Fine Print

Privacy Policies and Data Protection Laws

Ayesha Qamar
National University of Computer and Emerging Sciences
i160104@nu.edu.pk

Tehreem Javed
National University of Computer and Emerging Sciences
i160086@nu.edu.pk

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 often 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.

Keywords Privacy Policies, Data Protection Laws, GDPR, Privacy Policy Compliance

1. Introduction

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. 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 an application or software is doing with the user's data but that information alone is not adequate to judge a policy's transparency and its usefulness[4].

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

2. Motivation

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.

Privacy policies and data protection 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

CONF 'yy, Month d–d, 20yy, City, ST, Country.

http://dx.doi.org/10.1145/nnnnnnn.nnnnnnn

of Things there has been an escalation in the number and complexity of privacy policies themselves [3].

3. Related work

In 2016, Wilson et al[5] introduced a taxonomy for privacy policies, OPP-115, and made this corpus of 115 annotated policies publicly available. Since then, much work has been done to understand various aspects of privacy policies[12, 16]. Sarne et al[13] presented how using an unsupervised technique, Latent Dirichlet Allocation(LDA), can also provide a taxonomy for privacy policies that is much more fine-grained. 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. They also showed that the taxonomy obtained had a substantial overlap with that of OOP-115. The research also provides insight into the topics that are being addressed in privacy policies these days. Apart from that, hidden markov models[14] have been used previously to categorize privacy policies in an unsupervised way. The policies were 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.

Tesfay et al[20] presented an approach to summarize long privacy policies using Machine Learning and then check against GDPR aspects as a criteria. Work has also been done to visually represent policies, for that Harkous et al[12] developed a framework using Deep Learning techniques and the power of Convolutional Neural Networks to analyse policies on a finer level and developed a hierarchy to organise the information in privacy policies. They then presented the information in policies in a visual format and also provided a question answering interface where users' queries about a privacy policy are answered.

Recently, Zimmeck et al[16] compared the actual practice of a million apps with those stated in their privacy policies and flagged any discrepancies as compliance issues.

While work has been done to categorize, summarize and visualize privacy policies, none of the work has yet analysed the privacy policies to check their compliance with the very laws that regulate them.

4. Specific Proposal

We propose a system which, given a privacy policy checks its compliance with a data protection law. For this, we first collected a dataset of privacy policies and then labelled them. Data protection laws were also segmented and labelled. Finally, we checked for the compliance of the resulting chunks of policies with those of the law. The details are mentioned in the following sections.

4.1 Crawling 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 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 *60000* policies using these links.

It was also important to check if the extracted policy was indeed in English or if it was 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 is in English.** We used the 'langdetect' Python library.
- **The text belongs to a privacy policy.** We employed regular expressions to achieve this. We checked on keywords like "404 Error", "Webpage not found" etc.
- **The policy is of substantial length.** To ensure that the policy was legitimate we also removed any policy that had a length less than 80 characters.

4.2 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 cate-

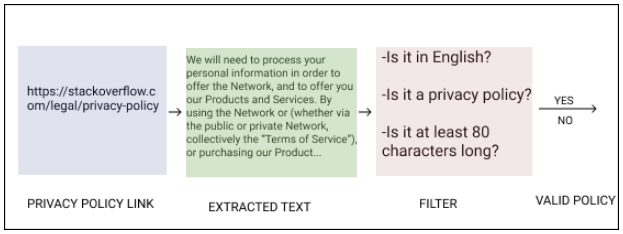

Figure 1. The pipeline employed to ensure the gathered policies are of high quality.

gories 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 0.75 overlap between the annotations, i.e., at least 2 of the 3 students have given the same label.

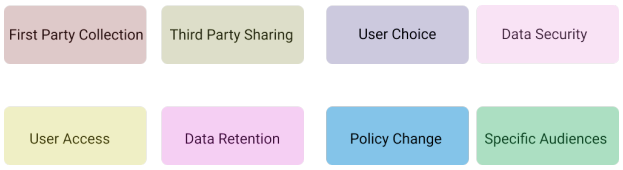

Figure 2. The 8 high level categories in the OPP-115 dataset. The other two categories not shown are *other* and *Do Not Track*. The latter category is not useful as it is no longer mentioned in policies.

To begin with, we extracted the 115 annotated policies from the OPP-115[5] 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 Text[19] as a starting point and trained a logistic regression model and a Support Vector Machine model for classification. In addition, we also used a fine tuned version of the BERT model. We trained classifiers for all the ten datasets and calculated their F1 scores.

4.3 Labelling Laws

For now we are only considering a single law i.e., the GDPR. The first step to labelling the laws is to segment them. 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.

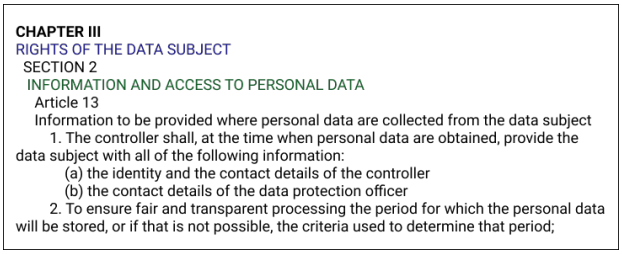

Figure 3. The structure of the GDPR. The hierarchy consists of Chapters, Sections, Articles and then points in those Articles.

We decided to use topic modelling for grouping together similar segments and thus creating a taxonomy of the law. We used Latent Dirichlet Allocation (LDA) to achieve that. The decision to use LDA was based on the promising results achieved by [13] to label privacy policies. 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 experimented with several values of k and found that setting it to 10 gave the most optimal results in our case.

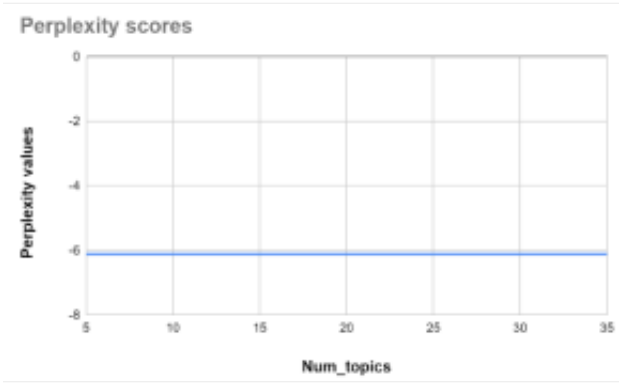

Figure 4. The perplexity score plotted against multiple values of the number of topics.

Perplexity scores did not give any insightful information to decide the value of k. Therefore, we used the coherence score as the deciding factor instead. The best coherence value was achieved when k was set to 5. But such a coarse labelling would not have served our purpose, since we know that the GDPR contains at least 10 different topics as those are the number of different chapters, so we went with the next favourable value of 15.

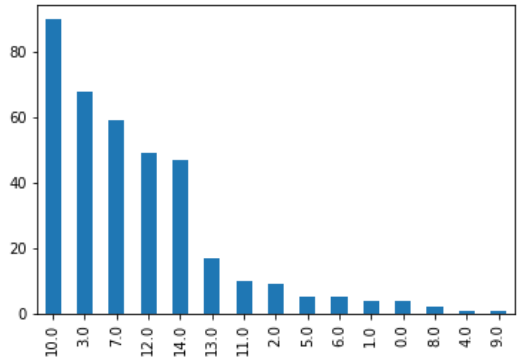

Figure 5. The number of segments belonging to each topic is depicted when k is set to 15. As shown, some of the topics only have one or two segments assigned to them.

But setting the number of topics to 15 gave rise to a few topics containing only one or two segments only and merging them seemed to be a sensible option. So in the end we decided to keep the number of topics to 10, doing so also opened up the possibility of doing a one-to-one mapping between laws and policies.

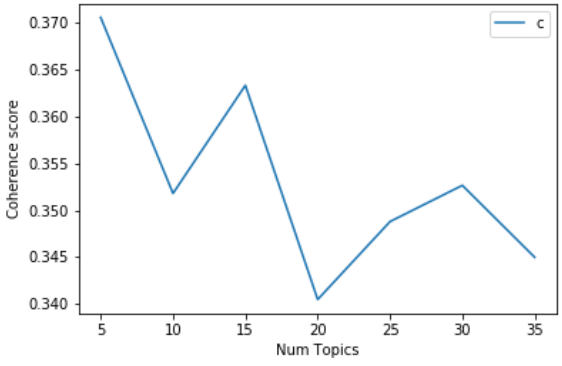

Figure 6. The best coherence score was achieved when k was set to 5.

Figure 7 shows the most occurring words in four 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.

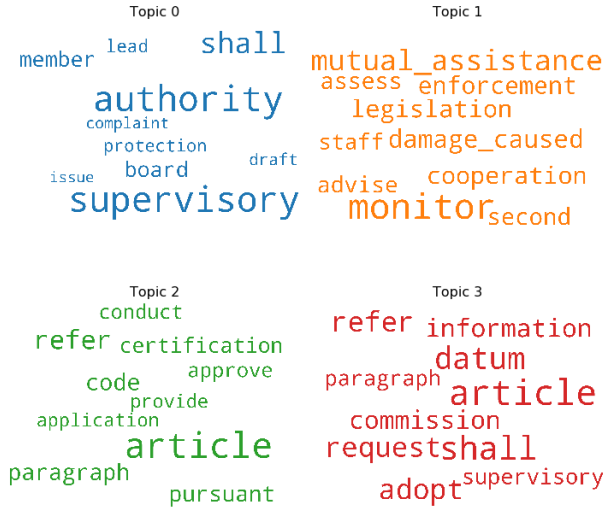

Figure 7. A visual representation of the most frequent words of some of the topics.

4.4 Finding Similarity

After allocating categories to segments of laws and policies, we find similarity between segments of the laws and policies which fall under the same category. This similarity is used as a measure to decide if the policy is in compliance with the law. We used BERT[17] word embeddings and Universal Sentence encoding[18] 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 pre trained BERT uncased model to get sentence embeddings by combining word embeddings through mean across layers of words. These embeddings are then used to find similarity between pairs of policy and law segment using cosine similarity and euclidean distance.

By using Universal Sentence Encoding[18], 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 in a 512 dimensional vector.

5. Experimental Evaluation

- **Labelling Policies:** To evaluate the labelling of privacy policies we used a held out dataset and checked the accuracy of our models(SVM, LR, BERT) on that data. The complete results can be seen in figure 9. BERT gave a better F1 score for most categories.
- **Labelling Laws:** For laws, we are going to have an expert verify the labelling and annotation since there is no labelled dataset of data protection laws available.
- **Finding Similarity:** Due to the unavailability of policy and law compliance dataset, we evaluate our similarity model by using it on the semantic textual similarity development dataset. The STS dataset comprises of sentence pairs from news, captions, and forums genre. 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 Pearson Correlation obtained by using BERT embedding and taking mean of all word vectors and sum of all vectors as well as the correlation obtained by using Universal Sentence encoding is shown in figure 8.

Model	Pearson Correlation
BERT with cosine similarity	0.55
BERT with euclidean distance	-0.57
Universal Sentence Encoder	0.76

Figure 8. The pearson correlation obtained on Universal Sentence Encoding outperforms BERT models.

Categories	F1 Score of Classifiers		
	LR	SVM	BERT
First Party Collection/Use	0.79	0.81	0.84
Third Party Sharing/Collection	0.80	0.71	0.84
User Choice/Control	0.65	0.47	0.71
User Access, Edit, and Deletion	0.44	0.16	0.41
Data Retention	0.32	0.23	0.0
Data Security	0.57	0.38	0.77
Policy Change	0.67	0.56	0.66
Do Not Track	1.0	1.0	1.0
International and Specific Audiences	0.74	0.63	0.90

Figure 9. The F1 score of the Logistic Regression, Support Vector Machine and BERT across all the categories.

6. Expected Results

(what do you hope your experiments will show?) We have made baseline models of all the use-cases of the proposed system. We aim now to integrate these models and deploy an application. The end result would enable an user to enter a privacy policy and select from a list of data protection laws. Our application would then generate a report which would show the overall compliance the policy has with the law. It would also highlight areas of the policy and the associated law that the policy is violating.

We are also going to further explore more complex models to associate and find correlation between privacy policy segments and corresponding chunks

of laws. We plan to employ Autoencoders or multi-perspective CNNs or a combination of both, whichever will give us better results. We are also going to look into the ways we can extract useful information about sentence/segment similarity from their embeddings in n-dimensional space. Also, we are going to expand the laws to incorporate multiple countries' data protection laws, like Canada's PIPEDA, US' Privacy Shield etc.

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] Tesfay, W. B., Hofmann, P., Nakamura, T., Kiyomoto, S., Serna, J. (2018, April). I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR. In Companion Proceedings of the The Web Conference 2018 (pp. 163-166). International World Wide Web Conferences Steering Committee.