Automated Compliance Checking for Chinese Privacy Policy: A New Task and Dataset

Anonymous ACL submission

Abstract

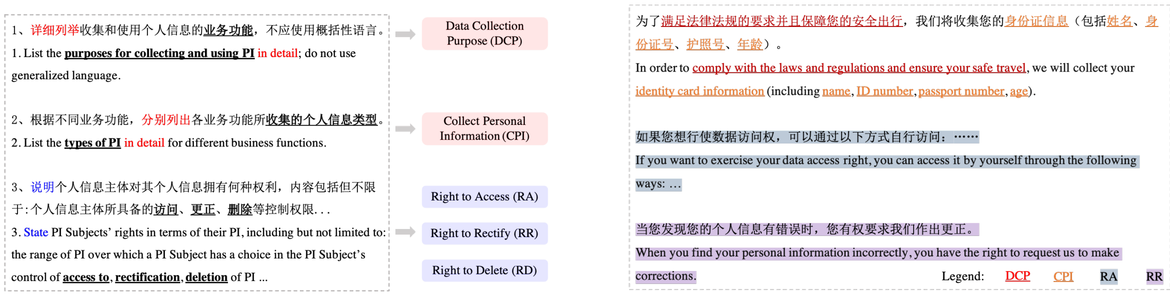
Privacy policy texts inform users about how their personal data is handled by online service providers. However, they may be long, complex, and non-compliant with laws and regulations. Therefore, automated compliance checking of privacy policy texts is needed. In this paper, we introduce the first dataset and task for automated compliance checking of Chinese privacy policy texts. Our dataset provides human experts' compliance annotation at both the document level and the fine-grained level. The fine-grained annotation includes both the existing named entity recognition (NER) task and 11 new sentence classification (SC) tasks for compliance checking. We treat the NER and classification subtasks as discriminative legal attributes that can help models to generate reliable compliance results and easy-to-understand explanations. Additionally, we further pretrain BERT-Chinese on a large corpus of compliance-related texts and evaluate it on all the tasks. Our results show that our further pre-trained BERT model outperforms the baseline models and demonstrates the potential of NLP techniques for automated compliance checking of privacy policies. Our dataset and the further pre-trained BERT model will be released soon.

1 Introduction

Web and mobile applications (apps) have become ubiquitous in recent years, enabling various services and functionalities for users. According to Statista (Statista, 2023), there were 254.94 billion app downloads worldwide in 2022, and China accounted for over 111.11 billion of them. However, these apps also collect a large amount of personal data from users, which poses privacy risks and challenges. To inform users about how their personal data are handled, software applications or websites provide privacy policies that describe their data collection, usage, and protection practices. On the other hand, regulators around the world have enacted laws and policies to govern the service providers and protect the user privacy, such as "General Data Protection Regulation"(GDPR)(GDPR, 2022) in the European Union and "Personal Information Protection Law"(PIPL)(PIPL, 2022) in China. However, both privacy policies and related regulations are often written in professional natural languages with many legal terms and software jargon that make them difficult to understand and even read for users. Therefore, it is desirable to use natural language processing (NLP) techniques to analyze privacy policies and help users understand them, which are essential for protecting user privacy and ensuring compliance with relevant laws and regulations. Furthermore, NLP techniques can also help legal professionals and clients verify the validity and legality of privacy policies and identify potential risks or violations.

However, existing research on NLP for privacy policy analysis is limited and mainly focuses on English privacy policies, which limits the applicability of these methods in regions with other languages. To the best of our knowledge, there is no previous work on NLP for Chinese privacy policy compliance checking. Moreover, existing open-source datasets for Chinese privacy policy only provide annotations for some aspects of privacy policy texts, such as named entities or key terms, but do not address the compliance issue at the document level. Conducting basic named entity recognition (NER) or sentence classification (SC) from several aspects is not sufficient to capture the compliance status of a privacy policy. Therefore, there is a lack of data and methods for automated compliance checking of Chinese privacy policy, which is a novel and urgent research problem, given the large number of app downloads and privacy-related regulations enacted in China.

This paper presents the first dataset and task for automated **Compliance Checking of Chinese Pri-**



(a) (b)

Figure 1: 1(a) Label Schema Construction. This figure illustrates how we construct the labels for PISS from the original text. The labels are divided into two categories: NER and SC. The NER labels have a red background and the SC labels have a blue background. The type of PISS expression determines the corresponding label category. 1(b) Annotation Examples. This figure shows a part of a Chinese privacy policy document, annotated with our label schema for both NER and SC subtasks. The NER entities are highlighted with underlines and the SC sentences are highlighted with background colors.

vacy Policy (C3P2), a novel document-level NLP task aimed at assessing whether a privacy policy text conforms to the compliance requirements and standards derived from relevant laws and regulations. Unlike existing tasks in the legal domain, which often involve complex reasoning or argumentation, C3P2 requires a straightforward yet challenging evaluation of the privacy policy text against a set of compliance points derived from relevant laws and regulations. The input for our task is a privacy policy text, and the output is a compliance result (yes or no) accompanied by a brief explanation. The compliance result indicates whether the privacy policy text satisfies all the compliance points, while the explanation provides evidence and justification for the compliance result. Based on a previous Chinese privacy policy dataset (Zhao et al., 2022), we construct the first automatic **compliance checking** dataset for **Chinese privacy policy**, named C3P2-483. Our dataset provides human experts' compliance annotations at both the document level and the fine-grained level. We annotate privacy policies from 14 aspects according to related laws and regulations, covering many more dimensions than previous work. To support our main task, C3P2, we introduce two subtasks: **Named Entity Recognition (NER)** and **Sentence Classification (SC)**. These subtasks aim to extract discriminative legal attributes from the privacy policy, enabling models to generate reliable compliance results and easy-to-understand explanations.

Moreover, we propose to further pre-train **BERT-Chinese**, a Chinese version of BERT pre-trained on general-domain corpora, on a large corpus of compliance-related texts. We hypothesize that this further pretraining can enhance BERT-Chinese's performance on our task by enabling it to learn the domain-specific vocabulary, concepts, and logic that are relevant for compliance checking. We also hypothesize that this further pretraining can help BERT-Chinese to adapt to the style and structure of privacy policy texts, which differ from general-domain texts. By further pretraining BERT-Chinese on compliance domain content, we aim to obtain a more robust and effective language model for our task and dataset. We then evaluate several baseline models and our further pre-trained BERT model, named **ComplianceBERT**, on the NER subtask, the SC subtask, and the document-level compliance task C3P2. Our results show that our **ComplianceBERT** model outperforms the baseline models on all the tasks.

We summarize our contributions as follows:

- We present the first dataset for automated compliance checking of Chinese privacy policy texts, based on a previous dataset (Zhao et al., 2022). Our dataset, named C3P2-483, provides human experts' compliance annotations at both the document level and the fine-grained level. The fine-grained annotation includes both the existing NER (Zhao et al., 2022) and 11 new SC subtasks for compliance checking.

- We treat the NER and SC subtasks as discriminative legal attributes that can help models generate reliable compliance results and easy-to-understand explanations. We consider many more aspects according to related laws and regulations than previous work, which either focused on coarse-grained levels (sentence or paragraph) or fine-grained levels (entity) only.

- We further pretrain BERT-Chinese on a large corpus of compliance-related texts. We evaluate several baseline models and our further pre-trained BERT model, named **Compliance-BERT**, on the NER subtask, the SC task, and the document-level compliance task. Our results show that our further pre-trained BERT model outperforms the baseline models on all tasks, demonstrating the feasibility and potential of applying NLP techniques to the automated compliance checking of privacy policies. Our dataset and further pre-trained BERT model will be released soon.

2 Related Work

Most previous work on compliance checking of privacy policies focuses on English policies and the EU General Data Protection Regulation (GDPR) (GDPR, 2022). For example, Liu et al. (Liu and Meng, 2021) annotate policy statements according to eleven items such as *Collection of Personal Info*, *Data Retention Period*, and *Data Processing Purposes*. They train several sentence classifiers, such as Support Vector Machine (SVM), Bidirectional Long Short Term Memory (BiLSTM) (Huang et al., 2015), and Bidirectional Transformer (BERT) (Devlin et al., 2018), and then employ a rule-based compliance analysis according to GDPR Article 13. Zimmeck and Bellovin (Zimmeck and Bellovin, 2014) propose an architecture for automatic privacy policy analysis powered by a rule classifier and a machine learning (ML) preprocessor. Zaeem et al. (Zaeem et al., 2018) present a free Chrome extension, PrivacyCheck, which automatically summarizes privacy policies and displays risk levels. They train ten classifiers, each answering a specific question about the privacy policy. Costante et al. (Costante et al., 2012) propose a solution to automatically assess the completeness of a policy using NLP and ML techniques, identifying six core elements such as *Choice and Access*, *Data Collection*, and *Data Sharing*. (Tesfay et al., 2018) tags policies on 10 compliance aspects derived from extensive GDPR analysis. Similarly, Sánchez et al. (Sánchez et al., 2021) annotate privacy policies according to seven elements for GDPR data protection goals and qualify the degree of compliance.

Zhao et al. (Zhao et al., 2022) annotate a NER dataset for Chinese privacy policy texts, covering data controller, data entity, collecting action, sharing action, condition, purpose, and data receiver. However, NER models only help users understand the policy content without evaluating the compliance level. Therefore, Zhao et al. suggest that detecting privacy compliance violations is an urgent and necessary future direction.

3 Dataset Construction

3.1 Label Schema

The Personal Information Protection Law (PIPL) was enacted in November 2021 as the general principle for personal information protection in China. While it covers various situations regarding privacy information usage, it may be too broad for specific privacy policy checking. To provide detailed and clear guidance on PIPL compliance, the National Information Security Standardization Technical Committee released the *Personal Information Security Standards* (PISS) (PISS, 2020). We consulted experts with substantial legal professional experience and manually extracted 14 labels representing the contents that should be included in a privacy policy. To the best of our knowledge, this is the most comprehensive privacy policy-checking framework with the most compliance labels. The extracted labels are as follows (see details in Appendix A:

- Collect Personal Information (CPI) `[PISS Art 5]`
- Policy Duration (PD) `[PISS Art 5]`
- Data Retention Period (DRP) `[PISS Art 6]`
- Data Retention Region (DRR) `[PISS Art 6]`
- Overdue Processing Method (OPM) `[PISS Art 6]`
- Data Collection Purpose (DCP) `[PISS Art 7]`
- User Portrait (UP) `[PISS Art 7]`
- Right to Access (RA) `[PISS Art 8]`
- Right to Rectify (RR) `[PISS Art 8]`
- Right to Delete (RD) `[PISS Art 8]`

Label	Frequency	Coverage	Avg.L	Fleiss' Kappa
Collect Personal Information (CPI)	8177	1.00	4.85	0.48
Policy Duration (PD)	586	0.49	25.27	0.58
Data Retention Period (DRP)	408	0.63	52.44	0.65
Data Retention Region (DRR)	360	0.71	42.81	0.60
Overdue Processing Method (OPM)	663	0.62	58.06	0.67
Data Collection Purpose (DCP)	7074	0.99	10.17	0.42
User Portrait (UP)	898	0.64	66.30	0.54
Right to Access (RA)	1199	0.76	46.13	0.59
Right to Rectify (RR)	1342	0.84	48.77	0.66
Right to Delete (RD)	1714	0.84	48.01	0.65
Right to Withdraw (RW)	1052	0.72	50.75	0.61
Right to Account Cancellation (RAC)	1484	0.71	45.94	0.64
Personal Information Sharing (PIS)	2190	0.93	4.45	0.46
Personal Information Protection (PIP)	3763	0.97	58.77	0.53
Avg	2208	0.78	40.19	0.58

Table 1: The details of the annotated corpus. The `Frequency` column indicates the total number of times each corresponding label appears in our corpus. `Coverage` shows the percentage of privacy policy documents that contain the corresponding label. The column `Avg.L` represents the average number of characters per annotation in our dataset. For fine-grained annotations, it is the average length of annotated entities, while for coarse-grained annotations, it is the average length of labeled sentences. The last column shows the `Fleiss' Kappa` of the annotation results (before merging).

- Right to Withdraw (RW) `[PISS Art 8]`
- Right to Account Cancellation (RAC) `[PISS Art 8]`
- Personal Information Sharing (PIS) `[PISS Art 9]`
- Personal Information Protection (PIP) `[PISS Art 11]`

3.2 Data Annotation and Statistics

We adopted two types of annotations for our compliance checking task, based on the requirements of PISS. For Collect Personal Information (CPI), Data Collection Purpose (DCP), and Personal Information Sharing (PIS), we used fine-grained annotations similar to NER annotation. For the remaining tasks, we used coarse-grained annotations similar to sentence classification annotation. We hired eight native participants, who were undergraduate and postgraduate students, to annotate the privacy policies. We compiled some common descriptions for each compliance label from 60 privacy policies with the help of compliance experts. We provided the participants with a description of the task, detailed instructions, and explanations with some common descriptions for each compliance label. We used a web-based annotation tool that allowed the participants to highlight the texts and select the labels from a drop-down menu. We also provided a feedback mechanism for the participants to report any difficulties or ambiguities they encountered during the annotation process.

For fine-grained annotations, we used the texts and annotations in `CA4P-483` dataset (Zhao et al., 2022) as a reference and reannotated CPI, DCP, and PIS labels to match our label schema. The requirements for annotating a pure NER task and a compliance checking task are not the same. For example, in `CA4P-483`, the label "Sharing Action" annotates any descriptions about sharing action corresponding to PIS. However, in our task, we also need to annotate any descriptions about not sharing personal information as PIS, since PIS requires to describe whether and how the personal information is shared. Therefore, we reannotated these compliance labels based on the previous annotation in `CA4P-483`. Another reason for reannotating the dataset is that the previous dataset is not fully annotated. They filtered possible sentences based on keywords and annotated them by humans, which may cause missing annotations. For example, some sentences that do not contain keywords

Dataset	# All	# Train	# Dev	# Test	Language	# Labels	Task Type
OPP-115	3792	2473	-	1319	English	12	NER
APP-350	7700	4136	1364	2200	English	18	SC
CA4P-483	18579	14678	2059	1842	Chinese	7	NER
Ours	91182	75312	8539	7331	Chinese	14	NER, SC

Table 2: Comparison with Other Privacy Policy Datasets

such as "collect", "use", or "share" may still contain relevant information for compliance checking. Therefore, in this work, we reannotated the dataset thoroughly without filtering any sentences. For coarse-grained annotations, we asked the participants to read each sentence in the privacy policy and assign one or more compliance labels to it, based on the definitions and examples of the labels.

Table 1 shows the details of the annotated corpus. We compare our dataset with other privacy policy datasets that are not necessarily for compliance checking, namely Chinese Android application privacy policy (CA4P-483) (Zhao et al., 2022), Online Privacy Policies (OPP-115) (Wilson et al., 2016), and Android app privacy policies (APP-350) (Zimmeck et al., 2019).

4 Task and Experiment Setup

4.1 Compliance Checking

Compliance checking is a task that verifies whether a privacy policy conforms to certain standards or regulations. It is a specific task in NLP that differs from more general tasks such as Named Entity Recognition (NER) or classification, which do not depend on specific regulations. However, to train and evaluate our models, we need privacy policies with their corresponding compliance judgments from regulators, which are difficult to obtain. One possible solution is to use a human-annotated dataset, where experts mark the privacy policies with compliance information, and train an end-to-end model based on this dataset.

However, this approach faces a significant limitation: most models cannot process the privacy policies in an end-to-end manner due to their length. This means that the models cannot take the entire policy as input and produce the compliance result as output directly. Therefore, we propose a more practical two-step approach:

First, we annotate the sentences or entities within the privacy policy with the labels introduced in Section 3. These labels represent discriminative legal attributes, such as data collection, data usage, data retention, etc. These attributes capture the essential information that influences the compliance status of the policy.

Second, we derive compliance rules based on the presence or absence of the corresponding attributes. For example, a rule that requires policy duration may be violated if the policy does not specify any attribute for this information. This way, our models can generate reliable compliance results and clear explanations, as we can use the attributes to justify why the policy is compliant or not.

4.2 Subtask Description

We label the sentences or entities in the privacy policy using Named Entity Recognition (NER) and Sentence Classification (SC) tasks. NER labels the types and categories of personal information, and the purposes and subjects of data collection and sharing. This is crucial because regulations require privacy policies to explicitly and individually state this information. For example, PISS (PISS, 2020) mandates that service providers inform data subjects of the specific types of personal information they collect and share, and obtain authorization for certain uses and disclosures. SC labels the sentences that describe the data collection and usage terms, as well as the rights and obligations of the data subjects and the service provider. By combining both tasks, we can better understand and explain the compliance status of the privacy policy text. We use both fine- and coarse-grained annotations as explained in Section 3. NER can also be used for further research, such as verifying if the services' actions match their privacy policies, and ensuring that the service provider collects and uses personal information only for the agreed purposes and minimally.

For the NER task, given a sentence $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ as a sequence of N tokens, the model aims to predict a label sequence $\mathbf{S} = (s_1, s_2, \ldots, s_N)$, where each label is a position in-

Model	Metrics	B-PIS	I-PIS	B-DCP	I-DCP	B-CPI	I-CPI	O	Avg
BiLSTM-CRF	P	**64.55%**	**72.44%**	40.75%	70.86%	63.35%	71.95%	95.58%	68.50%
	R	31.70%	38.32%	**18.60%**	36.60%	46.10%	60.86%	98.53%	47.24%
	F1	42.51%	50.12%	25.55%	48.27%	53.36%	65.94%	97.03%	54.68%
BERT	P	44.08%	61.95%	50.00%	**75.55%**	58.76%	68.19%	96.95%	55.27%
	R	**66.52%**	**76.54%**	18.24%	52.80%	60.78%	77.29%	97.92%	68.22%
	F1	53.02%	68.48%	**26.73%**	62.16%	59.75%	72.46%	**97.43%**	60.25%
ComplianceBERT	P	54.39%	67.67%	50.00%	74.03%	57.46%	65.64%	**97.19%**	55.20%
	R	55.36%	73.55%	17.99%	55.97%	**64.78%**	**82.27%**	97.64%	**70.43%**
	F1	**54.87%**	**70.49%**	26.46%	63.74%	**60.90%**	**73.02%**	97.42%	**61.37%**
BERT Multitask	P	54.94%	65.77%	**56.14%**	71.64%	**69.05%**	**85.32%**	96.17%	**71.29%**
	R	39.73%	41.12%	15.67%	57.77%	35.87%	51.68%	**98.63%**	48.64%
	F1	46.11%	50.60%	24.50%	**63.96%**	47.22%	64.37%	97.39%	56.31%
ComplianceBert Multitask	P	47.69%	64.93%	48.80%	68.13%	61.34%	79.34%	96.71%	66.70%
	R	45.98%	39.63%	17.38%	**62.36%**	46.00%	61.81%	98.11%	53.04%
	F1	46.82%	49.22%	25.63%	65.11%	52.58%	69.48%	97.41%	58.04%

Table 3: Precision/Recall/F1-score for NER Models

dicator (e.g., BIO schema).

For the SC task, given a sentence $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, the model aims to predict a set of labels $\mathbf{y} = (y_1, y_2, \ldots, y_k)$ that represents whether the sentence describes information about each of the k compliance labels (e.g., yes or no). A sentence can have multiple labels if it describes information about more than one compliance label.

4.3 Compliance Rules

In the second step of our approach, we apply compliance rules to privacy policies based on the presence or absence of the corresponding labels. Some labels imply conditional requirements, such as **Data Retention Period** (DRP), which is only required when the service provider **Collects Personal Information** (CPI). If they do not collect personal information, it is irrelevant to discuss the data retention period. Other labels imply unconditional requirements, such as **Policy Duration** (PD), which is always required regardless of whether the service provider collects personal information or not. We use these rules to check whether the policy is compliant and to provide explanations for the compliance result. The details of the rules will be presented in the A.2.

4.4 Model Summerization

4.4.1 Further Pretrain BERT

BERT (Devlin et al., 2019) is a pre-trained language model that can be fine-tuned for various natural language processing tasks. However, BERT is pre-trained on general-domain corpora, such as Wikipedia and BooksCorpus, which may not capture the specific vocabulary and semantics of a particular domain. Therefore, we propose **ComplianceBERT**, a further pre-trained BERT model on domain-specific corpora of privacy policy texts. To obtain such corpora, we collected 3.2 million texts from various sources, including legal websites, government websites, and online forums, containing information about personal information protection laws, regulations, and privacy policies. Following the approach of Liu et al. (Liu et al., 2019), we use only the masked language modeling (MLM) objective for further pretraining.

4.4.2 NER Model

For the Named Entity Recognition (NER) task, we compare three different models: **BiLSTM-CRF**, **BERT**, and **ComplianceBERT**. BiLSTM-CRF (Zhao et al., 2022) consists of a bidirectional LSTM (BiLSTM) encoder and a conditional random field (CRF) decoder. BERT and ComplianceBERT are both transformer-based models that use a linear layer and a softmax layer as decoders.

4.4.3 SC Model

Since all labels for the Sentence Classification (SC) task pertain to privacy policy compliance, we believe there are correlations among these labels that could enhance prediction performance. We adopt **CorNet** (Xun et al., 2020) for the output layer in our models. The CorNet layer consists of two sublayers: a correlation matrix layer and a correlation

Models	Metrics	PD	DRP	DRR	OPM	UP	RA	RR	RD	RW	RAC	PIP	Avg
BiLSTM-CRF	P	**96.15%**	89.66%	**100.00%**	57.14%	86.54%	**86.08%**	75.29%	57.34%	88.14%	**96.24%**	82.08%	83.15%
	R	73.52%	86.67%	60.00%	44.44%	66.17%	73.11%	57.14%	65.60%	78.79%	96.24%	58.39%	69.10%
	F1	83.33%	88.14%	75.00%	50.00%	75.00%	79.07%	64.97%	61.19%	83.20%	96.24%	68.23%	74.94%
BERT	P	88.57%	96.77%	96.15%	91.42%	**96.82%**	81.73%	80.87%	91.87%	95.59%	86.75%	89.35%	**90.54%**
	R	91.18%	100.00%	100.00%	88.89%	89.70%	91.40%	83.04%	90.40%	98.48%	98.50%	92.95%	93.14%
	F1	**89.85%**	98.36%	98.03%	90.14%	**93.13%**	86.29%	81.94%	91.13%	97.01%	92.25%	91.14%	91.75%
ComplianceBERT	P	82.50%	**100.00%**	96.15%	90.00%	90.00%	83.17%	81.67%	90.00%	**97.01%**	89.72%	**92.41%**	89.06%
	R	**97.06%**	**100.00%**	**100.00%**	**100.00%**	**92.65%**	90.32%	87.50%	**93.60%**	**98.48%**	**98.50%**	**93.96%**	**94.13%**
	F1	89.19%	**100.00%**	**98.04%**	**94.74%**	91.30%	86.60%	84.48%	**91.76%**	**97.74%**	93.90%	**93.18%**	**92.81%**
BERT Multitask	P	80.49%	96.67%	92.59%	83.33%	92.65%	86.02%	**88.10%**	**95.33%**	96.82%	93.48%	90.32%	90.53%
	R	97.06%	96.67%	100.00%	97.22%	92.65%	86.02%	66.07%	81.60%	92.42%	96.99%	93.96%	90.97%
	F1	88.00%	96.67%	96.15%	89.74%	92.65%	86.02%	75.51%	87.93%	94.57%	**95.20%**	92.10%	90.41%
ComplianceBert Multitask	P	80.00%	96.67%	96.15%	**91.67%**	93.33%	85.15%	82.93%	85.82%	96.83%	93.43%	92.23%	90.38%
	R	94.11%	96.67%	100.00%	91.67%	82.35%	**92.47%**	**91.07%**	92.00%	92.42%	96.24%	91.61%	92.78%
	F1	86.49%	96.67%	98.03%	91.67%	87.50%	**86.66%**	**86.81%**	88.80%	94.57%	94.81%	91.92%	91.45%

Table 4: Precision/Recall/F1-score for SC Models

enhancement layer. The correlation matrix layer learns a correlation matrix that captures the pairwise dependencies among the labels. The correlation enhancement layer uses the correlation matrix to enhance the raw label predictions by applying a nonlinear function to the predictions of other labels. The augmented label predictions are then used to compute the loss and update the model parameters. CorNet can learn and leverage label correlations to improve the predictions.

We also use three encoders for the Sentence Classification (SC) task: **BiLSTM** (Schuster and Paliwal, 1997), **BERT** (Devlin et al., 2019), and **ComplianceBERT**. Each encoder produces a sentence embedding from the input sentence, which is then passed to a fully connected layer to obtain the raw label predictions. These raw label predictions are subsequently enhanced by the CorNet layer, which generates the augmented label predictions by incorporating the compliance rules.

4.4.4 Multitask Model

In this work, we also propose a multitask model for NER and SC tasks, both of which require an encoder followed by an output layer. We use BERT and ComplianceBERT as encoders, which can learn shared representations from both tasks. The output layer is task-specific and can be adjusted according to the task objective. The multitask model adopts a sum loss of NER task and SC task, $L = \alpha L_{ner} + \beta L_{sc}$, where α and β are hyperparameters that control the relative weight of each task. The multitask model can optimize both tasks simultaneously and leverage the common information between them.

5 Evaluation

5.1 Experiment Result

Table 3 and Table 4 present the results for the NER and SC tasks, respectively. All results are the average results of multiple experiments with random seeds. The best values of precision, recall, and F1-score for each label are highlighted in bold. The row `Avg` displays the macro average of the 7 NER labels or the 11 SC labels. For the NER task, ComplianceBERT achieves the highest average recall and F1-score among all models, indicating its superior ability to identify and label personal information in privacy policy texts. For the SC task, ComplianceBERT outperforms other models in terms of average recall and F1-score, demonstrating its enhanced capability to classify sentences according to their compliance levels. ComplianceBERT effectively leverages the semantic and syntactic features of the text for sentence classification.

We also compare the performance of the multitask models with the single-task models. The multitask models utilize the same encoder parameters for both NER and SC tasks, whereas the single-task models employ separate encoder parameters for each task. The results show that the multitask models have a lower average F1 score than the single-task models for both tasks, particularly for the NER task. This suggests that the multitask models struggle to learn from both tasks simultaneously with shared parameters, and that the two tasks do not share substantial common information that benefits each other. In contrast, the single-task models can better capture task-specific features and independently optimize the parameters for each task.

	P	R	F1
BiLSTM	84.61%	66.67%	74.57%
BERT	97.05%	100%	98.50%
ComplianceBERT	**100.00%**	**100.00%**	**100.00%**
BERT Multitask	94.11%	96.96%	95.52%
ComplianceBERT Multitask	100%	96.96%	98.46%

Table 5: Precision/Recall/F1-score on C3P2 Task

5.2 Compliance Result

We evaluate the models' performance on the compliance checking task at the document level. This task involves determining whether a privacy policy document complies with a given regulation, based on the results of the Named Entity Recognition (NER) and Sentence Classification (SC) subtasks and the compliance rules. Precision, recall, and F1-score are used as evaluation metrics for this task. Table 5 presents the results for each model. Note that BiLSTM, BERT, and ComplianceBERT each refer to two models: one for the NER task and one for the SC task, using the same type of encoder. The results indicate that ComplianceBERT achieves the best performance across all metrics, demonstrating its accuracy and consistency in checking the compliance of privacy policy documents. While ComplianceBERT Multitask also performs well, it is slightly outperformed by ComplianceBERT. BERT and BERT Multitask exhibit high recall but low precision, indicating their ability to identify most relevant items but with a higher rate of false positives. BiLSTM shows lower precision and recall than the other models, suggesting its ineffectiveness for the compliance checking task. These results highlight the superiority of our ComplianceBERT model for the compliance checking task.

We can observe that the compliance checking performance score is higher than the subtasks' scores for most models. This discrepancy arises because the compliance checking results aggregate the outcomes of the subtasks at a document level, which helps mitigate the negative impact of errors in the subtasks. For instance, if a model incorrectly labels a single entity or sentence within a privacy policy, it may not significantly affect the overall compliance judgment of the document, provided that the majority of entities and sentences are correctly labeled. Consequently, the compliance checking task benefits from document-level aggregation, leading to higher performance compared to the subtasks.

6 Conclusion

In this paper, we address the issue of automated compliance checking of Chinese privacy policy texts. We make three primary contributions: First, we present the inaugural dataset and task for this problem, which includes compliance annotations by human experts at both the document level and the fine-grained level. Second, we introduce two subtasks to support our main task: Named Entity Recognition (NER) and Sentence Classification (SC), which aim to extract discriminative legal attributes from the privacy policies to aid models in generating reliable compliance results and clear explanations. Third, we further pre-train BERT on a large corpus of compliance-related texts, demonstrating that it outperforms baseline models across all tasks. Our work illustrates the feasibility and potential of applying Natural Language Processing (NLP) techniques to the automated compliance checking of privacy policy texts.

7 Limitations

However, we also encounter several limitations and challenges that we plan to address in our future work. These include: (1) developing more advanced methods to capture the complex requirements in the regulations that cannot be adequately addressed by Named Entity Recognition (NER) or Sentence Classification (SC) alone; (2) integrating dynamic analysis into our framework to verify the app's actual behaviors against the stated privacy policies; and (3) exploring multilingual methods that can adapt to different languages and regulations with minimal human intervention.

8 Ethics Statement

In conducting this research, we have adhered to the highest ethical standards to ensure the integrity and social responsibility of our work. The primary focus of our study is the automated compliance checking of Chinese privacy policy texts. This work is intended to improve the transparency and accountability of online service providers regarding the handling of personal data, thus contributing to the protection of user privacy.

Data Collection and Use The dataset comprises publicly accessible privacy policies from legal and government websites, ensuring no personal or sensitive information about individuals is included.

Data Annotations Annotators were fully informed about the task, compensated fairly, and provided with detailed instructions to ensure accuracy and consistency.

Impact and Use of Research The models developed and evaluated in this research are intended to assist in the compliance checking of privacy policies and are not designed to replace human judgment. These tools are meant to support legal professionals and regulatory bodies in their work. Our work aims to improve compliance with privacy regulations, protecting individual data and fostering trust in digital services. We advocate for the responsible use of these tools within legal and ethical guidelines.

By adhering to these principles, we aim to contribute positively to the field of natural language processing and the broader societal goal of safeguarding personal information.

References

Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry Den Hartog. 2012. A machine learning solution to assess privacy policy completeness: (short paper). In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, pages 91–96.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

GDPR. 2022. General Data Protection Regulation (GDPR) – Official Legal Text.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Shuang Liu and Guozhu Meng. 2021. Have you been properly notified? automatic compliance analysis of privacy policy text with gdpr article 13.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

PIPL. 2022. Translation: Personal Information Protection Law of the People's Republic of China - Effective Nov. 1, 2021 - DigiChina.

PISS. 2020. National Information Security Standardization Technical Committee Release Personal Information Security Standards.

David Sánchez, Alexandre Viejo, and Montserrat Batet. 2021. Automatic assessment of privacy policies under the GDPR. *Applied Sciences*, 11(4):1762.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Statista. 2023. App downloads 2021-2022, by country.

Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. ACM.

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 Hovy, Joel Reidenberg, and Norman 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 (Volume 1: Long Papers)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.

Guangxu Xun, Kishlay Jha, Jianhui Sun, and Aidong Zhang. 2020. Correlation networks for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1074–1082, New York, NY, USA. Association for Computing Machinery.

Razieh Nokhbeh Zaeem, Rachel L. German, and K. Suzanne Barber. 2018. PrivacyCheck. *ACM Transactions on Internet Technology*, 18(4):1–18.

Kaifa Zhao, Le Yu, Shiyao Zhou, Jing Li, Xiapu Luo, Yat Fei Aemon Chiu, and Yutong Liu. 2022. A fine-grained Chinese software privacy policy dataset for sequence labeling and regulation compliant identification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An architecture for automatically analyzing web privacy policies. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 1–16, San Diego, CA. USENIX Association.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies*, 2019:66–86.

A Appendix

A.1 Explanation of Labels

Collect Personal Information (CPI) This item describes information that can identify a natural person or reflect the activity of a natural person, such as name, phone number, email address, location, device information, etc. [PISS Art 5]

Policy Duration (PD) This item describes the date when the privacy policy was published, effective, or updated by the service provider, or personal information controller's (PI controller). [PISS Art 5]

Data Retention Period (DRP) This item describes the duration or criteria for which the personal information is retained by the PI controller. [PISS Art 6]

Data Retention Region (DRR) This item describes the geographic region or jurisdiction where the personal information is stored or processed by the PI controller. [PISS Art 6]

Overdue Processing Method (OPM) This item describes the method or procedure for disposing of or deleting the personal information when it is no longer needed for achieving the data collection purposes or when it exceeds the retention period. [PISS Art 6]

Data Collection Purpose (DCP) This item describes the specific and legitimate purposes for which PI is collected and used by the PI controller, such as to provide the service, to improve the service quality, to conduct market research, to send marketing messages, etc. [PISS Art 7]

User Portrait (UP) This item describes whether and how the personal information is used for creating a user portrait or a personalized display of the service. It also explains what benefits or risks may arise from such use and how the data subjects can opt-in or opt-out of such use. [PISS Art 7]

Right to Access (RA) This item describes the right of the data subjects to access their personal information that is held by the PI controller. [PISS Art 8]

Right to Rectify (RR) This item describes the right of the data subjects to rectify their personal information that is inaccurate or incomplete. [PISS Art 8]

Right to Delete (RD) This item describes the right of the data subjects to delete their personal information that is no longer necessary or relevant for achieving the data collection purposes. [PISS Art 8]

Right to Withdraw (RW) This item describes the right of the data subjects to withdraw their consent or authorization for collecting and using their personal information. [PISS Art 8]

Right to Account Cancellation (RAC) This item describes the right of the data subjects to cancel their account with the PI controller and terminate their use of the service.[PISS Art 8]

Personal Information Sharing (PIS) This item describes whether and how the personal information is shared, transferred or publicly disclosed by the PI controller to third parties, such as affiliates, partners, vendors, advertisers, etc. It also explains what types and categories of personal information are shared, transferred or publicly disclosed, for what purposes, and with whom. It also describes whether and how the PI controller uses third-party embedded code, plug-ins, or other tools to share personal information and what risks or benefits may arise from such use. [PISS Art 9]

Personal Information Protection (PIP) This item describes the technical and organizational measures that are taken by the PI controller to protect the personal information from unauthorized access, use, disclosure, modification, or deletion. It also describes the capabilities that are available for the data subjects to manage their personal information settings, such as encryption, anonymization, access control, notification, etc. [PISS Art 11]

A.2 The Details of Compliance Rules

The rules are as follows. We use the label name only to indicate that it is an unconditional label, meaning that it is always required for the policy to be compliant. We use a right arrow ($\rightarrow$) to indicate that it is a conditional label, meaning that it is required only when the condition on the left of the arrow is met. For example, Collect Personal Information (CPI) $\rightarrow$ Data Retention Period (DRP) means that if the policy has a label for CPI, it must also have a label for DRP. The rules are:

1. Policy Duration (PD)

2. User Portrait (UP)

3. Right to Account Cancellation (RAC)
4. CPI → Data Retention Period (DRP)
5. CPI → Data Retention Region (DRR)
6. CPI → Overdue Processing Method (OPM)
7. CPI → Data Collection Purpose (DCP)
8. CPI → Right to Access (RA)
9. CPI → Right to Rectify (RR)
10. CPI → Right to Delete (RD)
11. CPI → Right to Withdraw (RW)
12. CPI → Personal Information Sharing (PIS)
13. CPI → Personal Information Protection (PIP)

A.3 Implementation

To further pretrain BERT, we randomly mask 15% of the tokens in each text using the same strategy as BERT. We start from the BERT-base-chinese [1] model and fine-tune it on 3.2 million texts for 10 epochs. We use a batch size of 32, a learning rate of 5e-5, and a maximum sequence length of 512. We use the "BIO" schema for NER task, resulting in 7 types of NER labels and we have 11 labels for the SC task.

We split the dataset into three subsets: training, development, and test. We randomly select 40 documents for the development set and 40 documents for the test set, and use the remaining 403 documents for the training set. The number of sentences for each subset are shown in Table 2.

For BiLSTM, we set both the embedding size and the hidden size to 128, the learning rate to 1e-3, and we train the models for 30 epochs with a batch size of 64. We take α as 0.9 and β as 0.1 for multitask models. For BERT-base-chinese and our ComplianceBERT, we fine-tune them on our training data with a batch size of 32, 2 epochs and learning rates of 3e-5 for the encoder and 2e-4 for the output layers. For CorNet, we adopt same hyperparameters of the source code [2].

[1] https://huggingface.co/bert-base-chinese
[2] https://github.com/XunGuangxu/CorNet/blob/master/deepxml/cornet.py