Preprint: In proceedings of 7th International Workshop on Privacy and Security of Big Data (PSBD 2020), in conjunction with 2020 IEEE International Conference on Big Data (IEEE BigData 2020)

Measuring Semantic Similarity across EU GDPR Regulation and Cloud Privacy Policies

Lavanya Elluri
Information Systems Department
University of Maryland, Baltimore County
Baltimore, MD, USA
lelluri1@umbc.edu

Karuna Pande Joshi
Information Systems Department
University of Maryland, Baltimore County
Baltimore, MD, USA
karuna.joshi@umbc.edu

Anantaa Kotal
C.S.E.E.
University of Maryland, Baltimore County
Baltimore, MD, USA
anantak1@umbc.edu

***Abstract*—Data protection authorities formulate policies and rules which the service providers have to comply with to ensure security and privacy when they perform Big Data analytics using users Personally Identifiable Information (PII). The knowledge contained in the data regulations and organizational privacy policies are typically maintained as short unstructured text in HTML or PDF formats. Hence it is an open challenge to determine the specific regulation rules that are being addressed by a provider's privacy policies. We have developed a semantically rich framework, using techniques from Semantic Web and Natural Language Processing, to extract and compare the context of a short text in real-time. This framework allows automated incremental text comparison and identifying context from short text policy documents by determining the semantic similarity score and extracting semantically similar key terms. Additionally, we also created a knowledge graph to store the semantically similar comparison results while evaluating our framework across EU GDPR and privacy policies of 20 organizations complying with this regulation associated with various categories apply to Big Data stored in the cloud. Our approach can be utilized by Big Data practitioners to update their referential documents regularly based on the authority documents.**

***Index Terms*—General Data Protection Regulation; Document Similarity; Text Extraction; Semantic Web; Ontology; Big Data Categories; Organizations**

I. Introduction

Authorities have power or control in a domain that they monitor and administrate. For instance, law enforcement authorities, health authorities, bureaucracy, etc. exist to maintain law and order or provide a service to society. To make sure the efficient and secure operation of their domain, authorities formulate policies and rules governing the area in which the other organizations and individuals, operating in that sphere, must comply. These rules are available as large volume text documents and are often referred to in their documents by companies operating in that sphere. For instance, cloud privacy policies often include information in authority documents like Data Protection Regulations. However, this referred text is often short and concise, and the users have difficulty in extracting the appropriate context from these short paragraphs.

Organizations are analyzing large consumer datasets to determine behavior patterns related to market trends, fraud detection, or predicting consumer loyalty. Along with stated data, this analysis also uses obtained or inferred data and includes consumer's Personally Identifiable Information (PII) data. Moreover, prompt embracing of Cloud computing for big data analytics has also resulted in a large volume of PII data being handled and transmitted across the Internet. Security and Privacy of stated PII managed by vendors are of key concern to consumers. As a result, regulatory bodies throughout the sphere are releasing new data protection laws, like European Union's General Data Protection Regulation (EU GDPR) [40], that must be adhered to by Big Data Providers and Consumers. This surge in data protection regulations has resulted in vast legal compliance challenges of Big Data. GDPR [38] [39] [40] [41] identifies the responsibilities and policies for organizations using any EU customer data for their analytics. The GDPR has 99 articles in approximately 90 pages that categories rules for controller and processor. Also, it specifies the common obligations that both controller and processor should adhere to. GDPR covers all the rules that apply to any PII data related to EU individuals connected to their private, professional, or public life. It includes personal and sensitive information like person name, address, racial or ethnic origin, genetic data, biometric data, medical records, social media posts, sexual orientation.

GDPR and privacy policies that are committed to this regulation are accessible only in textual format and require considerable human time and effort to ensure compliance and thereby prevent data breaches. We have developed a semantically rich framework to extract and compare the context of a short text in real-time. This framework was built using techniques and technologies from Artificial Intelligence including Semantic Web, Text Extraction and Natural Language Processing. It allows automated incremental text comparison and identifying context from short text policy documents by determining the semantic similarity score and extracting semantically similar key terms. Additionally, we envision a semantically rich, machine-processable knowledge graph (or ontology) that captures the compliance comparison scores of the major organizational categories that apply to Big Data on the Cloud; this will substantially help in automating an organization's data compliance practices. In addition to saving compliance comparison scores and organizational resources dedicated to compliance adherence after developing privacy policy, it will also help to check the privacy policies while they are in development as well. By validating the privacy policies

while they are still in development helps the organizations not to pay huge penalties when the breach happens. As part of GDPR, if a data breach happens the organization has to pay fines of up to 10 million euros, or, in case of an undertaking, up to 2% of its entire global turnover of the preceding fiscal year, whichever is higher [40]. Another advantage of building this knowledge graph is that finding potential key terms that are in both regulation and privacy policies will be quicker. As a first step towards this idea of a data compliance comparison across the privacy policies knowledge graph, we have collected policies of twenty organizations that deal with Big Data and are committed to GDPR. We then retrieved the similarity scores and extracted the semantically similar key terms for validation. This Knowledge Graph is available in the public domain and can be used by Big Data practitioners to cross-check their compliance rate with GDPR.

The rest of this paper has been organized as follows: Section II covers the Related Work. Section III describes the methodology of our framework and the Knowledge graph that we have created to store the rules and semantically similar concepts. It also includes the details of our experimental results. Section IV describes the validation technique used to support our study. We conclude the paper in section V.

II. Related Work

Developing a compliance comparison knowledge graph, determining the similarity scores, and extracting the semantically similar keywords from short text are three major steps involved in this work. Therefore, we have investigated the papers that involve retrieving the similarity scores while comparing the privacy policy with any regulation and developing knowledge graphs for privacy policies. While researchers have used methods that are applied to solve extracting the rules from the privacy policies, they are not efficient in analyzing to what extent the policy is missing to address the regulation. To address this challenge and help several consumers and providers that deal with Big Data, we have come up with a methodology that helps in extracting the context and store the results obtained in a publicly available knowledge graph so that it is easy to refer to anyone throughout the globe and make decisions.

A. Semantic Web

In a services world, consumers and providers need to be competent to swap knowledge and requests with some assurance that they disclose a common meaning. This is crucial not only for the data but also for the data protection regulations monitored by consumers and providers. The processing of various policies is typically not present in a closed or centralized location but is a concern in the cloud environment. The requirement is not just for the data itself, but even for describing policies for data protection.

One possible method to this issue is to employ Semantic Web techniques for modeling and reasoning about data protection policies. We have used this approach to develop our knowledge graph. The Semantic Web-primarily deals with the data instead of documents. It allows data to be annotated with machine-understandable meta-data, allowing the automation of their retrieval and their usage in incorrect contexts. Semantic Web technologies include languages such as Resource Description Framework (RDF) [32] and Web Ontology Language (OWL) [8] for defining ontologies and illustrating meta-data using these ontologies as well as tools for reasoning over these descriptions. These technologies can be used to support common semantics of privacy policies enabling all agents who understand basic Semantic Web technologies to transmit and use each other's Services and data efficiently.

In our prior works, we developed an integrated Knowledge graph [5] to capture various regulations and also populated the Ontology with only a very few privacy policies that apply to Big Data. We extracted the rules based on the keywords listed in the glossary or appendix of any regulation. Also, we have developed ontologies to represent legal documents about cloud data like Service Level Agreements [33] and Data Privacy policies [2] [5]. As part of previous work we associated the rules to CSA controls [1] [4]. We also did a qualitative analysis of privacy policies based on their semantic vagueness [31]. We are now extending this work to check for the semantic similarity between GDPR and privacy policies. Also, we extracted the semantically similar key terms as part of this methodology which it helps with accurate validation. As part of this work, we mainly focused on GDPR as it has become crucial for all the organizations that deal with Big Data and penalties are massive in case of a data breach.

B. Text Extraction

Natural Language Processing techniques are popular with researchers to automate the extraction of relevant information from a large corpus of text documents. In the research, Rusu et al. [44] the authors suggested the technique to extract the information and relevant phrases in the form of subject-predicate-object triples. To do so, Parse Trees are generated from English sentences, and triplets were extracted from the parse trees [33] [44]. In the research work of Etzioni et al. [43], the author developed the KNOWITALL system which helped in the automation of extracting large collections of facts from the Web in an unsupervised, domain-independent, and scalable manner [44]. The author used the approach of Pattern Learning to address this challenge [33] [44]. Another important NLP technique applied for information extraction from unstructured text is 'Noun Phrase Extraction'. Author Rusu et al. in [44] showed the technique of creating triplets by considering 'Noun Phrases' obtained via various part-of-speech taggers. Different automated techniques have been used for extracting the permissions and obligations from legal documents [33]. Techniques such as text mining and semantic techniques have been explored and applied by various authors in the past [33] [34] [35]. In the research work of Kagal et al. [36] [37], the authors proposed an ontology-based framework to model conversation specifications and policies using obligations and permissions [33] [36] [37].

Category	Privacy Policies
Cloud Edge Service	Algolia, Upload care
Customer Relationship Management	Intercom, Support Bee
eCommerce	eBay, Recapture
Email Service Provider	Drip, SendGrid
File Sharing	Box, Dropbox
Infrastructure hosting	AWS, Digital Ocean
Manage Cloud	Pax8, Rackspace
Payments	FastSpring, PayPal
Search Engine	Google, Microsoft Bing
Social Media	Facebook, Snapchat

TABLE I: List of Big Data categories and Privacy Policies

C. Key components of GDPR

As part of our previous work [3] [5], we identified the key classes of a knowledge graph to represent the GDPR rules. We have referenced the GDPR available at [39] [40] for this. The regulation splits the responsibilities and obligations of consumers and providers, obligating consumers and providers that provide enough guarantees to implement appropriate technical and organizational measures to meet the regulation's policies and safeguard data subject's PII [39].

1) *Consumer is mainly obligated to below regulations:* Notifying about personal data Breach to Supervisory Authority and communicating about a personal breach to the data subject. Also, the controller is responsible for carrying the Data Protection Impact Assessment (DPIA). Consumers must Consult Supervisory Authority before processing if DPIA shows high risk. Besides, the controller should appoint a Data Protection Officer (DPO) while processing personal data on a large scale.
2) *Provider is mainly obligated to below regulations:* Support consumers during the data breaches and processing data as per consumer instructions. Maintaining records of all processing activities should be taken by the processor. The processor must provide sufficient data security and implement Privacy by Design / Default. Assisting consumers in DPIAs review / for risk processing and removing all the personal data after the end of the provision. Cooperating with the Supervisory Authority as needed and comply with certification requirements. Attain consent if any services are offered to a child company and follow the requirements for appointing and acting as a DPO. Also, the provider should fulfill the data subject rights and comply with the rules while transferring the data outside the EU.

III. Methodology

In this section, we illustrate our methodology to determine the document similarity score, extract semantically similar terms, develop an ontology to populate the result, and validate the results. We aim to determine the semantic similarity score and develop a semantically rich policy-based knowledge representation for GDPR comparison with the organizational policies. We used the Doc2Vec approach to get the similarity score in radians, extracted semantically similar keywords using NLP techniques, and created the Ontology using Protégé tool [45]. Our methodology is divided into four different phases after obtaining the controller, processor, and common obligations from GDPR repository and collecting privacy policies that are committed to GDPR, which fall under ten different categories that apply to Big Data stored in the cloud. We have selected twenty privacy policies for our analysis for the ease of representation. Fig. 1 is the representation of our architecture flow. The four phases of our methodology are:

- **Preprocessing stage:** After converting the source files from HTML/PDF formats to text files, for both the regulations and privacy policies, we have applied tokenization, stemming and lemmatization NLP word processing techniques and extracted all the appropriate keywords by avoiding stop words.
- **Determine Similarity Score:** In this stage, we check for the document similarity scores for all the twenty privacy policies and analyze the results of ten categories that we considered.
- **Knowledge Graph/Ontology Development:** We have created a comprehensive Data Compliance Comparison ontology that captures the similarity scores and semantically similar keywords details of both the regulation and organizational privacy policies. Detailed information is in subsection D. For creating the knowledge graph, we utilized the Protégé [45].
- **Validation:** We validated the knowledge graph by using the semantically similar keywords extracted from the privacy policies and compared them with the scores obtained. Policies that have a high score in radians should have less related keywords and vice versa. We validated across the twenty publicly available organization policies dealing with Big data PII.

A. Preprocessing stage

In the initial stage of our system, we went through all the privacy documents that were collected and segregated into the top 10 categories that apply to Big Data. Also, this process helped us to choose 20 organizations that are committed to GDPR and updated their privacy policy. The list of Big Data categories and the organizations that we choose are shown in TABLE I. As part of previous work [4], we already extracted the rules that apply to controllers and processors. We now extracted all the semantically similar key terms by applying NLP techniques from GDPR and considered the top 100 similar keywords to check across the privacy policies. The preview of all the extracted words is shown in Fig. 2.

To have a clear understanding of the key terms that were repeated several times in the regulation, we read all the 99 articles listed in GDPR and understood the purpose and scope of these keywords. The definition of these words is shown

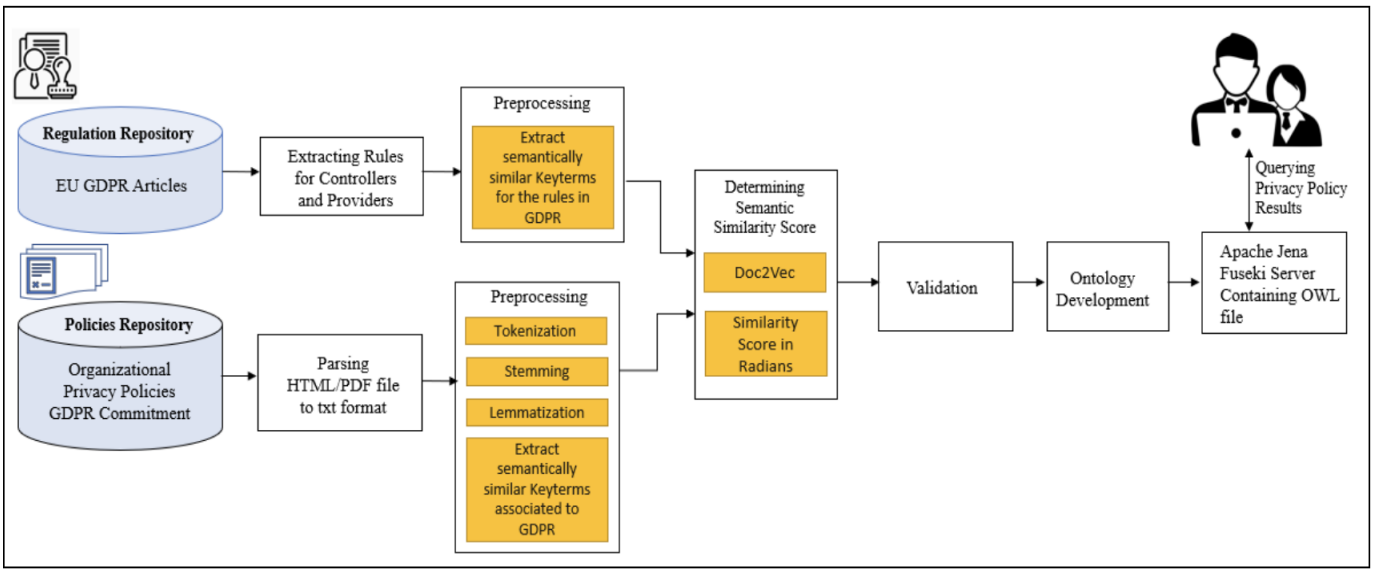


Fig. 1: Overall Architecture Flow

Fig. 2: Semantically Similar key terms from GDPR

Keywords	Definition
Data Subject	Indicates a person about whom Personal Information may be processed.
Controller	Represents a legal service provider alone or jointly decides the objectives processing of personal data.
Processor	Represents an agency alone or jointly decides the objectives to handle personal data on behalf of the controller.
Supervisory Authority	It is an agency in the EU country that is accountable for support with GDPR.
Data Protection Officer	DPO is the security leadership role required by the GDPR. They are responsible for administering a company's data protection policy and for making sure compliance with GDPR requirements.
Consent	Consent means a data subject has clearly stated and indicates agreement to the processing of personal data related to them.
Sensitive Information	Signifies the Data Subject's racial or ethnic origin, religious or philosophical beliefs, genetic data, biometric data, health data, or sexual orientation.
Process or Processing	Indicates an action performed on Personal Information of Data Subject.
Notification	This means that the controller has to inform the Supervisory Authority in case of any data breach.
Profiling	It is defined as automatic processing of data to analyze or to make predictions about individuals.
Article or Art	In GDPR, rules that apply to controllers and processors are listed in 99 articles.
Personal Data	Any information which is related to an identifiable natural person.

TABLE II: List of keywords and definitions

in TABLE II. These words play a vital role while we extract these words from privacy policies.

B. Determining similarity score and Key terms Extraction from Privacy Policies

For the corpus of organizational privacy policies that are committed to GDPR, we have determined the similarity scores using the document similarity approach by comparing it with GDPR, which has 99 Articles in approximately 90 pages [40]. We compared ten different categories of the privacy policies that captured user information and retrieved the similarity score in radians. Scores that are less in radians mean that the document is more in compliance with the regulation, which in our case, is EU GDPR. The reason for retrieving the scores for various categories is because we are curious to know what kind of organizations are more committed to the EU GDPR. Based on the results obtained, in Fig. 3, we can see that the categories managing Cloud and search engines have less score in radians that is 0.59 and 0.55, which means that these categories are adhering to the regulation much better than the other eight categories.

On the other hand, the file-sharing category has a similarity score of 0.91, which indicates less adherence to GDPR. Remaining seven categories (Social Media, Payments, Infrastructure hosting, Email Service Provider, eCommerce, Customer relationship management (CRM) and Cloud Edge Service have an average score of approximately 0.7 radians, indicating that categories are adhering much better than File Sharing category.

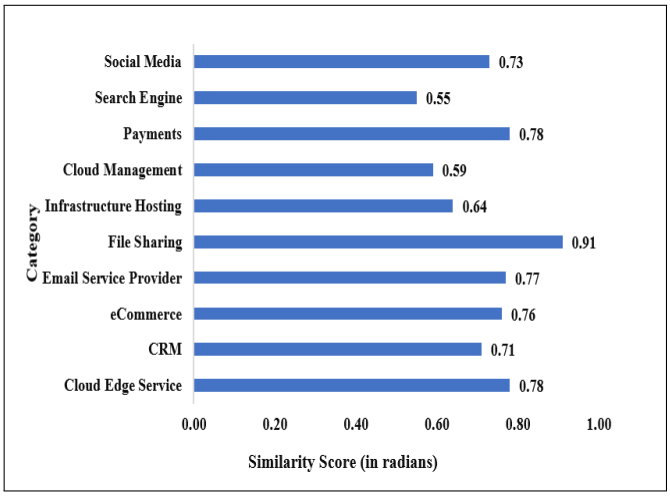


Fig. 3: Average similarity scores of categories

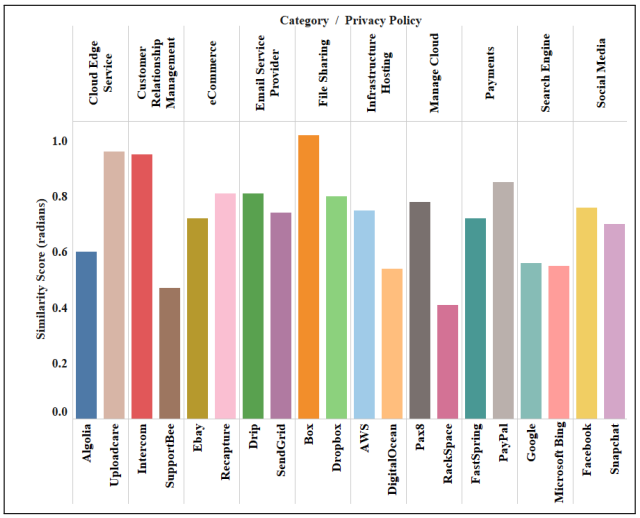


Fig. 4: Categories vs Privacy Policies similarity scores

After analyzing the results, we were curious to know if the policies under each category have scores close to average or not. Therefore, we created a graph as shown in Fig. 4 to analyze categories vs. privacy policy scores. Interestingly, five out of ten categories (Search Engine, Social Media, Email Service Provider, eCommerce, Payments) have individual privacy policies scores close to the category averages. The other five categories (Cloud Edge Service, CRM, File Sharing, Managing Cloud, Infrastructure hosting) have policy scores that are far away from their category averages. It means categories that carry massive data have their Privacy policies strictly adhering to GDPR. It is good to know that these categories are strictly following the regulation rules. Similarity scores obtained for all the privacy polices that are used as part of this work are shown in TABLE III.

Our results are helpful for organizations that are not following GDPR rules. Therefore, these companies can revisit the regulation to add the missing rules in their privacy policies. It will ensure to the end-users that they can rely on these organizations. This methodology can be applied while the privacy policy is still in the development stage before publishing it. By doing so, organizations can be confident enough for the fact that they are addressing and adhering to the regulation and are clearly stating the rules in the policies.

For GDPR, we also extracted the relevant key terms from the repository of controller processor. As defined in the EU GDPR [39] [40] repository, many key phrases should be taken into consideration when an organization is updating its privacy policy under GDPR. We have used Python language to develop the code for extracting the semantically similar keywords from the vast corpus of GDPR and privacy policies. In the code, we made a list of stop words that are not needed and are irrelevant to our context. Also, we did make sure that certain modal verbs like will', 'may' ,'should', 'can', 'must', 'could', 'shall' were not part of the stop words list since these words act towards defining permission obligations expressions. This approach helped us in separating any irrelevant terms. TABLE IV shows the list of generated semantically similar key terms from policies.

Category	**Privacy Policy**	**Similarity Score (in radians)**
Cloud Edge Service	Algolia [14]	0.60
Cloud Edge Service	Uploadcare [13]	0.96
CRM	Intercom [15]	0.95
CRM	SupportBee [29]	0.47
eCommerce	Ebay [28]	0.72
eCommerce	Recapture [27]	0.81
Email Service Provider	Drip [10]	0.81
Email Service Provider	SendGrid [26]	0.74
File Sharing	Box [25]	1.02
File Sharing	Dropbox [11]	0.80
Infrastructure hosting	AWS [22]	0.75
Infrastructure hosting	DigitalOcean [24]	0.54
Manage Cloud	Pax8 [16]	0.78
Manage Cloud	RackSpace [23]	0.41
Payments	FastSpring [18]	0.72
Payments	PayPal [17]	0.85
Search Engine	Google [12]	0.56
Search Engine	Microsoft Bing [21]	0.55
Social Media	Facebook [19]	0.76
Social Media	Snapchat [20]	0.70

TABLE III: Similarity scores of Privacy Policies

C. Permission & Obligations

Modal logic is a broad term used to include various other forms of logic, such as temporal logic and deontic logic [46]. Deontic logic statements including permissions and obligations illustrate time-based requirements. Deontic logic consists of four types of modalities:

Keyword	Occurrence
art	916
compliance	430
consent	178
controller	208
data protection officer	6
data subject	102
notification	74
personal data	822
power	20
processor	614
profiling	14
protection	528
rights	246
subject	168
supervisory	32
third	250

TABLE IV: Semantically similar GDPR Key terms from the organizational Privacy Policies

1) **Permissions / Rights:** Permissions are statements or rules that define the rights or consents for an entity.
2) **Obligations:** Obligations statements are the compulsory actions that an entity must accomplish.
3) **Dispensations:** Dispensations that refer to optional expressions and describe non-mandatory conditions.
4) **Prohibitions:** Prohibitions are the phrases that specify the actions which are prohibited.

To identify the data protection policies in terms of Permissions and Obligations, we extracted certain modal keywords like 'will', 'should', 'could', 'shall', 'must' 'can', 'could', 'shall', etc. These modal verbs assisted us in determining whether the sentence is to be categorized as a permission or an obligation. These permissions and obligations decide how the policies in GDPR affect consumer, provider, and end-user. Fig. 5 shows the aggregated counts of all the extracted modal from 10 categories. Type of modal verb and the occurrence in the privacy policies is listed in TABLE V.

We have used permissions obligations to separate sentences into any one of the categories. Sentences that have verbs like 'could' 'will','may', 'can' were considered as Permissions and sentences having verbs like 'must','shall', 'should' were categorized as Obligations. Below are some examples of our context:

Permissions (Rackspace):

"Customer may be either of the following (a) a Controller of Customer Personal Data, or (b) a Processor when it Processes Customer Personal Data on behalf of its End-users." [23].

Obligations (Rackspace):

"The duration of the Processing shall be from the date of this Addendum (or, if later, from the date Customer Personal Data is first Processed through the provision or use of the Services) until the Agreement expires or terminates in accordance with its terms" [23].

Verb Type	Modal Verb	Occurrence
Permission	can	179
Permission	could	10
Permission	may	241
Obligation	must	47
Obligation	should	37
Dispensation	might	10
Dispensation	shall	219

TABLE V: Frequency of occurrence of Modal verbs in privacy policies repository

Permissions (AWS):

"Customer agrees that AWS may use sub-processors to fulfill its contractual obligations under this DPA or to provide certain services on its behalf, such as providing support services." [22].

Obligations (AWS):

"The liability of the subprocessor shall be limited to its own processing operations under the Clauses." [22].

Permissions (Microsoft Bing):

"Give data subjects a copy of their personal data, together with an explanation of the categories of their data that are being processed, the purposes of that processing, and the categories of third parties to whom their data may be disclosed." [21].

Obligations (Microsoft Bing):

"We must implement the appropriate technical and organizational measures to assist you in responding to requests from data subjects exercising their rights as discussed above." [21].

D. Ontology Development

To build the data compliance comparison knowledge graph, we used Protégé software. This Ontology captures semantic similarity details between regulations and privacy policies. As part of this paper, we mainly focused on GDPR and took privacy policies that have been updated with GDPR rules. Fig. 6 demonstrates the high-level representation of all the classes. In our previous work, we developed the integrated semantically rich ontology [47] to capture various Data protection regulations and associated them with cloud threat and CSA controls [4]. We had manually identified the key terms of consumers, providers, and common obligations of various regulations like GDPR, PCI-DSS, HIPAA, MPAA, etc.

In this study, we automated the process of extracting semantically similar key terms from the regulations and privacy policies. We enhanced the integrated semantically rich ontology [47] to include the information on semantic similarity between organizational policies and GDPR. We checked for the document similarity scores with GDPR and captured all the metrics in the ontology. This Knowledge Graph that is available in the public domain [48] can be used to automate

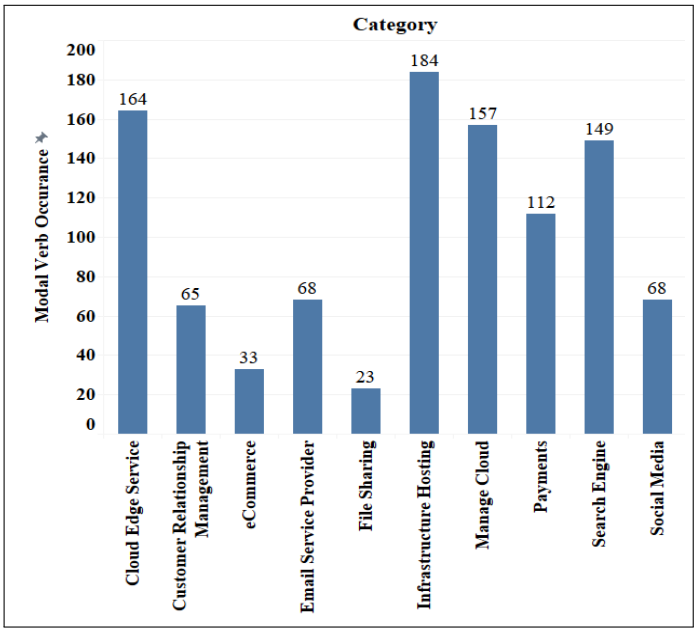


Fig. 5: Modal Verbs Occurrence in Big Data privacy policies Categories

Privacy Policy Name	Semantically Similar Key Terms	Keywords Occurrence
Dropbox	protection	38
Dropbox	compliance	28
Dropbox	personal data	8
Dropbox	law	6
Dropbox	information	10
Dropbox	controller	14
Dropbox	processor	10
Dropbox	legal	18
Dropbox	rights	8
PayPal	art	10
PayPal	controller	10
PayPal	data Subject	40
PayPal	personal data	24
PayPal	processor	10
PayPal	protection	12

TABLE VI: Dropbox and PayPal Semantically Similar Key terms associated with GDPR

data protection compliance in an organization significantly. The core classes of our knowledge graph are:

- The **Stake Holders** class is the main class that represents the crucial organizations that are affected by the rules of various regulations. This class has three main subclasses; they are regulators, consumers and providers.
- The **Regulators** class represents the authorities who generate regulations to protect user's data.
- The **Consumers** class represents the end users of the organization.
- The **Providers** class represents the organizations that have a responsibility of protecting consumer privacy.
- The **Regulations** class represents the regulations developed by Regulators to protect user's data. In Fig. 7a, we have shown a few examples of the most important regulations, like GDPR for European users, PCIDSS for mobile defense, GLBA for financial institutions, FERPA for student's information. Both Consumers and Providers Providers must comply with the regulations.
- The **Organization Privacy Policies** represents the privacy policies developed by the Providers outlining their privacy practices. Organizations have a set of rules that they have to adhere to for GDPR. Therefore, they should update their Privacy policies to show their commitment to GDPR. This class stores the details of the organizations that updated their policies to address GDPR rules. Instances of this class are twenty Privacy policies that we have chosen and are listed in the TABLE I. We have included the instances of the Organization Privacy Policies class in Fig. 7b.
- **Regulation Key Terms**: This class represents the semantically similar words for the key terms of the regulation. Also, we are capturing the occurrence of these words. Data Property hasRegulationKeyterms is linked to the class Regulations.
- **Privacy Policy Key Terms**: This class contains semantically similar words that are associated with the regulation. We only check for the similar words associated with the key terms extracted from the regulation. This class is associated with Organizations using the data property hasPolicyKeyterms.
- **Semantic similarity**: This class has regulation and the policy name that is measured for the similarity. Also, we populated the scores, key terms, and the count of occurrences. In TABLE VI we have shown the occurrence of similar keywords only for paypal and dropbox. Due to space constraint we have listed only two policies in the table. Fig. 8 shows the heatmap of all the categories and privacy policies. Size of the circle indicated the occurrence of the keyword in the respective organization policy document.

IV. Validation

As part of validation, we referenced data policies of major big data providers that have access to their customer's PII data. These include the ten categories and twenty US-based organizations that are listed in TABLE I. To validate, we compared the semantically similar key terms frequency from the privacy policies with the determined similarity score when associated with GDPR. Fig. 9 shows the summary of the organization policies used for validation with word counts and scores. We wanted to verify if generated semantically similar key terms and the scores obtained are appropriate to

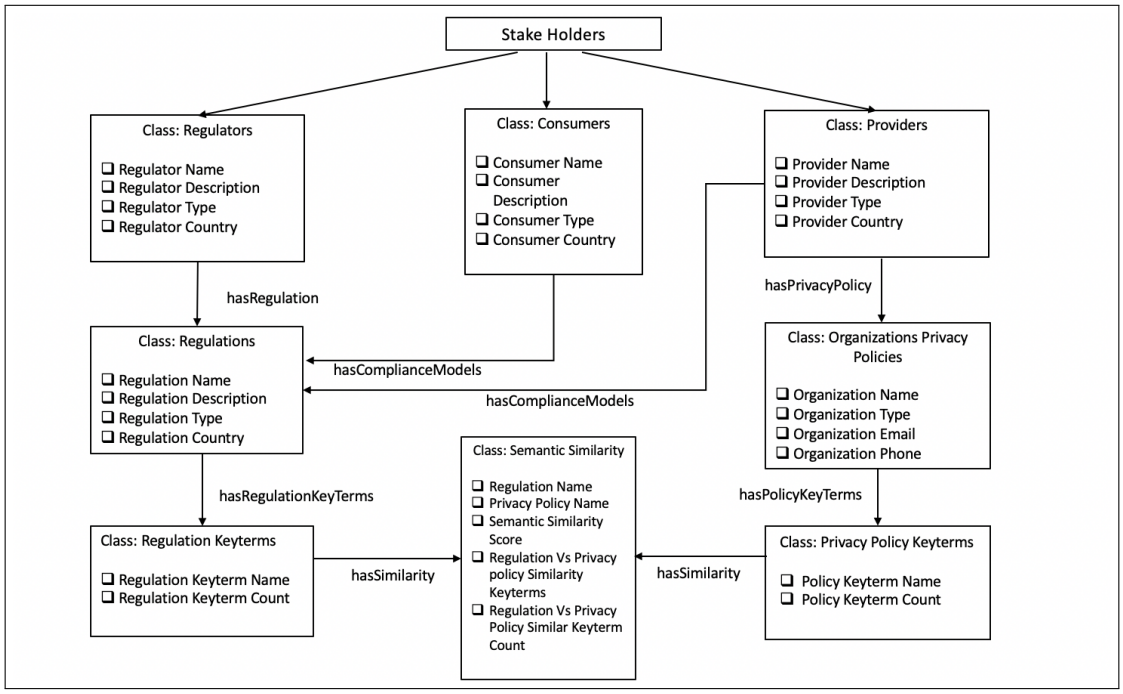
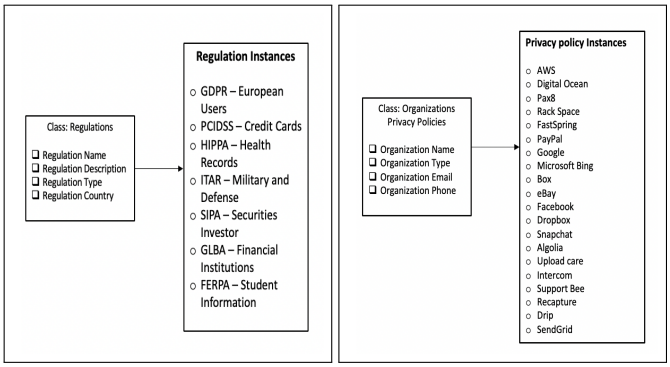


Fig. 6: Ontology for GDPR compliance comparison with Organizational Privacy Policies



(a) Instances of Regulations (b) Instances of Privacy Policies

Fig. 7

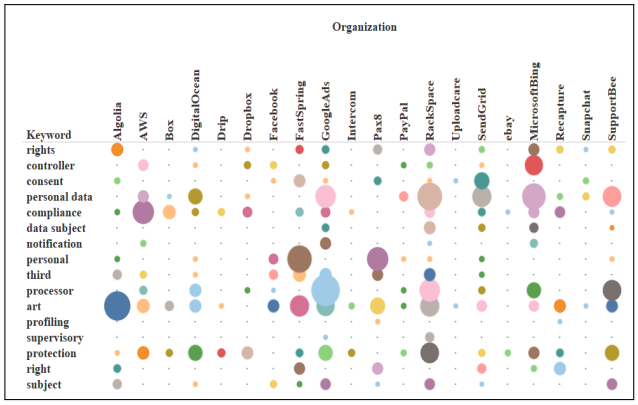


Fig. 8: Heatmap for privacy policies Vs. Semantically similar keywords

be populated as instances of our data compliance comparison across privacy policies knowledge graph.

After downloading the publicly available privacy policies that committed to GDPR and those that fall under ten different categories, we applied them to the pre-processing tools and the mechanisms to extract the semantically similar key terms and the similarity score. We have generated a trend comparison graph to see if the score in radians is going in the opposite direction to the trend of key terms. Policies that have several semantically similar key terms will have fewer radians and vice versa. We can see from the graph that the number of times a similar key term appeared in the organizational policies is inversely proportional to the score. By using this analysis, the organization's policies were populated as instances of our knowledge graph. This approach ensures that the GDPR committed data policies are now accessible as the RDF graph and are machine-processable.

This will now make it feasible to quickly check for the compliance score when compared with the actual rules listed in the regulation. Organizations can reduce human labor to mainly check for the rules that they are addressing in their privacy policies and can now quickly identify how the overall context is interpreted by the consumer. Based on that, they can revisit their policy to improvise and address most of the rules in their context. The automated approach can alert the consumers in case of any potential compliance violation.

V. Conclusions and Future Work

Organizations throughout the globe are updating their privacy policies to show commitment to GDPR compliance and are releasing a revised version of their policies by adding the

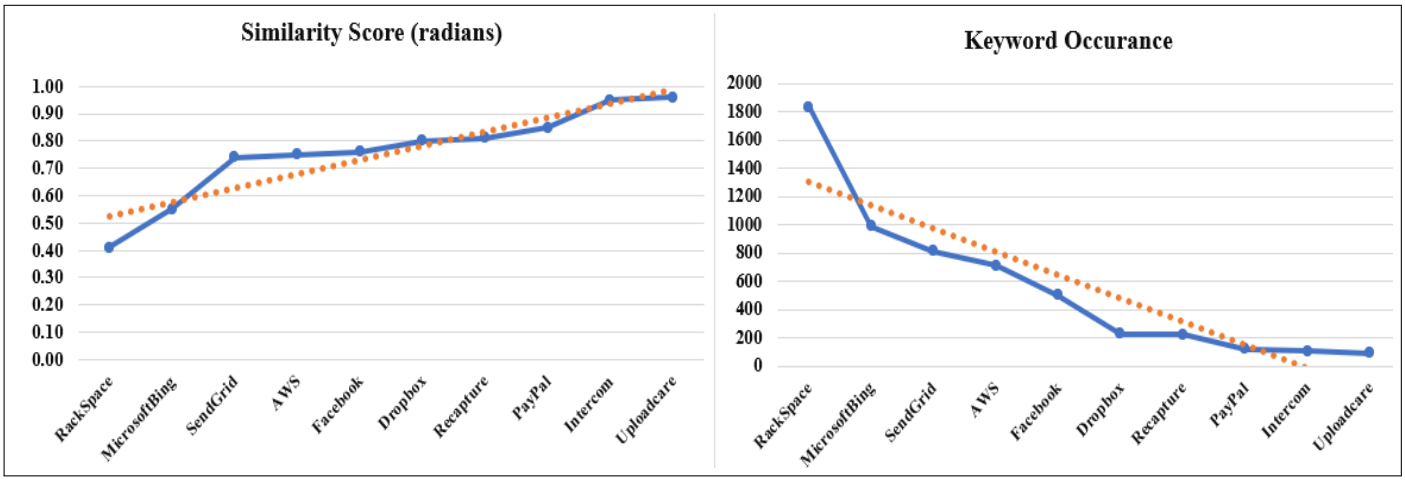


Fig. 9: Summary of document similarity scores vs semantically similar keywords occurrence

context related to GDPR. These privacy policies are so huge and are currently available only in textual format. It requires a substantial amount of human labor and effort to ensure compliance about context in actual regulation. We foresee that a semantically rich, machine-processable knowledge graph that captures the privacy policies, as they relate to Big Data on the Cloud, will substantially help in automating an organization's policy that is updating with any new regulation.

We have also identified the semantically similar keywords that are associated with GDPR regulation. We have used Natural Language Processing (NLP), text mining, and Semantic Web technologies to develop this approach. In this paper, we described the methodology used to determine the semantic similarity scores and keywords occurrences, and also we developed a semantically rich compliance comparison knowledge graph to capture the results obtained. We have validated this methodology against the ten different categories of major 20 vendors that deal with Big Data. Our methodology and the knowledge graph will help Big Data practitioners to get a clear view of the rate of the similarity and number of articles that were addressed in their privacy policy document. As part of future work, we aim to extract the context of the semantically similar keywords and classify them with the topics generated from the GDPR.

References

[1] Alliance, C. S. "Cloud Security Alliance Warns Providers of 'The Notorious Nine' Cloud Computing Top Threats in 2013."

[2] Karuna Pande Joshi, Yelena Yesha, and Tim Finin, "Automating Cloud Services Lifecycle through Semantic technologies", IEEE Computer Society Press, pp.109-122, Jan 2014.

[3] Elluri, L., Joshi, K. P. (2018, July). A knowledge representation of cloud data controls for eu gdpr compliance. In 2018 IEEE World Congress on Services (SERVICES) (pp. 45-46). IEEE.

[4] Elluri, L., Nagar, A., Joshi, K. P. (2018, December). An integrated knowledge graph to automate gdpr and pci dss compliance. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 1266-1271). IEEE.

[5] Joshi, K. P., Elluri, L., Nagar, A. (2020). An Integrated Knowledge Graph to Automate Cloud Data Compliance. IEEE Access.

[6] Ding, Li, et al. "Using ontologies in the semantic web: A survey." Ontologies. Springer US, 2007. 79-113.

[7] Understanding Semantic Web andOntologies: Theory and Applications

[8] Owl Web Ontology Language. Accessed: Aug. 2, 2019. [Online]. Available: http://WWW.W3.ORG/TR/OWL-FEATURES/

[9] General Data Protection Regulation — SendGrid. (2020). Retrieved 20 August 2020, from https://sendgrid.com/resource/general-data-protection-regulation-2/

[10] Drip. Retrieved 25 Jan 2020, from https://www.drip.com/blog/ecommerce/drip-general-data-protection-regulation-gdpr

[11] Dropbox. Retrieved 25 Jan 2020, from https://www.dropbox.com/security/GDPR

[12] Google. Retrieved 25 Jan 2020, from https://privacy.google.com/businesses/processorterms/

[13] Uploadcare. Retrieved 25 Jan 2020, from https://drive.google.com/file/d/1--SVUIZFQPcNA9dnDH1zekiIjGaqQd9D/edit

[14] Algolia. Retrieved 25 Jan 2020, from https://www.algolia.com/pdf/DPA-latest.pdf

[15] Intercom. Retrieved 25 Jan 2020, from https://www.intercom.com/help/en/articles/1385437-how-intercom-complies-with-gdpr

[16] Pax8. Retrieved 25 Jan 2020, from https://usc.pax8.com/resource/display/16383

[17] Paypal. Retrieved 25 Jan 2020, from https://www.paypal.com/us/webapps/mpp/gdpr-readiness-requirements

[18] FastSpring. Retrieved 25 Jan 2020, from https://m6u8p7c2.rocketcdn.me/wp/wp-content/uploads/2020/06/fastspring-privacy-policy-june-2020.pdf

[19] Facebook. Retrieved 25 Jan 2020, from https://www.facebook.com/legal/terms/businesstools/update

[20] Snapchat. Retrieved 25 Jan 2020, from https://businesshelp.snapchat.com/en-US/article/gdpr

[21] Microsoft Bing. Retrieved 25 Jan 2020, from https://docs.microsoft.com/en-us/microsoft-365/compliance/gdpr?view=o365-worldwide

[22] AWS. Retrieved 25 Jan 2020, from https://d1.awsstatic.com/legal/aws-gdpr/AWS_GDPR_DPA.pdf

[23] Rackspace. Retrieved 25 Jan 2020, from https://www.rackspace.com/sites/default/files/legal/RACKSPACE-%2350187-v2-Master_GDPR_DPA_JLF_220318-signed.pdf

[24] DigitalOcean. Retrieved 25 Jan 2020, from https://www.digitalocean.com/legal/data-processing-agreement/

[25] Box. Retrieved 25 Jan 2020, from https://cloud.app.box.com/v/getGDPRready

[26] SendGrid. Retrieved 25 Jan 2020, from https://sendgrid.com/resource/general-data-protection-regulation-2/

[27] Recapture. Retrieved 25 Jan 2020, https://recapture.io/gdpr

[28] Ebay. Retrieved 25 Jan 2020, from https://www.ebayinc.com/company/privacy-center/privacy-notice/

[29] SupportBee. Retrieved 25 Jan 2020, from https://d2bb5ika04lv1j.cloudfront.net/uploads/attachment/file/9179994/SupportBee_DPA_and_Standard_Clauses_Signed.pdf

[30] Zaiper. Retrieved 25 Jan 2020, from https://cdn.zapier.com/storage/files/46ac3128100f09a5eeda6ceb7bdb61aa.pdf

[31] Kotal, Anantaa, Karuna Pande Joshi, and Anupam Joshi. ”ViCLOUD: Measuring Vagueness in Cloud Service Privacy Policies and Terms of Services.” IEEE International Conference on Cloud Computing (CLOUD), 2020. 2020.
[32] Resource Description Framework (RDF). Accessed: Aug. 2, 2019. [Online]. Available: http://WWW.W3.ORG/RDF/
[33] K. P. Joshi and C. Pearce, ”Automating Cloud Service Level Agreements Using Semantic Technologies,” 2015 IEEE International Conference on Cloud Engineering, Tempe, AZ,2015, pp. 416-421, doi: 10.1109/IC2E.2015.63
[34] T. D. Breaux, M. W. Vail, and A. I. Anton, “Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations,” in RE’06: Proceedings of the 14th IEEE International Requirements Engineering Conference (RE’06), IEEE Society Press, September 2006.
[35] N. Kiyavitskaya, N. Zeni, T. D. Breaux, A. I. Anton, J. Cordy, L. Mich, and J. Mylopoulos, “Automating the extraction of rights and obligations for regulatory compliance,” in ER’08: Proceedings of the 27th International Conference on Conceptual Modeling (ER’08), Springer-V erlag, October 2008.
[36] L. Kagal and T. Finin, Agent Communication: International Workshop on Agent Communication, AC 2004, New York, NY, USA, July 19, 2004, Revised Selected and Invited Papers. Springer Berlin Heidelberg, 2005, ch. Modeling Communicative Behavior Using Permissions and Obligations.
[37] Kagal, L. Finin, T., “Modeling conversation policies using permissions and obligations,” Auton Agent Multi-Agent Syst (2007) 14: 187. doi:10.1007/s10458-006-0013-z
[38] GDPR Resource Center. (n.d.). Retrieved March 07, 2018, from https://gdpr.cloudsecurityalliance.org/
[39] General Data Protection Regulation (GDPR) – Final text neatly arranged.” General Data Protection Regulation (GDPR), gdpr-info.eu/.
[40] Resources about the GDPR. Retrieved March 07, 2018, from https://gdpr.eu/tag/gdpr/
[41] GDPR. (2018). General Data Protection Regulation. European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
[42] K. Barker and N. Cornacchia, “Using noun phrase heads to extract document keyphrases,” in Advances in Artificial Intelligence. Springer, 2000, pp. 40–52.
[43] Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., ... Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence, 165(1), 91-134.
[44] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenic, “Triplet extraction from sentences,” in Proceedings of the 10th International Multiconference” Information Society-IS, 2007, pp. 8– 12.
[45] Protégé Editor- Protégé Tool http://protege.stanford.edu
[46] Modal Logic: http://plato.stanford.edu/entries/logic-modal/
[47] PURL. Accessed: May 20, 2020. [Online]. Available: http://purl.org/csc/ontologyfiles
[48] PURL. Accessed: May 20, 2020. [Online]. Available: http://purl.org/csc/policysemanticsimilarity