 technical articles, filed and been granted several patents, and has received research support from NSF, NASA, DARPA, the U.S. DoD, NIST, IBM, Microsoft, Qualcom, Northrop Grumman, and Lockheed Martin. His primary focus has been on data management and security/privacy in mobile/pervasive computing environments, and policy-driven approaches to security and privacy. His research interests include the broad area of networked computing and intelligent systems.