Is Your Policy Compliant? A Deep Learning-based Empirical Study of Privacy Policies' Compliance with GDPR

Tamjid Al Rahat
University of California, Los Angeles
tamjid@g.ucla.edu

Minjun Long
University of Virginia
ml6vq@virginia.edu

Yuan Tian
University of California, Los Angeles
yuant@ucla.edu

ABSTRACT

Since the *General Data Protection Regulation* (GDPR) came into force in May 2018, companies have worked on their data practices to comply with the requirements of GDPR. In particular, since the privacy policy is the essential communication channel for users to understand and control their privacy when using companies' services, many companies updated their privacy policies after GDPR was enforced. However, most privacy policies are verbose, full of jargon, and vaguely describe companies' data practices and users' rights. In addition, our study shows that more than 32% of end users find it difficult to understand the privacy policies explaining GDPR requirements. Therefore, it is challenging for the end users and law enforcement authorities to manually check if companies' privacy policies comply with the requirements enforced by GDPR. In this paper, we create a privacy policy dataset of 1,080 websites annotated by experts with 18 GDPR requirements and develop a Convolutional Neural Network (CNN) based model that can classify the privacy policies into GDPR requirements with an accuracy of 89.2%. We apply our model to automatically measure GDPR compliance in the privacy policies of 9,761 most visited websites. Our results show that, even after four years since GDPR went into effect, 68% of websites still fail to comply with at least one requirement of GDPR.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; **Privacy protections**; • **Computing methodologies → Active learning settings**.

KEYWORDS

Privacy Policy, GDPR, Compliance Check, Active Learning, Deep Learning, CNN

ACM Reference Format:
Tamjid Al Rahat, Minjun Long, and Yuan Tian. 2022. Is Your Policy Compliant? A Deep Learning-based Empirical Study of Privacy Policies' Compliance with GDPR. In *Proceedings of the 21st Workshop on Privacy in the Electronic Society (WPES '22), November 7, 2022, Los Angeles, CA, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3559613.3563195

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WPES '22, November 7, 2022, Los Angeles, CA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9873-2/22/11.
https://doi.org/10.1145/3559613.3563195

1 INTRODUCTION

Privacy regulations are introduced to protect the personal data of individuals that are collected by public or private companies, governments, and other individuals. Among many privacy regulations, GDPR (*General Data Protection Regulation*) has been considered as one of the strictest privacy regulations [41, 49]. The primary purpose of GDPR is to give more control to individuals over their personal data and to ensure their rights regarding those data. GDPR applies to any individual or company that collects, stores, or processes any personal information in connection to services or goods offered in European Union (EU) countries. It significantly affects millions of websites, applications, and online services.

Since GDPR came into effect in May 2018, companies have been focusing on reforming their data practices, including changing the privacy policies [4], in view of the fact that non-compliance with GDPR could be fined up to 4% of the company's total annual revenue or 20 million Euros, whichever is higher (article 83 [3]). However, even with such stiff fines and penalties, it appears that many companies are not yet fully compliant with GDPR. For example, the Information Commissioner's Office (ICO) of the United Kingdom fined British Airways 183 million Euros [6] for failing to protect users' personal data and violating GDPR. The Irish Data Protection Commission (DPC) has opened 19 investigations [10] into big companies' potential privacy breaches (e.g., Google [8], Facebook [7], and Twitter [9]) since May, 2019.

One of the most important GDPR requirements is that companies must disclose to users their personal data handling process and certain rights of the individuals, such as disclosing the information about what personal data they collect and how they collect, store, process, and share it. However, previous studies show that internet users find most privacy policies inscrutable and read privacy policies in rare cases [14, 27]. Privacy policies written in unstructured natural languages mostly suffer from insufficient readability [16, 21, 31, 32]. According to the most recent literacy survey conducted by the National Center for Education Statistics [22], over half of the Americans may find it hard to understand large and complex texts. Thus, a vast majority of these complicated privacy policies can not be comprehended by users. However, these studies mostly represent the general privacy policies and do not focus on specific privacy laws. Consequently, we conducted a user study (more details in Section 5) that shows approximately 32% of users find it difficult to comprehend the policies describing GDPR requirements. Therefore, it is critical to automatically analyze the privacy policies' compliance with privacy laws.

Additionally, since the privacy policy is the *de facto* standard mechanism for communicating companies' data practices, studying the privacy policies provides useful insights to understand the

GDPR compliance. Existing works are focused on partial GDPR compliance for certain data practices such as cookie usage [19], erasure of personal data [40], and opt-out choices [24]. Some works [30, 34] adapt privacy policy corpus [48] created before GDPR and study the data practices in privacy policies that are also addressed in GDPR. The major limitation of these works is that they are focused on a limited number of fine-grained data practices in privacy policies and do not address the key requirements of GDPR such as users' rights regarding the collection and processing of personal data. More recently, Torre et al. [45] introduced a classification model to check privacy policies against GDPR. However, their model is trained only on 234 privacy policies collected from the fund management companies, which as our experimental study shows (Section 5), does not perform well for the privacy policies other than financial domains.

In this paper, we explore the research question: *Do privacy policies of the companies fully comply with GDPR requirements to communicate their data practices and users' rights?* To answer the question, we build a Convolutional Neural Network (CNN) based classifier to determine whether privacy policies are fully compliant with GDPR. One of the major challenges in training such a classifier is to get an annotated dataset and currently, no such dataset is publicly available. Therefore, we create a new dataset of 1,080 privacy policies annotated with GDPR requirements. To build the annotated dataset, we first create a corpus of 9,761 privacy policies from the most visited websites in Europe. Further, to understand what is required to be disclosed in the privacy policies by GDPR, we identify 18 requirements (Table 2) from the GDPR legislation [1] with the help of two legal experts, who also train four human annotators to annotate privacy policies. These 18 categories of information must be disclosed in the privacy policy when companies (i.e., data controllers) collect personal data from human subjects or other sources.

Another challenge is that privacy policy segments of different GDPR requirements may contain overlapping features that can lead to misclassification of the categories. To overcome this challenge, we use the Convolutional Neural Network (CNN), which can detect the set of features that contribute most towards a class regardless of their position in the input. Our initial trained model achieves an accuracy of 80.5%. To further improve accuracy, we require more labeled data. However, collecting labeled data is expensive and time-consuming, whereas unlabeled data are free. Therefore, we leverage the pool-based *active learning* technique to improve our model's accuracy using less labeled data. Our final classification model achieves an accuracy of 89.2%, which is a 10.8% relative increase compared to the initial performance.

We apply our classification model to determine the compliance of 9,761 top websites' privacy policies with the 18 GDPR requirements. Since these websites are selected based on visit frequency, they span various domains, such as search engines, social networks, streaming service, etc. We find that there are only 32% of websites fully comply with GDPR in their privacy policies. We also identify six major requirement categories of GDPR that are implemented by only 37% of websites. For example, GDPR requires companies to disclose any information regarding profiling or any other automated decision-making process by using users' personal data. Surprisingly, only 28% of websites disclose information about this in their privacy policies. In summary, we make the following contributions:

- We build a new dataset containing 1,080 privacy policies annotated by expert annotators, using labels representing 18 privacy disclosure requirements of GDPR. We make our annotated dataset publicly available[1].
- We develop a CNN-based privacy policy classifier to classify privacy policy segments into 18 GDPR requirements. Our model achieves an accuracy of 89.2% with an average F1-score of 0.88.
- We further utilize our model to investigate the GDPR compliance scenario of 9,761 most visited websites, and our findings show that 68% of websites do not fully comply with GDPR.

The rest of the paper is organized as follows. In Section 2, we describe GDPR terminologies and 18 GDPR privacy policy disclosure requirements. We present an overview of our classification model and user study design in Section 3. In Section 4, we detail our methodology for building privacy policy corpus, training neural-network-based classification model, and user study design. We present our experiments and findings in Section 5. Potential limitations and promising future directions of our study are discussed in Section 6. Finally, we present related works in Section 7 and conclude the paper in Section 8.

2 BACKGROUND

In this section, we provide definitions of terminologies used in GDPR requirements and describe the key requirements that are compulsory for privacy policies.

2.1 GDPR terminologies

Data Subject. GDPR (Art. 4) defines a data subject as any person whose personal data is collected, stored, or processed by data controllers. Personal data can refer to anything from a person's name, address, or social media information. Thereby, anyone can become a data subject when they book a flight, apply for a job, or use a credit card for a purchase. For the convenience of the readers, we use *'user'* to refer to the data subject throughout the paper.

Personal data. GDPR (Art. 4) defines personal data as any information that relates to an identified or identifiable individual (data subject). Information such as name, address, ethnicity, gender, and web cookies can be considered as personal data.

Data Controller. GDPR (Art. 4) defines a data controller as any organization, person, or body that controls personal data and determines the purposes and means of processing data, and is responsible for it, alone or jointly. In short, the data controller controls the procedures and purpose of data usage. For the convenience of the readers, we use *company* to refer data controller throughout the paper.

Data Processor. GDPR (Art. 4) defines a data processor as any third party that processes personal data on behalf of a data controller. Throughout the paper, we use *third party* to refer data processor.

[1] https://github.com/tamjidrahat/gdpr-dataset

2.2 GDPR: Privacy Disclosure Requirements

We have gleaned through the official GDPR legislation [1], with the help of two legal experts, to find out the information that must be provided in the privacy policy. We find that when personal data are collected from a human subject or other sources, data controllers have to provide 18 categories of information in the privacy policy. More details about these requirements can be found in GDPR regulation (Chapter 3, Art. 12–23) [3]. In the following, we describe the 18 categories of GDPR requirements with examples of privacy policy segments that companies provide to comply with the requirements.

(1) Data Categories: No matter whether data is collected from users or any other sources, privacy policies should provide information about the categories of personal data collected, stored, or processed by any organization or third parties. For example, *Apple* discloses certain categories of personal information they collect as the following - "We may collect a variety of information, including your name, mailing address, phone number, email address, contact preferences, device identifiers, IP address, location information, credit card information and profile information where the contact is via social media."

(2) Processing Purpose: Privacy policies should include the purposes of processing user's personal data as well as a lawful basis to process those data. Before any organization begin to process users' data, they must determine their lawful basis, which in most cases requires that the data processing is necessary for specific purposes. For example, *Apple.com* describes one of the purposes for collecting users' personal information as the following - "We also use personal information to help us create, develop, operate, deliver, and improve our products, services, content and advertising, and for loss prevention and anti-fraud purposes."

(3) Data Recipients: An organization should disclose the information about the recipients to whom users' personal data have been or will be shared, including the recipients from third countries or international organizations. For example, *Msn.com* provides the information regarding recipients of users' personal data as the following - "We may share your personal information with third parties, such as advertisers, sponsors, and other promotional and business partners."

(4) Source of Data: Privacy policies should mention information about which source personal data originate from if they are not obtained directly from users. For example, *TheGuardian.com* discloses the source of some personal information as the following - "We may obtain information about you from partners so that we can make our online advertising more relevant."

(5) Provision Requirement: Privacy policies should include whether providing personal data is required, or if the user is obligated to provide the personal data and the consequences for not doing so. For example, *Gettyimages.com* discloses the existence of this right in their privacy policy as the following - *"You may always choose not to provide personal data, but if you so choose, certain products and services may not be available to you."*

(6) Data Safeguards: If the controller intends to transfer data to a third country or an international organization in the absence of an adequacy decision, reference to appropriate safeguards of personal data should be provided in the privacy policy. Additionally, privacy policy also needs to provide the means to obtain a copy of the available safeguards to protect users' personal data. For example, *The Guardian* discloses this requirement in their privacy policy as the following - "Whenever we transfer your personal data out of the European Economic Area (EEA), we ensure similar protection and put in place at least one of these safeguards ..."

(7) Profiling: GDPR has provisions for profiling or other automated decision-making. The automated decision-making is defined as making a decision solely by automated means without any human involvement, and profiling, which can be a part of automated-decision making, is defined as the automated processing of personal data to evaluate certain things about an individual. According to (Art. 22) of GDPR, data controllers can carry out such decision-making if they have a lawful basis and explicit consent of users for the relevant processing of personal data. The privacy policy should include the existence of any automated-decision making, including profiling and what information is used for such decision-making. For example, *Dropbox* discloses information regarding profiling as the following - "Dropbox collects and processes your personal information using automated decision-making (including machine learning) to provide, improve, and market the dropbox services in furtherance of its legitimate interests or based on your consent when appropriate."

(8) Storage Period: Privacy policies should disclose the period for which personal data will be stored, or if not possible, the criteria used to determine that period. GDPR mandates that personal data shall not be kept longer than is necessary for the purposes for which it is processed. The period of personal data stored by a data controller should be limited to a strict minimum and a time limit for the deletion of the data should be determined and disclosed in the privacy policy by the data controller. For instance, *Academia.edu* discloses the personal data storage period in the following manner - "academia.edu retains the personal information we receive as described in this privacy policy for as long as you use the academia.edu service or as necessary to fulfill the purpose(s) for which it was collected."

(9) Adequacy Decision: When controllers intend to transfer personal data to a third country or international organization, the existence or absence of an adequacy decision by the European Commission must be disclosed in the privacy policy. Adequacy decision made by the EU Commission is based on whether a country outside of the EU offers an adequate level of data protection. For example, Oracle discloses this requirement as the following - *"If personal information is transferred to an Oracle recipient in a country that does not provide an adequate level of protection for personal information, Oracle will take measures designed to adequately protect information about you, such as ensuring that such transfers are subject to the terms of the EU model clauses."*

(10) Controller's Contact: Data controllers shall provide their identity and contact details along with the contact details of any controller's representative, if applicable. For example, *The Guardian* news provides the data controller's contact information as the following - "The data controller for our sites and apps is Guardian News & Media Limited, Kings Place, 90 York Way, London N1 9GU.4".

(11) DPO Contact: Every organization that collects and processes the personal information of EU citizens has to appoint a Data Protection Officer (DPO) and the contact information of the DPO has to be provided in the privacy policy. For example, *Microsoft* provides the information about their Data Protection Officer as the following - "If you have a privacy concern, complaint, or question for the Microsoft Chief Privacy Officer or EU Data Protection Officer, please contact us by using our web form."

(12) Withdraw Consent: Privacy policies should include the existence of users' right to withdraw consent at any time. For example, *Nextdoor.com* discloses the existence of this right as the following - *"If you wish, you can access the content of the (electronic) consent as well as revoke the consent with effect for the future at any time."*

(13) Lodge Complaint: Privacy policies should include the existence of users' right to lodge a complaint with a supervisory authority. For example, *Gofundme.com* discloses the existence of this right in their privacy policy as following - *"You also have the right to lodge a complaint with the local data protection authority if you believe that we have not complied with applicable data protection laws."*

(14) Right to Access: Data controllers should disclose the existence of users' right to request the controller to access or rectify the personal data. According to GDPR (Art. 15), users have the right to obtain information about whether their personal data is being collected, used, or stored; the categories of the data; with whom the data have been or will be disclosed; whether the data has been or will be transferred to a third country or organization; and how long the data will be stored or processed. For example, legacy.com discloses that *"you may access the information we hold about you anytime via your profile/account."*

(15) Right to Erase: Data controllers must disclose the existence of users' right to erase personal data. As mentioned in GDPR (Art. 17), an organization shall have an obligation to erase personal data when data are no longer necessary for the purposes they were collected or stored. In addition to that, if users withdraw their consent at any time, corresponding personal data must be erased. For example, *Ranker.com* describes this right as following - "If you no longer want us to use your information during the provision of the Services to you, you can request that we erase your personal information and close your Account."

(16) Right to Restrict: Privacy policy should include the existence of users' right to restrict or suppress the processing of their personal data. For example, *Politico.com* discloses this right as the following - "You may also have rights to restrict our processing of your personal data"

(17) Right to Object: Privacy policies should include the existence of users' right to object to the processing of their personal data. For example, users have the absolute right to stop their data from being used in direct marketing. For example, *Ranker.com* discloses the existence of users' right to object as the following - "If you object to such processing we will no longer process your personal information for these purposes."

(18) Right to Data Portability: Privacy policies should include the existence of users' right to receive personal data in a structured, commonly used, and machine-readable format and the right to transfer those data to another data controller. For example, *Oclc.org* discloses this right as the following - "You may receive personal data that you have provided to us in a structured, commonly used, and machine-readable format and have the right to transmit it to other data controllers."

3 OVERVIEW

Figure 1 gives a high-level overview of our approach. We first create a privacy policy dataset with the GDPR requirements and then train a CNN-based classification model using our dataset. We further improve the performance using active learning. Additionally, we conduct a comprehensive user study to measure the effectiveness of privacy policies from users' perspectives.

3.1 Classification Model

To train our supervised model for classifying GDPR requirements, we create a labeled privacy policy dataset containing 9,510 policy segments from 1,080 privacy policies. To build the dataset, we collect plain privacy policies from 9,761 most visited websites by using UK's IP addresses as the UK is one of the top English-speaking countries in the EU. From 9,761 plain privacy policies, we randomly select 1,080 policies to be annotated by trained annotators. Upon consolidating the annotation from four annotators, we finally annotate 9,510 privacy policy segments with 18 GDPR requirements, which we use to train our CNN-based classification model.

We train a CNN-based model to classify the segments in a privacy policy into 18 classes, each of which represents a disclosure requirement of GDPR. Before training the model, we represent the input texts with sparse low-dimensional vectors (embeddings), which gives the CNN classification model more generalization power. To make the embeddings specific to privacy policies, we initially train an unsupervised word embedding model using FastText [39] with our entire privacy corpus that contains 9,761 privacy policies. Then, for training the classification model, we split our labeled dataset containing 1,080 privacy policy documents into 864 (80%) as training data and 216 (20%) as test data. Our trained CNN-based model initially achieves an accuracy of 80.5% with an average F1-score of 0.79. We find the primary reason for misclassifications is the overlapping features between classes. For example, *"You have the right to object, in relation to specific processing of your personal data"-Selectminds.com* represents an instance from *Right to Object* class, whereas *"You have the right to request that we delete your personal data"-Spotify.com* represents an instance from *Right to Erase* class. Both privacy policy segments describe users' rights and personal data, although they represent two different classes. To improve the accuracy, the model needs to learn to extract the distinguishable features between the classes that may have overlapping features. One way to resolve this problem is to increase the amount of training data. However, since obtaining more training data is expensive and time-consuming, we use the *active learning* technique to further improve the performance of the model while using less amount of training data.

The primary hypothesis in active learning is that if a learning algorithm, instead of learning from a large pool of randomly sampled data, can choose the data it can learn most effectively from, it can perform better than traditional learning methods with substantially less amount of training data. We use an iterative pool-based active learning method since obtaining unlabeled privacy policy

data is free in our case. In particular, in each iteration, we randomly select a set of unlabeled privacy policy documents from the corpus and feed the unlabeled privacy policy segments into the trained classification model. Then, we use margin sampling to select the instances (queries) that the model predicted with the least confidence. We manually label those ambiguous instances by following the annotation approach similar to what we did to build the original dataset. Finally, we re-train the classification model with the newly labeled training data, which helps the model discriminate between the classes with overlapping features more effectively. We repeat this iterative active learning approach until the overall performance of the model no longer improves. Overall, we achieve an accuracy of 89.2% with an average F1-score of 0.88, while adding only 10.5% additional training data.

With the compliance classification model, we run a measurement study for the compliance of the most visited 9,761 websites' privacy policies.

4 METHODOLOGY

In this section, we first describe our privacy policy corpus and our labeled dataset. Then, we explain the details of our classification model and the active learning process.

4.1 Privacy Policy Extraction

To build a classifier to help us perform a compliance measurement study, we use a web crawler to extract a privacy policies corpus and label a part of the corpus following a systematic approach with the help of two legal experts and four trained annotators. We collect the privacy policies from the top 10,000 websites listed in Quantcast Top UK websites [5]. Since we are not located in EU countries, we use Europe-based VPNs to capture the privacy policies of the EU. Moreover, it is worth mentioning that some companies (e.g., Facebook, Inc.) provide additional information in the privacy policy for the EU countries to comply with GDPR. To collect the privacy policies, we first use Yandex Search API [13] to search for the URL of the privacy policy for each website. During the search, we append the keywords "privacy policy" with the domain name of the website. From the top 5 search results, we select the URL containing keywords such as privacy, and policy. During this search operation, we limit our search results to the English language only. After fetching the URLs, we scrape the HTML pages to extract the privacy policies in plain text, while removing unimportant information such as images and navigation links. However, we exclude URLs that are broken or do not link to the privacy policy page of the corresponding website. For example, our filtered search results for Wikipedia.com leads to an article [12] on *Privacy Policy* from Wikipedia instead of Wikipedia's own privacy policy. We manually filter out such cases from the corpus. Finally, we successfully extract 9,761 privacy policies.

4.2 Annotation Process

With the help of two legal experts, we first identify 18 categories of information (shown in Table 2) that are required to be disclosed in the privacy policy to fully comply with GDPR. These categories include information about how companies should handle users' personal data and what are users' rights regarding their personal data.

Privacy Policy Documents	1,080
Total Words	2,83,374
Total Classes	18
Annotated Segments	9,510
Additional Segments for Active Learning	1,001
Total Annotated Segments	10,511
Annotators Per Document	4
Total Annotators	6

Table 1: Statistics on the privacy policy dataset annotated with GDPR requirement categories.

Note that GDPR requires the companies to present all the 18 requirements even if the companies don't have certain data practices (e.g., companies should still clearly indicate they don't share user data rather than not saying anything). To create a privacy policy dataset labeled with GDPR requirements, we randomly select 1,080 privacy policies from our privacy policy corpus. Our legal experts train six human annotators to annotate the privacy policies with GDPR requirements. Each privacy policy document is annotated by four annotators. Each annotator works independently during the annotation process. We design a detailed *annotation schema* to advise our annotators on how to label the instances of data correctly and update the schema based on the feedback from the annotators after each annotation phase. We develop a privacy policy annotation tool in Java to help our expert annotators label privacy policies with 18 GDPR disclosure requirements. The tool allows annotators to load a privacy policy and read the policy segment by segment. We remove all HTML tags and non-ASCII characters to improve readability. At each step, annotators read a segment from the privacy policy and mark it with any of the 18 categories if the segment represents any GDPR requirement. Privacy policy segments that do not represent any particular GDPR requirement are marked as *other* category, which we exclude from the dataset we use to train our model. We finally receive 11,271 annotated privacy policy segments, each of which is annotated by four annotators. Table 2 shows category-wise statistics for the labeled dataset, such as the number of labeled privacy policy segments, and the mean and median number of words across the segments in each category.

To consolidate the labels, we set a 0.75 agreement threshold. For each privacy policy segment, if at least three of the four skilled annotators agree to a label, we consider it as the true label for the corresponding privacy policy segment. Out of 11,271 labeled segments, 8,826 (78.3%) belong to this threshold. For the privacy policy segments having an agreement threshold of 0.5 (13.4% segments), annotators discuss together to find if an agreement over the true label can be reached. We accept 684 segments (6%) for which annotators can reach an agreement and reject 829 segments (7.3%) for which they cannot. We discard the segments that have less than a 0.5 agreement threshold: no skilled annotators can reach an agreement on the true labels. As a result, we accept the true labels of in total of 9,510 privacy policy segments from 1,080 privacy policies. Table 1 shows the descriptive statistics for the labeled dataset after consolidation.

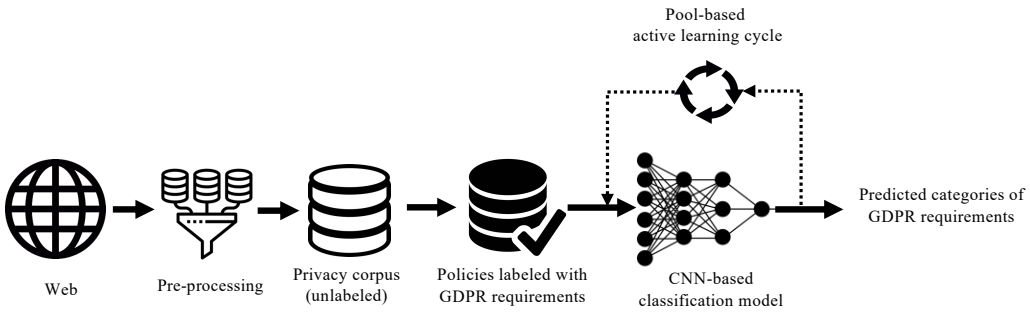

Figure 1: Overview of the CNN-based classification model and user perceived effectiveness measurement for the privacy policies relevant to GDPR requirements.

Requirement Categories	GDPR ref.	Freq.	Words	
			mean	med.
1. Data Categories	14(1.d)	1443	24	20
2. Processing Purpose	13(3)	1926	26	22
3. Data Recipients	13(1.e)	875	26	22
4. Source of Data	14(2.f)	595	26	20
5. Provision Requirement	14(5.b)	542	27	25
6. Data Safeguards	14(1.f)	331	24	23
7. Profiling	14(2.g)	381	27	16
8. Storage Period	13(2.a)	486	29	27
9. Adequacy Decision	13(1.f)	202	41	37
10. Controller's Contact	13(1.a)	308	16	14
11. DPO Contact	13(1.b)	530	26	25
12. Withdraw Consent	13(2.c)	510	27	25
13. Lodge Complaint	13(2.d)	345	31	27
14. Right to Access	14(2.c)	388	17	15
15. Right to Erase	14(2.c)	197	19	13
16. Right to Restrict	14(2.c)	507	17	12
17. Right to Object	14(2.c)	847	33	27
18. Right to Portability	14(2.c)	559	29	28

Table 2: 18 categories of GDPR privacy policy requirements and their corresponding statistics in our dataset. Mean and median were calculated across the population of words in the segments for each categories.

4.3 Privacy Policy Classifier

Our classification technique consists of two main parts: (1) an unsupervised training to build *Word Embedding* vectors for privacy policies and (2) a CNN-based supervised training as a classifier for GDPR disclosure requirements. We use Word Embedding and CNN-based classification model due to their great success in recent works of text classification [28, 36, 50]. Figure 2 illustrates the major components of our model.

4.3.1 Privacy policy specific word embeddings. Classical text classification models represent texts in a dataset using metrics such as words and their frequencies. However, such text representation techniques (e.g., Bag of Words) do not consider the position of the words in a sentence and are not capable of capturing the context and semantic meaning of a sentence. For example, *personal information* and *private data* are often used in the same context in privacy policies. Bag of Words considers them as two different phrases, which fails classification tasks in domains such as privacy policy where semantic context is critical. In addition to this, Bag of Words encodes every word in the dataset as a one-hot-encoded vector with the size of the entire vocabulary, which may result in the curse of dimensionality. To solve this problem, we choose Word Embeddings to represent the words in our dataset. Word embedding is a sparse low-dimensional vector representation of a text, which is learned in an unsupervised manner. Words that appear in the same context in the corpus are represented by similar vectors. This gives the neural network model more generalization power since it allows the model to recognize words that are not in the training set, as long as they are in the large corpus used for training the word embedding model.

To capture the privacy-specific semantic context, we train a custom word embedding using our privacy policy corpus containing 9,761 privacy policies. For training word embedding for this corpus, we use *fastText* [39] skip-gram model. To capture the semantic context, the skip-gram model predicts nearby words given a source word. However, instead of learning vectors for words directly like well-known word embedding model Word2Vec [33] or GloVe [35], *fastText* represents each word as an n-gram of characters in addition to the word itself. For example, fastText's representation for the word *privacy*, for n=3, is $< pr, pri, riv, iva, vac, acy, cy >$, where the angular brackets represent the beginning and end of the word. This representation allows fastText to assign vectors to the words that were not even seen during the training, which is an advantage for the domain-specific text classification task.

4.3.2 Model architecture. We design a CNN-based multi-class classifier to predict the probability of the classes, given an input sentence taken from privacy policies. As shown in Figure 2, an embedding layer is followed by a one-dimensional CNN layer in our classifier model. The CNN layer allows us to use the pre-trained word embeddings, which provides the capabilities to capture semantic meaning from the input sentence. In addition, a CNN layer

with filter size k can recognize the features (i.e., set of words) that represent a certain class, regardless of the position of the features in input texts. This layer applies a rectified linear unit (ReLU) as the activation function. Since we use a specific pre-trained privacy policy word embedding, we do not train the embedding layer in our model so that the embedding weights do not get updated while the classifier is learning. To prevent our model from over-fitting, we apply a dropout layer with a 0.1 probability for regularization. A max-pooling layer is applied to extract the most important features from the input. Output vector from the CNN layer is applied to a fully connected layer, which is again followed by a dropout layer with a probability of 0.5. This fully connected layer also applies ReLU as the activation function and is followed by another fully connected layer containing the number of units the same as the total number of classes. Finally, we use *softmax* function to determine the probability for each class. In our problem scenario, collecting

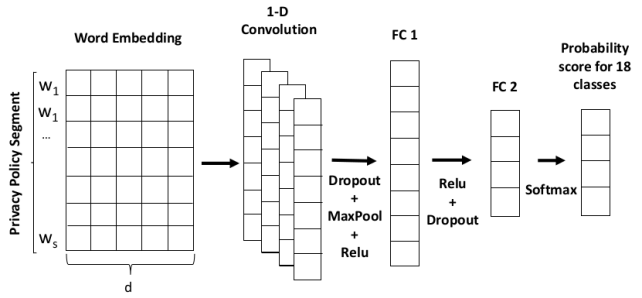

Figure 2: Architecture of our Convolutional Neural Network (CNN)-based classification model.

labeled data is time-consuming and expensive, whereas collecting unlabeled data (privacy policies) is free. Thereby, to improve the performance of our trained classification model, we leverage pool-based active learning to achieve higher accuracy with less amount of training data.

4.4 Pool-based Active Learning

Active learning has been successfully used to improve the accuracy of the machine learning tasks [20, 42, 43, 51] where unlabeled data can be easily obtained and labeled data is difficult to collect. The intuition behind using active learning is that the learning model can perform better even with substantially less amount of training data if the learning model can choose the data it needs instead of learning from a large set of randomly sampled data (passive learning). Since privacy segments generally disclose information about users' privacy and rights, it is common for segments from different categories to have overlapping features, which results in misclassification by the model. To overcome this, we use active learning to let our model learn from the segments that the model initially predicted with the least confidence between the two classes. To select such segments (queries), we first sample instances from our large pool of unlabeled privacy corpus. We classify the sampled instances using our trained classifier. Finally, based on the results from the classifier, we use *margin sampling* to decide whether to label an instance or not.

Figure 3 illustrates the key components of our active learning framework. Since a large pool of unlabeled segments can be obtained at once, we use pool-based selection sampling to sample unlabeled segments from our corpus. We sub-sample a small set of unlabeled privacy policies and pass the segments as input to our classification model. Based on the output probabilities from the model, our active learning framework decides whether to query a segment for labeling or discard it. To this end, we discard any instance that receives a maximum probability score of 0.5 or less because segments with a 0.5 or less probability score are less likely to be relevant to any of the GDPR disclosure requirements. For other instances with more than 0.5 probability scores, we use margin sampling to decide whether to label it or not. Margin sampling considers the difference in prediction score between the first and second most probable classes, based on the probability score of the classification model. We select a query instance that has a minimum difference, which means the model is less confident between the two classes.

$$x_m = argmin(P(\hat{y_1}|x) - P(\hat{y_2}|x))$$

Here, $\hat{y_1}$ and $\hat{y_2}$ represent the probabilities of the first and second most probable classes for segment x predicted by the model. Thus, x_m represents the instances that the classification model predicted with the least confidence. Intuitively, instances with a large margin of prediction scores are easy since the model has little doubt in differentiating between the two most likely classes. On the other hand, instances with small margins are ambiguous. Therefore, knowing the true labels would help the model discriminate between those two classes more efficiently.

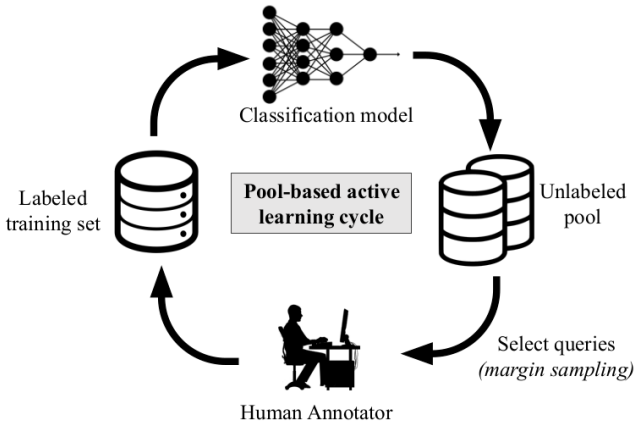

Figure 3: Active learning framework to improve classification accuracy.

5 EXPERIMENTS AND RESULTS

In this section, we present details of our model development and performance evaluation. We then present the measurement results of using our model to analyze how well the companies are following the GDPR requirements. To determine the effectiveness of our privacy policy classifier, we evaluate it on the privacy policy of the most visited websites to answer the following research questions:

- **RQ1**: Can our privacy policy classifier detect GDPR categories effectively?
- **RQ2:** Can we use the policy classifier to assess the current GDPR compliance scenario?
- **RQ3:** Is it necessary for users to have an automated privacy policy classifier to check compliance?

5.1 Model Development

5.1.1 Pre-processing. We remove all punctuation marks, special characters, digits, and white spaces from the privacy policy documents. We also remove all stop words since these words are not useful for the learning model and may cause additional memory overhead.

5.1.2 Vector representation. We train a fastText word embedding model to encode each word in our privacy policy corpus into a 300-dimensional vector. During the training, we use a minimum length of n-gram as three and a maximum length of six, which means the fastText model uses all the substrings contained in a word between three to six characters. This allows the embedding model to encode any new word that is unseen during the training. Overall, we train five epochs with a learning rate of 0.05.

5.1.3 Initial training and testing. We use 864 privacy policies (80% of the dataset) to train our model. We divide each privacy policy into segments and encode the segments to build a CNN layer, in which we use 400 filters with a kernel size of four. We use zero-padding with stride one for the 1-D convolution operation in this layer. The CNN layer uses the ReLU activation function and is followed by a dropout layer with a rate of 0.1 and a max-pooling layer. Output vectors from the CNN layer are fed into a fully connected (FC) layer with 256 units. This layer is followed by another FC layer with 18 units. We apply the softmax function on the output vector of this layer to determine the probability score for each of the 18 classes in our dataset. After training 50 epochs with a learning rate of 0.001, we test our initial trained model on 216 (20% of the dataset) privacy policies and achieve an accuracy of 80.5% with an F1-score of 0.79.

5.1.4 Performance improvement. We manually investigate some misclassified cases and find that the primary reason for a privacy segment being misclassified is because of the existence of overlapping features between the segments. We further use pool-based active learning to make our model more effective in discriminating between the segments from two classes having overlapping features. In each iteration of the active learning framework, we randomly sample 100 privacy policies from our large pool of unlabeled privacy policies. Segments from these privacy policies are then fed into our trained classification model. Based on the predicted probability score of the model, we use margin sampling to decide whether to label a segment or not. In each iteration, we select 250 segments that the model predicted with the least confidence, particularly with the minimum margin between the predicted score of two classes. However, we discard any segments with a probability score of 0.5 or less, since such segments are more likely to be irrelevant to GDPR requirements. For annotating the new segments, we follow the similar annotation process and consolidation strategy as discussed in Section 4. After annotating the new segments, we feed them as training data to re-train the classification model. Overall, we conduct seven iterations until when feeding new training data no longer improve the average performance of our model. Figure 4 illustrates the improvement of performance (average F1-score) during each iteration.

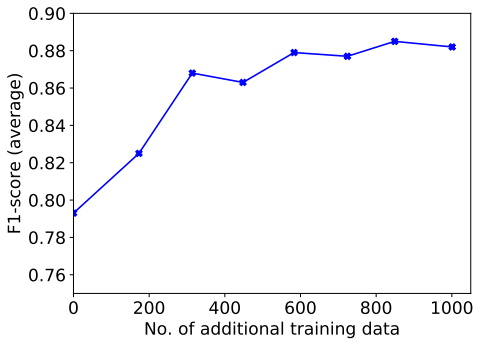

Figure 4: Improvement of model's performance during active learning iterations. In total, we conduct seven iterations until when feeding new training data no longer improves the average performance of our model.

5.2 Performance Evaluation

To answer the research question **RQ1**, we evaluate our final model using 216 privacy policies (20% of the labeled dataset) as the test set, which are never seen by the classification model during any phase of the training. We achieve an overall accuracy of 89.2% with an average F1-score of 0.88, while adding in total of 1,001 additional training data, which is only 10.5% of the initial training data. Table 3 presents the precision, recall, and F1-score (macro-average per label) of the evaluation on the test set. As evident in the table, our CNN-based classifier can predict the GDPR requirements from a given privacy policy segment with high accuracy. On average, we achieve 90% precision, 86% recall, and 0.88 F1-score. These metrics are higher than the other automated privacy policy analysis tool presented in [26, 44, 48]. The primary reason for the misclassification in our final model is that the privacy segments might be too vague or might cover multiple requirements with a very brief description.

5.3 Compliance Analysis in Large Scale

To answer the research question **RQ2**, we apply our classifier to assess the compliance scenario with GDPR requirements of the most visited 9,761 websites. In particular, to detect how well websites comply with GDPR privacy disclosure requirements, our model predicts the GDPR disclosure requirements for each of the privacy policies in the entire corpus of 9,761 policies. Figure 5 illustrates the number of websites meeting the GDPR privacy disclosure requirements for each of the classes. We find that many companies still do not follow the requirements introduced by GDPR. For example, among the 9,761 websites, only 28.5% follow the GDPR requirement of *profiling*, which requires that companies have to disclose how users' personal information is used for automated decision making or profiling purposes. Also, requirements such as users' *right to*

Classes	Prec.	Recall	F1	Support
1. Data Categories	0.87	0.75	0.81	289
2. Processing Purpose	0.86	0.80	0.83	373
3. Data Recipients	0.77	0.88	0.82	191
4. Source of Data	0.87	0.77	0.82	119
5. Provision Requirement	0.99	0.92	0.95	112
6. Data Safeguards	0.84	0.83	0.83	70
7. Profiling	1.00	0.88	0.94	77
8. Storage Period	1.00	0.99	0.99	83
9. Adequacy Decision	0.89	1.00	0.94	32
10. Controller's Contact	0.84	0.80	0.82	66
11. DPO Contact	0.97	0.98	0.98	114
12. Withdraw Consent	0.95	0.96	0.96	85
13. Lodge Complaint	0.94	0.94	0.94	78
14. Right to Access	0.82	0.77	0.80	63
15. Right to Erase	0.90	0.76	0.82	37
16. Right to Restrict	0.90	0.78	0.83	116
17. Right to Object	0.95	0.83	0.89	167
18. Right to Portability	0.86	0.92	0.89	116
Average	*0.90*	*0.86*	*0.88*	

Table 3: Classification results for 18 privacy disclosure requirements of GDPR. With using active learning technique, we achieved 89.2% classification accuracy with an average F1-score of 0.88.

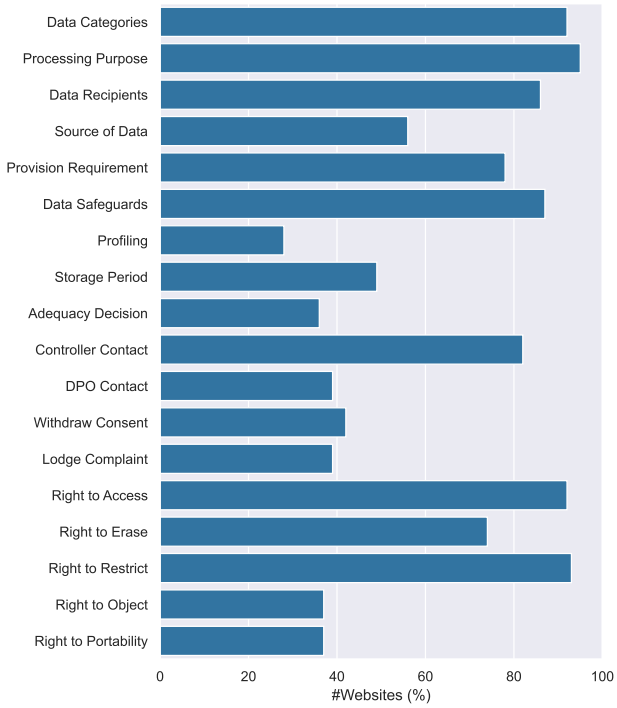

Figure 5: Visual representation of the overall compliance scenario for the most visited websites. The measurement result is produced by our CNN-based privacy policy classifier across 9,761 privacy policies for 18 GDPR requirements.

portability and *right to object* are covered by 36.2% and 37.4% websites, respectively. Other primary requirements, such as users' *right to withdraw consent* and disclosing the *storage period* are covered by 40-45% websites. Figure 6 shows the number of companies complying with 0-18 GDPR requirements. It appears that only 32% of websites fully comply with 18 requirements. These findings indicate that many websites still do not follow the requirements of GDPR.

5.4 Comparison with existing model

We compare our model with Torre et al.'s [45] work that proposes an SVM-based sentence classification model to check the completeness of privacy policies against GDPR. Their model has trained over 234 privacy policies collected from financial (i.e., fund management) websites. Since the model and the dataset are not publicly available, we collect a new labeled dataset of 100 privacy policies from the fund management and other financial websites. To label the dataset, we follow the similar annotation process described in Section 4. Similar to Torre et al., we also use the SVM model from [47] and use its default hyper-parameters for sentence classification. We also train the SVM classifiers with positive examples representing the sentences that have been labeled with a certain class in our dataset and negative examples labeled with other classes. Table 4 shows the comparative performance between the SVM-based and our CNN-based classification model. The average precision score obtained from the SVM classification model is 0.64, which is 29% less than our CNN-based classification model. On the other hand, the recall value from the SVM model is 22% less than the CNN model, showing the efficacy of our model in detecting GDPR requirements in privacy policies.

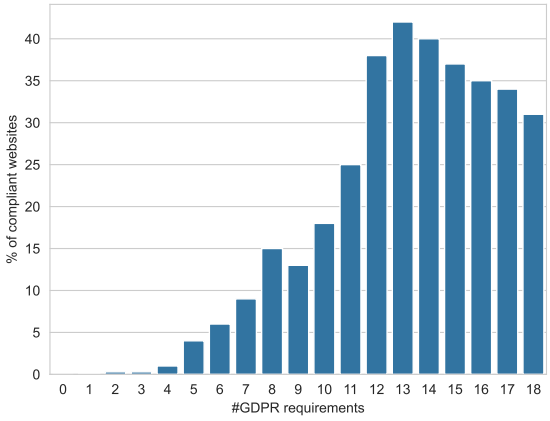

Figure 6: Number of websites' policies complying with 0-18 GDPR requirements. Here, only 32% websites fully comply with all of the 18 requirements.

5.5 User-perceived Privacy Utilization

To answer the research question **RQ3**, we conduct a user study to find how users comprehend privacy policies' compliance with

Category	Financial dataset		Baseline dataset	
	precision	recall	precision	recall
1. Data Categories	0.61	0.58	0.87	0.75
2. Processing Purpose	0.67	0.70	0.86	0.80
3. Data Recipients	0.70	0.62	0.77	0.88
4. Source of Data	0.53	0.70	0.87	0.77
5. Provision Requirement	0.60	0.82	0.99	0.92
6. Data Safeguards	0.53	0.58	0.84	0.83
7. Profiling	0.66	0.71	1.00	0.88
8. Storage Period	0.77	0.85	1.00	0.99
9. Adequacy Decision	0.55	0.69	0.89	1.00
10. Controller's Contact	0.78	0.80	0.84	0.80
11. DPO Contact	0.81	0.86	0.97	0.98
12. Withdraw Consent	0.67	0.58	0.95	0.96
13. Lodge Complaint	0.55	0.62	0.94	0.94
14. Right to Access	0.63	0.71	0.82	0.77
15. Right to Erase	0.77	0.76	0.90	0.76
16. Right to Restrict	0.61	0.73	0.90	0.78
17. Right to Object	0.54	0.44	0.95	0.83
18. Right to Portability	0.65	0.58	0.86	0.92

Table 4: Performance comparison between two models: 1) SVM-based model [45] trained with 100 financial privacy policies, 2) CNN-based model trained with 864 privacy policies from our baseline dataset. The precision and recall values are computed for 216 privacy policies that were unseen by both models.

GDPR without any help from automated tools. Since privacy policy works as a medium between users and organizations for communicating users' privacy and rights, it is critical for privacy policies to be understandable and perceivable to users. Precisely, privacy policies should clearly convey the information that is required to be disclosed to users. In order to get insights into how users perceive the privacy policy disclosures described in companies' websites or applications, our user study is designed to assess the effectiveness of current privacy policies available online. Towards that goal, in our study, we first explain (in plain English) the GDPR requirements to the participants and ask them if a given privacy policy segment is compliant with the requirement. Participants can indicate the degree of compliance and non-compliance or can indicate if the policy is too vague to understand. Since the original description extracted from the GDPR legislation can be too complex to understand for users, with the help of experts, we rephrase the requirements to make them comprehensible for regular users. However, if participants do not comprehend any particular requirement, we skip showing the corresponding privacy policy and do not record the responses for those requirements.

5.5.1 Participants. We recruit 102 participants on Amazon's Mechanical Turk (MTurk), requiring them to have at least a 95% HIT approval rating with at least 100 tasks completed on MTurk. Our participants must be at least 18 years of age, fluent in English, and users of at least one online service where they provided their personal information. Additionally, we require our participants to be located in the U.S. or EU countries. To avoid bias, we target random users and do not require them to know about GDPR before the study. At the beginning of the study, we explain to the participants what GDPR is about, why it is important for users, and how it is designed to protect users' rights and privacy over personal data. Our study takes approximately 30 minutes to complete, and we pay 15$ to each participant. We conduct our user study in ten batches on MTurk and randomly assign ten – eleven participants to each batch. In total, we have 95 participants after removing seven participants that do not meet our requirements for the study. Our sample is fairly diverse. The median age is 30, with a range of 18 – 59. Among the participants, 56 (58.9%) are male, 37 (38.9%) are female, and two prefer not to disclose their gender. Their education levels range from high school degrees to graduate degrees.

5.5.2 Study design. We design a user study asking the participants about the compliance of the given privacy policies with GDPR requirements. Before starting the survey, we first describe the GDPR law and its purpose in terms of protecting users' privacy. We also ask a set of demographic questions regarding age, race, education level, etc. For each GDPR privacy disclosure requirement, we first explain the requirement in plain English. With the help of legal experts, we simplify the description and avoid using convoluted legal jargon. Once the participant indicates that they understand the requirement, we randomly select a privacy policy segment corresponding to the category. Since our goal is to measure whether privacy policies describing the GDPR requirements are understandable and effective from human perception, we ask participants to which degree they believe the given policy complies with the requirement. Specifically, participants can indicate whether a policy is fully compliant, moderately compliant, somewhat compliant, or not compliant at all in terms of the given requirement. On the other hand, if the participant cannot comprehend the given policy, they can indicate the option 'I don't understand the policy'. This allows us to find how much of the privacy policy can user comprehend and whether it is necessary to use an automated tool to detect the requirements.

5.5.3 User study results. Figure 7 summarizes the responses from 95 participants regarding the compliance of privacy policies in terms of 18 GDPR requirements. To our surprise, on average, 32% of participants were unable to understand the privacy policies. Even worse, for the core GDPR requirements such as profiling, data safeguards, and users' right to object, 43% participants were unable to understand the privacy policies. It is worth mentioning that these privacy policies are taken from the most visited websites and they are written for general users. These findings indicate the necessity of using automated tools to help users to understand the privacy policies and check their compliance. In what follows, we explain more details about our findings for a selected number of GDPR requirements used in our user study.

For the profiling requirement, we provide a privacy policy segment describing the profiling policy from a randomly selected website's privacy policy and ask the participants how compliant the given policy is in explaining the processing of users' personal information for automated decision making. As shown in Figure 7, our study finds that 41.27% participants did not understand the policy that explains how user's data is used by the underlying company for automated decision making. As the participants indicated, vague, complex, and long descriptions were the primary reasons for not understanding the policy description. GDPR requires the companies to disclose that the users have the right to withdraw their consent to process the data at any time. However, this is one of the GDPR requirements that is explained in a vague manner. In our study, 46.7% of participants did not understand the policy

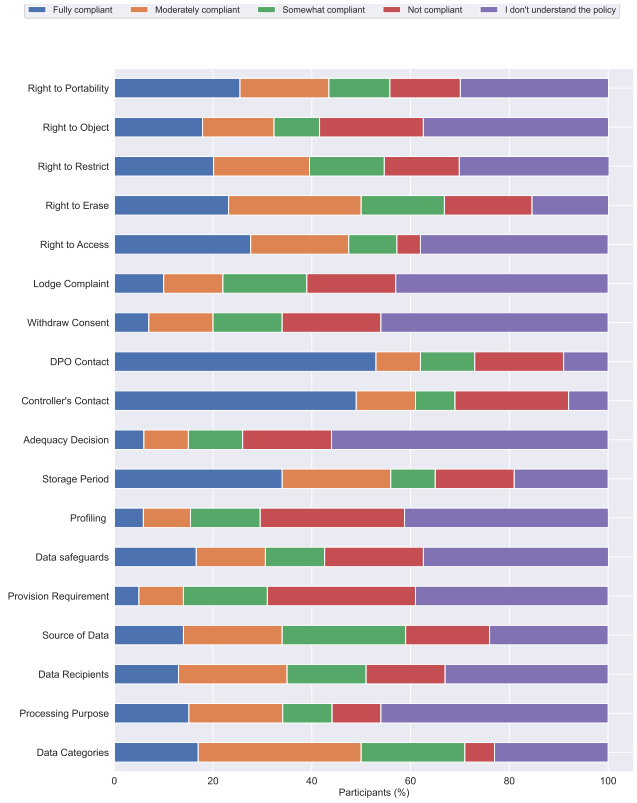

Figure 7: Category-wise response from the participants for the utilization of the current privacy policies' compliance with GDPR. Here, more than 40% participants were unable to fully understand policies that explain critical rights such as withdrawing consent, lodging complaint against the data processor, and disclosing information about profiling.

that explains their right to withdraw consent at any time. Even if they understand the policy, only 7.2% participants found the policies were fully compliant with the corresponding requirement and 20.1% found the policy was not compliant at all. These findings indicate that the privacy policies used by the companies to fulfill the disclosure requirements are not very helpful or effective for the end users.

Moreover, users have the right to lodge a complaint with a single supervisory authority regarding their data processing or transfer handled by the companies. According to our findings from the study, this requirement is also vaguely explained in the privacy policies, as 43% of our total participants were unable to comprehend the disclosure regarding this right of the users. In addition, 18% of participants found the policies not compliant at all. Privacy policies from the provision requirement, which enforces the companies to disclose whether providing the data is required, were not understandable to 39.7% users. To our surprise, only 5% of participants found the policies are fully compliant with the provision requirement. However, requirements relevant to the common privacy rights such as the right to access and the right to data portability seem to be more compliant with GDPR, as indicated by the participants of our study. For example, 27.6% found the corresponding policies compliant to the requirement of users' right to access data and 25.5% participants found the policies were compliant with users' right to data portability. On the other hand, 30% users indicated that they did not understand the policies describing these rights.

6 DISCUSSION

In this work, we contribute a new privacy policy dataset labeled with GDPR requirements and propose a CNN-based privacy policy classifier that can detect the compliance of privacy policies against GDPR with an accuracy of 89.2%. In our training model, currently, we only consider privacy policies in the English language since our team is not familiar with other languages in the EU. However, we are able to perform in-depth analysis to make up for language coverage. In addition, while we consider only privacy policy as it is a dominating way to communicate privacy disclosures to users [38], in practice, there might be other ways (e.g., newsletter) that companies could use to disclose information to users. Although analyzing other sources of the communication might be useful in some cases, we currently focus on the most widely used communication method of the privacy policy.

Our proposed classification model can be adapted to any future changes in the privacy laws, as we can add additional classes with new labeled data. Also, our trained model can be extended to other privacy laws such as California Consumer Privacy Act 2018 [2]. Although GDPR is primarily focused on EU countries and their citizens, it has become a standard for online privacy law, and legislators from other countries such as USA and UAE are planning to adopt GDPR for implementing new privacy protection laws [11]. Thus, our model, which is trained with GDPR requirements, can be extended by using learning methods such as *transfer learning* for detecting compliance with other privacy laws on large-scale.

Our dataset and the trained classifier can be utilized to explore several future research directions. For example, further research can be conducted to predict the usefulness score of human perception. It would allow not only to identify the compliance but also to design effective privacy policies for end users. In addition, analyzing the *flow-to-policy* consistency in terms of GDPR requirements would be an interesting research problem to explore in the future. Our dataset and model can be utilized to determine whether the actual data flow of any application or website is consistent with the information that is disclosed in the privacy policy. For example, similar to [15, 37], we can define privacy-specific queries to check if an application's data handling policies are compliant with its privacy policies. This will also help to determine the compliance of data flows with GDPR requirements.

Suggestions for better privacy policies. According to our analysis, 68% companies still fail to comply with at least one GDPR requirement. Also, many disclosure requirements of GDPR such as profiling, adequacy decision, DPO contact, lodge complaint, withdraw consent, and right to object are covered by only less than 40% website. Thus, we recommend companies be aware of these categories of requirements while implementing their privacy policy. On the other hand, for the companies that implement the requirements in their privacy policies, our study finds that 32% of users do not find the policies easy to understand. Therefore, we recommend

companies write privacy policies in more readable, understandable, and effective ways to improve user privacy. For example, by analyzing the privacy policy segments that are rated as helpful for users, we find these segments are usually written in a lay form, and sometimes provide easy to understand examples. We also find the less helpful and privacy-violating policy segments are usually vague or claim substantial control of users' data.

7 RELATED WORK

Privacy policies are one of the most common ways to disclose how companies collect, store or process users' personal data. Researchers have worked on identifying data practices from privacy policies by leveraging both manual and automated approach. Tesfay et al. [44] built a privacy policy summarization tool based on 11 privacy aspects of GDPR. However, since their tool was built upon pre-defined keywords-based features and trained on only 45 privacy policies, it resulted in lower accuracy and could not provide a reliable large-scale analysis. Degeling et al. [19] performed a longitudinal study on the privacy policies and cookie consent notices of 6,579 websites representing the 500 most popular websites in the EU countries. They showed that 72.6% of websites updated their privacy policies close to the date of May 25, 2018, when GDPR came into effect. However, their work primarily focused on cookie consent notices and reflection of GDPR terminologies in privacy policies without considering cross-checking compliance with GDPR requirements. Basin et al. [17] proposed a theoretical methodology to decompose GDPR compliance for auditing with the idea of identifying a business process and a purpose. Torre et al. [46] encoded GDPR and its compliance mechanisms into a UML representation so that they can be machine-analyzable, and leave the automated methods of assessing compliance to future researchers. Unlike their work, we contribute an automated tool to assess GDPR compliance in privacy policies using machine learning techniques.

Researchers have also performed privacy policy analysis using natural language processing and machine learning. Wilson et al. [48] introduced a corpus of 115 privacy policy documents annotated with fine-grained data practices (OPP-115). Harkous et al. [26] leveraged OPP-115 to build an automated QA model (Polisis) for regular data practices in privacy policies. Linden et al. [30] designed seven queries adapted from Polisis to represent a limited number of GDPR requirements and performed compliance analysis using a filtering and scoring approach. Chang et al. [18] proposed an automated privacy policy extraction system to predict and extract an app's privacy policies based on users' concerns under different contexts. However, these papers focused on extracting fine-grained data practices from privacy policies and did not consider disclosure requirements of the privacy laws. Hence, they cannot be easily extended for assessing the compliance of privacy policies with GDPR, which introduces more rules than regular data practices along with an unprecedented number of privacy rights regarding data collection and processing. In terms of automated GDPR compliance checking, Hamdani et al. [25] used NLP techniques to extract data practices from privacy policies and encoded GDPR rules in another module to check the presence of mandatory information. However, they only evaluated the model on 30 privacy policies. Torre et al. [45] created 20 metadata types by analyzing policies in GDPR and extracted metadata from the company's privacy policies automatically with the help of ML and NLP to check for compliance. However, they also had a limited dataset on evaluation, where they did a case study over 24 privacy policies. In addition, all 234 privacy policies they used for training were collected from financial fund management firms, which may infuse their model to be biased toward a particular domain, specifically to the financial domain in this case.

Researchers have also studied how users perceive online privacy policies. Linden et al. [29] conduct a user study to understand changes in a visual representation of privacy policies after GDPR. Their study finds that, after GDPR came into effect, EU websites made significant visual improvements in their privacy policies in terms of attractiveness, whereas websites outside of the EU did not have noticeable improvements. McDonald et al. [31] perform a comparative analysis to understand how well-standardized policies work in practice. Also, Reidenberg et al. [38] investigate the differences in interpretation of the privacy policies among different user groups. Gluck et al. [23] conducted multiple user studies to determine an effective design format for short-term privacy notices and participants' awareness regarding privacy practices. However, since companies have been changing their privacy policies since the implication of GDPR, these studies do not reflect the new changes.

8 CONCLUSION

In this paper, we provide insights into the compliance and effectiveness of privacy policies regarding the requirements of GDPR. We first identify comprehensive categories of GDPR requirements for privacy policies and create a privacy policy dataset annotated with 18 requirements by legal experts. We build a CNN-based automated tool to classify the privacy policies into GDPR requirements with an accuracy of 89.2%. We run our model to analyze privacy policies on large scale, which shows that only 32% of websites fully comply with GDPR in terms of their privacy policies.

ACKNOWLEDGEMENTS

We are grateful to the anonymous reviewers for their insightful and constructive feedback and suggestions. This work is supported in part by National Science Foundation (NSF) under the award numbers 1829004, 1920462, 1943100, 2002985 and 2114074, by Facebook Faculty Fellowship and by Google Research Scholar Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

REFERENCES

[1] 2018. 2018 reform of EU data protection rules. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en. Accessed: Feb 13, 2019.
[2] 2018. California Consumer Privacy Act 2018. https://en.wikipedia.org/wiki/California_Consumer_Privacy_Act. Accessed: July 19, 2019.
[3] 2018. General Data Protection Regulation (GDPR). https://gdpr-info.eu. Accessed: Feb 13, 2019.
[4] 2018. Privacy policies of tech giants 'still not GDPR-compliant'. https://www.theguardian.com/technology/2018/jul/05/privacy-policies-facebook-amazon-google-not-gdpr-compliant. Accessed: Feb 13, 2019.
[5] 2018. Top UK Websites. https://www.quantcast.com/top-sites/GB. Accessed: Feb 13, 2019.
[6] 2019. British Airways faces record £183m fine for data breach. https://www.bbc.com/news/business-48905907. Accessed: July 8, 2019.

[7] 2019. Data Protection Commission opens statutory inquiry into Facebook. https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-opens-statutory-inquiry-facebook-0. Accessed: July 1, 2019.

[8] 2019. Data Protection Commission opens statutory inquiry into Google Ireland Limited. https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-opens-statutory-inquiry-google-ireland-limited. Accessed: July 1, 2019.

[9] 2019. Data Protection Commission opens statutory inquiry into Twitter. https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-opens-statutory-inquiry-twitter-0. Accessed: July 1, 2019.

[10] 2019. Data Protection Commission reflects on the first year of the GDPR. https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-reflects-first-year-gdpr. Accessed: July 1, 2019.

[11] 2019. UAE data protection law, similar to GDPR, likely landing this year. https://www.techradar.com/amp/news/uae-data-protection-law-similar-to-gdpr-likely-landing-this-year. Accessed: Feb 13, 2019.

[12] 2019. Wikipedia: Privacy Policy. https://en.wikipedia.org/wiki/Privacy_policy. Accessed: July 28, 2019.

[13] 2019. Yandex Search API. https://yandex.com/support/search/robots/search-api.html. Accessed: Feb 15, 2019.

[14] Alessandro Acquisti and Ralph Gross. 2006. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. In *Privacy Enhancing Technologies, 6th International Workshop, PET 2006, Cambridge, UK, June 28-30, 2006, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 4258)*. Springer, 36–58. https://doi.org/10.1007/11957454_3

[15] Tamjid Al Rahat, Yu Feng, and Yuan Tian. 2019. Oauthlint: an empirical study on oauth bugs in android applications. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 293–304.

[16] Annie I. Antón, Elisa Bertino, Ninghui Li, and Ting Yu. 2007. A roadmap for comprehensive online privacy policy management. *Commun. ACM* 50, 7 (2007), 109–116. https://doi.org/10.1145/1272516.1272522

[17] David Basin, Søren Debois, and Thomas Hildebrandt. 2018. On Purpose and by Necessity: Compliance Under the GDPR. In *Financial Cryptography and Data Security*, Sarah Meiklejohn and Kazue Sako (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 20–37.

[18] Cheng Chang, Huaxin Li, Yichi Zhang, Suguo Du, Hui Cao, and Haojin Zhu. 2019. Automated and Personalized Privacy Policy Extraction Under GDPR Consideration. In *Wireless Algorithms, Systems, and Applications*, Edoardo S. Biagioni, Yao Zheng, and Siyao Cheng (Eds.). Springer International Publishing, Cham, 43–54.

[19] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2018. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. *CoRR* abs/1808.05096 (2018). arXiv:1808.05096 http://arxiv.org/abs/1808.05096

[20] Thomas Drugman, Janne Pylkkönen, and Reinhard Kneser. 2016. Active and Semi-Supervised Learning in ASR: Benefits on the Acoustic and Language Models. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*. ISCA, 2318–2322. https://doi.org/10.21437/Interspeech.2016-1382

[21] Tatiana Ermakova, Benjamin Fabian, and Eleonora Babina. 2015. Readability of Privacy Policies of Healthcare Websites. In *Smart Enterprise Engineering: 12. Internationale Tagung Wirtschaftsinformatik, WI 2015, Osnabrück, Germany, March 4-6, 2015*. 1085–1099. http://www.wi2015.uni-osnabrueck.de/Files/WI2015-D-14-00051.pdf

[22] National Center for Education Statistics. [n. d.]. The Program for the International Assessment of Adult Competencies (PIAAC) 2012/2014 Results. https://nces.ed.gov/surveys/piaac/results/summary.aspx. Accessed: July 19, 2019.

[23] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman M. Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How Short Is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *Twelfth Symposium on Usable Privacy and Security, SOUPS 2016, Denver, CO, USA, June 22-24, 2016*. USENIX Association, 321–340. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/gluck

[24] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2019. An empirical analysis of data deletion and opt-out choices on 150 websites. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*.

[25] Rajaa El Hamdani, Majd Mustapha, David Restrepo Amariles, Aurore Troussel, Sébastien Meeùs, and Katsiaryna Krasnashchok. 2021. A combined rule-based and machine learning approach for automated GDPR compliance checking. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 40–49.

[26] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *27th USENIX Security Symposium, USENIX Security 2018, Baltimore, MD, USA, August 15-17, 2018*. 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[27] Carlos Jensen, Colin Potts, and Christian Jensen. 2005. Privacy practices of Internet users: Self-reports versus observed behavior. *International Journal of Man-Machine Studies* 63, 1-2, 203–227. https://doi.org/10.1016/j.ijhcs.2005.04.019

[28] Rie Johnson and Tong Zhang. 2015. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. The Association for Computational Linguistics, 103–112. http://aclweb.org/anthology/N/N15/N15-1011.pdf

[29] Thomas Linden, Hamza Harkous, and Kassem Fawaz. 2018. The Privacy Policy Landscape After the GDPR. *CoRR* abs/1809.08396 (2018). arXiv:1809.08396 http://arxiv.org/abs/1809.08396

[30] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The privacy policy landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (2020), 47–64.

[31] Aleecia M. McDonald, Robert W. Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. A comparative study of online privacy policies and formats. In *Proceedings of the 5th Symposium on Usable Privacy and Security, SOUPS 2009, Mountain View, California, USA, July 15-17, 2009 (ACM International Conference Proceeding Series)*. ACM. https://doi.org/10.1145/1572532.1572586

[32] Gabriele Meiselwitz. 2013. Readability Assessment of Policies and Procedures of Social Networking Sites. In *Online Communities and Social Computing - 5th International conference, OCSC 2013, Held as Part of HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013. Proceedings (Lecture Notes in Computer Science, Vol. 8029)*. Springer, 67–75. https://doi.org/10.1007/978-3-642-39371-6_8

[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 3111–3119.

[34] Najmeh Mousavi Nejad, Damien Graux, and Diego Collarana. 2019. Towards Measuring Risk Factors in Privacy Policies. In *Proceedings of the First Workshop on AI in the Administrative State*.

[35] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[36] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2227–2237. https://aclanthology.info/papers/N18-1202/n18-1202

[37] Tamjid Al Rahat, Yu Feng, and Yuan Tian. 2022. Cerberus: Query-driven Scalable Vulnerability Detection in OAuth Service Provider Implementations. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*.

[38] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. 2015. Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ* 30 (2015), 39.

[39] Facebook Research. [n. d.]. FastText. https://github.com/facebookresearch/fastText/. Accessed: July 28, 2019.

[40] Subhadeep Sarkar, Jean-Pierre Banatre, Louis Rilling, and Christine Morin. 2018. Towards Enforcement of the EU GDPR: Enabling Data Erasure. In *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 222–229.

[41] Adam Satariano. [n. d.]. What the G.D.P.R., Europe's Tough New Data Law, Means for You. https://www.nytimes.com/2018/05/06/technology/gdpr-european-privacy-law.html. Accessed: Aug 22, 2019.

[42] Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=H1aIuk-RW

[43] Yanyao Shen, Hyokun Yun, Zachary Chase Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep Active Learning for Named Entity Recognition. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017*. Association for Computational Linguistics, 252–256. https://aclanthology.info/papers/W17-2630/w17-2630

[44] Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, IWSPA@CODASPY 2018, Tempe, AZ, USA, March 19-21, 2018*. ACM, 15–21. https://doi.org/10.1145/

3180445.3180447

[45] Damiano Torre, Sallam Abualhaija, Mehrdad Sabetzadeh, Lionel Briand, Katrien Baetens, Peter Goes, and Sylvie Forastier. 2020. An ai-assisted approach for checking the completeness of privacy policies against gdpr. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE, 136–146.

[46] Damiano Torre, Ghanem Soltana, Mehrdad Sabetzadeh, Lionel C Briand, Yuri Auffinger, and Peter Goes. 2019. Using models to enable compliance checking against the GDPR: an experience report. In *2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems (MODELS)*. IEEE, 1–11.

[47] Sida I Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 90–94.

[48] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard H. Hovy, Joel R. Reidenberg, and Norman M. Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. http://aclweb.org/anthology/P/P16/P16-1126.pdf

[49] Ben Wolford. [n. d.]. What is GDPR, the EU's new data protection law? https://gdpr.eu/what-is-gdpr. Accessed: Aug 22, 2019.

[50] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. 649–657. http://papers.nips.cc/paper/5782-character-level-convolutional-networks-for-text-classification

[51] Ye Zhang, Matthew Lease, and Byron C. Wallace. 2017. Active Discriminative Text Representation Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. AAAI Press, 3386–3392. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14174