DETECTING TEXTUAL SALIENCY IN PRIVACY POLICY

DS5500 PROJECT PROPOSAL - PHASE 1
NORTHEASTERN UNIVERSITY

Kaushik Nishtala
nishtala.k@northeastern.edu

Shubhanshu Gupta
gupta.shubh@northeastern.edu

October 8, 2021

Github https://github.com/Kau5h1K/ds5500-userprivacy

1 Summary

Websites, mobile apps, and other product and service providers share how they gather, use and manage customers' data in the form of privacy policy documents. However, due to their lengthy and complex nature, a majority of people tend to ignore these documents [9]. To counteract the risks posed by this ignorance, we aim to develop language models to extract and deliver salient policy information to end-users. To that end, we propose two use case-specific modules, namely the INFORM module and the QUERY module, that collectively enable users to understand their rights better.

The primary objective of the INFORM module lies in communicating key insights by exploiting a dataset containing 115 privacy policies defined by US companies. Additionally, we seek to build a selection of models that can annotate pre-defined categories to privacy policy paragraphs using supervised machine learning. The QUERY module can enable users to ask policy-related questions using a freeform question-answering system for privacy policies. By retrieving the most relevant information from the policy document given a question, this system facilitates searching the target information from a lengthy policy document. To this purpose, we make use of a Question-Answer dataset that provides 714 human-annotated questions written for a wide range of privacy practices. Finally, we attempt to build on earlier studies [5, 7, 14] in the literature and make the functionality publicly available using an interactive web framework.

2 Proposed plan

We reckon that the curation of the OPP-115 corpus of privacy policies and its annotation scheme produced by domain experts [15] is central to our efforts. An individual data practice in the corpus belongs to at least one of the ten data practice categories (segments), further classified into a set of practice attributes. We expect to build classifiers to annotate paragraphs of a document with high-level categories by framing this as a multi-label classification problem. This segment-based classification is particularly appropriate because a segment can contain information about multiple categories, such as the first-party collection of data, third-party sharing, and data security. However, due to the intricate nature of the dataset, a moderate level of data preparation may be necessary prior to any downstream analyses.

The first step of the INFORM module is to build a visualization framework to convey high-level insights from the corpus. The produced plots may answer holistic questions such as "How is my health information collected/stored in general?". This functionality can be achieved by integrating a SQLAlchemy database holding the cleaned corpus with flask web framework and Dash to serve interactive plots. Next, we conduct experiments using traditional machine learning methods and neural networks to predict one or more categories for each paragraph of a given privacy policy. More specifically, in the case of traditional models, we seek to build multiple category-specific classifiers that accept word vectors transformed using Paragraph2Vec [6] and the GENSIM toolkit [13]. Inspired by prior work [12], we also intend to explore a sequence labeling approach to apply hidden Markov models (HMMs) to privacy policy text.

Given the wide range of applicability of deep neural networks in natural language processing, we want to leverage Google Cloud Platform to build Convolutional Neural Networks (CNN) with general-purpose pre-trained embeddings such as Word2vec [8], and GloVe [10]. We also look to utilize a corpus of 441k privacy policies collected from mobile apps [17] to train custom word embeddings for the privacy-policy domain using fastText [2], which can then be used to seed the CNN's embedding layer. Furthermore, we wish to evaluate Bidirectional Encoder Representations from Transformers (BERT) [4] that proved very useful in language processing. We choose to employ out-of-the-box $BERT_{BASE}$, fine-tuned BERT, and LEGAL-BERT [3], a recently proposed BERT model trained on legal domain-specific corpora. We hope to provide a thorough ablation analysis through these methods to uncover any critical factors that affect category classification.

Although not concrete enough, we have been tossing ideas around the best way to frame the QA task (Information Retrieval QA or NLP QA) in the QUERY phase. We are interested in exploiting a recently released PolicyQA dataset [1] curated from the OPP-115 corpus. 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 [16]). We presume that a similar dataset SQuAD (Stanford Question Answering Dataset) [11] curated from Wikipedia articles could help provide more insight into the process. Finally, after training the classifiers and QA models, we wish to test the best-performing models by integrating them with a web framework, which could help the end-users to annotate and query previously unseen privacy policies scraped using a web crawler.

3 Preliminary results

The curated OPP-115 corpus includes privacy policies from sixteen different sectors according to DMOZ.org's top-level website sectors. Figure 1 shows the distribution of policies in each of these sectors. We also note that the data collection team has limited the "Regional" sector to the "U.S." sub-sector to ensure consistency in legal and regulatory requirements. In Figure 2, We identify the ten different data practice categories for each policy that act as the prediction labels for our supervised learning models. This frequency distribution conveys the usage/coverage of these labels in each policy document. Data Practices such as "First Party Collection/Use" are extensively used in a majority of the policies, while "Do Not Track" is only used a few times in some documents.

Finally, Figures 3 and 5 show the policy's probability density functions of data practices and word count, respectively. We find that we have policies ranging from 80 words to 6000 words in length. Similarly, the number of data practices marked by annotators ranges from 9 to 600, indicating a need to normalize the data.

4 Project milestones

Given the scope of our proposed work, we believe it meets the requirements of a full-semester project. Phase 1 of our study focuses solely on the implementation of the INFORM module. Concretely, by mid-phase (10/19/2021), we expect to complete data preparation, produce detailed visualizations using Tableau and obtain results for the traditional models that help annotate the policy documents. Also, we anticipate that we will have made some progress on the front-end and back-end development of the visualization framework from the INFORM module. By the end of Phase 1 (11/02/2021), we imagine being finished with the visualization framework and will also have obtained category classification results for the neural network models, excluding BERT. The above schedule gives us ample time to learn and apply our knowledge meticulously and enables us to evaluate the progress and success of our methods.

5 Figures

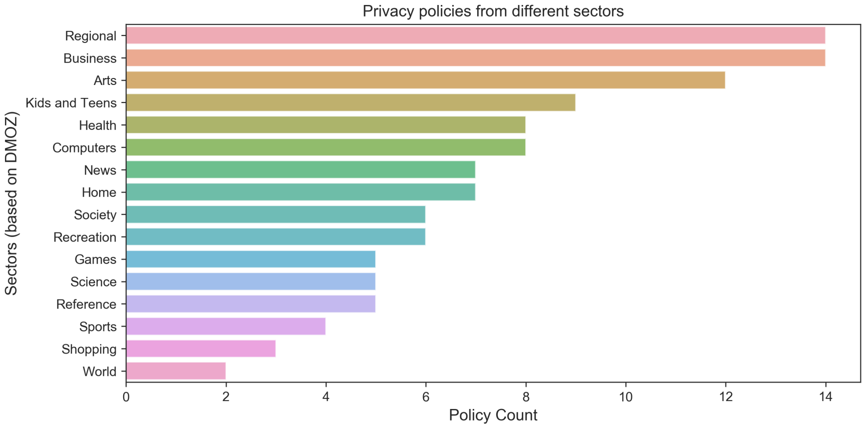

Figure 1: Distribution of the sectors (based on DMOZ) in the corpus

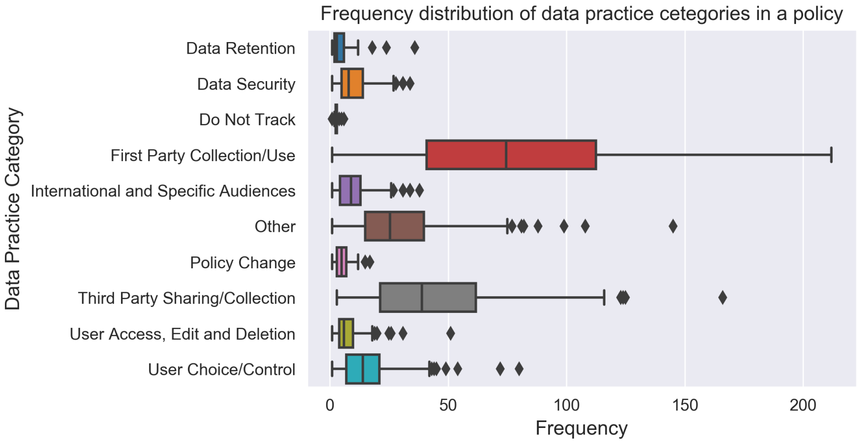

Figure 2: Frequency distribution of data practice categories in policies

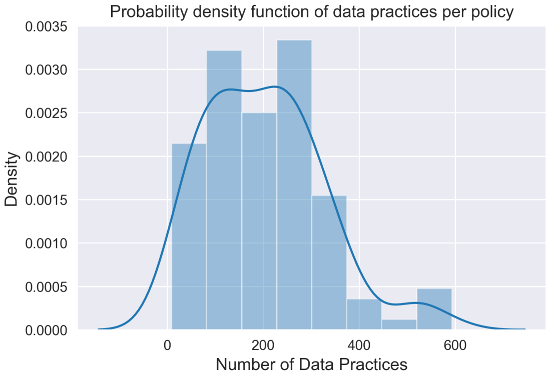

Figure 3: PDF of number of data practices per policy

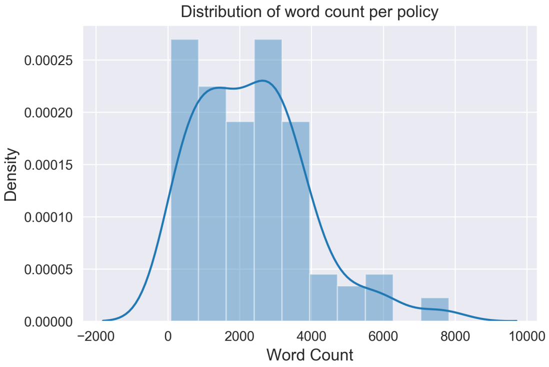

Figure 4: Word count distribution per policy

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] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, *Legal-bert: The muppets straight out of law school*, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019.

[5] 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.

[6] Q. V. Le and T. Mikolov, *Distributed representations of sentences and documents*, 2014.

[7] F. Liu, S. Wilson, P. Story, S. Zimmeck, and N. M. Sadeh, *Towards automatic classification of privacy policy text*, 2017.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013.

[9] J. A. Obar and A. Oeldorf-Hirsch, *The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services*, Information, Communication & Society, 23 (2020), pp. 128–147.

[10] J. Pennington, R. Socher, and C. D. Manning, *Glove: Global vectors for word representation*, in Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[11] 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.

[12] R. Ramanath, F. Liu, N. Sadeh, and N. Smith, *Unsupervised alignment of privacy policies using hidden markov models*, vol. 2, 06 2014, pp. 605–610.

[13] R. Rehurek and P. Sojka, *Gensim–python framework for vector space modelling*, NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3 (2011).

[14] N. Sadeh, A. Acquisti, T. Breaux, L. Cranor, A. McDonald, J. Reidenberg, N. Smith, F. Liu, N. Russell, F. Schaub, S. Wilson, J. Graves, P. Leon, R. Ramanath, and A. Rao, *Towards usable privacy policies: Semi-automatically extracting data practices from websites' privacy policies (poster)*, 07 2014.

[15] 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.

[16] 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.

[17] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. Reidenberg, N. Russell, and N. Sadeh, *Maps: Scaling privacy compliance analysis to a million apps*, Proceedings on Privacy Enhancing Technologies, 2019 (2019), pp. 66–86.