Have You been Properly Notified? Automatic Compliance Analysis of Privacy Policy Text with GDPR Article 13

Shuang Liu
College of Intelligence and Computing, Tianjin University
China

Baiyang Zhao
College of Intelligence and Computing, Tianjin University
China

Renjie Guo
College of Intelligence and Computing, Tianjin University
China

Guozhu Meng
SKLOIS, Institute of Information Engineering
School of Cyber Security, University of Chinese Academy of Sciences
China

Fan Zhang
School of New Media and Communication, Tianjin University
China

MeiShan Zhang*
School of New Media and Communication, Tianjin University
China

ABSTRACT

With the rapid development of web and mobile applications, as well as their wide adoption in different domains, more and more personal data is provided, consciously or unconsciously, to different application providers. Privacy policy is an important medium for users to understand what personal information has been collected and used. As data privacy protection is becoming a critical social issue, there are laws and regulations being enacted in different countries and regions, and the most representative one is the EU General Data Protection Regulation (GDPR). It is thus important to detect compliance issues among regulations, e.g., GDPR, with privacy policies, and provide intuitive results for data subjects (i.e., users), data collection party (i.e., service providers) and the regulatory authorities. In this work, we target to solve the problem of compliance analysis between GDPR (Article 13) and privacy policies. We format the task into a combination of a sentence classification step and a rule-based analysis step. We manually curate a corpus of 36, 610 labeled sentences from 304 privacy policies, and benchmark our corpus with several standard sentence classifiers. We also conduct a rule-based analysis to detect compliance issues and a user study to evaluate the usability of our approach. The web-based tool AutoCompliance is publicly accessible [1].

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

Privacy, Compliance Analysis, Natural Language Processing

ACM Reference Format:
Shuang Liu, Baiyang Zhao, Renjie Guo, Guozhu Meng, Fan Zhang, and MeiShan Zhang. 2021. Have You been Properly Notified? Automatic Compliance Analysis of Privacy Policy Text with GDPR Article 13. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442381.3450022

1 INTRODUCTION

Recent years have witnessed a rapid development and wide adoption of web and mobile applications in our daily life. As a result, more and more personal data are provided to different application providers. There have been many reports on privacy invasions in the past few years. One of the most influential cases was with Alipay. When generating the 2017 annual bill for users, Alipay pre-checked the agreement tickbox of the "Zhima Credit Service Agreement" without notifying users explicitly[2]. This violates the conditions for consent in the General Data Protection Regulation (GDPR)[2] (Recital 32). Another case, shown in Figure 2(b), is the privacy policy of a Chinese deepfake-like APP named "ZAO", which explicitly states that user must agree to grant ZAO and its affiliates irrevocable, permanent rights (to their personal data)[3]. This offends users' right to data rectify, erase and object to processing (GDPR Article 13.2), as is shown in Figure 2(a).

In addition to GDPR, there have been other laws and regulations enacted in different areas and countries, e.g., the California Consumer Privacy Act in America [3] and the Data Protection Act [4]. Among them, GDPR is the most well-known, due to its large territorial scope, as well as some well-known punishment cases happened recently. For instance, there was an investigation towards Mobike by the Berlin commissioner of data protection and freedom of information (based on GDPR) [4].

Although there have been laws and regulations aiming at protecting personal data, it is hard to know or check whether they are properly obeyed by companies/agencies which collect, process or store users' personal data. The difficulties mainly comes from two

[*] Meishan Zhang is the corresponding author.

[1] www.ppvisual.site

[2] https://www.scmp.com/tech/china-tech/article/2126772/chinas-ant-financial-apologises-over-alipay-user-data-gaffe

[3] The information is obtained from news snapshot, the privacy policy of ZAO has been updated after the report.

[4] https://medium.com/@a.hanff/chinas-surveillance-social-credit-system-alive-kicking-in-berlin-6c2b3b10b197

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450022

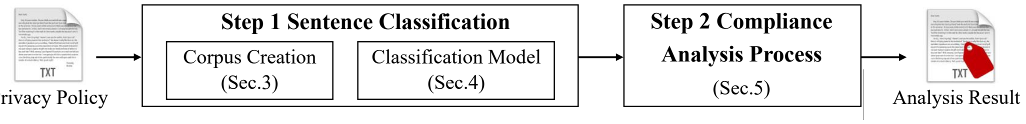
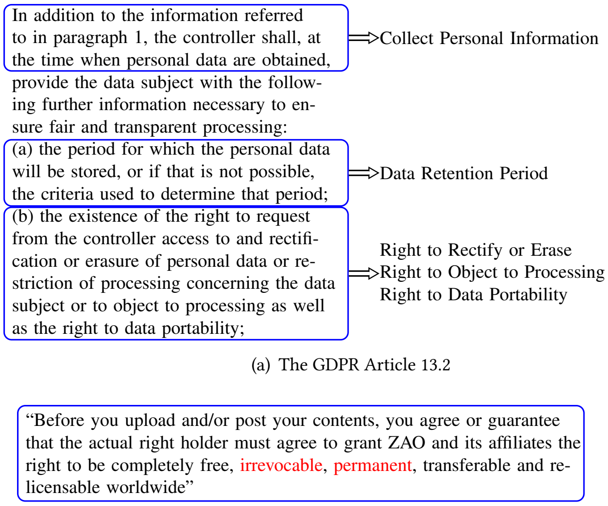

TXT
Privacy Policy
Step 1 Sentence Classification
Corpus Creation (Sec.3)
Classification Model (Sec.4)
Step 2 Compliance Analysis Process (Sec.5)
TXT
Analysis Result

Figure 1: The Overview of Our Approach



In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: ⟹ Collect Personal Information

(a) the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period; ⟹ Data Retention Period

(b) the existence of the right to request from the controller access to and rectification or erasure of personal data or restriction of processing concerning the data subject or to object to processing as well as the right to data portability; ⟹ Right to Rectify or Erase, Right to Object to Processing, Right to Data Portability

(a) The GDPR Article 13.2

"Before you upload and/or post your contents, you agree or guarantee that the actual right holder must agree to grant ZAO and its affiliates the right to be completely free, irrevocable, permanent, transferable and re-licensable worldwide"

(b) The ZAO privacy policy Excerpt

Figure 2: The Motivating Example

aspects. First, similar to the other laws and regulations, GDPR is written in natural language and contains a large number of law-specific terms, which is hard for users without domain knowledge to understand. Second, privacy policies are usually long documents written in natural language, which are time-consuming for App users to read. A previous study [21] concludes that an average of 40 minutes per day is required from American citizens if they read all privacy policies displayed to them.

As a result, it is hard for Application users to discover an infringement towards themselves. Moreover, in some cases, the service developers/providers violate the laws/regulations unintentionally due to a lack of related knowledge. Therefore, it is desirable to detect the compliance issues of privacy policies with privacy protection laws and regulations, e.g., GDPR, automatically.

In this work, we propose an approach to automatically analysis the privacy policy contents and report violations against GDPR (Article 13). The reason for choosing GDPR Article 13 is that it regulates the aspects, which contain the regulations on what and how information must be provided to the data subjects [11], and are the most suitable to be reflected in privacy policies. We devise a classification scheme based on GDPR (Article 13) and annotate a corpus of 304 privacy policies. We then benchmark our corpus with several standard classification models and conduct a thorough analysis on the results. Lastly, we check whether the privacy policy is compliant with GDPR Article 13 based on the rules extracted from GDPR and the classification results. Our method detects $1,180$ compliance issues in 304 privacy policy documents.

We implement our method in a web-based tool named AutoCompliance. AutoCompliance is of high importance to web and mobile application users, service providers as well as the regulatory authority. It is powered by a sentence classifier and a rule-based compliance analysis module. We also conduct a user study to evaluate the usability of AutoCompliance, and the results show that our tool reduces the reading cost of 55% on average among all volunteers. The user interview shows that all volunteers in the experiment group confirm the usefulness of the tool. AutoCompliance as well as the labeled corpus, are publicly available[5].

The rest of the paper is organized as follows: Section 2 introduces the overall framework of our approach. Section 3 reports the process of corpus creation. The details of the sentence classification models and the compliance analysis process are introduced in section 4 and section 5, respectively. The evaluations on the sentence classification and compliance analysis are reported in section 6. We present the implementation details of AutoCompliance and the user study details in section 7. Related work is discussed in section 8 and section 9 concludes the paper.

2 OVERVIEW OF OUR APPROACH

The overview of our approach is shown in Figure 1. Our approach contains two main steps, including the sentence classification step and the rule based compliance analysis step. Given a privacy policy document, we first conduct sentence classification to predict a label for each sentence, which is conducted with supervised learning approaches. To achieve the goal, we first create a corpus of $36,610$ sentences from 304 privacy policies based on the label scheme devised from GDPR Article 13, which is introduced in detail in section 3. Then we train classification models based on the labeled data for the sentence classification task, the details of which are discussed in section 4.

After obtaining the label for each sentence, we conduct the rule-based compliance analysis to identify compliance issues, as detailed in section 5. The detected compliance issues are reported. To better assist users in finding and understanding the compliance issues, we implement our approach in a web-based tool named AutoCompliance, and conduct user study to evaluate the usability of our approach. The details are introduced in section 7.

3 CORPUS CREATION

3.1 Label Scheme based on GDPR

The General Data Protection Regulation (GDPR) was enacted since May, 2018. It aims to protect all EU citizens from privacy and data

[5] https://github.com/ppcompliance/PPGDPR

breaches, and the territorial scope is all companies processing the personal information of data subjects residing in the EU, regardless of the location of the companies. GDPR is written in natural language. It contains 11 chapters and 99 articles.

As is discussed by [11], Articles 12-14 of GDPR contain the regulations on what and how information that must be provided to the data subjects. Article 12 focuses on practices of exercising the rights of the data subjects. Article 13 describes the information which the data controller has to provide to the data subject when personal data is collected, and are suitable to be reflected in privacy policies. Article 14, on the other hand, describes what information has to be provided when personal data has not been collected.

Since natural language descriptions are inheritably ambiguous, we focus on the aspects which are objective and clearly stated, without case-specific context information to be considered/referred. For instance, a common description in GDPR is using "where applicable" in the clauses, which makes the corresponding clauses context-dependent and subjective. Our work focuses on analyzing whether the data controllers violate GDPR when they intend to collect personal data. Therefore, we focus on the clauses stated in Article 13 of GDPR. By reading through the contents of Article 13, we manually extract 10 labels, which are contents regarding personal information collection and can naturally be presented in privacy policies. During the label extracting process, we consult the experts from law school on the meanings of the concepts in GDPR to have a concise understanding. The details of the extracted labels and their correspondences with GDPR clauses are explained in the following:

1. *Collect Personal Information*: Collect data subjects' information which can identify their personal IDs. `[GDPR Art 13.1]`
2. *Data Retention Period*: Retention period of personal information. `[GDPR Art 13.2(a)]`
3. *Data Processing Purposes*: The purposes of processing personal data. `[GDPR Art 13.1(c)]`
4. *Contact Details*: The contact details of the controller or the Data Protection Officer. `[GDPR Art 13.1(a)(b)]`
5. *Right to Access*: The right (of the data subject) to request from the controller to access their personal information. `[GDPR Art 13.2(b)]`
6. *Right to Rectify or Erase*: The right (of the data subject) to request from the controller to rectify or erase of their personal information. `[GDPR Art 13.2(b)]`
7. *Right to Restrict of Processing*: The right (of the data subject) to request from the controller to restrict processing concerning the data subject. `[GDPR Art 13.2(b)]`
8. *Right to Object to Processing*: The right (of the data subject) to request from the controller to object to processing. `[GDPR Art 13.2(b)]`
9. *Right to Data Portability*: The right (of the data subject) to receive and transmit his/her personal data to another controller. `[GDPR Art 13.2(b)]`
10. *Right to Lodge a Complaint*: The right (of the data subject) to lodge a complaint with a supervisory authority. `[GDPR Art 13.2(d)]`

Table 1: The Totalized Statistics on the Annotated Corpus

No. Documents	304
No. Words	926, 858
Annotated Sentences	36, 610
Annotators per Sentence	3
Total Annotators	22

3.2 Privacy Policy Collection

We collect privacy policies of APPs from Google Play [6], one of the most popular application stores. We use the Scrapy web framework[7] and Selenium[8] for data crawling. In our work, we aim at collecting a set of high quality privacy policies with diverse application categories. We use the following rules to collect privacy policies: (1) the privacy policies of applications which are in the top list of Google Play; and (2) the privacy policies should be from diverse categories since different categories may have different requirements on accessing user information. The privacy policies we collect cover 22 application categories, such as Communication, Game, and Business.

Note that there may be some noise in the crawled privacy policies. To ensure the quality of the collected privacy policy documents, we filter the crawled privacy policies with the following criteria: (1) the privacy policy is written in English; (2) the contents of the privacy policy are not too short. To be specific, 2KB[9] is set as the lower bound size of the privacy policy documents; and (3) duplicated privacy policies, which are usually from different apps of the same company, are removed.

We crawl $1,313$ privacy policies and after filtering, 304 valid privacy policies are left for annotation. The statistics of our corpus are shown in Table 1, which contains 304 privacy policies of more than $926K$ words and $36K$ sentences. The average length of privacy policies in our corpus is $3,049$ words, among which 10% have less than $1,000$ words, with the shortest one having 154 words.

3.3 Data Annotation

We adopt and customize YEDDA [35], an open source text span annotation tool, for our annotation task.

We recruit 22 volunteers, who are undergraduate and postgraduate students major in law and computer science, and are good at English, to annotate the privacy policies. To control the annotating quality, we first train the volunteers on the annotation task. We give a brief tutorial and provide labeled example sentences to clarify the meaning of each label. After the training, we require the volunteers to label a small set of privacy policies and check on the quality of their annotation results, then we clarify the misunderstandings if any. With such a process, we ensure that all volunteers have a clear understanding of the meaning of the labels, and thus have control on the annotation quality. Each sentence is labeled by 3 volunteers

[6] https://play.google.com/store/apps
[7] https://scrapy.org/
[8] https://docs.seleniumhq.org/
[9] The 2KB limit is set based on a pre-analysis on the privacy policies we crawled from Google Play, in which we find that the files of size smaller than 2KB are usually noise data with an average of 38 words per file.

Table 2: The Categorized Statistics on the Annotated Corpus

Label	Frequency	Coverage (%)	Avg.W	Fleiss' Kappa
Collect Personal Information (CPI)	1,542	94.41	31.61	0.45
Data Retention Period (DRP)	448	61.51	30.50	0.45
Data Processing Purposes (DPP)	1,839	93.75	25.76	0.51
Contact Details (CD)	721	85.20	24.13	0.47
Right to Access (RA)	115	29.28	25.32	0.47
Right to Rectify or Erase (RRE)	562	70.07	23.61	0.49
Right to Restrict of Processing (RRP)	127	29.28	23.03	0.51
Right to Object to Processing (ROP)	245	40.46	23.24	0.47
Right to Data Portability (RDP)	167	35.53	26.30	0.57
Right to Lodge a Complaint (RLC)	145	36.84	24.77	0.57
Other	30,699	100.00	24.98	0.48

independently. Each volunteer is assigned with a set of privacy policies and they annotate the assigned tasks independently. It takes an average of 40 minutes for each volunteer to annotate one privacy policy. After all volunteers finish their own annotation task, we ask the 3 volunteers who label the same privacy policies to meet and merge the labels. Following the standard process, if all 3 volunteers give the same label, then the label is used as the final label of the sentence. Otherwise, they will discuss until a consensus is reached.

The details of the annotated corpus are shown in Table 2. The `Frequency` column shows the collective counts of the corresponding label in our corpus. `Coverage` indicates the coverage of the corresponding label, i.e., the percentage of privacy policy documents which contain that label. The column `Avg.W` is the average number of words per sentence in our corpus. The last column is the Fleiss' Kappa of the annotation results (before merging). The labels *Data Processing Purposes (DPP)* and *Collect Personal Information (CPI)* appear the most frequently in privacy policies. Other categories, such as the data subjects' *Right to Access (RA)*, *Right to Restrict of Processing (RRP)*, *Right to Data Portability (RDP)* and *Right to Lodge a Compliant (RLC)*, are much less frequently mentioned in privacy policies. Note that there is a special label, i.e., *Other*, which contains all sentences that are not in the 10 labels. Since our corpus focuses on annotating the entire privacy policy document, we explicitly annotate all sentences in a privacy policy document. The `Other` category counts for 84% of the total sentences.

The Fleiss' Kappa value ranges from 0.45 to 0.57, which falls in the moderate agreement level according to Landis et al. [17]. Through a thorough analysis of the labeled sentences, we identify three potential reasons for the moderate level of agreement: (1) The data is imbalanced. The portion of sentences with the GDPR-related labels, i.e., the first 10 labels in Table 2, count for 16% of the total number of sentences. (2) There are various sentence descriptions for each single label; and (3) For some categories, such as *Collection Personal Information (CPI)*, there are sentences which are ambiguous for the annotators since there is no clear boundary on personal and non-personal information. According to GDPR Article 4(1), "personal data are any information which are related to an identified or identifiable natural person". There is no quantitative definitions and thus annotators may give subjective decisions. One example is the following sentence, where some annotators mislabel it as the *Collect Personal Information (CPI)* category: *"When you use our service, our servers automatically record certain log file information, including your web request, browser type, referring / exit pages and URLs, number of clicks and how you interact with links on the Service, domain names, landing pages, pages viewed, and other such information".*

4 THE CLASSIFICATION MODEL

In this work, we exploit three different kinds of models to benchmark the performance of the sentence classification task, the accuracy of which is critical to the compliance analysis task. We adopt support vector machine (SVM) [8] with traditional one-hot manually-crafted features. Due to the advances neural models have achieved on natural language processing tasks, we also investigate two representative neural models, i.e., (1) embedding-based inputs with bi-directional long short term memory networks (BiLSTM) [12], and (2) contextualized BERT [10] representations as inputs for sentence-level classification.

4.1 SVM

SVM [8] takes linear human-designed discrete features as inputs, which has shown to be a strong baseline for a number of NLP classification tasks. Therefore, we adopt it as one baseline of our classification task. We follow the majority of previous work to use n-gram [33] and TF-IDF [25] features in this work. The TF-IDF values are calculated with the training corpus.

4.2 BiLSTM

We adopt a standard BiLSTM [12] model for sentence classification, which consists of four layers:

(1) The word representation layer converts discrete words $\boldsymbol{w}_1\boldsymbol{w}_2 \cdots \boldsymbol{w}_n$ into low-dimensional vectors $\boldsymbol{x}_1\boldsymbol{x}_2 \cdots \boldsymbol{x}_n$.
(2) The BiLSTM layer takes word representations as inputs, and compose high-level representations which can capture implicit long-distance connections between words: $\boldsymbol{h}_1\boldsymbol{h}_2 \cdots \boldsymbol{h}_n = \textsc{BiLSTM}(\boldsymbol{x}_1\boldsymbol{x}_2 \cdots \boldsymbol{x}_n)$.
(3) The Pooling layer aggregates word-level features for sentence classification. We exploit max pooling in our approach. The process is formalized as: $\boldsymbol{s} = \textsc{MaxPool}(\boldsymbol{h}_1 \cdots \boldsymbol{h}_n)$.
(4) The Classification layer conducts final predictions through a feed-forward neural (FFN) layer: $\boldsymbol{o} = \text{FFN}(\boldsymbol{s})$.

Table 3: The Precision/Recall/F1-score for Classification Models

	SVM			**LSTM**						**BERT**					
				$\mathcal{L}$			$\mathcal{L}_{\mathcal{W}}$			$\mathcal{L}$			$\mathcal{L}_{\mathcal{W}}$		
Category	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
CPI	**75.96**	5.12	9.60	59.15	32.68	42.11	49.13	49.35	49.24	56.27	43.97	49.36	55.72	**55.78**	**56.73**
DRP	**83.62**	33.04	47.36	67.13	43.30	52.65	61.90	49.33	54.91	66.59	68.53	67.55	68.63	**72.77**	**70.64**
DPP	**82.35**	3.05	5.87	69.76	32.25	44.11	60.56	45.84	52.18	58.74	44.05	50.34	64.78	**56.61**	**60.42**
CD	**85.71**	46.60	60.38	83.37	59.08	69.16	75.76	68.93	72.19	76.18	66.99	71.29	84.78	**73.37**	**78.66**
RA	**70.69**	35.65	47.40	61.25	42.61	50.26	66.28	49.57	56.72	55.04	**61.74**	58.20	65.42	60.87	**63.06**
RRE	**81.95**	40.39	54.11	72.73	58.36	64.76	71.81	67.08	69.37	73.36	61.74	67.05	69.50	**69.75**	**69.63**
RRP	**84.21**	50.39	63.05	74.32	43.31	54.73	77.55	59.84	67.56	83.18	70.08	76.07	83.62	**76.38**	**79.84**
ROP	**88.98**	46.12	60.75	73.33	49.39	59.02	76.47	63.67	69.49	75.65	59.59	66.67	77.83	**64.49**	**70.54**
RDP	**83.94**	68.86	75.66	80.77	62.87	70.71	75.16	70.66	72.84	80.50	76.65	78.53	81.76	**83.23**	**82.49**
RLC	**91.30**	72.41	80.77	84.92	73.79	78.97	81.34	75.17	78.14	83.85	75.17	79.27	82.78	**86.21**	**84.46**
Avg	**82.87**	40.16	50.50	72.67	49.76	58.65	69.60	59.94	64.26	70.94	62.85	66.43	73.48	**70.15**	**71.65**
Other	86.98	**99.39**	92.77	90.05	97.04	93.41	91.98	94.23	93.09	91.86	94.97	93.39	**93.61**	94.51	**94.06**

Table 4: The Compliance Analysis Rules

1. *Collect Personal Info* ⟶ *Data Retention Period*
2. *Collect Personal Info* ⟶ *Data Processing Purposes*
3. *Collect Personal Info* ⟶ *Contact Details*
4. *Collect Personal Info* ⟶ *Right to Access*
5. *Collect Personal Info* ⟶ *Right to Rectify or Erase*
6. *Collect Personal Info* ⟶ *Right to Restrict of Processing*
7. *Collect Personal Info* ⟶ *Right to Object to Processing*
8. *Collect Personal Info* ⟶ *Right to Data Portability*
9. *Collect Personal Info* ⟶ *Right to Lodge a Complaint*

4.3 BERT

BERT [10] is a well-known model for contextualized word representations, which has achieved state-of-the-art performance on a wide range of NLP tasks [31]. It accepts a full sentence as input, and outputs a sequence of hidden vectors based on a well pre-trained model. Following the standard setting in Devlin et al. [10], we use the vectorial output of the sentence start symbol [CLS] in the last-layer as the full sentence representation, and then use an FFN layer to score each candidate label. The process can be formalized as:

$$\begin{aligned} \boldsymbol{s} &= \mathrm{BERT}(w_1 \cdots w_n) \\ \boldsymbol{o} &= \mathrm{FFN}(\boldsymbol{s}) \end{aligned} \tag{1}$$

We follow the standard process, which has been demonstrated to be effective in several tasks [10, 14], to fine tune the BERT parameters along with our task objective.

4.4 Training

For SVM, we use the standard max-margin objective function for model optimization. For the two neural models, we exploit the cross-entropy loss as the final objective:

$$\mathcal{L} = -\log \frac{\exp(\boldsymbol{o}_g)}{\sum_l \exp(\boldsymbol{o}_l)}, \tag{2}$$

where g is the golden standard label and the denominator is a normalization factor.

We observe that there exists significant imbalance problem towards targeted labels, we make slight adaptations on the cross-entropy objective aiming to better train our models. Concretely, we add one weight for each label when computing the loss:

$$\mathcal{L}_{\mathcal{W}} = -\lambda_g \log \frac{\exp(\boldsymbol{o}_g)}{\sum_l \exp(\boldsymbol{o}_l)}, \tag{3}$$

where λ_g is the inverse proportion to the label frequency in the training corpus, and a normalization is applied to make $\lambda_g \in (0, 1)$.

The details on the training process and hyper parameters settings are introduced in section 6.1.

5 THE COMPLIANCE ANALYSIS PROCESS

We observe that the clauses defined in GDPR Article 13 follow the pattern of "if A holds, then B must be satisfied", where "A" represents operations on the data subjects' personal information and "B" represents the information that the data controller should provide to the data subject, which are captured by our label scheme. With the 10 labels extracted from Article 13 of GDPR, we obtain 9 rules, which are listed in Table 4. Each rule indicates a kind of information that the data controller should provide to the data subject if the data subject's personal information is collected.

As an example, We list a part of the `GDPR Article 13.2` in Figure 2(a). From this article, we can extract 4 rules according to the given pattern. Rule 7 in Table 4 can be extracted, where "A" is reflected by the sentence "*controller collects personal information (from the data subject)*" and is captured by the label *Collect Personal Information (CPI)*; "B" is reflected by the sentence "*the controller shall provide the data subject the existence of right to...object to processing*" and is captured by the label *Right to Object to Processing (ROP).*

Since the compliance analysis rules follow the implication format, based on properties of propositional logic:

$$A \rightarrow B \equiv \neg A \vee B \tag{4}$$

we only need to check if "B" appears in a privacy policy document, or if "A" never appears ($\neg A$) in the privacy policy. Both cases gave a true evaluation, i.e., the GDPR clause is not violated. Otherwise, i.e., $A \wedge \neg B$ is true, a violation is reported. With this observation, our compliance analysis task is further decomposed into the sentence classification task to check whether a privacy policy contains sentences with the required labels, i.e., the privacy policy does not have any sentences describing collecting personal information ($\neg A$), or there are sentences properly describe the required user rights (B).

6 EXPERIMENTS

6.1 Experiment Settings

For the sentence classification task, we conduct the ten-fold cross-validation experiments, where the whole corpus is evenly divided into ten folds, and each fold is tested by regarding the remaining 8 folds as the training set and one fold as the development set. We run all experiments 5 times, and report the median-performance results. Standard precision (P), recall (R) and F1-score (F) are used as the evaluation metrics.

We exploit the SciKit-learn 0.22 toolkit to implement the SVM model, and the Linear kernel function is adopted. Pytorch 1.2.0 is used to implement the neural classification models including BiLSTM and BERT. For BiLSTM, we adopt the Glove [24] embedding as the input vector and the vector size is set to 100. For Bert, we adopt the Google released BERT-Base, Uncased version [10] in our evaluation. Online learning with mini-batching (i.e., batch size is 4) is exploited to train the neural classification models. We exploit the Adam algorithm for optimization, with the initial learning rates of 2e-4 and 5e-5, for BiLSTM and BERT, respectively. The maximum training epoch is set to 16, where the best-performance model on the development set is selected as the final model.

6.2 Classification Result

The classification results are shown in Table 3. The best P, R, F values of each category are highlighted in bold. The row Avg shows the macro average value of the 10 labels extracted from GDPR. We explicitly report the classification results of the *Other* category, as this category is not related to our compliance analysis task, yet it contains a dominant number of sentences which affect the classification results of the other 10 labels.

From the results, we can observe that, in general BERT shows the best average F1-score, followed by BiLSTM. SVM shows the worst average F1-score among all models. In particular, SVM shows the best precision for all 10 labels and the lowest average recall.

There are some categories, such as *Collect Personal Information (CPI)*, which is more difficult than the other categories due to ambiguities. This is due to the vague descriptions on personal information. According to GDPR Article 4(1), "personal data is any information which are related to an identified or identifiable natural person". However, there exist descriptions in privacy policies which do not explicitly specify what information they collect, other cases collect information such as user browser version, which is not personal information according to GDPR. The most common mis-classified cases happen between the *Collect Personal Information (CPI)* and the *Other* categories. There are sentences which describe the concept of "personal data", which are labeled as the *Other* category. Those sentences may share similar keywords and context, which confuse the classification model. This also happens for human annotators as we discussed in Section 3.3. Specifically, sentences of label *Collect Personal Information (CPI)* and *Data Processing Purposes (DPP)* share similar keywords, such as " *Collect Personal Information*", where the *Data Processing Purposes (DPP)* focuses on describing the purposes, and the *Collect Personal Information (CPI)* focuses on the way of collection or contents being collected. These two labels are not well recognized by all models.

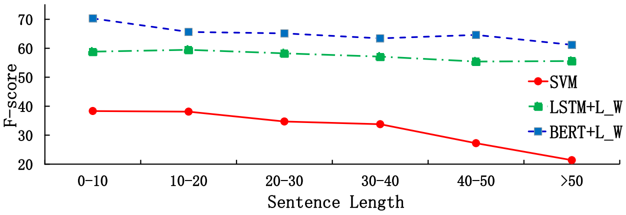


Figure 3: F1-Score Against Sentence Length for SVM, LSTM + $\mathcal{L}_{\mathcal{W}}$ and BERT + $\mathcal{L}_{\mathcal{W}}$

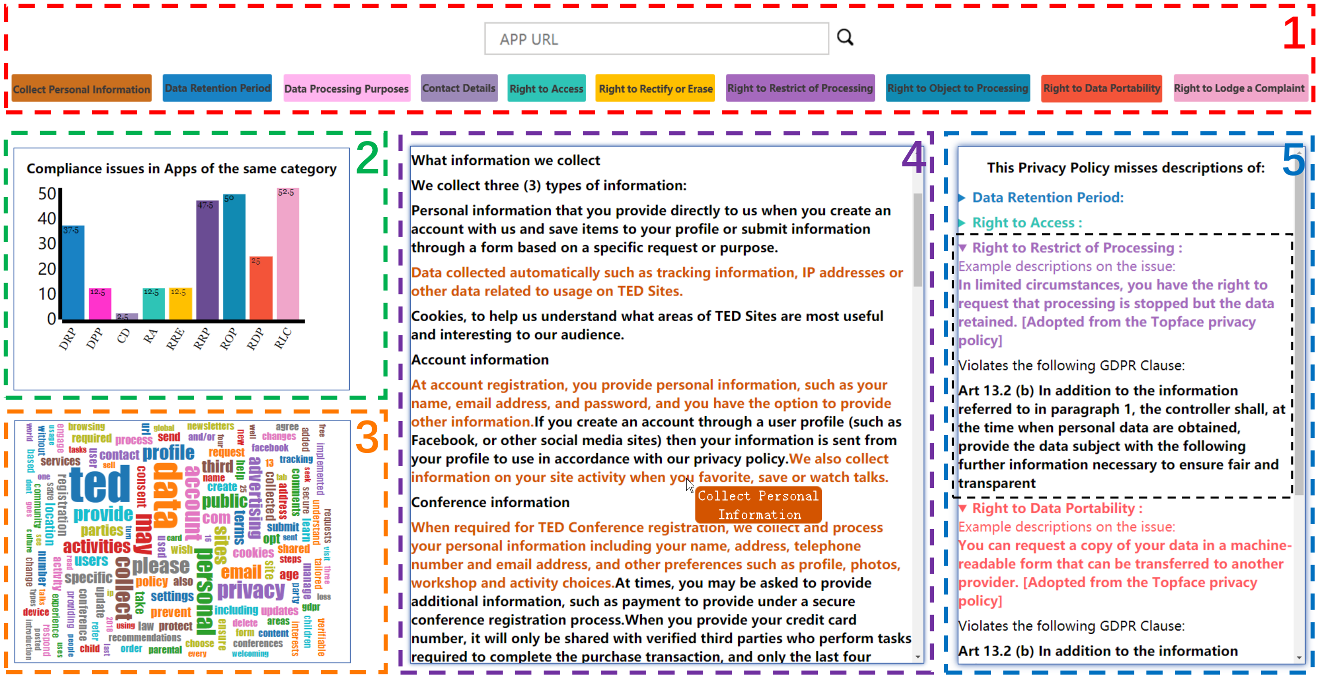
We can also observe that the weighted loss function contributes to the performance improvement of both BiLSTM and BERT. In particular, more than 5% increase of F1-score is achieved for both BiLSTM and BERT.

While for the BiSLTM model, no pre-trained sentence-level semantic information is involved, all semantic knowledge is learned through the annotated corpus, and thus the label distribution could offer less prior knowledge, which can balance the semantic information for the GDPR classification. Moreover, BERT also provides a stronger baseline result, which makes it more difficult to achieve large improvements.

Note that the *Other* category, which occupies 84% sentences in the corpus, yet is not related to our compliance analysis task, affects the classification results on the other 10 labels. Our classification task favors more on recall than on precision, since we would like to recognize all sentences which belong to the 10 categories. Therefore, models of high recall are more preferable in our task. We can observe that the SVM model shows the highest average precision, and the lowest average recall on the 10 labels. The weighted loss function contributes to the improvement of recall value for both BiLSTM and BERT. In particular, an improvement of more than 10% on the recall is achieved for BiLSTM and more than 7% for BERT. Since the BERT+$\mathcal{L}_{\mathcal{W}}$ model achieves the best recall and F1-score, we adopt it for the compliance analysis tasks in the rest of this paper.

To further observe the performance of different models affected by the sentence length (in terms of the word count in a sentence), we compare the F1-score[10] against sentence length for SVM, LSTM + $\mathcal{L}_{\mathcal{W}}$ and BERT + $\mathcal{L}_{\mathcal{W}}$, and the results are shown in Figure 3. We can observe that SVM shows a decreasing performance with the increase of the sentence length, and the decreasing trend is sharp after sentence length of 40. The main reason might be that features used in SVM are only able to reflect local information. Neural network models, i.e., BiLSTM + $\mathcal{L}_{\mathcal{W}}$ and BERT + $\mathcal{L}_{\mathcal{W}}$, outperform SVM on the 10 label classification results. They can achieve stable performance over all length distributions, due to the fact that the network structures of BiLSTM and BERT can

[10] We use the Micro average and remove the *Other* category.


1
APP URL
Collect Personal Information
Data Retention Period
Data Processing Purposes
Contact Details
Right to Access
Right to Rectify or Erase
Right to Restrict of Processing
Right to Object to Processing
Right to Data Portability
Right to Lodge a Complaint
2
Compliance issues in Apps of the same category
50
40
30
20
10
0
37.5
12.5
2.5
12.5
12.5
47.5
50
25
52.5
DRP
DPP
CD
RA
RRE
RRP
ROP
RDP
RLC
3
ted
data
personal
account
provide
privacy
collect
profile
please
may
4
Collect Personal Information
5

What information we collect

We collect three (3) types of information:

Personal information that you provide directly to us when you create an account with us and save items to your profile or submit information through a form based on a specific request or purpose.

Data collected automatically such as tracking information, IP addresses or other data related to usage on TED Sites.

Cookies, to help us understand what areas of TED Sites are most useful and interesting to our audience.

Account information

At account registration, you provide personal information, such as your name, email address, and password, and you have the option to provide other information.If you create an account through a user profile (such as Facebook, or other social media sites) then your information is sent from your profile to use in accordance with our privacy policy.We also collect information on your site activity when you favorite, save or watch talks.

Conference information

When required for TED Conference registration, we collect and process your personal information including your name, address, telephone number and email address, and other preferences such as profile, photos, workshop and activity choices.At times, you may be asked to provide additional information, such as payment to provide under a secure conference registration process.When you provide your credit card number, it will only be shared with verified third parties who perform tasks required to complete the purchase transaction, and only the last four

This Privacy Policy misses descriptions of:

▸ Data Retention Period:

▸ Right to Access :

▾ Right to Restrict of Processing :
Example descriptions on the issue:
In limited circumstances, you have the right to request that processing is stopped but the data retained. [Adopted from the Topface privacy policy]

Violates the following GDPR Clause:

Art 13.2 (b) In addition to the information referred to in paragraph 1, the controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent

▾ Right to Data Portability :
Example descriptions on the issue:
You can request a copy of your data in a machine-readable form that can be transferred to another provider. [Adopted from the Topface privacy policy]

Violates the following GDPR Clause:

Art 13.2 (b) In addition to the information

Figure 4: The Screenshot of AutoCompliance

capture sentence-level global features, since long-distance word connections are properly modeled.

One example of the detected compliance issues is in privacy policy of the Discord APP[11]. The compliance issue is about *Right to Lodge a Complaint (RLC)*, which violates Clause 2 (d) in GDPR Article 13. The reason is that there are no explicit descriptions on user right to lodge a complaint in the privacy policy when the *Collect Personal Information (CPI)* is explicitly stated by the sentence *“Information we collect may include but not be limited to username, email address, and any messages, images, transient VOIP data (to enable communication delivery only) or other content you send via the chat feature.”* LSTM + $\mathcal{L}_W$ and BERT + $\mathcal{L}_W$ successfully label that sentence as *Collection Personal Information (CPI)*, while SVM fails to correctly label the sentence.

All three models detect missing of descriptions on *Right to Lodge a Complaint (RLC)*. However, since SVM fails to label the sentence of collecting personal information, it fails to report the compliance issue. LSTM + $\mathcal{L}_W$ and BERT + $\mathcal{L}_W$ successfully report the compliance issue following the compliance analysis rule 9 in Table 4. The details of the compliance analysis results are discussed in section 6.3.

6.3 Compliance Analysis Results

We conduct compliance analysis with the BERT + $\mathcal{L}_W$ model, which shows the best and most stable performance. The compliance analysis is based on the rules in Table 4. Recall that the rule is evaluated to false, i.e., a compliance issue is reported, if $A \wedge \neg B$ holds. Our method reports 1,180 compliance issues, out of which 1,164 are real compliance issues, and there are 107 issues that are undetected. This gives the compliance analysis an accuracy of 90% and recall of 91%.

There are 73 in the 107 missed issues are due to classification errors, i.e., our model does not correctly classify the label and thus fails to report the $\neg B$ part of the rule. In particular, the *Data Retention Period (DRP)* and *Contact Details (CD)* labels account for most of the cases.

Among the successfully detected issues, the most frequently occurred issues are due to missing of *Right to Lodge a Complaint (RLC)* label, which occurs more than 20 times. Among the successfully detected issues, the most frequently occurred issues are due to missing of *Right to Access (RA)* and *Right to Restrict of Processing (RRP)* labels, both of which occur more than 180 times. Note that even though our method has a recall of only 56% on the sentence classification of the *Collect Personal Information (CPI)* label, it is sufficient for the compliance analysis task, as there are usually more than 2 sentences describing the meaning of *Collect Personal Information (CPI)*, and our approach can correctly identify the compliance issue if any CPI sentence is correctly recognized.

7 USABILITY EVALUATION

To evaluate how practical and useful is our method in relieving users efforts in checking the compliance of privacy policies in accordance with GDPR Article 13, we implement our algorithm as a web application named AutoCompliance, which has been made publicly accessible [12]. The audiences of our tool include normal App users, App developers as well as auditing authorities. For normal

[11] https://discord.com/privacy

[12] www.ppvisual.site

Table 5: The Survey Questions and Results

No.	Question	Exp. Group	Cont. Group
Q_1	How concerned are you about privacy information? (0 for not concerned, 5 for very concerned)	3.3	3.5
Q_2	Have you been troubled by the privacy related issues when using applications? (Y/N)	Y	Y
Q_3	Do you read privacy policies when encountered? (Y/N)	N	N
Q_4	Rate the difficulty of the given task. (0 for very easy, 5 for very difficult)	2.17	2.5
Q_5	Does the tool help complete the task faster? (0 for not helpful, 5 for very helpful)	4.33	-
Q_6	How does the tool help in completing the task?	-	-
Q_7	Put down your recommendations for the tool.	-	-
Q_8	What difficulties do you encounter when reading the privacy policy?	-	-
Q_9	What suggestions do you have to assist complete the task faster?	-	-

Table 6: Time Spent for the Tasks

Task		Average Time (s)	
Privacy Policy	Label	Control Group	Experiment Group
Ted	DRP	648	256
Ted	ROP	348	186
Opera	RRP	754	282
Opera	RLC	319	207
average		517.25	232.75

App users and developers, our tool serves as an alerting purpose and assists identifying compliance issues in privacy policies with GDPR. For auditing authorities, our tool can be a probing tool that automatically crawls and detects potential violations.

We then design a user study and employ 10 volunteers to use AutoCompliance, and then interview them to understand their experience.

7.1 Implementation

We use the JavaScript library D3 [1] to implement the visualization functionality. D3 is known to have better performance and platform portability as compared with other visualization libraries. Moreover, it provides better interactive experience. The screenshot of AutoComplianceis shown in Figure 4.

There are in total 5 parts in our tool, as dash-boxed and highlighted with index numbers in Figure 4. The first part is composed of a search box and the label buttons. Given an APP link, our tool can automatically crawl the privacy policy, conduct compliance analysis and display the analysis results accordingly in the rest parts of the interface. The buttons below the search box represent the labels defined in our task. Clicking on each label, the sentences belong to the category will be highlighted in the corresponding color. This functionality helps users find the sentences of the targeted labels. The second part shows the compliance issues detected in APPs of the same category. For instance, 37.5% of the privacy policies in APPs of the education category, which the TED APP belongs to, have the compliance issue of missing descriptions on data retention period, as is illustrated by the first bar in the histogram. This function provides statistics of compliance issues on privacy policies of APPs in the same category, and helps users evaluate the quality of the privacy policy of interest. The third part is a word cloud of the current privacy policy, which provides an intuitive view of keywords involved. We adopt the most common TF-IDF [25] algorithm for this purpose. The forth part shown in the middle of the screen is the text of the privacy policy. AutoCompliance highlights the sentences of each category in the corresponding color. We also add the floating window which shows the label of the current sentence. All detected compliance issues are listed in the right-most part. We also provide example descriptions adopted from other similar privacy policies in case of missing related descriptions. For example, the example TED APP privacy policy[13] misses descriptions on *Right to Restrict of Processing*, AutoCompliance highlights this compliance issue in purple, provides a description from another similar APP as an example and also lists the GDPR clause that is violated.

7.2 User Study

Since our approach can benefit both the individual users and companies hosting web and mobile services, the volunteers are purposely hired based on the application scenario of our approach. In particular, we hire 2 managers of startup companies, which have their own websites and thus privacy policies, as representatives of industry volunteers. We also hire 8 graduate students, who frequently browse web and mobile applications, as individual user volunteers. All recruited volunteers are comfortable with reading in English, and are familiar with the applications from which the privacy policy is adopted for user study. We divide the volunteers into a control group and an experiment group following the standard process. Both groups are provided with the same tasks, i.e., to find the required contents in the privacy policy. For fair comparisons, we conduct preliminary English reading test, with a sentence labeling task, to the volunteers, and assign each volunteer to a corresponding group based on their English sentence labeling accuracy. In this way, we ensure that volunteers in the control and experiment groups are of similar English reading capability.

We design 4 tasks related to 2 privacy policies. For each privacy policy, we design 2 individual tasks, each of which corresponds to one clause in GDPR. The applications we selected are the TED talk APP and the Opera browser APP[14], which are among the popular applications in the APP store. For the TED talk APP, the two tasks are finding statements on the *Data Processing Purposes (DPP)* and *Right to Object to Processing (ROP)*. For the Opera browser APP, the

[13] https://www.ted.com/about/our-organization/our-policies-terms/privacy-policy
[14] https://www.opera.com/privacy

two tasks are finding statements on the *Right to Restrict of Processing (RRP)* and *Right to Lodge a Complaint (RLC)*.

Each volunteer is given one task from the TED talk privacy policy and one from the Opera browser privacy policy. For each task, the volunteers are asked to read through the privacy policy, then either find the sentences describing the contents in the task, or report no such sentences in the privacy policy, which indicates a violation of the GDPR clauses. The experiment group reads the privacy policy with the assistance of AutoCompliance, and the control group read the privacy policy without any assistance. We record the time used for each volunteer on each task and their answer. After each volunteer finishes the task, we interview them individually to understand the experience of completing the task. In particular, they are asked the several questions shown in Table 5. The first 4 questions are common for both the control group and the experiment group, questions 5-7 are specific for the experiment group, and the last two questions are specific for the control group.

The time spent on each task is shown in Table 6. We can observe that volunteers in the experiment group spend less time on all four tasks than volunteers in the control group. AutoCompliance helps the experiment group achieves an average of 55% time reduction in completing the tasks.

The results of the user interview are shown in the last two columns of Table 5. We can observe from the results that both the control and the experiment group show relatively high level of concern about privacy information. All of them state that they have been troubled by privacy related issues, yet none of them reads privacy policies when encountered. The control group rates the tasks as more difficult.

Through the interview, we find that all volunteers report they have encountered privacy information related issues or concerns when using the applications. Two volunteers expressed their concern of illegal personal data collection or the unawareness of personal data collection. Three volunteers reported that they are worried about personal data being sold to some third party companies/agencies for commercial usage. A volunteer also states his concern of using APPs developed by small companies, due to the reason that those APPs usually have very short and unclear privacy polices, which escape critical statements about collecting, sharing and processing personal information. Half of the volunteers reported they have received jam messages/calls due to the reason that their mobile numbers were collected/sold to various platforms without their consent.

Questions Q_8 and Q_9 are presented specifically to the control group. For the question of difficulties encountered when reading the privacy policy, seven volunteers reported that they usually did not read the privacy policies, especially when they were in a rush installing the applications. There are specific terms in the privacy policy, which prevent deep reading and understanding. Moreover, vague descriptions on processing of personal information further increases the difficulty of reading. As for suggestions to complete the task faster (Q_9), two volunteers suggested that they can make use of the subtitle information to skip unrelated contents. They also suggested that highlighting semantic related contents could be very useful.

Questions Q_5 to Q_7 are shown specifically to the experiment group. We can observe that volunteers in the experiment group provide an average score of 4.33 for this question, which indicates that they find AutoCompliance to be helpful in assisting completing the task faster. For question Q_6, all volunteers in the experiment group reported that the labels provided by the tool assist reading and understanding of the contents, and reduce the time spent to finish the task. As for suggestions on the tool (Q_7), most of them suggested a better coloring template on different labels. There are also suggestions on incorporating labeling functionality to enable crowd source labeling.

In addition to the individual APP users, we also recruit two managers from two IT startup companies, which maintain their own website in the company. During the interview, both of them agree that the task is practical for their company, and that AutoCompliance is able to help them identify missing contents, and improve the clarity of their privacy policy. Especially the recommended descriptions on the missing labels, which are of practical usages to their companies.

8 RELATED WORK

8.1 Privacy Policy Corpus Creation and Analysis

Wilson et al. [34] create a website privacy policy corpus named OPP-115, which contains 23K fine-grained data practices, based on crowd-sourcing. Sathyendra et al. [29] extend the OPP-115 corpus to label fine-grained information of opt-out choices. They focus on the task of automatically identifying user choices in privacy policy text. Zimmeck et al. [37] build a corpus which contains 350 privacy policies of mobile Apps. They provide a scalable pipeline to analyze potential compliance issues of APP executable with privacy policy.

Kaur et al. [16] study frequent/ambiguous keywords in privacy policies, as well as the impact laws/regulations have on privacy policies. They draw conclusions statistically from analyzing a large number of privacy policies, and recommend to use the analysis results to make a good template of privacy policies. Lebanoff et al. [18] propose an approach to automatically detect vague words and sentences in privacy policies. They created a vague words corpus through crowd sourcing, an AC-GAN method is then proposed to predict vague words and vague sentences. Tesfay et al. [32] take one step forward to create a corpus including 45 manually labeled privacy policies. The corpus concentrates on the risk levels of the privacy policies, which are defined by experts.

Harkous et al. [15] design an approach for automatically annotating previously unseen privacy policies, then they provide three web services to make privacy policies visible. Liu et al. [20] annotate pairwise privacy policy paragraphs for the privacy alignment task by manual annotation assisted with clustering methods. Sarne et al. [27] adopt unsupervised learning techniques to extract topics from a large number of privacy policies.

Wilson et al. [5] use automatic text classification methods to answer simple classification questions about privacy policies, aiming at making it easier for people to understand the privacy policy. Sathyendra et al. [28] propose to find out the sentences about Opt-Out choices in the privacy policy. The OPP-115 dataset is used for the model training purposes. Kumar et al. [6] develop a tool that uses text classification methods to automatically find various opt-out links from the privacy policy, reducing people's reading and searching time. Story et al. [30] conduct a preliminary study with 1 million APPs and find that around half of the studied APPs do not have a privacy policy link. They then propose a task of predicting the possibility of APP having a privacy policy link based on manually designed features.

None of the existing work focuses on the automated compliance analysis task between privacy policy and regulations, and there is no corpus created for this purpose. We propose a new task, i.e., compliance analysis of privacy policies with GDPR Article 13. We devise a classification scheme based on GDPR and manually curate a corpus of 304 privacy policies for this purpose.

8.2 GDPR Related Analysis

Degeling et al. [9] conduct a survey and analysis on the compliance of cookie consent implementations with GDPR. Their findings show that there are lack of functionalities as well as mechanisms to allow users actively consent or deny consent, as required by GDPR. Linden et al. [19] conduct a study on the privacy policy changes after GDPR was put into practice. They analyze 6, 278 unique English privacy policies by comparing pre-GDPR and post-GDPR versions. They found quality improvement in the privacy policies in the EU area. Gerl et al. [11] analysis the Art.12-Art.14 in GDPR and propose a systematic approach to create and present privacy policies in a unified way utilizing the Layered Privacy Language. There are also some approaches which create ontology for privacy policies [23]. The proposed ontology is consistent with our label scheme.

Nejad et al. [22] investigate semantic text matching techniques that map privacy policy segments with relevant GDPR articles. KnIGHT checks the consistencies between GDPR and privacy policies, whereas our approach checks the inconsistencies. The difficulty in checking inconsistencies is that, if data is missing, it is infeasible to use the text matching based approaches proposed by Nejad. Our approach conducts a rule-based checking to solve the problem. Reyes et al. [26] conduct compliance checking on 5, 855 most popular free children Apps with Children's Online Privacy Protection Act (COPPA)[15]. Their results show that majority of the analyzed APPs violate COPPA, mainly because of the usage of third party SDKs. Chang et al. [7] adopt the OPP-115 corpus to train models which automatically predict APP's privacy policies with personalized privacy concerns provided by users. Gruschka et al. [13] discuss the status of legislation and regulations. They also analyze the different data protection and privacy protection technologies in the context of big data analysis, and discuss the type of information which may become privacy risks.

There are also approaches targeting different aspects of privacy issues with APPs, for instance, Yu et al. [36] propose to automatically generate natural language privacy policy descriptions from system behaviors extracted from the source code. Our work focuses on the task of compliance checking of GDPR (Article 13) with privacy policy documents. Our purpose is orthogonal to the previous research works. The evaluation results show the effectiveness of our approach in identifying compliance issues (with GDPR Article 13) in privacy policies.

9 CONCLUSION

In this work, we propose a new task of compliance analysis between GDPR (Article 13) and privacy policies. We design a label scheme based on Article 13 of GDPR and manually create a corpus of 304 privacy policies. We benchmark our corpus with standard sentence classifiers and then conduct rule based compliance analysis based on the classification results. Our approach successfully detects 1, 180 compliance issues in 304 privacy policy documents. We implement our approach into a web-based tool named AutoCompliance, and conduct a user study with 10 volunteers. The results confirm the usability of AutoCompliance, which successfully reduces the user reading time by 55%.

The work can be further improved in the following aspects. The corpus suffers from the imbalanced data problem, which greatly affects the classification accuracy. More efforts could be put in this direction to improve the compliance analysis results. More specific text visualization model could be designed to assist user quickly comprehend the presented information.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China 61802275, U1836214, 61902395 and Innovative Fund of Tianjin University 2020XRG-0022.

REFERENCES

[1] 2011. *Data-Driven Documents.* https://d3js.org/ (access on 2020.10.17).

[2] 2016. *General Data Protection Regulation.* https://gdpr-info.eu/ (access on 2020.10.19).

[3] 2018. *California Consumer Privacy Act in America.* https://oag.ca.gov/privacy/ccpa (access on 2020.10.19).

[4] 2018. *Data Protection Act.* https://www.gov.uk/data-protection (access on 2020.10.19).

[5] Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. 2012. Automatic categorization of privacy policies: A pilot study. *School of Computer Science, Language Technology Institute, Technical Report CMU-LTI-12-019* (2012).

[6] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. 2020. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *Proceedings of The Web Conference 2020.* 1943–1954.

[7] Cheng Chang, Huaxin Li, Yichi Zhang, Suguo Du, Hui Cao, and Haojin Zhu. 2019. Automated and Personalized Privacy Policy Extraction Under GDPR Consideration. In *WASA 2019.* Springer, 43–54.

[8] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.

[9] 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. *arXiv preprint arXiv:1808.05096* (2018).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Armin Gerl and Bianca Meier. 2019. The Layered Privacy Language Art. 12–14 GDPR Extension–Privacy Enhancing User Interfaces. *Datenschutz und Datensicherheit-DuD* 43, 12 (2019), 747–752.

[12] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE workshop on automatic speech recognition and understanding.* IEEE, 273–278.

[13] Nils Gruschka, Vasileios Mavroeidis, Kamer Vishi, and Meiko Jensen. 2018. Privacy issues and data protection in big data: a case study analysis under GDPR. In *2018 IEEE International Conference on Big Data (Big Data).* IEEE, 5027–5033.

[15] https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule

[14] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and Understanding the Effectiveness of BERT. In *EMNLP/IJCNLP*.

[15] 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 *USENIX Security 18*. 531–548.

[16] Jasmin Kaur, Rozita A Dara, Charlie Obimbo, Fei Song, and Karen Menard. 2018. A comprehensive keyword analysis of online privacy policies. *Information Security Journal* 27, 5-6 (2018), 260–275.

[17] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[18] Logan Lebanoff and Fei Liu. 2018. Automatic detection of vague words and sentences in privacy policies. In *EMNLP 2018*.

[19] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2018. The privacy policy landscape after the GDPR. *arXiv preprint arXiv:1809.08396* (2018).

[20] Fei Liu, Rohan Ramanath, Norman Sadeh, and Noah A Smith. 2014. A step towards usable privacy policy: Automatic alignment of privacy statements. In *COLING 2014*. 884–894.

[21] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *ISJLP* 4 (2008), 543.

[22] Najmeh Mousavi Nejad, Simon Scerri, and Jens Lehmann. 2018. Knight: Mapping privacy policies to gdpr. In *European Knowledge Acquisition Workshop*. Springer, 258–272.

[23] Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, and Livio Robaldo. 2018. PrOnto: Privacy ontology for legal reasoning. In *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 139–152.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP 2014*. 1532–1543.

[25] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. 133–142.

[26] Irwin Reyes, Primal Wijesekera, Joel Reardon, Amit Elazari Bar On, Abbas Razaghpanah, Narseo Vallina-Rodriguez, and Serge Egelman. 2018. "Won't Somebody Think of the Children?" Examining COPPA Compliance at Scale. *Proceedings on Privacy Enhancing Technologies* 2018, 3 (2018), 63–83.

[27] David Sarne, Jonathan Schler, Alon Singer, Ayelet Sela, and Ittai Bar Siman Tov. 2019. Unsupervised Topic Extraction from Privacy Policies. In *WWW 2019*. ACM, 563–568.

[28] Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman M Sadeh. 2016. Automatic Extraction of Opt-Out Choices from Privacy Policies.. In *AAAI Fall Symposia*.

[29] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *EMNLP 2017*. 2774–2779.

[30] Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Which apps have privacy policies?. In *Annual Privacy Forum*. Springer, 3–23.

[31] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune BERT for text classification?. In *China National Conference on Chinese Computational Linguistics*. Springer, 194–206.

[32] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR. In *Companion Proceedings of the The Web Conference*. 163–166.

[33] Sida 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: Short papers-volume 2*. Association for Computational Linguistics, 90–94.

[34] 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, et al. 2016. The creation and analysis of a website privacy policy corpus. In *ACL 2016 (Volume 1: Long Papers)*. 1330–1340.

[35] Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2017. YEDDA: A lightweight collaborative text span annotation tool. *arXiv preprint arXiv:1711.03759* (2017).

[36] Le Yu, Tao Zhang, Xiapu Luo, Lei Xue, and Henry Chang. 2016. Toward automatically generating privacy policy for android apps. *IEEE Transactions on Information Forensics and Security* 12, 4 (2016), 865–880.

[37] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling privacy compliance analysis to a million apps. *PoPETs* 2019, 3 (2019), 66–86.