FYP-1 Mid Evaluation Report

FINE PRINT

Privacy Policies and Cyber Laws

Team Members:

Tehreem Javed 16i-0086

Ayesha Qamar 16i-0104

Supervisor:

Dr. Mirza Omer Beg

PLAGIARISM STATEMENT

We certify that this report is our own work. We have acknowledged all research items used in this work and have not copied in part or whole or otherwise the work of any other individual.

Name: Ayesha Qamar

Signature:_____

Name: Tehreem Javed

Signature:_____

Table of Contents

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.

Motivation

A reliable privacy policy validation tool can be of value to companies, consumers and regulators alike. Companies can use it to help in privacy policy modeling during product launches. It can also be of help when introducing the product to a foreign country as it will significantly ease the process of verifying the policies in light of the new data protection law. Consumers can avail the service to better educate themselves regarding the policies they are agreeing to as well as their legal rights without having to go through an ambiguous legal document. Departments or

organizations dedicated to cyber security can also use this service to find loopholes in any given policy and take relevant action.

Literature Review

Research Item 1

Unsupervised Topic Extraction from Privacy Policies

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 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 merging 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 the 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 method 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. Where as 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

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 not 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

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 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 embeddings 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 analysis 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 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

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 pairs, 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

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 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.

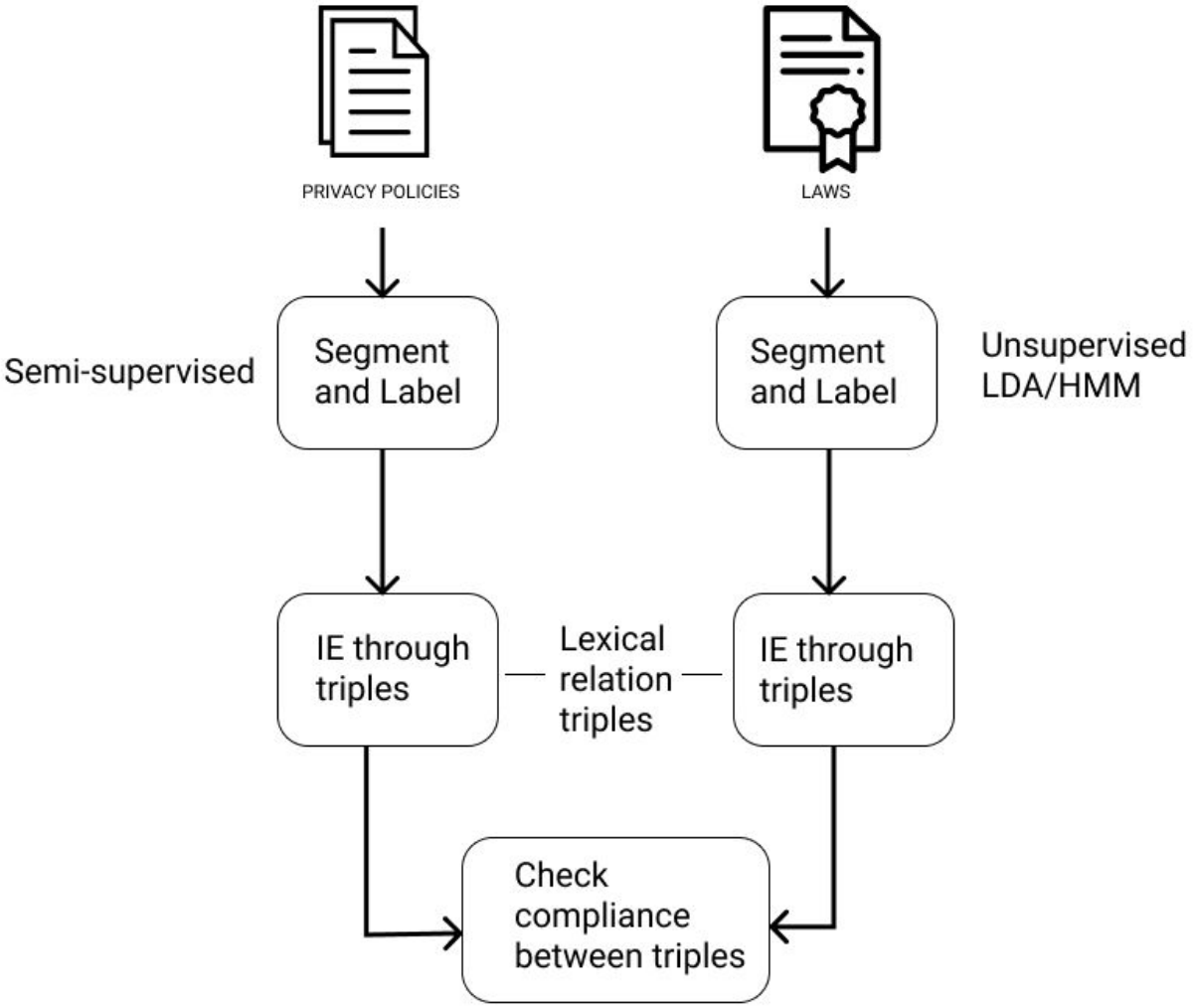
Research Gap

There have been many studies on privacy policies. Some have highlighted their relation and dynamics with changing laws [8]. Others have explored their usability [9] and readability[10]. While some have also analysed their content [11].

All the recent works on privacy policies have failed to highlight their correspondence with data protection laws. Even though there have been some studies about the relation of laws with policies[8], none has catered to the problem of seeing if the policies are conforming to the laws. Furthermore, some research has aimed to find ways to segment and analyse privacy policies [12, 13, 14] for their better understanding; but no such effort has been made to analyse the laws regulating them.

There clearly is a need to correlate policies with data protection laws. Since those are the very laws that dictate what the policies address. With the recent trend of countries amending their cyber laws to keep up with the ever growing and dynamic nature of the data collection through the internet, there is a greater need now more than ever to have a tool to automate this process.

Proposed Approach



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 or Hidden Markov Model. Afterwards, relation triples will be used to extract information from the policies and laws' segments. And lastly those triples of policy and law's chunks will be checked for conformance.

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.