Toward Domain-Guided Controllable Summarization of Privacy Policies

Moniba Keymanesh
keymanesh.1@osu.edu
The Ohio State University

Micha Elsner
elsner.14@osu.edu
The Ohio State University

Srinivasan Parthasarathy
parthasarathy.2@osu.edu
The Ohio State University

ABSTRACT

Companies' privacy policies are often skipped by the users as they are too long, verbose, and difficult to comprehend. Identifying the key privacy and security risk factors mentioned in these unilateral contracts and effectively incorporating them in a summary can assist users in making a more informed decision when asked to agree to the terms and conditions. However, existing summarization methods fail to integrate domain knowledge into their framework or rely on a large corpus of annotated training data. We propose a hybrid approach to identify sections of privacy policies with a high privacy risk factor. We incorporate these sections into summaries by selecting the riskiest content from different privacy topics. Our approach enables users to select the content to be summarized within a controllable length. Users can view a summary that captures different privacy factors or a summary that covers the riskiest content. Our approach outperforms the domain-agnostic baselines by up to 27% in ROUGE-1 score and 50% in METEOR score using plain English reference summaries while relying on significantly less training data in comparison to abstractive approaches.

ACM Reference Format:
Moniba Keymanesh, Micha Elsner, and Srinivasan Parthasarathy. 2020. Toward Domain-Guided Controllable Summarization of Privacy Policies. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop, 24 August 2020, San Diego, US.* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION AND RELATED WORK

Privacy policy and terms of service are unilateral contracts by which companies are required to inform users about their data collection, processing, and sharing practices. Users are required to agree to abide by the terms before they can use any service. However, many users do not read or understand these contracts [1]. Thus, they often end up consenting to terms that may not be aligned with legislation such as the General Data Protection Regulation (GDPR)[1] [2]. This behavior is often because these contracts are too long and difficult to comprehend [3]. Summarization is an intuitive way to assist users with conscious agreement by generating a condensed equivalent of the content. Broadly, there are two main lines of summarization systems: *abstractive* and *extractive*. The abstractive paradigm [4–10] aims to create an abstract representation of the input text and involves various text rewriting operations such as paraphrasing, deletion, and reordering. The extractive paradigm [11, 12] on the other hand, creates a summary by identifying and subsequently concatenating the most important sentences in the document. The abstractive systems are more flexible while the extractive models enjoy better factuality [13]. However, existing summarization techniques perform poorly on contracts. Unsupervised methods [14, 15] rely on structural features of documents, such as lexical repetition, to identify and extract important content. These heuristics work poorly on the legal language used in contracts [16]. Supervised methods [7, 9, 17] can learn to cope with the features of a particular domain. However, training these complex neural summarization models with thousand of parameters requires a large corpus of documents and their summaries. Currently existing corpora in the legal domain are not large enough to train such models. We propose a hybrid approach for extractive summarization of privacy contracts: using existing annotated resources, we train a classifier to predict which pieces of content are most relevant to users [1]. In particular, we identify parts of the contract which place users at risk by imposing unsafe data practices on them, such as selling email addresses to third parties or allowing the company to appropriate user-generated content. Next, we use this risk classifier for content selection within an extractive summarization pipeline. The classifier is substantially less expensive than learning to summarize directly but enables our approach to outperform a selection of domain-agnostic unsupervised summarization methods.

Prior computational work on privacy policies has used information extraction and natural language processing methods to classify segments of these documents into different data practice categories [18–20]. Another trajectory of work has focused on presenting a graphical "at-a-glance" description of the privacy policies to the user. For example, PrivacyGuide [21] and PrivacyCheck [22] define a few privacy factors and map each factor to a risk level using a data mining model. Relying on these "at-a-glance" description methods raises several concerns. First, there is no way for the user to check the factuality of the predicted risk classes or interpret the reasoning behind them. Moreover, users tend to have an easier time comprehending the content when provided in natural language. Researchers also have focused on assigning a risk factor–green, yellow, or red–to each segment of the privacy policies [23, 24]. However, summarizing the text may benefit users more than directly presenting the classifier output. We draw on these approaches in building our own classifier. The first module of our framework extends prior work [23, 24] to highlight segments of privacy policies that have a higher risk. We employ a pre-trained encoder and convolutional neural network to classify sentences of the contracts into different risk levels. To address the limitations of previous work, we incorporate the domain information predicted by the classifier in the form of a summary by comparing a risk-focused and a coverage-focused content selection mechanism. The coverage-focused selection mechanism aims to reduce

[1] https://eugdpr.org/

NLLP @ KDD 2020, August 24th, San Diego, US

the information redundancy by covering the riskiest sentence from each privacy topic. We evaluate the effectiveness of employing a classifier on identifying the domain knowledge for summarization. We also evaluate the quality of summaries extracted by our two content selection criteria. Using our approach users can view a summary that captures different privacy factors or a summary that covers the riskiest content. We release our dataset of 151 privacy policies annotated with risk labels to assist future research.

2 METHODOLOGY

Given a privacy policy document D consisting of a sequence of n sentences $\{s_1, s_2, ...s_n\}$ and a sentence budget m such that $m < n$ our summarization model extracts a risk-aware summary with m sentences. For each sentence $s_i \in D$ we predict a binary label y_i (where a value of 1 means s_i is included in the summary). We achieve this by computing an inclusion probability $p(y_i|s_i, D, \theta)$ for each sentence s_i. θ are the model's parameters. We aim to maximize the inclusion probability for risky sections of the privacy policies and minimize it for non-risky sections. We also would like to cover different privacy factors within the sentence budget m by reducing the redundancy. The main intuition behind our proposed approach is that users when going through the privacy policies are most interested in knowing how their information can potentially be abused [1]. Thus, a condensed equivalent of the terms should include such risky sections. Next, we explain the architecture or our risk prediction model and our content selection mechanisms.

2.1 Risk Prediction

Given the content of privacy policies, the first step in our framework is to identify the associated risk class with each sentence of the contract. We rely on a crowd-sourcing project called TOS;DR[2] to automatically annotate 151 privacy contracts. TOSDR has annotated several snippets of privacy contracts based on the average Internet user's perception of risk. We explain our dataset extraction in section 3. We use this dataset to train our risk classifier. Prior research has exploited word embeddings and Convolutional Neural Networks (CNN) for sentence classification [25–28]. These simple architectures achieve strong empirical performance over a range of text classification tasks. Our model is a slight variant of the CNN architecture proposed in [25].

2.1.1 *Model architecture*.

Let $s_j = \{t_1, t_2, ...t_n\}$ be the j-th sentence in the contract D and $v_i \in R^d$ be the d-dimentional vector representation of token t_i in this sequence. Word representations are output of a pretrained encoder [29] and will be discussed in Section 2.1.2. We build the sentence matrix $A \in R^{n \times d}$ by concatenating the word vectors v_1 to v_n:

$$A_{1:n} = v_1 \oplus v_2 \oplus ...v_n$$

Following [25] we apply convolution filters to this matrix to produce new features. The length of the filters is equal to the dimensionality of the word vectors d. The height or region size of the filter is denoted by h and is the number of rows (word vectors) that are considered jointly when applying the convolution filter. The feature map $c \in R^{n-h+1}$ of the convolution operation is then obtained by repeatedly applying the convolution filter w to a window of tokens $t_{i:i+h-1}$. Each element c_i in feature map $c = [c_1, c_2, ...c_{n-h+1}]$ is then obtained from:

$$c_i = f(w \ . \ A[i : i + h - 1] + b)$$

where $A[i : j]$ is the sub-matrix of A from row i to j corresponding to a window of tokens t_i to t_j and "." represents the dot product between the filter w and the sub-matrices. $b \in R$ represents the bias term and f is an activation function such as a rectified linear unit. We use multiple kinds of filters by using various region sizes. This extracts various types of features from bigrams, trigrams, and so on. The dimensionality of the feature map c generated by each convolution filter is different for sentences with various lengths and filters with different heights. We apply a max-over-time [25] pooling operation to downsample each feature map c by taking the maximum value over the window defined by a pool size p. The max-pooling operation naturally deals with variable sentence lengths. The outputs generated from each filter map are concatenated to build a fixed-length feature vector for the penultimate layer. This feature vector is then fed to a fully connected softmax layer that predicts a probability distribution over the risk level categories. We apply dropout [30] as a means of regularization in the softmax layer. Our objective is to minimize the binary cross-entropy. The trainable model parameters include the weight vectors w of the filters, the bias term b in the activation function, and the weight vector of the softmax function. We minimize the loss using *Stochastic gradient descent* and back-propagation [31].

2.1.2 *Pretrained Word Vectors*.

Prior research indicates that better word representations can improve performance in a variety of natural language understanding (NLU) tasks [32]. We use ELMo [29]-a deep contextualized word representation model-to map each token t_i in sentence s_i in contract D to its corresponding contextual embedding v_i with length 1024 [3]. ELMo uses a bidirectional LSTM [34] for language modeling and considers the context of the words when assigning them to their embeddings[4].

2.2 Content Selection and Redundancy Reduction

Given the probability distributions over the risk categories, we apply two content selection mechanisms to account for the summarization budget m and minimize the information redundancy. The first mechanism focuses on including the most "risky" sections while the second mechanism focuses on covering diverse privacy factors. Next, we explain these two variations of our model.

2.2.1 *Risk-Focused Content Selection:*

Given a privacy policy contract D with sentences $\{s_1, ...s_n\}$, a summarization budget m, and risk score $p(y_i = 1|s_i, D, \theta)$ predicted for s_i by the risk classifier, the risk-focused selection mechanism assembles a summary by extracting the top m sentences that have the highest risk score.

[2] https://TOS;DR.org

[3] Model was trained on the One billion word benchmark [33] and was obtained from https://github.com/allenai/allennlp

[4] BERT [35] as the current state-of-the-art for language model pretraining has achieved amazing results in many NLU tasks with minimal fine-tuning. However, our preliminary results of fine-tuning bert did not outperform our results from Elmo word vectors and task-specific architecture explained in Section 2.1.1.

2.2.2 *Coverage-Focused Content Selection:*

Given a privacy policy contract D with sentences $\{s_1, ...s_n\}$, a summarization budget m, and risk scores $p(y_i = 1|s_i, D, \theta)$, the coverage-focused selection method finds m privacy factors by clustering sentences for which the risk score is larger than a predefined value of α. Next, the riskiest sentence from each privacy factor cluster is selected to be included in the summary. Note that if less than m sentences have a risk score greater than α the summary will have less than m sentences. To find privacy topics of a contract, we apply k-means [36] to sentence representations. Sentence representations are obtained through concatenating the word vectors. Number of clusters is set to $min(m, |r|)$ where $r = \{s_i \mid p(y_i = 1) > \alpha\}$.

3 DATASET EXTRACTION

In this section, we explain the dataset that we compiled from the TOS;DR website and privacy contracts of 151 companies. TOS;DR is a website dedicated to rating and explaining privacy policy of companies in plain English. Members of the website's community classify specific sections of privacy policies into "bad", "good", "blocker", and "neutral" categories and provide summaries for them. We collected the user agreement contracts of 151 services that were annotated on TOS;DR from the companies' websites. Some companies have several such contracts e.g. privacy policy, terms of service, and cookie policy. In this case, all the contracts were merged into a single document. Next, we compared each sentence of the contract with specific snippets that were annotated on TOS;DR. If the corresponding sentence or a very similar sentence was annotated by the TOS;DR contributors, the same label was used. Otherwise, it was annotated as "neutral". The assumption behind our annotation schema is that, if a section was not annotated by the contributors, it most likely does not include a privacy risk and thus, is considered neutral. NLTK was used to segment the contracts into sentences. Jaccard similarity of the vocabulary was used to measure the similarity of the sentences. Two sentences from the same contract were considered similar if the Jaccard similarity of their tokens was more than 50%. We combined the "bad" and "blocker" sections to build the "risky" class. The "good" and "neutral" classes were also combined to build the "non-risky" class. This dataset is highly imbalanced with 61674 non-risky sentences and only 719 risky sentences. To build the ground truth risk-aware summary of each privacy policy we concatenate the plain English summaries of the snippets that have a "risky" label. The dataset statistics of the 151 privacy policies and their corresponding summaries are presented in Table 1. Our dataset is available online [5].

Dataset	Min	Max	Median	Mean
Privacy Policies	61	1707	350	411.6
Plain English Summaries	1	53	1	3.5

Table 1: The min, max, median, and average number of sentences in 151 privacy contracts and their summaries.

4 EXPERIMENTS

In this section, we discuss our data augmentation mechanism to reduce the data imbalance problem, our hyper parameter choice for designing the risk classifier, and the training details. We discuss our evaluation criteria in Section 4.2.

4.1 Hyperparameters and Training Details

For the CNN model, we use two filter region sizes 3 and 4 each of which has 50 output filters. We use rectified linear unit as the activation function of the convolution layer. The pool size in the max pooling operation is set to 50. We apply dropout with a rate of 20%. We optimize the binary cross-entropy loss using stochastic gradient descent with a learning rate of 0.01. To account for the class imbalance problem, we randomly under-sampled the majority class (non-risky) with a rate of 10%. We also apply *SMOTE* over sampling [37] on the minority class (risky) with rate 50%. We train our model on this resampled dataset for 20 epochs and weight the loss function inversely proportional to class frequencies in the input data. To set the value of risk threshold α in the content selection module, we used the ROC curve of the validation set of each fold. We set α for each fold to the threshold value that achieves 80% true positive rate.

4.2 Evaluation Metrics

In our experiments, we seek to answer two questions: i. how well does our model identify the risky sentences in the contracts? and ii. what content selection method leads to more "human-like" summaries? To answer the first question we report the Macro-F1 and Micro-F1 score of our classifier. To answer the second question, we evaluate the quality of the extracted summaries by our model by computing the average F1-score for ROUGE-1, ROUGE-2, and ROUGE-L [38] metrics (which respectively measure the unigram-overlap, bigram-overlap, and longest common sequence between the reference summary and the summary to be evaluated). ROUGE metrics fail to capture semantic similarity beyond n-grams [39]. Thus, we also report the METEOR score [40] which goes beyond the surface matches and accounts for stems and synonyms while finding the matches.[6] We evaluate our model using 5-fold cross-validation. In each fold, contracts of 96 companies are used for training, 24 contracts are used for validation, and the rest is used for testing. We explain our baselines in Section 4.3 and our experimental results in Section 5.

4.3 Summarization Baselines

We compare the performance of our domain-aware extractive summarization model with the following unsupervised baselines. Unlike the evaluation setup in [16], we run the models on the entire contract. For methods that require a word limit as the budget, a compression ratio r is multiplied by the average number of tokens in all contracts (10488.7) to compute the word limit. Similarly, the compression ratio of r is multiplied by the average number of sentences in all contracts (413.1) to build a sentence limit.

- **TextRank**: An algorithm introduced in [14] that uses page rank to compute an importance score for each sentence. Sentences with the highest importance score are then extracted to build a summary until a word limit is satisfied.

[5] www.github.com/senjed/Summarization-of-Privacy-Policies

[6] We use pyrouge and NLTK python packages for computing ROUGE and METEOR values respectively.

	Compression Ratio = 1/64				Compression Ratio = 1/16			
	P	R	Macro-F1	Micro-F1	P	R	Macro-F1	Micro-F1
CNN + RF	22.40	28.13	61.94	98.01	9.86	59.74	56.65	93.10
CNN + CF	19.64	24.06	60.26	97.95	12.19	52.65	58.51	94.94

Table 2: Precision(P), Recall(R), Macro-F1, and Micro-F1 of the CNN classifier with two different content selection mechanisms risk-focused(RF) and coverage-focused(CF) at two different compression ratios $\frac{1}{16}$ and $\frac{1}{64}$.

- **KLSum**: Introduced in [15], KLSum aims to minimize the Kullback-Lieber (KL) divergence between the input document and proposed summary by greedily selecting sentences.
- **Lead-K**: A common baseline in news summarization that extracts the first k sentences of the document until a word limit is reached.
- **Random**: This baseline picks random sentences of the document until a word limit is satisfied. For this baseline, we report the average results over 10 runs.
- **Upper Bound Baseline**: This baseline picks all the sentences in a contract with ground truth label "risky". This baseline indicates the performance upper bound of an extractive method on our dataset.

5 RESULTS

In this section, we discuss our experiments conducted using 5-fold cross-validation. We shared our training details in Section 4.1. As an example, summaries extracted by our model and the baselines from privacy policy of Brainly [7] is displayed in Figure 1. It can be seen that both of the summaries generated by our method indicate that third party advertising companies will be able to collect information about use of Brainly. KLSum misses this information and the traditional lead-k heuristic which is very effective for news performs poorly on the contracts. This indicates the advantage of injecting domain-specific knowledge into content selection.

5.1 Classification Results:

In this section, we evaluate the performance of our model discussed in Section 2.1.1 and study the effect of different content selection mechanism on the risk prediction task. We evaluate our summaries at two compression ratios of $\frac{1}{64}$ and $\frac{1}{16}$. The summarization budget m at each compression ratio r is achieved by multiplying r in the average number of sentences(or words) in the contracts. Thus, at the compression ratio of $\frac{1}{64}$, summaries are restricted to the maximum length of 6 sentences or 164 words. Similarly, at the compression ratio of $\frac{1}{16}$, summaries are limited to the maximum length of 29 sentences or 656 words. We report the precision, recall, Micro-F1, and Macro-F1 of our risk classifier with two different content selection mechanisms namely risk-focused (RF) and coverage-focused (CF) in Table 2. As can be seen in the table, the Micro-F1 scores of both content selection methods are quite high. However, the best Macro-F1 value is achieved by the risk-focused approach and is 61.94. The large gap between the two values is due to the high level of class imbalance in our dataset (1 positive sample for every 100 negative samples). At $\frac{1}{64}$ compression ratio, risk-focused performs more than two times better in terms of recall. When the compression ratio is $\frac{1}{16}$, the risk-focused method captures many more risky sections and achieves a recall of 59.74. However, with this increase in recall, the false positive rate also increases. On the other hand, the coverage-focused method is better at preserving the precision at higher budgets (only 7.45 drop in precision with a 28.59 points increase in recall). This observation is caused by extracting sentences with a risk score greater than α in coverage-focused content selection. This naturally puts an upper bound on the false positive rate. We conclude that both mechanisms are moderately successful at identifying the risky sections of contracts. We also conclude that at higher compression ratios, the risk-focused mechanism can be used where recall is more essential while the coverage-focused mechanism can be used when precision is more of interest. In the next section, we examine whether the domain information given by the risk classifier can improve the quality of summaries in comparison to domain-agnostic extractive summarization baselines.

5.2 Summarization Results:

In this section, we evaluate the quality of the summaries extracted by our model and the baselines. We introduced our evaluation metrics in Section 4.2 and our baselines in Section 4.3. We compare the summaries against two type of reference summaries. The first type of summary is built by assembling all the sentences that have ground truth "risky" label. These sentences are derived directly from text of the contract. We will refer to this reference summary as "quote text" reference. The second type of summary is derived by assembling the plain English summary of the "risky" sections written by the TOS;DR contributors. The summarization results using the quote text summaries is presented in Table 3. The summarization results using the plain English reference summaries is presented in Table 4.

5.2.1 ***Extracting the risky content:*** As it can be seen in Table 3, at both compression ratios, both variation of our model outperform the baselines. At compression ratio of $\frac{1}{64}$, the CNN + RF, achieves the best ROUGE and METEOR results with 49.8% improvement in ROUGE-1, 124.6% improvement in ROUGE-2, 56.3% improvement in ROUGE-L, and 65.6% improvement in METEOR in comparison to the best performing domain-agnostic baseline for each metric. At compression ratio of $\frac{1}{16}$ the CNN + CF achieves the best ROUGE results by improving ROUGE-1 by 12.2%, ROUGE-2 by 30.2%, ROUGE-L by 8.8%, and METEOR by 23.7% in comparison the the best performing baseline for each metric. The improvement in METEOR score is found to be statistically significant using Wilcoxon signed ranked test [41] with p-value < 0.01 (Bonferroni corrected [42] to account for multiple testing). Similar to our observation in classification task, we find that the risk-focused content

[7] https://Brainly.com

Plain English Summary: The Privacy Policy states, "We and our third party partners may also use cookies and tracking technologies for advertising purposes.". In the Privacy Policy, it states that, "Although we do our best to honor the privacy preferences of our users, we are unable to respond to Do Not Track signals set by your browser at this time." The Privacy Poilicy says Brainly can track usage information and personal information "through a variety of tracking technologies, including cookies, web beacons, Locally Stored Objects (LSOs such as Flash or HTML5), log files, and similar technology (collectively, “tracking technologies”)." If Brainly aims to "preserve all content posted on the site," then we can conclude that such personal data is still necessary for the purpose of the site. There are places on the site where answers without usernames or profile pictures are visible. The Cookie Policy states, "Service oparator [sic] informs that restricting the use of cookies may affect some of the functionalities available on the Website." For users not in europe, brainly reserves the right, in its sole discretion, to immediately modify, suspend or terminate your account, the brainly services, your brainly subscription, and/or any products, services, functionality, information, content or other material. *<truncated>*

CNN + RF: We participate in interest-based advertising and use third party advertising companies to serve you targeted advertisements based on your online browsing history and your interests. We permit third party online advertising networks, social media companies and other third party services, to collect, information about your use of our service over time so that they may play or display ads on our service, on other websites, apps or services you may use, and on other devices you may use. We may share a common account identifier (such as an email address or user id) or hashed data with our third party advertising partners to help identify you across devices. Brainly reserves the right to moderate the Brainly services and to remove, screen, or edit your content from the Brainly services at our sole discretion, at any time, and for any reason or for no reason, with no notice to you. Brainly reserves the right, in its sole discretion, to immediately modify, suspend or terminate your account, the Brainly services, your Brainly subscription, and/or any products, services, functionality, information, content or other materials available on, through or in connection with the Brainly services and/or your Brainly subscription, including, but not limited to, the mobile software, and/or your access to some or all of them without cause and without notice. In the event that Brainly suspends or terminates your account, the Brainly services or your Brainly subscription, you acknowledge and agree that you shall receive no refund or exchange for any unused time on a Brainly subscription or any subscription fees or anything else.

CNN + CF: We participate in interest-based advertising and use third party advertising companies to serve you targeted advertisements based on your online browsing history and your interests. We permit third party online advertising networks, social media companies and other third party services, to collect, information about your use of our service over time so that they may play or display ads on our service, on other websites, apps or services you may use, and on other devices you may use. We may share a common account identifier (such as an email address or user id) or hashed data with our third party advertising partners to help identify you across devices. To the fullest extent permitted by applicable law, no arbitration or claim under these terms shall be joined to any other arbitration or claim, including any arbitration or claim involving any other current or former user of the Brainly services or a Brainly subscription, and no class arbitration proceedings shall be permitted. We may modify or update this privacy policy from time to time to reflect the changes in our business and practices, and so you should review this page periodically. If you object to any changes, you may close your account. Continuing to use our service after we publish changes to this privacy policy means that you are consenting to the changes.

Lead-K: Welcome to Brainly!. Brainly operates a group of social learning networks for students and educators. Brainly inspires students to share and explore knowledge in a collaborative community and engage in peer-to-peer educational assistance, which is made available on www.Brainly.com and any www.Brainly.com sub-domains(the “website”) as well as the Brainly.com mobile application (the “app”) (the “website” and the “app” are collectively the “Brainly services”. We have two sets of terms and conditions: part(a) sets out the terms that apply to our users unless you are based in Europe and part (b) sets out the terms that apply to our users in Europe. It is important that you read and understand the terms that apply to you when you use the Brainly services before using the Brainly services. Part (a): terms and conditions applicable to users unless you are based in Europe. This part and the documents referred to within it set out the terms and conditions that apply to your use of Brainly services if you access Brainly services from within the united states or other countries except Europe. The Cookie Policy states, "Service oparator [sic] informs that restricting the use of cookies may affect some of the functionalities available on the Website."

KLSum: Brainly reserves the right, in its sole discretion, to immediately modify, suspend or terminate your account, the Brainly services, your Brainly subscription, and/or any products, services, functionality, information, content or other materials available on, through or in connection with the Brainly services and/or your Brainly subscription, including, but not limited to, the mobile software, and/or your access to some or all of them without cause and without notice. Brainly makes no warranty that the Brainly services and/or any products, services, functionality, information, content or other materials available on, through or in connection with the Brainly services or your Brainly subscription, including, but not limited to, the mobile software, will meet your requirements, or that the Brainly services or Brainly subscriptions will operate uninterrupted or in a timely, secure, or error-free manner, or as to the accuracy or completeness of any information or content accessible from or provided in connection with the Brainly services or Brainly subscriptions, regardless of whether any information or content is marked as “verified”. You must not: use Brainly services other than for its intended purpose as set out in the terms of use; *<truncated for presentation purpose. Rest of the summary includes examples of misuse of the Brainly services.>*

Figure 1: The summaries extracted by our model (CNN + RF and CNN + CF) and the baselines from the privacy policy and cookie policy of Brainly at compression ratio of $\frac{1}{64}$.

selection achieves more recall and thus, achieves a better METEOR score in comparison to the coverage-focused mechanism. On the other hand, by increasing the summarization budget, the ROUGE values for this method slightly drop. This is because, in most of the contracts, the number of risky sentences is smaller than the budget at ratio of $\frac{1}{16}$ (29 sentences).

5.2.2 *Building Human-like summaries:*

We present our summarization results using the plain English summaries as reference summaries in Table 4. At compression ratio of $\frac{1}{64}$, both variations of

	Compression Ratio = 1/64				Compression Ratio = 1/16			
	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
CNN + RF	**43.09**	**31.21**	**36.80**	**41.98**	34.0	24.96	24.83	40.03
CNN + CF	40.45	28.69	34.01	41.55	**37.93**	**28.82**	**29.23**	**43.91**
Textrank	28	13.89	22.06	22.4	33.78	22.12	26.85	35.49
KLSum	28.75	13.14	23.53	25.34	24.74	11.36	18.86	26.95
Lead-k	25.57	9.09	20.25	19.54	25.67	11.33	19.77	26.85
Random	24.26	6.45	18.78	18.11	24.43	9.85	18.08	27.01

Table 3: ROUGE-1, ROUGE-2, ROUGE-l, and METEOR score of our model (highlighted in light gray) in comparison to the baselines in compression ratios $\frac{1}{64}$ and $\frac{1}{16}$. RF refers to the risk-focused content selection while CF refers to the coverage-focused content selection. The quote text of the risky sections was used to build the reference summaries.

	Compression Ratio = 1/64				Compression Ratio = 1/16			
	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Upper Bound	22.45	13.7	18.27	22.32	22.56	13.95	18.49	23.03
CNN + RF	**13.97**	**6.08**	**9.83**	**16.58**	9.07	3.94	5.53	12.07
CNN + CF	12.39	4.81	8.51	14.93	**10.18**	**4.54**	**6.58**	**13.16**
Textrank	10.94	2.78	7.51	11.2	10.08	3.37	6.37	12.47
KLSum	10.96	2.43	7.34	12.54	8.37	1.92	5.26	11.06
Lead-k	11.21	1.9	7.9	11.04	9.33	2.44	5.96	11.87
Random	11.44	1.87	8.03	12.02	9.13	2.32	5.73	12.45

Table 4: Performance of our model (highlighted in light gray) in comparison to the baselines in compression ratios $\frac{1}{64}$ and $\frac{1}{16}$. RF refers to the risk-focused content selection while CF refers to the coverage-focused content selection. The plain English summaries of risky sections was used to build the reference summaries.

our model outperform the baselines. Our CNN + RF model, increases the METEOR score by 32.2% over KLSum and 48% over textrank. This improvement is found to be statistically significant (with p-value < 0.01). The CNN + CF outperforms the baselines over all evaluation metrics. However, the improvement is not statistically significant. At compression ratio of $\frac{1}{16}$, CNN + RF outperforms all domain-agnostic baselines. This improvement however, is not statistically significant. At this compression ratio, CNN + RF achieves comparable result with textrank. We conclude from our experiments that our domain-aware extractive model does moderately better than the baselines at lower compression ratios, however, due to high level of abstraction in plain English summaries of TOS;DR [16], a fully-extractive approach cannot mimic the human-like qualities in the plain English summaries. This can also be seen by looking at the performance of the upper bound baseline.

6 CONCLUSION AND DISCUSSION

In this paper, we proposed a domain-aware extractive model for summarizing the privacy contracts. Our model, employs a convolutional neural network to identify risky sections of the contracts. We build summaries by using a risk-focused and a coverage-focused content selection mechanism. Our approach enables users to select the content to be summarized within a controllable length while relying on substantially less training data in comparison to the existing supervised summarization methods. Our two different content selection mechanisms enable users to build budgeted summaries of contracts based on their preference of coverage vs risk. In spite of the moderate success in classification of our realistically imbalanced dataset, we observed a noticeable improvement in ROUGE and METEOR metrics in comparison to domain agnostic baselines. We believe the summaries generated by our method can be improved in multiple ways. First, the classifier itself, and the redundancy reduction system, could be improved, bringing content selection performance closer to the upper bound scores derived using a perfect classifier. Secondly, our summaries would be more accessible if written in plain English rather than legalese [2]. An abstractive system could be used to rewrite the contract text in this way. However, the abstractive summaries should not change the legal interpretation of the content and should be linkable to the original content to be considered binding. In addition to improving the system, it is also necessary to conduct more extensive evaluation experiments, involving human readers as well as automated metrics. This will help determine the most effective ways to present information from click-through contracts so that users can understand their terms and make a more informed decision. We are planning to explore if the risk classifier module can be used independently to enhance the productivity of annotators by identifying the sections that need to be summarised. This can potentially facilitate annotating larger resources for training abstractive models.

ACKNOWLEDGEMENT

We are immensely grateful to Prof. Junyi Jessy Li, Prof. Bryan H. Choi, Dr. Daniel Preoţiuc-Pietro, Mayank Kulkarni, and three anonymous reviewers for valuable discussions.

REFERENCES

[1] Lorrie Faith Cranor, Praveen Guduru, and Manjula Arjula. User interfaces for privacy agents. *TOCHI*, 2006.

[2] Jonathan A Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *ICS*, 2020.

[3] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Isjlp*, 2008.

[4] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv:1509.00685*, 2015.

[5] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv:1602.06023*, 2016.

[6] Qian Chen, Xiao-Dan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. Distraction-based neural networks for modeling document. In *IJCAI*, 2016.

[7] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv:1704.04368*, 2017.

[8] Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. Abstractive document summarization with a graph-based attentional neural model. In *ACL*, 2017.

[9] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

[10] Ritesh Sarkhel*, Moniba Keymanesh*, Arnab Nandi, and Srinivasan Parthasarathy. Transfer learning for abstractive summarization at controllable budgets. *arXiv:2002.07845*, 2020.

[11] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, 2017.

[12] Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. Graph-based neural multi-document summarization. *arXiv preprint arXiv:1706.06681*, 2017.

[13] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. In *AAAI*, 2018.

[14] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *EMNLP*, 2004.

[15] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *NAACL*, 2009.

[16] Laura Manor and Junyi Jessy Li. Plain english summarization of contracts. *arXiv:1906.00424*, 2019.

[17] Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*, 2018.

[18] Frederick Liu, Shomir Wilson, Peter Story, et al. Towards automatic classification of privacy policy text. 2018.

[19] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, et al. The creation and analysis of a website privacy policy corpus. In *ACL*, 2016.

[20] Sebastian Zimmeck and Steven M Bellovin. Privee: An architecture for automatically analyzing web privacy policies. 2014.

[21] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. Privacyguide: Towards an implementation of the eu gdpr on internet privacy policy evaluation. In *IWSPA*, 2018.

[22] Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. Privacycheck: Automatic summarization of privacy policies using data mining. *TOIT)*, 2018.

[23] Najmeh Mousavi Nejad, Damien Graux, and Diego Collarana. Towards measuring risk factors in privacy policies. In *ICAIL*, 2019.

[24] Hamza Harkous, Kassem Fawaz, Rémi Lebret, et al. Polisis: Automated analysis and presentation of privacy policies using deep learning. 2018.

[25] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *JMLR*, 12(Aug):2493–2537, 2011.

[26] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv:1408.5882*, 2014.

[27] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[28] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv:1510.03820*, 2015.

[29] Matthew E Peters, Mark Neumann, Mohit Iyyer, et al. Deep contextualized word representations. *arXiv:1802.05365*, 2018.

[30] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[31] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 1986.

[32] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

[33] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

[34] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45, 1997.

[35] Jacob Devlin, Ming-Wei Chang, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[36] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE TPAMI*, 2002.

[37] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *JAIR*, 2002.

[38] Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *ACL*, 2002.

[39] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. Recent advances in document summarization. *Knowledge and Information Systems*, 2017.

[40] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, 2014.

[41] Frank Wilcoxon, SK Katti, and Roberta A Wilcox. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1970.

[42] Charles W Dunnett. New tables for multiple comparisons with a control. *Biometrics*, 1964.