DETECTING TEXTUAL SALIENCY IN PRIVACY POLICY

DS5500 PROJECT PROPOSAL - PHASE 2
NORTHEASTERN UNIVERSITY

Kaushik Nishtala
nishtala.k@northeastern.edu

Shubhanshu Gupta
gupta.shubh@northeastern.edu

November 16, 2021

1 Summary

Data privacy regulations have increased over the last twenty years, making these documents almost double in length and complexity. As initially proposed, the primary objective of this project lies in communicating a high level and a more granular breakdown of privacy policy practices to counteract the risks posed by the ignorance of these documents. We aim to develop language models to extract and deliver salient policy information to end-users in the form of the INFORM and the QUERY modules.

The functionality of the INFORM module is twofold. The first part enables users to explore the general trends and patterns in privacy policy practices using Tableau dashboards. In contrast, the second part facilitates the analysis of an individual user-specified policy. To that end, we classify privacy policy segments into a set of pre-defined privacy practice categories using machine learning algorithms. The first phase of the project dealt majorly with implementing the INFORM module, where we produced baseline language models with promising results.

Motivated by the need to add model and data complexities, we hope to extend our prior work by developing sequence models and transformers and using contextual word embeddings in this project's final phase. Subsequently, we implement the QUERY module that enables users to ask policy-related questions within a specified policy. By retrieving the most relevant information from the policy document for a given query, this system reduces the burden of searching the target information from a lengthy policy document. Finally, we also expect a significant chunk of this phase to involve the implementation/integration of the machine learning models with an interactive public-facing web framework.

2 Proposed plan

We believe that the project continues its desired trajectory, and our overall goals and objectives remain unchanged as we move on to the second phase. In the previous phase of the project, we developed baseline models with extensive pre-processing and tuning and obtained results with reliable predictions. However, the models fail to capture essential distinctions between the categories effectively due to the limitations in how the data is represented (TF-IDF and random embeddings) and the models themselves. Thus, we believe there is a need to scale up the data and model complexities further to gain improved performances on the privacy policy text classification. Therefore, we plan to supplement the model performances by utilizing word embeddings and complex neural architectures.

However, using context-free embeddings would generate a single embedding representation for each word in the vocabulary irrespective of the context ("minute" would have the exact representation in "one minute" and "minute fraction of an inch"). We instead plan to adopt the approach of training our word embeddings specific to the domain of privacy policies. To that end, we seek to utilize a corpus of around 1M raw privacy

policies [7] using fastText [2] to provide more domain-specific contextual embeddings for the models to learn from. We would also like to extend this further by employing sequence model architectures (such as the Long Short Term Memory (LSTM), Gated Recurrent Units (GRU)), as well as transformer models like BERT (Bidirectional Encoder Representations from Transformers) [3] in an effort to achieve competitive scores. Intuitively, deep bi-directional contextual models like BERT generate a representation of each word based on the other words in the sentence (on either side of the word, hence bi-directional).

Regarding the implementation of the QUERY module, we plan to use the recently released PolicyQA dataset [1] curated from the OPP-115 corpus [8]. The dataset enables us to model the task as predicting the answer in the given policy segment using existing neural approaches from literature (utilizing Hugging Face transformer API [9]). PolicyQA is a reading-comprehension style dataset that includes information about the annotated spans, corresponding policy segments, and the associated Practice, Attribute, Value triples derived from the OPP-115 corpus to form the examples in the dataset. Two annotators were involved in the annotation process to form questions for each specific triple, producing 714 individual questions from their 258 corresponding triples. We presume that a similar dataset SQuAD (Stanford Question Answering Dataset) [5] curated from Wikipedia articles could help provide more insight into the process. Finally, after training the QA models, we wish to integrate them with the web framework, which could help the end-users query and extract information from previously unseen privacy policies given as input to the model.

3 Preliminary results

The previous phase of the project focused solely on implementing the INFORM module, where we presented a series of processes, decisions, and artifacts involved. Utilizing the OPP-115 corpus [8], we performed a slew of pre-processing tasks such as iterative splitting [6], cleaning, encoding, and tokenizing the policy texts. We posed the objective as a multi-label classification problem since each segment can contain information for multiple categories. Our primary focus was to establish baseline models while motivating the need for adding complexity from both the dataset and model architecture standpoints. Leveraging the idea of "the strength of weak learners", we developed traditional one-vs-rest classifiers, namely Logistic Regression, Support Vector Machines (RBF), Random Forest, and XGBoost. In addition, we employed Convolutional Neural Networks (CNN) for multi-label text classification [4] with randomly initialized word embeddings.

These baselines enabled us to perform an ablation study via hyper-parameter tuning and helped discover code issues (such as changes in the architecture). Besides producing fairly decent performance results, they helped promote interpretability and data-driven decision-making that helped impact our workflow in this final phase to a great degree. Concretely, understanding the data structure and the label annotation scheme led us to apply effective pre-processing methods to account for data issues, namely the data imbalance and multi-label splitting. Interpreting model results allowed us to identify any categories that the models are consistently underperforming on. Prioritizing critical tradeoffs such as the performance metrics, latency, size, compute requirements will aid us with practical considerations to develop and deploy the web framework. Performing an organized ablation study helped us track and identify multiple facets and parameters of the modeling process that may prove helpful in building complex architectures in this phase. Finally, we believe that understanding the strengths and weaknesses of each baseline model in predicting an aspect (category) of the data may eventually help us in the pursuit of building ensembles of these weak learners.

4 Project milestones

The final phase of our project focuses on the extension of the INFORM module to some degree before moving on to the implementation of the QUERY module. Concretely, by mid-phase (11/30/2021), we expect to produce the results for sequence models and transformers for the multi-label classification problem in the INFORM module. This would also include training these architectures using our custom-trained contextual word embeddings. Also, we anticipate that we will have made some progress on the front-end and back-end development of the category classification of privacy policies from the INFORM module. By the end of the semester (12/15/2021), we imagine being finished with implementing the QUERY module. We will also have obtained question-answering performance results for the transformer models, including BERT. We believe

that the above schedule would enable us to evaluate our methods' progress and success effectively. Finally, we expect to make adequate progress on bringing together the functionality of the INFORM and the QUERY modules to present the final deliverable as an interactive public-facing web framework.

Github The code used for analysis, experiments and the web framework is provided in a repository at `https://github.com/Kau5h1K/ds5500-userprivacy`.

References

[1] W. AHMAD, J. CHI, Y. TIAN, AND K.-W. CHANG, *PolicyQA: A reading comprehension dataset for privacy policies*, in Findings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov. 2020, Association for Computational Linguistics, pp. 743–749.

[2] P. BOJANOWSKI, E. GRAVE, A. JOULIN, AND T. MIKOLOV, *Enriching word vectors with subword information*, arXiv preprint arXiv:1607.04606, (2016).

[3] J. DEVLIN, M.-W. CHANG, K. LEE, AND K. TOUTANOVA, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019.

[4] H. HARKOUS, K. FAWAZ, R. LEBRET, F. SCHAUB, K. G. SHIN, AND K. ABERER, *Polisis: Automated analysis and presentation of privacy policies using deep learning*, 2018.

[5] P. RAJPURKAR, J. ZHANG, K. LOPYREV, AND P. LIANG, *SQuAD: 100,000+ questions for machine comprehension of text*, in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, Nov. 2016, Association for Computational Linguistics, pp. 2383–2392.

[6] K. SECHIDIS, G. TSOUMAKAS, AND I. VLAHAVAS, *On the stratification of multi-label data*, in Machine Learning and Knowledge Discovery in Databases, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, eds., Berlin, Heidelberg, 2011, Springer Berlin Heidelberg, pp. 145–158.

[7] M. SRINATH, S. WILSON, AND C. L. GILES, *Privacy at scale: Introducing the privaseer corpus of web privacy policies*, 2020.

[8] S. WILSON, F. SCHAUB, A. A. DARA, F. LIU, S. CHERIVIRALA, P. GIOVANNI LEON, M. SCHAARUP ANDERSEN, S. ZIMMECK, K. M. SATHYENDRA, N. C. RUSSELL, T. B. NORTON, E. HOVY, J. REIDENBERG, AND N. SADEH, *The creation and analysis of a website privacy policy corpus*, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, Aug. 2016, Association for Computational Linguistics, pp. 1330–1340.

[9] T. WOLF, L. DEBUT, V. SANH, J. CHAUMOND, C. DELANGUE, A. MOI, P. CISTAC, T. RAULT, R. LOUF, M. FUNTOWICZ, J. DAVISON, S. SHLEIFER, P. VON PLATEN, C. MA, Y. JERNITE, J. PLU, C. XU, T. L. SCAO, S. GUGGER, M. DRAME, Q. LHOEST, AND A. M. RUSH, *Transformers: State-of-the-art natural language processing*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Oct. 2020, Association for Computational Linguistics, pp. 38–45.