Automated Privacy Policy Annotation with Information Highlighting Made Practical Using Deep Representations

Abdulrahman Alabduljabbar
University of Central Florida
jabbar@knights.ucf.edu

Ahmed Abusnaina
University of Central Florida
ahmed.abusnaina@knights.ucf.edu

Ulku Meteriz
University of Central Florida
meteriz@knights.ucf.edu

David Mohaisen
University of Central Florida
mohaisen@ucf.edu

ABSTRACT

The privacy policy statements are the primary mean for service providers to inform Internet users about their data collection and use practices, although they often are long and lack a specific structure. In this work, we introduce TLDR, a pipeline that employs various deep representation techniques for normalizing policies through learning and modeling, and an automated ensemble classifier for privacy policy classification. TLDR advances the state-of-the-art by (i) categorizing policy contents into nine privacy policy categories with high accuracy, (ii) detecting missing information in privacy policies, and (iii) significantly reducing policy reading time and improving understandability by users.

ACM Reference Format:
Abdulrahman Alabduljabbar, Ahmed Abusnaina, Ulku Meteriz, and David Mohaisen. 2021. Automated Privacy Policy Annotation with Information Highlighting Made Practical Using Deep Representations. In ,. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The privacy policy statements are legal statements that inform Internet users about websites and businesses data collection and information usage practices. Those policies are long and complex, and it is argued that the ordinary user may not thoroughly understand the context of the policy, nor the website's actual practices. It has been estimated that it would take the average user 201 hours to read the privacy policies encountered per year [6].

A key challenge in this space that the prior work tried to address the ambiguity and lack of a standard format in privacy policies. Especially, users would be overwhelmed by both the policies breadth (i.e., many policies to deal with every year) and depth (i.e., complexity of the individual policies). While several attempts are made to improve their readability [2], privacy policies still lack a standard format, as of the moment of writing this work. Motivated by that, recent studies [4, 5, 8, 9] leveraged natural language processing and deep learning techniques to effectively annotate the privacy policies contents.

To improve upon those studies, this work investigates several annotation techniques for a practical automation of policy annotation. The goal of our annotation is to provide users with high-level annotations on whether various privacy policies they encounter in their daily life meet certain requirements with respect to a broad set of privacy and security expectations. In particular, we built an ensemble of classifiers using six word representation techniques and learning algorithms for automating policy annotation.

Our pipeline, called TLDR, operates at the paragraph level, and is trained on nine categories highlighting different uses typically found in the privacy policies. The ensemble outputs a binary decision for each category, positive (*i.e.,* a paragraph contains information on the privacy policy category) or negative (*i.e.,* a paragraph does not contain such information).

Through experiments on a widely used dataset, we show that TLDR achieves high performance in categorizing privacy policy practices, with an average F_1 score of 91%, and can highlight important paragraphs within a privacy policy. Through a user study, we show that TLDR reduces the reading time by 39.14%. Moreover, by eliminating unnecessary information in the policy statements, TLDR improves their understandability by 18.84%. TLDR also is shown effective in highlighting critical information of the privacy policy, and its extracted statements are shown to be preferred over the original policies in 67% of the times, per our user study.

Contribution. In this work, we advance the state-of-the-art on privacy policy annotation by delivering the following contributions:

(1) We propose TLDR, a pipeline that employs various deep privacy policy representation techniques and an automated ensemble of privacy policy classifiers. TLDR achieves a state-of-the-art average F_1 score of 91%.
(2) Using TLDR, we analyze the privacy practices in Alexa top-10,000 websites, unveiling major issues in reporting user tracking and data security practices by those websites.
(3) We develop a paragraph highlighting mechanism to reduce the number of paragraphs that a user needs to read in order to uncover certain privacy practices in a privacy policy.
(4) We conduct a user study to understand the effectiveness of TLDR in highlighting important paragraphs within the privacy policy. In this user study, 67% participants expressed that they prefer to read the privacy policy with only the highlighted paragraphs over the original policies.

CCS '21, The ACM Conference on Computer and Communications Security (CCS)

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

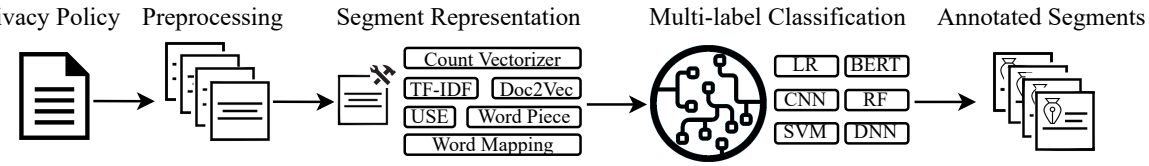

Figure 1: The pipeline of TLDR, with preprocessing, segmentation and representation, and segment annotation.

2 THE TLDR PIPELINE

The TLDR pipeline is shown in Figure 1. In the following, we describe the process of implementing each of TLDR's steps.

Ground Truth and Key Terminology. As a baseline for privacy policy annotation, we used the Online Privacy Policies (OPP-115) dataset [8]. The dataset consists of privacy policies collected from 115 websites, manually annotated to 9 categories, including paragraph-level associated privacy policies, such as *first-party data collection*, *third party information sharing*, and *user tracking* practices.

Privacy Policy Preprocessing. For each paragraph, the *stopwords*, common words that do not add meaning, are removed. Then, WordNet Lemmatizer is used for words *lemmatization* and *stemming*. This process removes the generic words and words/sub-words that do not contribute to the meaning or context of the paragraph, making the learning process more efficient and accurate.

Paragraph Representation. To find a suitable highly discriminative representation for each category, various text representation techniques are used in TLDR: word mapping, count vectorizer, TF-IDF, Doc2Vec, Universal Sentence Encoder (USE), and WordPiece. Preprocessing and feature representation are critical in TLDR implementation to unveil the hidden patterns within each paragraph without the need for human labor or manual annotation.

Learning Algorithms. TLDR trains an ensemble of learning algorithms for associating paragraphs with their corresponding privacy policy categories. Doing so reveals the abstract content of the privacy policy without the need for reading such content. We leverage six machine and deep learning algorithms for privacy policy detection, including, Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Convolutional Neural Networks (CNN), Deep Neural Network (DNN), and Bidirectional Encoder Representations from Transformers (BERT), evaluating their effectiveness in detecting various paragraph-level categories.

Acknowledging that privacy policy categories are unique, TLDR leverages the best performing data representation for category classification. The wide range of explored representations and learning algorithms is due to the diversity of categories, where treating them indiscriminately results in reduced performance.

3 EVALUATION AND DISCUSSION

3.1 Annotation Results of TLDR

Experimental Setup. We used *document-based splitting*, where 80% of the documents in OPP-115 dataset are used for training the ensemble while the remaining 20% of the documents are used for validation. For deep learning architectures, we adopt the architecture by Harkous *et al.* [4] for our CNN model, and replace the convolutional layers with fully connected layers to build the DNN model. We configure the BERT model with a 512 maximum number of words, with the number of features in the range of [1,000, 5,000] with an increment of 1,000.

Table 1: The F_1 score of the TLDR using best performing word representations and learning algorithms on OPP-115.

Category	TLDR	Wilson [8]	Harkous [4]	Liu [5]
First party	0.94	0.75	0.79	0.81
Third party	0.89	0.7	0.79	0.79
User choice	0.85	0.61	0.74	0.70
User access	0.91	0.61	0.80	0.82
Data retention	0.87	0.16	0.71	0.43
Data security	0.88	0.67	0.85	0.80
Policy change	0.95	0.75	0.88	0.85
Do not track	1.00	1.00	0.95	1.00
Specific audiences	0.94	0.70	0.95	0.85
Overall	0.91	0.66	0.83	0.78

Annotation Results. Table 1 shows the best performing architecture for each category used for analysis. Similarly, we report the best performing evaluation results of Wilson *et al.* [8], Harkous *et al.* [4], and Liu *et al.* [5], on the OPP-115 dataset using the F_1 score. As shown, except for "Specific Audiences" and "Do Not Track" categories, TLDR outperforms its counterparts by a large margin, particularly for "Data Retention".

3.2 Alexa Top-10,000 Websites Analysis

We used TLDR to analyze the Alexa [1] top-10,000 websites privacy policies practices. The Alexa top-10,000 represents the most visited websites by users worldwide. Analyzing such websites uncovers the common practices of popular websites and their service providers, targeting a large portion of the Internet users.

Privacy Policy Extraction. We start by obtaining the privacy policies of the websites among the Alexa top-10,000 websites list. Using Selenium [7], the privacy policies were crawled between Nov. 4 and Nov. of 2020. Then, the privacy policies are extracted by searching the webpages within a website for terms, e.g., *privacy policy*, and *privacy*. Among the top-10,000 websites, we successfully extracted the policies of 5,598 websites. The remaining websites are either (1) non-English or (2) do not directly link to their privacy policy on the website. **Validation.** To validate the correctness of the extraction process, we manually inspect 1,000 extracted privacy policies, and verified that 95.8% of them are correctly extracted. The process of website crawling and cleaning is shown in Figure 2.

Data Preprocessing & Representation. The extracted paragraphs (345,920) are then preprocessed and represented in a way similar to OPP-115 [8]. We removed the stop-words, then the words in the paragraph were lemmatized and stemmed. We limit the hyper-parameters configurations to the best performing within the feature representations and learning algorithms referenced in Table 1.

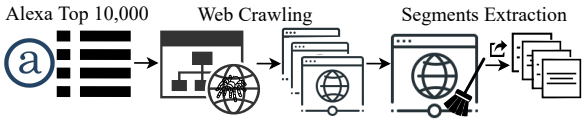

Figure 2: Alexa Top-10,000 data collection and paragraph extraction pipeline.

Evaluation & Discussion. We used TLDR to extract existing privacy practices within Alexa top-10,000 websites. Figure 3 shows the percentage of websites containing information regarding the policy categories for both OPP-115 and Alexa top-10,000 websites.

Overall, the "first-party use" and "third party sharing" categories are the most common within the privacy policies, with 95% of the websites containing first-party use information, and 90% of them including information regarding third-party sharing. On the other hand, the "do not track" category is the least common within the privacy policies, with only 20% of the websites reporting information associated with it. Given that the ensemble achieves an F_1 score of 100% on this category, the results are of high confidence.

Missing Information. By examining the ensemble results, we found that a large number of websites' privacy policies miss key information and attributes, by not covering important areas including data security and user tracking. This comes as a surprise, given that the extracted privacy policies are from the top-visited websites as of 2020, which are potentially the subject of great interest, and their policies are the subject of great scrutiny.

Future Direction: Per Topic Analysis. To better understand the privacy policy practices disclosure by the website provides, the next step is to conduct per topic analysis of the reported practices. This can be done using Webshrinker [3], a machine learning-powered domain data and threat classifier, to obtain the categorization of the domains of the websites. This step is essential to draw deep observations on the trends among the most popular websites.

3.3 User Study: Information Highlighting

We conduct a user study with a total of 20 participants to investigate the effectiveness of TLDR in information highlighting. Removing unnecessary policy information reduces the efforts required to understand the privacy practices by the service providers. However, omitting important privacy policy aspects can be critical by reducing the users awareness of the reported practices.

Each participant is assigned three policies of different lengths. For each policy, two instances were initiated, the original and the TLDR filtered policies (after removing all paragraphs with no associated privacy policy information). The participant then read the two instances of each privacy policy in a random order, and was not aware of the filtering process nor the objective of this study. The participant is expected to keep a record of the reading time of each policy, and answer a survey after reading each privacy policy.

Results. According to 82% of the participants, the privacy policies are understandable and suitable upon applying the TLDR filtering process. This is surprising, as only 69% of the participants indicated the original privacy policies are understandable. This may be a result of removing unnecessary (legal) information that does not necessarily contribute to the practices. The participants also reported a reduced reading time for TLDR filtered policies, *i.e.,* an average of 39.14% time reduction in comparison with the original policy. As noted, removing important information may result in missing critical data collection and sharing practices. According to the participants' responses, only 23% of the policies include critical information that is not reported in the TLDR filtered instance.

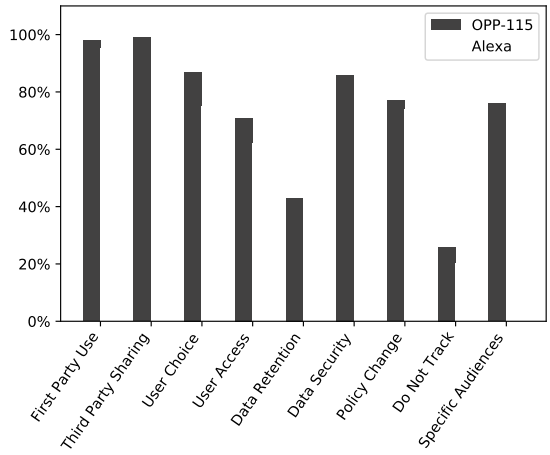

Figure 3: Percentage of websites with positive paragraphs per category for Alexa top-10,000 and OPP-115 websites.

4 CONCLUDING REMARKS

In this work, we revisit the automated privacy policy annotation problem by exploring improving the accuracy of annotation through various learning and representation techniques. In particular, we propose TLDR, a pipeline employing deep representation techniques and an ensemble of machine and deep learning-based models to automatically and accurately categorize each paragraph in the privacy policy to its corresponding high-level content category, achieving an average F_1 score of 91%. Our user study shows the effectiveness of TLDR in highlighting privacy policy practices to reduce the reading time and increase the understandability.

REFERENCES

[1] Amazon, "Alexa top websites," November 2020. [Online]. Available: https://www.alexa.com/topsites
[2] L. F. Cranor, *Web privacy with P3P - the platform for privacy preferences.* O'Reilly, 2002.
[3] Developers. (2020) Webshrinker. https://www.webshrinker.com/.
[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," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 531–548.
[5] F. Liu, S. Wilson, P. Story, S. Zimmeck, and N. Sadeh, "Towards automatic classification of privacy policy text," *School of Computer Science Carnegie Mellon University*, 2018.
[6] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *Isjlp*, vol. 4, p. 543, 2008.
[7] Selenium, "Seleniumhq browser automation," November 2020. [Online]. Available: https://www.selenium.dev/
[8] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell *et al.*, "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)*, Aug. 2016.
[9] R. N. Zaeem, R. L. German, and K. S. Barber, "Privacycheck: Automatic summarization of privacy policies using data mining," *ACM Trans. Internet Techn.*, vol. 18, no. 4, pp. 53:1–53:18, 2018.