FYP-2 Final Evaluation Report

FINE PRINT

Privacy Policies and Cyber Laws

Team Members:

Tehreem Javed 16i-0086

Ayesha Qamar 16i-0104

Supervisor:

Dr. Mirza Omer Beg

Anti-Plagiarism Declaration

This is to declare that the above FYP report produced under the Title: *FINE PRINT* is the sole contribution of the authors and no part hereof has been reproduced on as it is basis (cut and paste) which can be considered as Plagiarism. All referenced parts have been used to argue the idea and have been cited properly. We will be responsible and liable for any consequence if violation of this declaration is determined.

Date = 10th - June - 2020

Student 1

Name: Ayesha Qamar

Student 2

Name:Tehreem Javed

Supervisor (Faculty)

Name: Dr. Omer Beg

Table of Contents

Abstract

Privacy Policies are the legal documents that describe the practices that an organization or company has adopted in the handling of personal data of its users. About 201 hours on average are needed to read all the privacy policies encountered by a user in a year[1]. But as policies are a legal document, they are of-ten written in extensive legal jargon that is difficult to understand[2]. Though a lot of work is being carried out to analyse[5,12,14], understand[16] and better represent[12] privacy policies, none of the work targets to relate privacy policies with data protection laws.We aim to bridge that gap by providing a framework that will analyse privacy policies in the light of varies data protection laws, such as the General Data Protection Regulation (GDPR). Firstly, we segment and label both the privacy policies and laws. Then we present a framework to relate them and check the compliance of privacy policies with laws.

Introduction

Problem Domain

In recent times in the field of Natural Language Processing, work has been done on privacy policies but none that caters to the problem of verifying if a given privacy policy adheres to the data protection laws of a given country or state. A possible solution is to create a system powered by machine learning to review the privacy policy and see if it is in accordance to the laws of the country (or countries) and identify any areas where a violation between them is detected.

Privacy policies and cyber laws regulating these policies are both highly extensive and full of legal jargon. In fact, it is estimated that about 201 hours on average are needed by any average user just to read all the privacy policies encountered in a year [1]. As a result, consumers don’t fully understand what they are signing up for [2] and often do not know whether the policies that they are agreeing to are infringing on their legal rights.

Moreover, a company’s legal department spends hours to review its privacy policies to see if it is compatible with a given country’s laws. This is a rigorous process because each country has its own data protection laws and also because with the upsurge of Internet of things there has been an escalation in the number and complexity of privacy policies themselves [3].

Research Problem Statement

The automation of checking compliance of privacy policies with laws can be of great value. It will arm users to understand policies with respect to laws without getting into the apprehension of legal jargon and details.

The analysis of privacy policies on their own is not enough. There needs to be a mechanism to relate those policies with laws. The policies dictate what they are doing with the user’s data and how they are doing it but that information alone is not adequate to judge a policy’s transparency and its usefulness. [4]

Using such an automation tool, a user can have a deeper understanding of what’s happening with their data in legal light.

Literature Review

Research Item 1

Unsupervised Topic Extraction from Privacy Policies [6]

Summary

The paper focuses on labelling privacy policies using topic modeling, which is an unsupervised approach. The research provides insight into the topics that are being addressed in privacy policies these days.

The privacy policies of mobile apps were collected from the Google play store. 4982 privacy policies were left after data pre-processing.The policies were segmented based on paragraphs which resulted in 45,622 paragraphs in total. The Latent Dirichlet Allocation (LDA) method for topic modeling was used. LDA doesn't require the data to be pre-labelled. It works by randomly grouping words together into topics and then iteratively improving the grouping till convergence. The method is based on the assumption that both words belonging to a specific topic and topics in a document are few. After 600 iterations of LDA with the number of topics set to 100, words were grouped into topics along with their probabilities. Those probabilities were then used to assign paragraphs to topics by summing the probabilities of each word of a paragraph appearing in a topic. The topic with the maximum score was assigned the paragraph. The topics were then manually checked and merged by an expert. The merging process involved selecting 30 paragraphs from each topic, the expert then gave a one sentence summary of each topic. The topics were merged together according to their redundancy and relevance with one another. The merger left 36 topics.

The merged topics reveal the underlying structure of privacy policies. Some topics have thousands of paragraphs associated with them, in part because those topics were created after merging several sub-topics together. On the other hand, some topics were created after merging only one or two sub-topics but still have thousands of paragraphs mapped to them. Those topics represent the areas that privacy policies address the most. Amongst them are privacy policy change notifications, contact information and the option to opt-out of privacy policies. The topics were also validated against the OPP-115[5], which is a data set of 115 privacy policy annotated into 22 topics by legal experts. The mapping showed that the current method of extracting topics revealed more fine-grained details from privacy policies.

Critical Analysis

- Strengths:
    - This method can be used to analyze privacy policies and their evolution over time. As it is not dependent on any labelled data set like the OPP-115 which was created in 2016. This is of crucial importance because many firms are updating their policies to comply with the ever stricter laws coming into place like Europe's GDPR.

    - It also provides more finer details about privacy policies. The number of paragraphs being mapped to a certain topic and the number of sub-topics under one topic. Insights like these can be helpful to determine what the makers of privacy policies are considering crucial and addressing the most.
    - It is one of the first unsupervised methods of annotating privacy policies.
    - The results obtained are compared with the standard OPP-115 dataset as a means of validation.
- Weaknesses:
    - The segmentation of privacy policies was done on the basis of paragraphs. This was done on the assumption that different paragraphs describe different legal aspects. Whereas this may not be true in all cases.
    - The method is not completely independent of human annotator as it requires a domain expert to merge the topics.
    - The topics were summarised by the expert based on a sampling of 30 paragraphs only from each topic. There is no proof that those samples were a good representation of the topics.

Relationship to the proposed research work

The initial part of our problem is to label laws and privacy policies. While there is a corpus of labelled privacy policies, there is none for data protection laws. We can use the unsupervised methodology proposed in this paper to label laws. As the terminology used in laws and policies overlap and the method has performed well on privacy policies.

Research Item 2

Leveraging Linguistic Structure For Open Domain Information Extraction [7]

Summary

The paper describes a method which can be used in open domain information extraction for extracting relation tuples from sentences. These tuples can be used in natural language processing for question-answering, information retrieval and relation extraction. The tuples are extracted in two stages. In the first stage, the sentence is broken down into self-contained clauses to reduce false triples. A classifier is used to create clauses which are logically in accordance with the original sentence. A greedy search approach is used in which a sentence is traversed using a dependency tree. The traversal is recursive and at each edge it is decided if an independent clause should be yielded. This decision is taken by using a multinomial logistic regression classifier which predicts whether an edge should be recursed with or without yielding a clause or if the recursion should stop.

In the second stage natural logic is used to obtain the most specific triple form the clauses by removing superfluous information. These triples are of the form subject-verb-object and retain the essential semantics which the original sentence had. Natural language formalism is used to find operators such as all, no and many and to determine from these if a proposed triple can be turned into something more general or specific.

Critical Analysis

- Strengths:
  - Incomplete utterances are avoided by allowing a sub-clause whose subject is controlled by the governing clause's subject to inherit from the governing clause. By doing so, the long-range dependencies of a sentence can be captured.
  - Removing non-subsective adjectives is prohibited as doing so would lead to loss of information.
  - A better generalization is done by splitting sentences into clauses which is useful for working with out-of-domain texts.
- Weaknesses:
  - The errors made in splitting the clauses manifest themselves across an array of sentences.
  - Complex assertions are not interpreted correctly. There is no mechanism to determine if the assertion in a sentence is only conditionally true or hypothetical in nature.

Relationship to the proposed research work

The relation triples produced as the result of this paper can be used to extract information from the laws. These triples can then be used to compare them with the privacy policy more efficiently to find if the policies comply with them.

Research Item 3

Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning [12]

Summary

The paper presents a framework for automated analysis of privacy policies. It uses the OPP-115[5] dataset for labelling of data followed by a hierarchy of neural network classifiers. The framework is manifested in the form of two applications, automating assignment of icons to privacy policies and a question answer system.

The framework is based on three layers: Application layer, Data layer and Machine Learning layer. A privacy policy is first split into smaller segments. The application layer consists of a query and a class comparison module. It allows users both structured and free-form querying. The responses are segments of the privacy policy that satisfy the query. The policy from the application layer is passed on to the Data Layer and the query to the Machine

Learning. The Data Layer crawls a privacy policy from the website URL. It then segments the policy first on the basis of its representation in <div> and <p> tags in HTML format. Then a more detailed segmentation is done using custom word embeddings generated by using a corpus of 130K privacy policies. The high level along with the fine-grained segments are then passed to the Machine Learning layer. This layer also has two components: query analyzer and segment classifier. The ML layer first generates a custom word embedding as mentioned above. These word embeddings are then used to train an array of neural network classifiers based on the OPP-115[5] dataset. The segments are assigned from 10 high level categories and several low level attributes. The classifiers assign class labels to the segments in two stages. In the first stage, the classifier predicts one or more than one high level categories for the paragraph segments. In the second stage, the classifier predicts values for the attributes under each high level category. Thus, the ML layer analyzes and assigns labels to the policy segments at a much detailed level using CNN. In total 22 multi-class classifiers are trained at the ML layer. The output from this layer is in the form of class-value pairs for both query and the segments of policy which are then passed back to the Application layer's class comparison module. This module finally matches the labels of the query with those of segmented policy and gives results to the user.

Critical Analysis

- Strengths:
    - The word embeddings are trained using fastText. It allows it to be trained on subwords. This is particularly useful in the case of spelling mistakes when querying the question answering system.
    - The framework's accuracy is tested in the form of two applications. Both are rigorously validated against previous work and through human annotation.
    - Leverages the OPP-115 dataset's labels of does and doesn't, indicating the presence or absence of a category.
- Weaknesses:
    - Is dependent on the OPP-115 dataset for labelling of policies and queries. The dataset was revealed in 2016 and since then there has been a radical change in the way privacy policies are being made. [6 , 7].
    - The custom word embedding doesn't take advantage of the already present ones.

Relationship to the proposed research work

The above paper not only segments and labels policies but also correlates query segments with policy segments. It also provides insight into using CNNs to segment and annotate privacy policies. All the steps are an integral part of our research problem.

Research Item 4

Unsupervised Alignment of Privacy Policies using Hidden Markov Models [14]

Summary

This paper presents an approach to align privacy policies. Many privacy policies are similar to each other as they address the same issues and therefore can be aligned using an unsupervised approach. A corpus of 1000 privacy policies was collected for this task manually. This is because despite attaining the URLs of the policies it was difficult to extract the policy with its structure intact as each website is different and presents challenges of its own. The policies were then segmented based on their section headings by crowdworkers. A Hidden Markov Model like approach is then used to align the segments such that an issue (addressed in the policy) corresponds to a hidden state. This correspondence is based on the bigrams in the segment of the policy and its distribution of words.For each state '*t*', a bag of terms is drawn from the section '*t*' of the policy unlike classic Hidden Markov Models where only a single term is drawn each time.

To evaluate the results, the paper presents two evaluation techniques which are reusable. These techniques approached the problem as one of grouping rather than alignment. The first technique was to evaluate the results by creating an answer set. Nine questions were created by domain experts. Then the domain experts not involved in the process of creating the questions selected the segments of policies they thought best answered each of the nine questions. They did this for thirty policies. The model is then evaluated by calculating precision and recall using the answer sets as a gold standard. The second evaluation technique is by direct judgement in which 994 policy segment pairs are selected from the 1000 policies across four ranges of cosine similarity. For each section pair, crowdworkers are asked if the pairs are talking about the same thing, broadly related to each other or not identical at all. The results of this were then used to calculate precision and recall as before.

Critical Analysis

- Strengths:
  - Created and used a new dataset of 1000 manually segmented privacy policies.
  - Evaluation benchmarks created are better than previous naïve methods. They also do not require a pre-labelled dataset.
- Weaknesses:
  - There is a lot of human effort involved in data gathering.
  - In the first evaluation technique a very small number of policies are selected which may likely be biased

Relationship to the proposed research work

The laws or data protection acts that we will use are not labelled. Therefore we need an unsupervised technique to align them into classes. As the approach mentioned in the paper is using privacy policies and since the laws and policies have similar legal jargon, we can use it to align the laws.

Research Item 5

Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks [15]

Summary

The paper proposed a convolutional neural network technique to find similarity in sentences using multiple perspectives. Recent work in sentence similarity is moving towards using distributed representations along with neural networks rather than hand crafted features. This paper also takes a step forward in this direction. Given sentences A and B, the proposed approach finds a measure of similarity sim (A, B) between the two sentences. This would be done in two steps. Firstly, the sentences are input into identical sentence models where a convolutional neural network is used to extract information from different perspectives and multiple pooling types. Then the outputs from these models act as input for the similarity measurement layer. This layer computes similarity based on multiple distance functions. At the end, two fully connected layers with an activation function in between and a final log-SoftMax layer is used to get the overall similarity score.

The sentence model layer uses two different filters; holistic and per-dimension. The holistic filter is used to extract temporal information. They convolve the entire word vector for the number of words specified by the sliding window width at a time. The per-dimension filter is used to extract information at a finer spatial level. These filters convolve for each dimension of the word vector for the number of words specified by the sliding window width at a time. The holistic filters block then uses min, mean and max pooling whereas the per-dimension filter only uses max and min pooling. The window widths can be of multiple sizes to learn different features as is done in n-gram models. A window width of infinity is added in a holistic filter block to ensure that the original word embeddings are also included. The Similarity measurement layer is then used to compare local regions using cosine, euclidean and element-wise distance functions. The local regions for comparison are selected on the basis that they are either from convolution layers with the same type of filter, window size or pooling.

Critical Analysis

- Strengths:
  - The approach does not rely on resources such as parsers or wordnet as high-quality parsers are not readily available for specialized domains
  - The use of multiple filters leads to better information extraction and makes richer sentence models.

    - The need of hand-crafted features of traditional NLP approaches is removed through this approach
    - The information loss from flattening the output from a convolved layer is rectified by using structured comparisons over certain areas of the sentence representations in the similarity measurement layer.
- Weaknesses:
    - The architecture engineering of the model is complex as it compensates for hand-picked features.
    - The model may not be able to compete with a simple but deeper neural network which is trained using a large set of data.

Relationship to the proposed research work

The privacy policy and law segments belonging to the same category are compared to find if the policies are semantically similar, that is if the policies comply with the laws. The papers approach for finding sentence similarity can be used for this step of our research work.

Research Item 6

BERT:Pre-training of Deep Bidirectional Transformer for Language Understanding [17]

Summary:

In this paper, the authors have presented a novel language representation model, Bidirectional Encoder Representations from Transformers(BERT). Their proposed model learns representations by concurrently accounting for both left and right context. This enables the representations to be used with some task specific fine-tuning without altering the model architecture.

The framework is divided into two parts; pre-training and fine-tuning. The model architecture comprises a multi-layer bidirectional Transformer encoder. They have experimented with two model sizes: base and large. The former consists of 110M parameters and the later 340M. The first token of sentences is a special token, [CLS]. As the input can consist of pairs of sentences, so to distinguish between them a special token [SEP] is used.

To pre-train BERT, two supervised tasks are used. First, a Masked Language Model is used. To do this, 15% of the input tokens are masked at random and then the task is to predict those. The vectors of those masked tokens obtained from the final hidden layer are then passed to a softmax layer. Second, the Next Sentence Prediction task is used to pre-train the model. The input consists of two sentences, where 50% of the time the 2nd sentence proceeds the 1st and 50% of the time two random sentences are passed.

For fine-tuning, a classification layer is added on top of the pre-trained model and all the learnt parameters are fine-tuned at the same time. Fine-tuning is computationally inexpensive as only the output layer weights are added. No layers are frozen during fine-tuning. Only K x H new parameters are added at the output layer, where K is the number of labels and H is the size of the hidden state. And then a standard softmax loss is calculated.

Critical Analysis:

- Strengths:
  - The pre-trained representations are built using both left and right context of words.
  - Bert produces state of the art results in 11 NLP tasks.
  - Also, the architecture remains the same for both pre-training and fine-tuning, the only difference being the output layer. The model works effectively for both single text and text pairs.
  - Although the model mostly focuses on fine-tuning, feature-based techniques can also be used to adapt the model for another task.
- Weaknesses:
  - As with all big models, it is computationally expensive to pre-train BERT.
  - Also, the transformer architecture puts equal weight on each word surrounding the target word but that might not be beneficial in all cases, as words closer to target word do in some cases convey greater meaning.
  - There is a difference in the training and testing data due to the masked language model as in training data there will be no [MASK] tokens.

Relationship to the proposed research work

BERT word embeddings can be used for various downstream tasks. Therefore we can use them with fine-tuning and an additional classification layer for labelling policies. In addition, the embeddings can also be used to find similarity between segments of laws and policies using cosine similarity.

Research Item 7:

How to Make Privacy Policies both GDPR-Compliant and Usable[20]

Summary:

The paper provides a template for privacy policy makers in terms of the design they should use and the practices they should adopt in order to make their privacy policies in accordance with the rules and regulations of GDPR and also to make them understandable and accessible to a general user.

The starting point of the author's research is GDPR itself. They identify and state six major GDPR requirements. These include specification of data being collected, justification for collection, information regarding how the data will be processed, the amount of time for which data will be stored, who to contact to remove/change data and finally to communicate this privacy information to the end user. The last requirement entails that making privacy policies usable is not a choice but an obligation under GDPR.

Then to identify where and how the current privacy policies fail, the authors analyze the privacy policies of the UK's top ten most visited websites. They found only one out of the ten websites seemed to fulfill all the six GDPR requirements. To test the usability of the policies, the authors found the Gunning Fog Index Score (GFI) of these policies. The score shows how many years of education a person would require in order to understand the policies.It was found that the text of all these policies requires eleven or more years of schooling which is quite high.

After this, the authors carried out a literature review of the previously existing privacy policy design research. They found no design specifications for the duration for which the data will be retained and how the data will be processed. Specifying data collection and data access/edit was briefly touched by some publications. The “Communication of Privacy Information” requirement as already mentioned, focuses on making the privacy policy more clear and concise. This has been covered thoroughly in the literature. The findings for this requirement can be divided into two areas; content and delivery. In case of the content of the privacy policies it was found that dividing the privacy policies into modules, making them personal to the user by addressing them, giving them control and choice to opt out, highlighting important information and maximising understandability by minimizing the use of legal jargon are the ways in which the content can be added suitable for a general user. For the delivery of the policies, the design practices cover the time, location and the appearance of the policy. It was found that using practices such as neutral background colors and speedy recognition tactics are preferred by the user.

Finally, using this research the authors created a privacy policy template. They simplified the text to ensure that someone with less than high school education could understand them.

Critical Analysis:

- Strengths:
    - It summarizes the content of the GDPR into six requirements which are far easier to understand than to go through the law itself.
    - It compares the privacy policy landscape before GDPR, thereby giving a thorough insight to where they are lacking in compliance.
    - The proposed template has a very low Gunning Fog Index and therefore can be understood by a layman.
- Weaknesses:

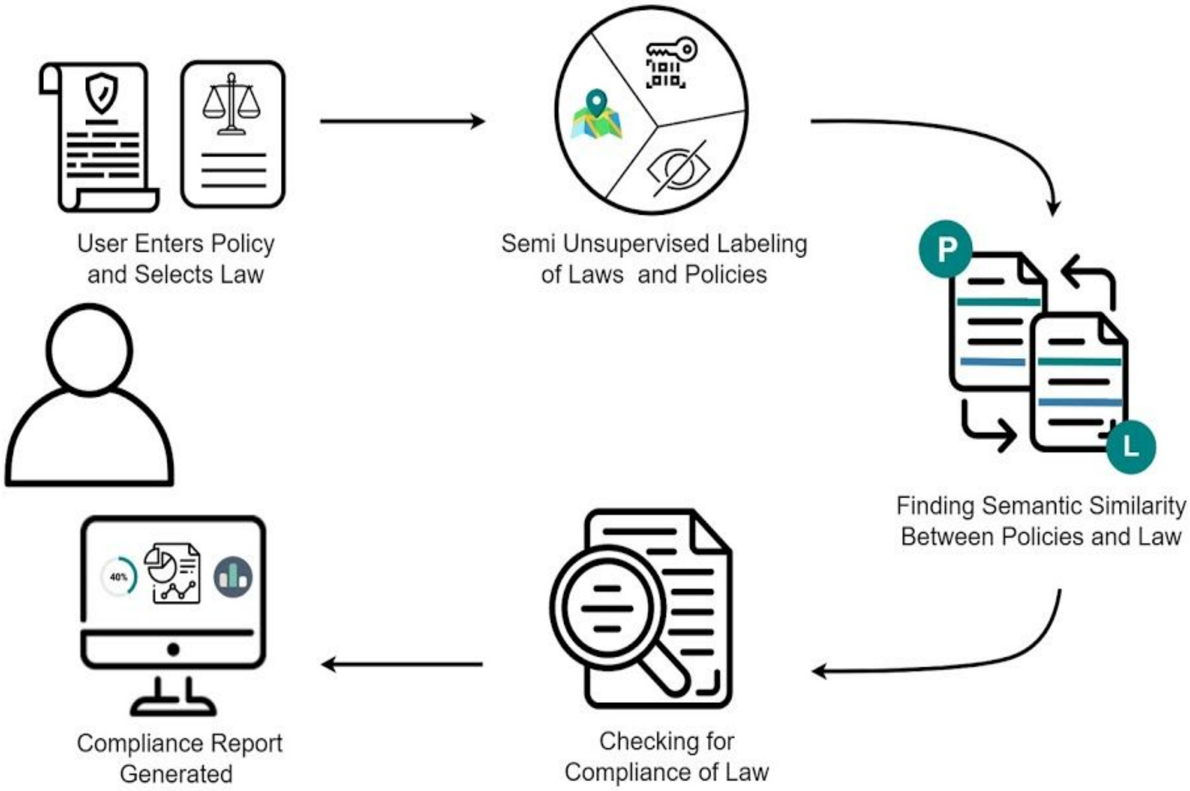
- There is no general privacy policy template provided. Instead an example of the template is given. A general template would have been more useful to privacy policy makers.

- The privacy policies examined before GDPR are all of well established companies. The policies of small websites have been left out in the analysis of the policies with regard to GDPR.

Relationship to the proposed research work

The GDPR requirements provided in the paper act as a guideline to correlate the categorized policies with GDPR law segments. This correlation then allows the segments of GDPR related to a certain policy to be checked for similarity and compliance.

Proposed Approach


User Enters Policy and Selects Law
Semi Unsupervised Labeling of Laws and Policies
P
L
Finding Semantic Similarity Between Policies and Law
Checking for Compliance of Law
40%
Compliance Report Generated

Our approach is to first segment policies, annotating them labels according to the OPP-115 dataset[5]. Laws will also get segmented and labelled using an unsupervised technique- either Latent Dirichlet Allocation. Afterwards, the linked segments of policy and law pertaining to the same category are compared for compliance. And lastly a final compliance score is calculated across all policy segments and law segments by aggregating their individual scores.

Implementations

1. Crawling Privacy Policies

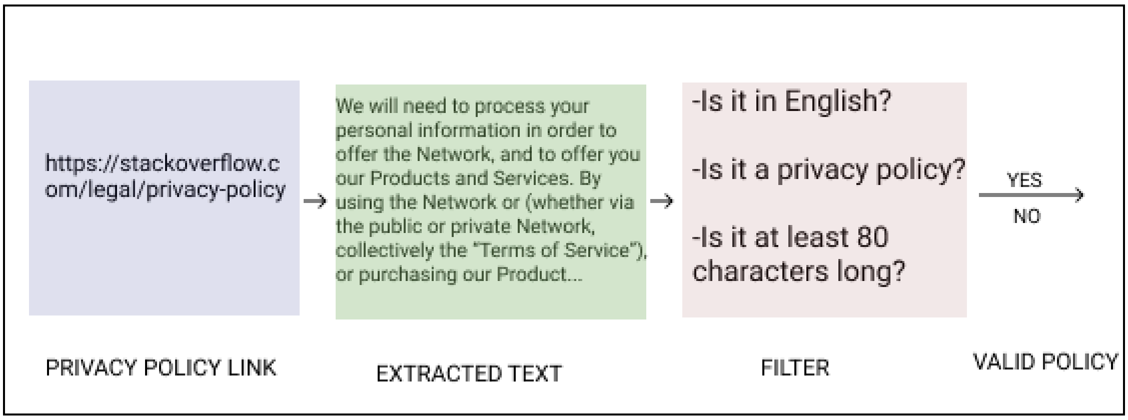
We have crawled around *60,000* privacy policies from different android applications from GooglePlay store. To do so, we wrote two different scrapers. The first one extracts the privacy policy links itself and then extracts text from them. The crawler visited around *15000* web pages of android applications. Some of the apps did not have privacy policy's link mentioned in their details. Apart from that, most apps coming from the same developer or company shared a single policy. So, in the end we were able to get around *4000* unique links. The second one makes use of the MAPS dataset [16] of around 4 million links to privacy policies. We found that most of the links provided were redundant, after eliminating those we were left with around *150000* unique links. Again, many links from this set turned out to be broken links or policies not in English. We gathered around *56,000* policies using these links.

To ensure the final quality of privacy policies extracted was upto a certain mark, we wrote two different scrapers to extract text from policy links. One of them was more resilient to the way webpages were structured but the text extracted from it contained a lot of gibberish symbols. The other one, though failed to recognise some special symbols, and whenever that happened would not scrape the whole policy, but extracted much neater policy texts. In the end, we went with the latter for the sake of higher quality of policies over their quantity.

It was also important to check if the extracted policy was indeed in English or if it was a privacy policy or just the main web page of the website. To do so, we passed the extracted text through a pipeline that checked the following conformities:

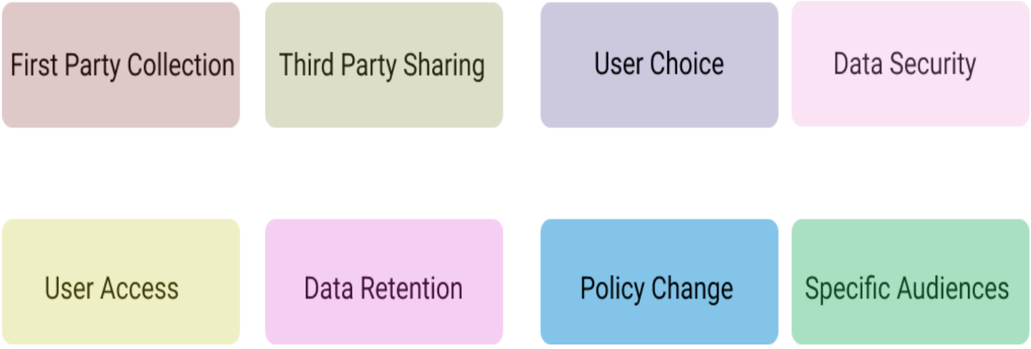
- The privacy policy was in English. To do so, we used the 'langdetect' Python library.
- The text was of privacy policy. We employed regular expressions to achieve this. We checked on keywords like "*404 Error", "Webpage not found"*.. Etc
- The policy was of substantial length. To ensure that the policy was legitimate we also removed any policy that had a length less than 80 characters.

The whole process can be visualised as follows:


https://stackoverflow.com/legal/privacy-policy
We will need to process your personal information in order to offer the Network, and to offer you our Products and Services. By using the Network or (whether via the public or private Network, collectively the "Terms of Service"), or purchasing our Product...
-Is it in English?
-Is it a privacy policy?
-Is it at least 80 characters long?
YES
NO
PRIVACY POLICY LINK
EXTRACTED TEXT
FILTER
VALID POLICY

The pipeline employed to ensure the gathered policies are of high quality.

An interesting insight we got from using our own scraper to gather privacy policy links and the MAPS dataset was that, unlike a lot of redundancy of links we saw because of top apps appearing in multiple categories, our extracted links and those from the MAPS dataset had no overlap. This further highlights that our pool of privacy policies is heterogenous, and contains policies written for top apps of several regions and it also has less popular apps.



The 8 high level categories in the OPP-115 dataset.

Eight of the ten categories are represented here. The other two categories not shown are *other* and *Do Not Track*. Do not tract, which is no longer provided in privacy policies, and the category other that captures all the miscellaneous information in policies. We discarded both of these categories as they were of no use to us.

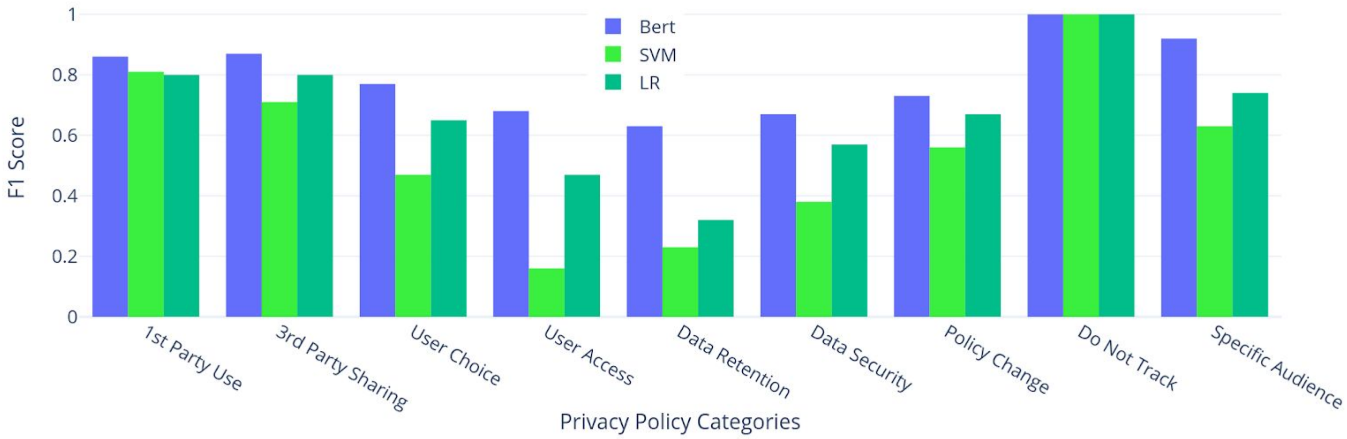
2. Labelling Policies

We have labelled policies based on the taxonomy provided by Wilson et al[5]. The taxonomy is based on a hierarchy of labelling and consists of 10 broad categories and 112 fine grained categories. The policies are segmented at paragraph level and each segment gets assigned multiple labels. The annotations were done by 3 graduate law students; there are three versions of the annotations. We have used the annotations in which there is a .75 overlap between the annotations, i.e., at least 2 of the 3 students have given the same label.

To begin with, we extracted the 115 annotated policies from the OPP-115[13] dataset, and only relevant information like the text segment of policies themselves and the assigned labels were kept. Then to cater for these multi class labels, we made binary models for classification for the 10 broad categories. Thus, for training the classifiers the dataset was divided into ten subsets where each set corresponded to one category and had binary labels (0 if the text segment did not belong to the category and 1 otherwise).

Then for the classification, we used Towards Automatic Classification of Policy Tex[19] as a starting point and trained a logistic regression model and a SVM model for classification. In addition, we also used a fine tuned version of the BERT model. We took the pre-trained BERT model for classification and fine tuned it using a low learning rate for each policy category.

We tested the three models for each category on a held out test set and calculated their F1 scores. The BERT Classifier as shown below gave better results than others, so we saved the trained model to use for privacy policy categorization at run time in the final product.



3. Labelling Laws

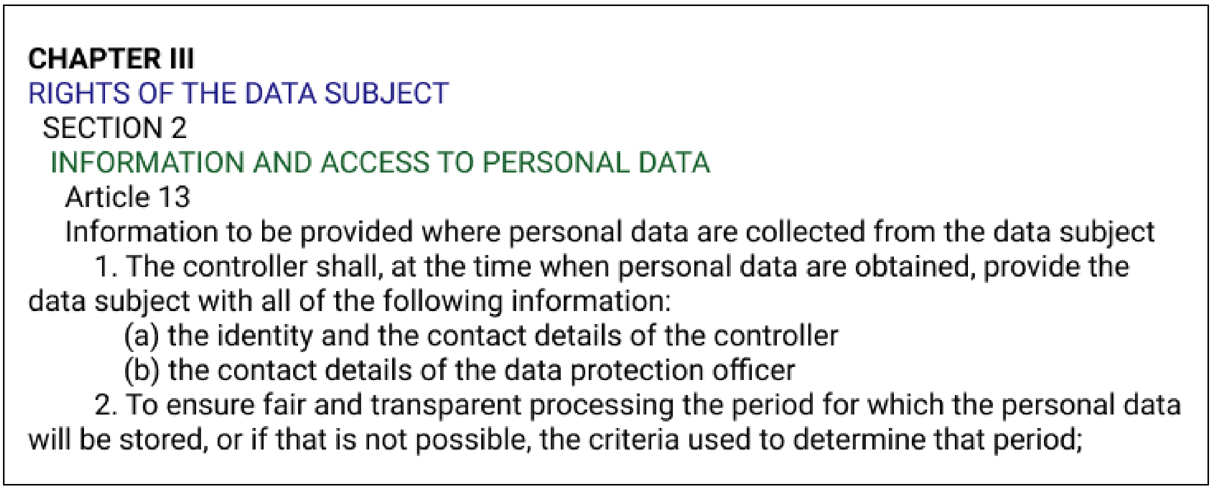
Since there is no available taxonomy of labelled laws along with the fact that we are not dealing with a specific country's law but rather creating a system that would be universal and would work on several countries' law, we decided to use an unsupervised technique to label laws in order to account for this adaptability. Leveraging the insights from the work of Sarne et al [13], we chose to use Topic modelling and specifically Latent Dirichlet Allocation, LDA, to annotate laws. This not only frees us from the need of having an expert in the field to label them,but we can also easily extend this technique to label multiple laws.

For our initial work, we are only focusing on GDPR and PDPA and then we plan to incorporate other laws too. GDPR, which came into effect in 2016, is the Data Protection law applicable to organizations processing data of EU citizens, even when the data being processed is out of Europe. PDPA is the singaporian Data Protection law that came into effect in 2012. Firstly the GDPR details are provided followed by PDPA.

General Data Protection Regulation

The GDPR is hierarchically structured in chapters, subjects and finally articles. The articles are themselves organised in numbered bullets. We have leveraged this and have split the law on the basis of this organisation. To do this, we have used regular expressions, as the GDPR does not have newlines between different paragraphs and has a unique structure; so, we could not use a built in function or library to achieve the segmentation as we desired.

For the GDPR, we followed the natural hierarchy in which it is written and segmented it according to the Articles, with one segment consisting of all the subpoints of an Article. By following this segmentation scheme we were left with 371 segments with an average word count of 75.11 words per segment. After that, we removed stopwords, punctuations and lemmatized the words.



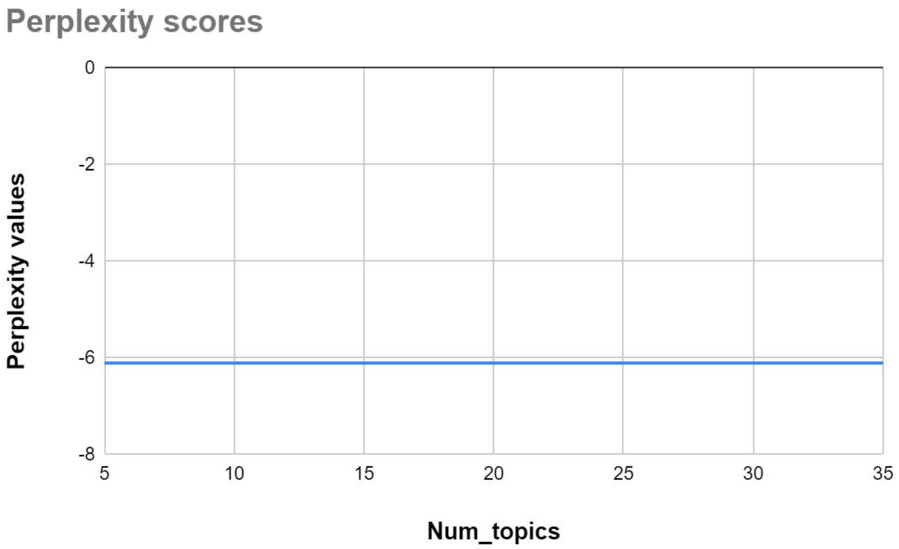
CHAPTER III
RIGHTS OF THE DATA SUBJECT
SECTION 2
INFORMATION AND ACCESS TO PERSONAL DATA
Article 13
Information to be provided where personal data are collected from the data subject

1. The controller shall, at the time when personal data are obtained, provide the data subject with all of the following information:
   (a) the identity and the contact details of the controller
   (b) the contact details of the data protection officer
2. To ensure fair and transparent processing the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;

The structure of the GDPR. The hierarchy consists of Chapters, Sections, Articles and then points in those Articles.

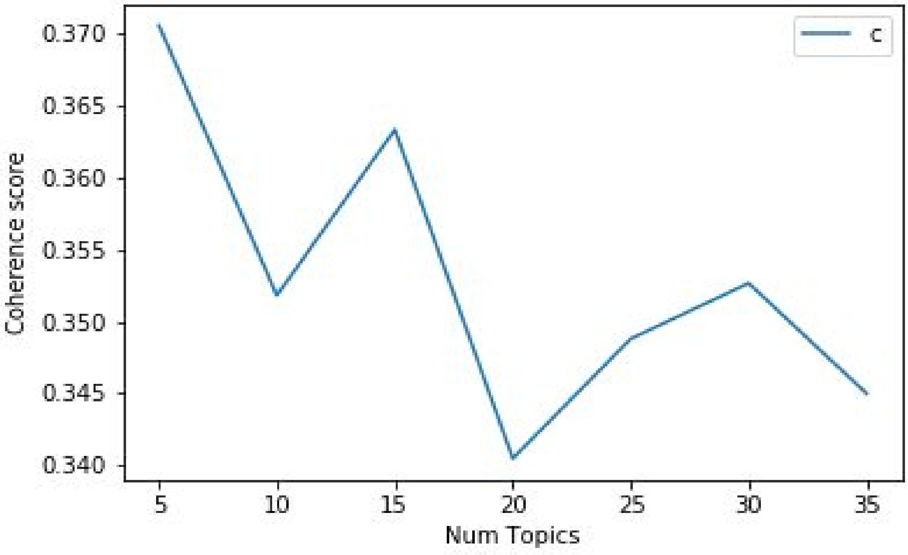
Next, we used Genism's implementation of LDA to perform the topic modelling. LDA works by assuming that topics in a document and words in a topic follow some specific distribution. Since it's an unsupervised technique, we only need to provide the number of topics, k, the document has.

Since it's a hyperparameter, we decided to set it to 50, so that it's broad enough to capture all the finer details of different topics and narrow enough that merging them is not a daunting task; but we later performed some experiments and found that 50 is not a suitable value. Firstly, we used perplexity scores to determine this hyperparameter, but those scores didn't tell us much about the data as is evident from the below graph:



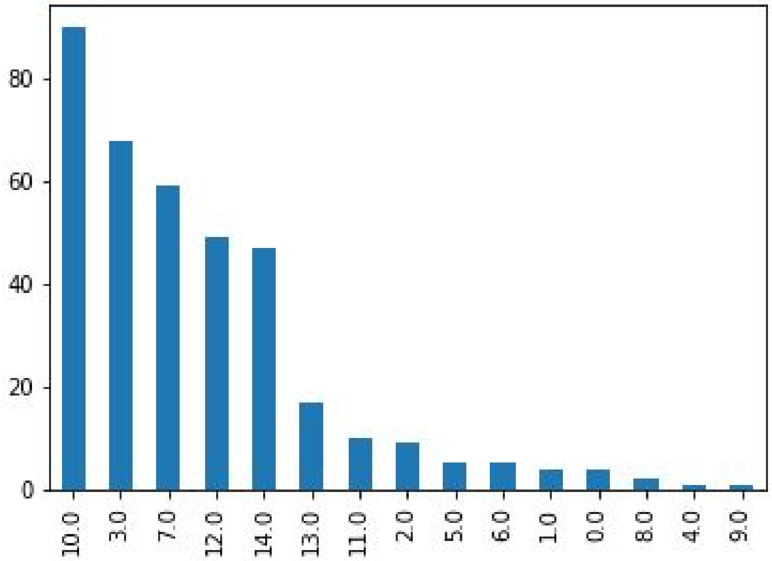
The perplexity score plotted against multiple values of the number of topics

Next we tried coherence scores to give us a better idea about the underlying topic distribution.



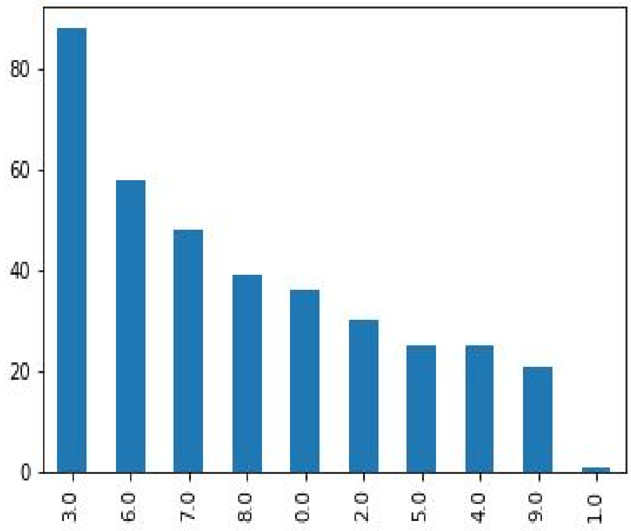
The best coherence score was achieved when k was set to 5.

We know from the domain knowledge that setting the number of topics to 5, and getting a coarse labelling, would not give us the desired results as the number of chapters are 10 in GDPR, so there are at least 10 different topics in the document. The next best option was to choose k to be 15 but, as is visible from the graph below, two of the categories only had one segment belonging to them, which clearly indicates that those segments could just be put under one category- other.



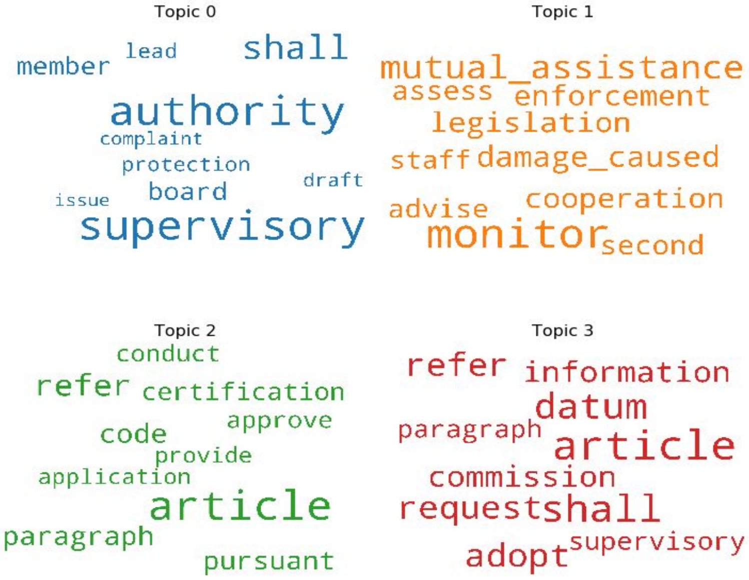
The number of segments belonging to each topic is depicted when k is set to 15. The number of topics are plotted on x-axis and number of segments on y-axis.

Finally, after these experiments coupled with our domain knowledge about GDPR and the fact that the number of categories in privacy policies is also 10, we decided to keep the hyperparameter to 10. The segment distribution seems much more coherent with k set to 10, doing so also opened up the possibility of doing a one-to-one mapping between laws and policies.



The segment distribution is depicted when k is set to 10. With the number of topics on x-axis and number of segments on y-axis.

As an example, the word distribution in four of the topics is visually represented below:



A visual representation of the most frequent words of some of the topics.

Most of the words are non-overlapping i.e., do not occur in multiple topics and hence show that the labelling is efficient.

Personal Data Protection Commission

The PDPA has two main provisions:

- Data Protection (DP) provisions

  These provisions are directly concerned with the handling and collection of users' personal data.

- Do Not Call (DNC) provisions

  Do Not Call Registry is not applicable to privacy policies as that part is only concerned with how to handle the telephone numbers of singaporian users but doesn't in particular detail how the phone number should be collected. Including video or voice calls or text messages. But since these requirements aren't directly linked with privacy policies, we skipped those divisions.

As the PDPA is a relatively shorter law, we did not feel the need to label it through any unsupervised method to obtain the segment categorizations. Upon manual reading only PART II to V of the law were relevant to personal data and we extracted appropriate text from these parts.

Titles of the parts are mentioned below:

- PART II: *PERSONAL DATA PROTECTION COMMISSION AND ADMINISTRATION*
- PART III: *GENERAL RULES WITH RESPECT TO PROTECTION OF PERSONAL DATA*
- PART IV: *COLLECTION, USE AND DISCLOSURE OF PERSONAL DATA*
- PART V: *ACCESS TO AND CORRECTION OF PERSONAL DATA.*[22]

4. Extracting Relevant Law Segments

General Data Protection Regulation

Leveraging the work done by Karen et al[20], where they provide a template and lay out the requirements that policies must follow in order to be GDPR compliant, the GDPR requirements that customers must be informed about are:

1. GDPR1: What Data will be collected and why
2. GDPR2: How Data Will Be Processed
3. GDPR3: How Long Data Will Be Retained
4. GDPR4: Who Can Be Contacted to Have Data Removed or Produced

A brief description of each GDPR requirement is mentioned below:

- **Specify Data Being Collected**
  Users and customers should be made aware of the information that is gathered and stored about them.
- **Justification For Data Collection**
  Organizations need to detail in their privacy policies not only their right to collect any particular data but also explain and justify *why* they need to collect it.
- **How Data Will Be Processed**
  Organizations need to inform customers of all the lawful rights the company has to process user's personal data legally. According to GDPR, processing is only legal when the customer has explicitly given consent to handle their data. Also, the user should have the option to opt out of any processing or profiling at any time. For sensitive data- including a person's religious, political or health related data- separate justification should be provided for processing.
- **How Long Data Will Be Retained**
  Organizations must specify the exact amount of time the data collected is going to be retained for.
- **Who Can Be Contacted to Have Data Removed or Produced**
  Costumes have the right to get all their data from the organization upon request. The organization has to, in turn, provide all the data they have collected of EU customers within a reasonable time frame. The customers also have the right to ask organizations to erase all their data. For both these purposes, the contact information of data handlers should be provided. The policies should also detail who the Data Protection Officer (controller) is along with their contact details.
- **Communication of Privacy Information**
  According to GDPR, the information on all above points should be provided in a succinct manner. In the words of the law itself: "*requires the information to be provided in concise, easy to understand and clear language*" [20]

Next we were left with the task of manually extracting all the text from the GDPR that pertained to the specific category of our interest. Because the law was already categorized using LDA, this step became easier. Only some portion of the law was useful for our purpose i.e., the articles related to personal data processing and not the chapters about how the law itself should be implemented or where it is applicable such as the *Chapter X: DELEGATED ACTS AND IMPLEMENTING ACTS* .

For example, the *Article 14* of GDPR on *Information to be provided where personal data have not been obtained from the data subject* states "*Processing shall be lawful if the data subject has given consent to the processing of his or her personal data for one or more specific purposes the categories of personal data concerned the recipients or categories of recipients of the personal data where applicable*"[21], this was categorized as a GDPR segment belonging to "What Data will be Collected and Why" and so will be mapped to the *First Party* and *Third Party* category of privacy policies. In total, ten such law segments were made and given one of the four above mentioned requirements.

Personal Data Protection Act

According to the Personal Data Protection Commission (PDPC), there are three broader categories and then further subcategories of the PDPA [23]. They are discussed below:

- **Collection, use and disclosure of personal data**
    - *Consent*
      An organization should first ask customers to give consent to collect, use or disclose their personal data. Users should also have the ability to withdraw consent.
    - *Purpose and Notification*
      Consent should only be taken for data that is essential to provide a given service to users. Users' data can only be obtained or disclosed for the purposes for which the user was informed about. Users should also be made aware of the reasons for data collection.

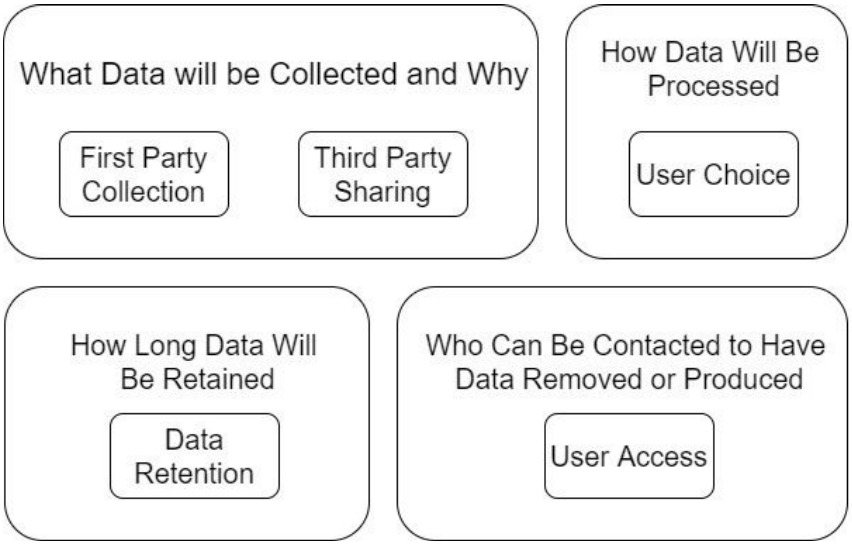
- **Accountability to individuals**
    - *Access and Correction*
      If customers request, they should be provided with their collected personal data along with the ways in which the data was collected and used in the timeframe of a year. Users can also request to get their data updated to fix any errors.
- **Care of personal data**
    - *Retention*
      Data should be deleted once the purpose it was collected for has been fulfilled. Keeping data longer than needed for business reasons is prohibited.

5. Correlating Policy and Law Segments

Now if a new policy is entered to check for compliance, first it is segmented and then each segment gets labelled one or more of the labels. At this stage, we have categorized policy segments and labelled law segments.

General Data Protection Regulation

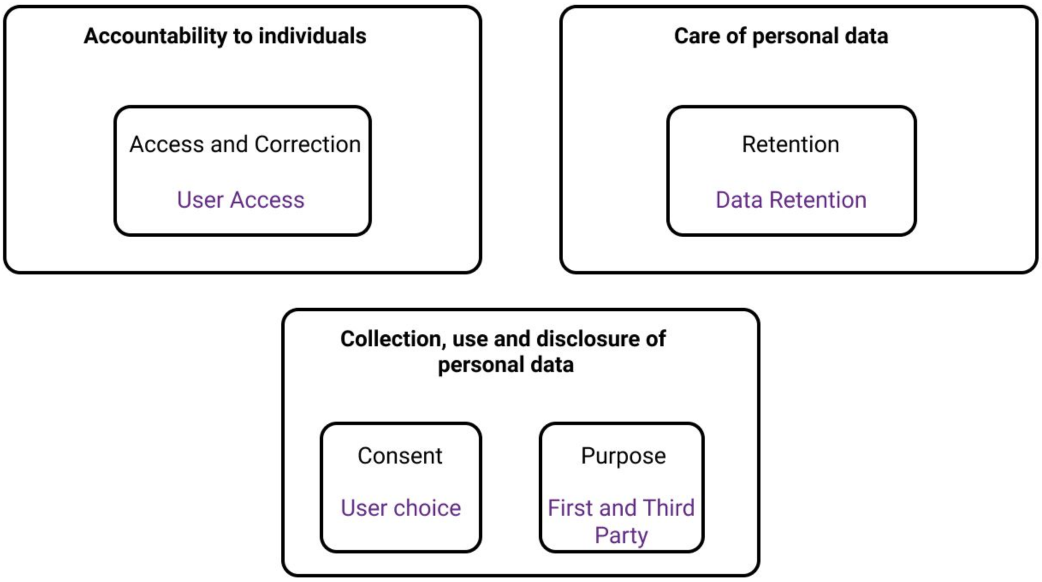
So, for GDPR, we take the labelled segments of policy and map the five privacy policy categories to the four GDPR requirement categories.


What Data will be Collected and Why
First Party Collection
Third Party Sharing
How Data Will Be Processed
User Choice
How Long Data Will Be Retained
Data Retention
Who Can Be Contacted to Have Data Removed or Produced
User Access

The GDPR law segments categories represented by the outer box and the policy categories linked with each.

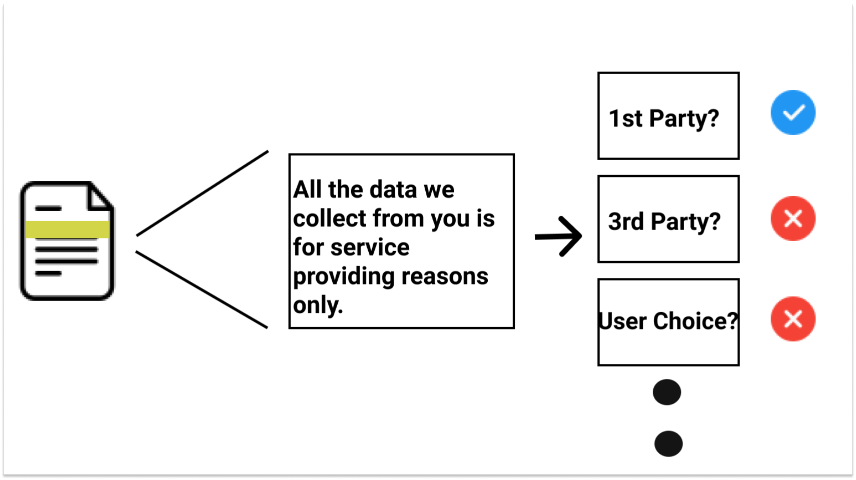
Personal Data Protection Act

In much the same way, for PDPA, we map the policy segments to three broader PDPA categories and four finer categories.

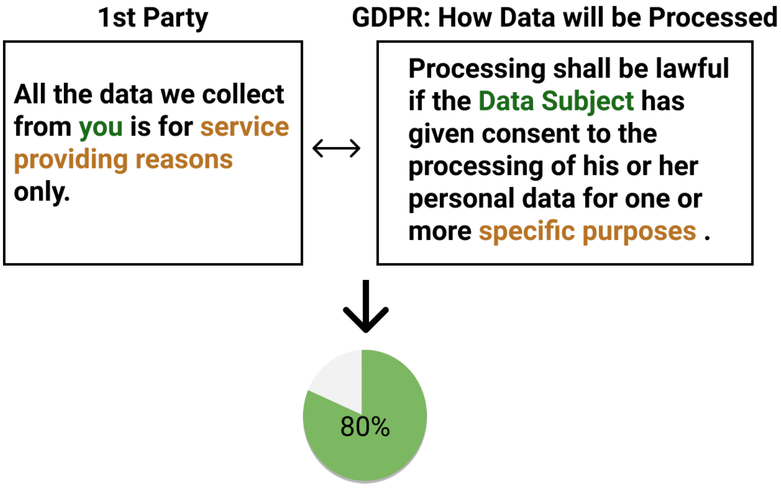


The law segments categories represented by the outer box and the policy categories written in purple.

An example is detailed below:


All the data we collect from you is for service providing reasons only.
1st Party?
✓
3rd Party?
✗
User Choice?
✗

Firstly, the policy is categorized. It can have multiple labels.


1st Party
GDPR: How Data will be Processed
All the data we collect from you is for service providing reasons only.
Processing shall be lawful if the Data Subject has given consent to the processing of his or her personal data for one or more specific purposes .
80%

The law segments with categories related to the policy categories is checked for compliance using the policy segment to find a compliance score.

So following the above example, this policy segment will be compared for compliance with law segments pertaining to *What Data Will be Collected and How*. After comparing the policy text with all relevant law segments a compliance score is generated for each category. The details on computing the score are given in the next section.

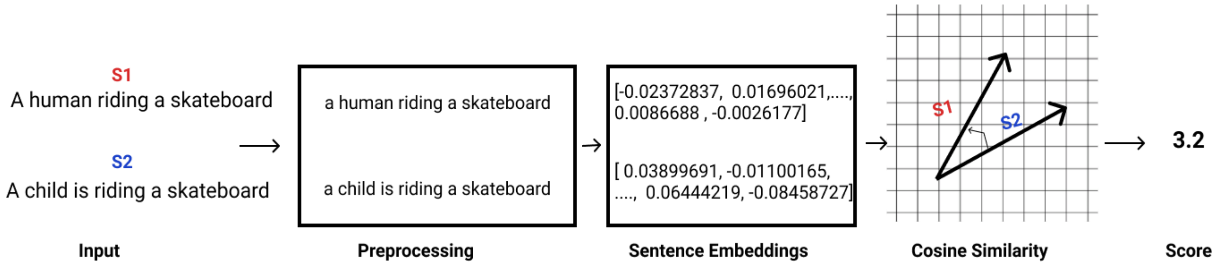
6. Finding Similarity

After allocating categories to segments of laws and policies, we find similarity between segments of the laws and policies which are related. This similarity is used as a measure to decide if the policy is in compliance with the law.

We used BERT[17] word embeddings to find the similarity. Word embeddings such as word2vec and Glove have been useful in improving accuracy across NLP tasks. BERT word embeddings improve upon these methods because it is the first unsupervised, deeply bidirectional system for pre-training NLP. Context-free models such as word2vec or GloVe generate a single "word embedding" representation for each word in the vocabulary, so “bank” would have the same representation in bank deposit and river bank.

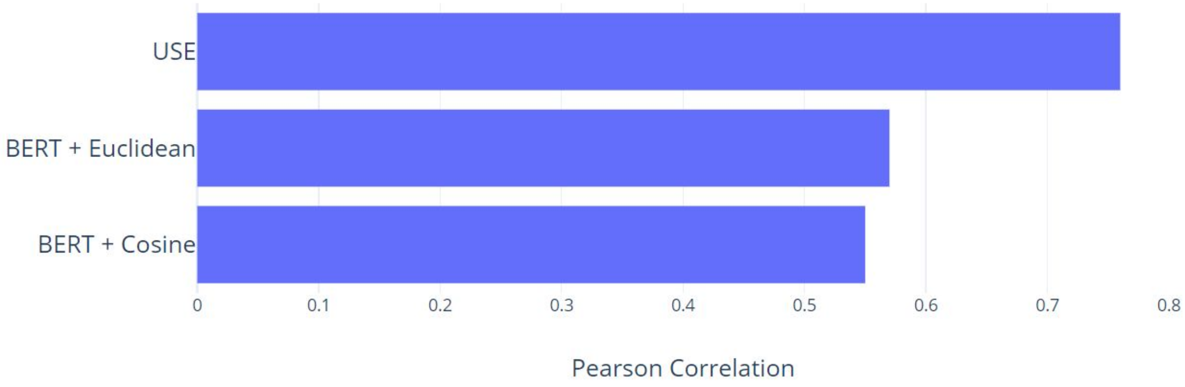
We use the pre-trained BERT uncased model to get sentence embeddings by combining word embeddings through sum and mean across layers of words. These embeddings are then used to calculate cosine similarity of the pair of sentences.

In addition, we also used Universal Sentence Encoding[18]. Here also we obtained sentence embeddings and then used cosine similarity. Universal Sentence Encoder uses transformers and attention based mechanisms to capture the context of the words in a sentence. There are two model architectures present, one uses transformer architecture and gets higher quality embeddings while requiring greater resources and computing power and the second one uses less resources but at the cost of slightly lesser accuracy. Considering that we had to eventually run the whole pipeline on our own systems, we went with the latter one to utilize resources optimally. The architecture we used is the Deep Averaging Network (DAN), first word embeddings along with bi-grams are averaged and then used as input to feedforward deep neural networks (DNN) to get sentence embeddings of 512 dimensions.



An example of similarity score being computed between two sentences.

Out of both sentence encodings, Universal sentence encoding worked better as shown in the graph below and so they were used in the end.



7. Compliance Score

Using the similarity score between the policy segments and the law segments which are related to each other, as the starting point we calculate the compliance score using the formula shown below:

$$\text{Compliance} = (\text{Max} - \text{Score}) / (\text{Max} - \text{Min})$$

As we don't have a labelled dataset for compliance score between law and policy segment, we created a small set to find the required compliance thresholds(max and min) .To decide on where to set the threshold for compliance and non-compliance from the cosine similarity score obtained from Universal Sentence Encodings of policy and law segments, we created a dummy dataset. The dataset consists of a law segment, for both PDPA and GDPR, and policy segment along with a score from 0 to 5; 1 being the least compliant, 5 being completely compliant and 0 showing absolute irrelevance between texts. Then the problem simply reduced to identifying the correct value of thresholds to turn the similarity score into compliance score.

For the GDPR, the threshold was found to be max 0.6 and minimum 0.25, that is, when a policy segment was in complete compliance of a law segment the similarity score was 0.6 and when it had zero compliance the score was 0.25. Using these thresholds, we find the compliance score for the four GDPR requirements; what data will be collected and why, how data will be processed, how long it will be retained and who can be contacted to have data removed or produced.

For the PDPA, the thresholds that gave the optimal results were max .09 for total compliance and min .21 for ½ compliance, with the compliance decrementing as the score increased towards .50.

For each law category we find the overall compliance score by finding the score through the above given formula between each policy segment(under that category) and law segment. We then take an average of these individual scores to find the overall compliance score for each category.

The compliance score of the entire privacy policy is calculated by taking an average across the four categories.

8. Integration

After carrying out the required research on various ways to categorize policies and laws and finding similarity, we integrated our findings to create a deployable application.

We created a web based front end using flask as it is compatible with python. A user can enter their privacy policy and select the appropriate law they want to check the policy against. Then the user is displayed the total compliance score along with the compliance score of each sub category.

The privacy policy is segmented and categorized using a pre trained BERT[17] classifier into five categories. Then the similarity score of the policy segments and the corresponding law segments is calculated using USE[18]. After this, the compliance score of each category of law and overall score is calculated and shown to the user.

Results and Discussions

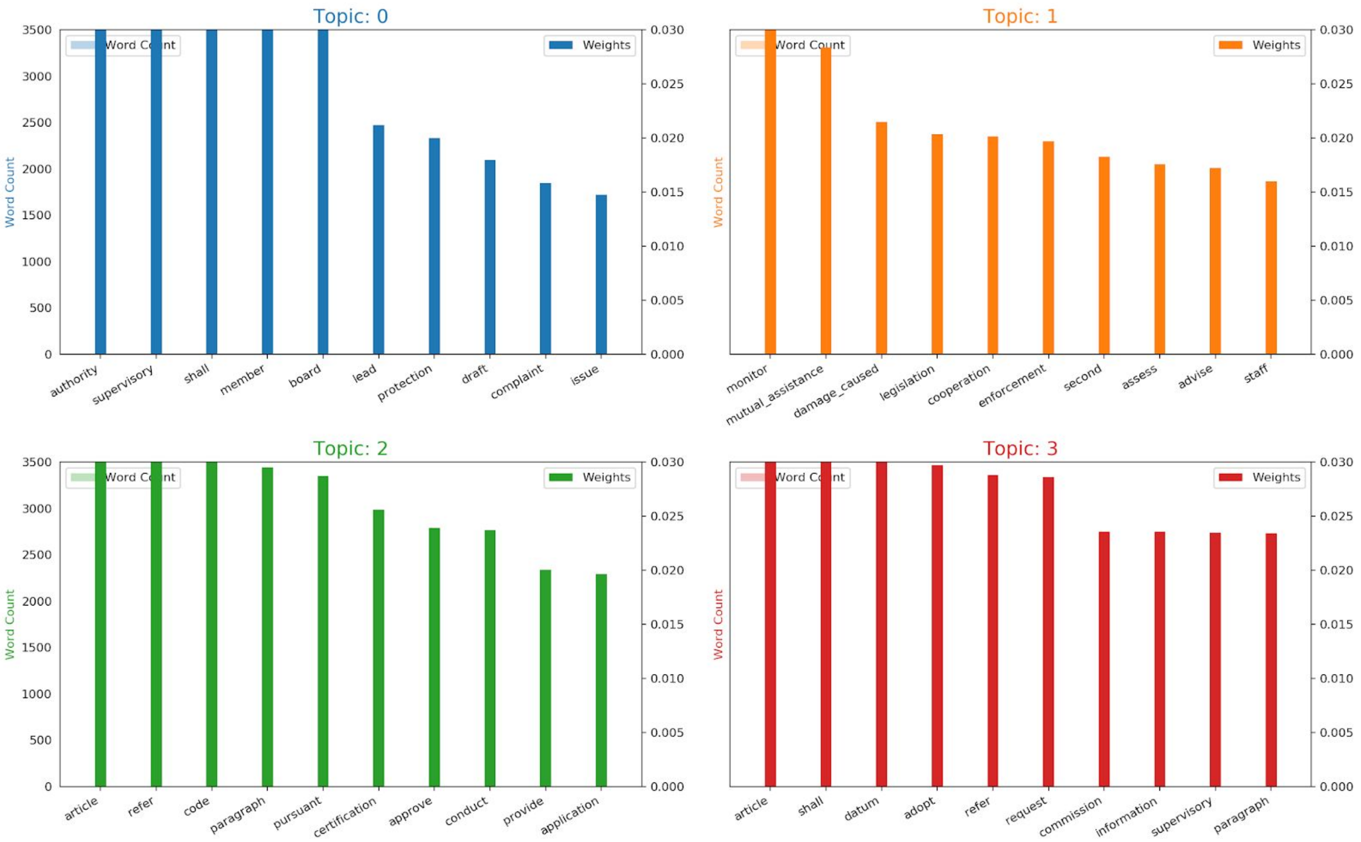
1. Labelling Policies

Categories	F1 Score of Classifiers		
	LR	SVM	BERT
First Party Collection/Use	0.80	0.81	**0.86**
Third Party Sharing/Collection	0.80	0.71	**0.87**
User Choice/Control	0.65	0.47	**0.77**
User Access, Edit, and Deletion	0.47	0.16	**0.68**
Data Retention	0.32	0.23	**0.63**
Data Security	0.57	0.38	**0.67**
Policy Change	0.67	0.56	**0.73**
Do Not Track	1.0	1.0	1.0
International and Specific Audiences	0.74	0.63	**0.92**

BERT binary classifiers outperform SVM and LR for all categories.

2. Labelling Laws

By using LDA we got labelled law segments. The number of topics was set to 10 for GDPR. LDA essentially identifies the word distribution in each topic and then each segment, based on the words it contains, is assigned to a particular group of segments.



The words and their counts are displayed in four of the categories for the GDPR.

For example, a segment containing words mutual assistance occurring together will likely be assigned to topic number 1.

For the PDPA, annotated segments were manually extracted.

3. Finding Similarity

We evaluate our similarity model by using it on the semantic textual similarity dataset as we cannot use it on our own datasets since it is not labelled. The STS dataset comprises sentence pairs from news, captions, and forums. These sentence pairs are labelled for similarity on a scale of 0 to 5 where 5 means complete similarity and 0 means no similarity at all.

The table below shows the pearson correlation obtained on the STS development set:

Model	Pearson Correlation
BERT with cosine similarity	0.55

BERT with euclidean distance	-0.57
Universal Sentence Encoder	0.76

The Universal Sentence Encoder gave better results as compared to BERT.

4. Test Case

We tested our system by using the Nestlé privacy policy. This policy contains a clause about data retention which states *“Nestlé will only retain your personal data for as long as it is necessary for the stated purpose, taking into account also our need to answer queries or resolve problems, provide improved and new services, and comply with legal requirements under applicable laws.This means that we may retain your personal data for a reasonable period after your last interaction with us. When the personal data that we collect is no longer required in this way, we destroy or delete it in a secure manner.*”

We first run the privacy policy against GDPR as it is and the system gives a 99.7% data retention compliance score as it should. Then we replace this section with “*Nestle will store the data for as long as we want*”. When this altered policy is run against GDPR, the compliance report gives a score of 0% Retention.

Conclusion and Future Work

Our work proves that automated compliance check with regard to legal documents gives plausible results. This opens the possibility of using such techniques to find legal compliance in contracts etc. Further work can be done in this area by adding further data protection laws such as Canada’s PIPEDA and US’ Privacy Shield. Another improvement that can be done is to try more complex architectures and models to correlate laws and policies.

References

[1] A. M. McDonald and L. F. Cranor, “*The cost of reading privacy policies*,” ISJLP, vol. 4, p. 543, 2008.

[2] A. Rao, F. Schaub, N. Sadeh, A. Acquisti, and R. Kang, “*Expecting the unexpected: Understanding mismatched privacy expectations online*,” in *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. Denver, CO: USENIX Association, 2016, pp. 77–96.

[3] F. Schaub, R. Balebako, and L. F. Cranor, “*Designing effective privacy notices and controls*,” *IEEE Internet Computing*, vol. 21, no. 3, pp. 70–77, 2017.

[4] Cranor, Lorrie Faith. "Giving notice: why privacy policies and security breach notifications aren't enough." *IEEE Communications Magazine* 43.8 (2005): 18-19.

[5] Wilson, Shomir, et al. "The creation and analysis of a website privacy policy corpus." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016.

[6] Sarne, D., Schler, J., Singer, A., Sela, A. and Bar Siman Tov, I., 2019, May. Unsupervised Topic Extraction from Privacy Policies. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 563-568). ACM

[7] Angeli, Gabor, Melvin Jose Johnson Premkumar, and Christopher D. Manning. "Leveraging linguistic structure for open domain information extraction." *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2015.

[8] Linden, Thomas, et al. "The privacy policy landscape after the GDPR." *arXiv preprint arXiv:1809.08396* (2018).

[9] Jensen, Carlos, and Colin Potts. "Privacy policies as decision-making tools: an evaluation of online privacy notices." *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2004.

[10] M. Hochhauser (2001). *Lost in the fine print: Readability of financial privacy notices*. Retrieved September 30, 2019 from http://www.privacyrights.org/ar/GLB-Reading.htm.

[11] Antón, Annie I., Julia Brande Earp, and Angela Reese. "Analyzing website privacy requirements using a privacy goal taxonomy." *Proceedings IEEE Joint International Conference on Requirements Engineering*. IEEE, 2002.

[12] Harkous, Hamza, et al. "Polisis: Automated analysis and presentation of privacy policies using deep learning." *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 2018.

[13] Sarne, D., Schler, J., Singer, A., Sela, A. and Bar Siman Tov, I., 2019, May. Unsupervised Topic Extraction from Privacy Policies. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 563-568). ACM

[14] Ramanath, Rohan, et al. "Unsupervised alignment of privacy policies using hidden markov models." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2014.

[15] He, Hua, Kevin Gimpel, and Jimmy Lin. "Multi-perspective sentence similarity modeling with convolutional neural networks." Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015.

[16] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. "*MAPS: Scaling Privacy Compliance Analysis to a Million Apps.*" Privacy Enhancing Technologies Symposium 2019.

[17] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[18] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Sung, Y. H. (2018). Universal sentence encoder. arXiv preprint arXiv:1803.11175.

[19] Liu, F., Wilson, S., Story, P., Zimmeck, S. and Sadeh, N., 2017. Towards Automatic Classification of Privacy Policy Text.

[20] Renaud K, Shepherd LA. How to make privacy policies both GDPR-compliant and usable. In2018 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA) 2018 Jun 11 (pp. 1-8). IEEE.

[21] General Data Protection Regulation GDPR. Retrieved March 5, 2019 from https://gdpr-info.eu/

[22] PERSONAL DATA PROTECTION ACT 2012. Retrieved June 10, 2020 from https://sso.agc.gov.sg/Act/PDPA2012

[23] Data Protection Starter Kit. Retrieved June 4, 2020 from https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/dp-starter-kit---171017.pdf