Analysis of Privacy Compliance by Classifying Multiple Policies on the Web

Keika Mori, Tatsuya Nagai, Yuta Takata, and Masaki Kamizono
Deloitte Tohmatsu Cyber LLC
{keika.mori, tatsuya.nagai, yuta.takata, masaki.kamizono}@tohmatsu.co.jp

***Abstract*—Companies and organizations inform users of how they handle personal data through privacy policies on their websites. Particular information, such as the purposes of collecting personal data and what data are provided to third parties is required to be disclosed by laws and regulations. An example of such a law is the Act on the Protection of Personal Information in Japan. In addition to privacy policies, an increasing number of companies are publishing security policies to express compliance and transparency of corporate behavior. However, it is challenging to update these policies against legal requirements due to the periodic law revisions and rapid business changes. In this study, we developed a method for analyzing privacy policies to check whether companies comply with legal requirements. In particular, the proposed method classifies policy contents using a convolutional neural network and evaluates privacy compliance by comparing the classification results with legal requirements. In addition, we analyzed security policies using the proposed method, to confirm whether the combination of privacy and security policies contributes to privacy compliance. In this study, we collected and evaluated 1,304 privacy policies and 140 security policies for Japanese companies. The results revealed that over 90% of privacy policies sufficiently describe the handling of personal information by first parties, user rights, and security measures, and over 90% insufficiently describe the data retention and specific audience. These differences in the number of descriptions are dependent on industry guidelines and business characteristics. Moreover, security policies were found to improve the compliance rates of 46 out of 140 companies by describing security practices not included in privacy policies.**

***Index Terms*—Privacy Compliance, Privacy Policy, Security Policy, Convolutional Neural Network**

I. Introduction

The rapid development of internet services using personal data has raised awareness of privacy. Several laws, such as the EU General Data Protection Regulation (GDPR) and the Japanese Act on the Protection of Personal Information (APPI), require companies to disclose their personal data-handling practices. Companies use privacy policies to comply with legal requirements and express privacy transparency to users. However, privacy researchers reported that over 50% of corporate privacy policies in the EU did not disclose the information categories they collected, although required by the GDPR [1].

In Japan, the APPI requires companies to disclose information related to the handling of personal information. Companies work toward definition and publication of their privacy policies, as required by government institutions. In addition to privacy policies, an increasing number of companies are publishing various policies, such as privacy guidelines, privacy statements, and security policies, to express compliance and transparency of corporate behavior. However, it is difficult to update these policies in accordance with legal requirements due to periodic law revisions and rapid business changes [2]. Therefore, the following three questions were addressed in this study. **RQ1.** *What types of contents and volume are described in the privacy policies of Japanese companies?*, **RQ2.** *To what extent do privacy policies comply with Japanese laws?*, and **RQ3.** *Do security policies contribute to privacy compliance?*

To answer these questions, we developed a method to analyze privacy and security policies and evaluate their compliance with Japanese laws. The proposed method collects and extracts the contents of policies and classifies them using a convolutional neural network (CNN). Based on the classification results, they are compared with legal requirements. We updated the privacy practice categories introduced by Wilson et al. [3], annotated Japanese privacy policies with these categories, and trained the CNN models using the annotated data. Moreover, we designed logical expressions for legal comparison based on existing work [1].

First, we analyzed 1,304 privacy policies for Japanese companies in this study. We found that over 90% of the privacy policies have descriptions of "1st-party Collection," "Access, Edit, and Deletion," and "Data Security." In addition, we revealed that the compliance rate of the APPI provisions that require companies to disclose information was 71.9%, and clarified the differences between the rates of various industries. For example, the compliance rate is high in the financial, wholesale, and telecommunication industries. Thereafter, we analyzed 140 security policies for Japanese companies with privacy policies. We found that several companies provide a more detailed description of security in security policies than in privacy policies. We demonstrated the importance of analyzing various policies on the Web, in addition to privacy policies, to achieve accurate compliance checks.

II. Background

A. Act on the Protection of Personal Information (APPI)

The APPI [4] was enforced in Japan in 2005. In light of the social and economic changes, the law is reviewed every three years by the Personal Information Protection Commission (PPC). Under the current law, business operators who handle personal information are required to inform users of how they handle their information. Article 18 requires the operators to disclose the utilization purpose, and Article 36

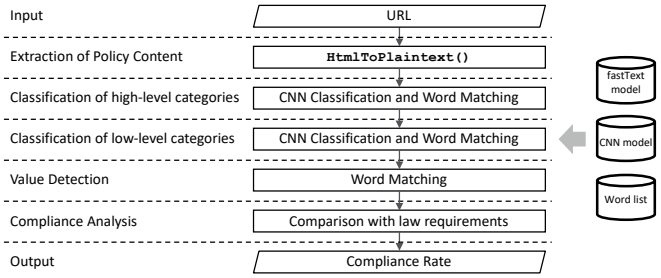

Fig. 1. Analysis pipeline of the proposed method.

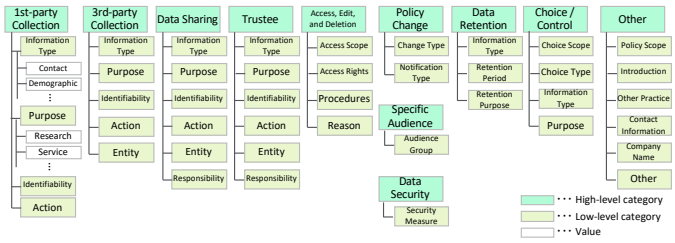

Fig. 2. Category scheme for classification.

requires the operators to disclose the categories of information contained in the anonymously processed information when it is produced.

B. Guideline

The PPC published guidelines for companies to meet the provisions set in the APPI. In addition to the guidelines, government institutions published guidelines for companies in specific industries. They recommend that companies publish privacy policies. The guidelines for companies in the financial and medical industries specifically describe items that should be disclosed in privacy policies [5], [6].

C. Privacy Policy

Companies disclose their data processing and protection practices in their privacy policies. For example, privacy policies include practices of handling personal information, such as collecting information, providing information to third parties, and sharing data. These policies are used to comply with the legal requirements. However, previous studies revealed that it is difficult to meet the requirements of the GDPR with respect to privacy policies [1], [7]. Given that the law is revised regularly and personal data utilization has accelerated, Japanese privacy policies may be in the same scenario as foreign privacy policies.

D. Other Policy

Several companies publish various policies to disclose specific privacy practices for the improvement of privacy transparency. For example, privacy guidelines or statements are mainly used to express the privacy attitudes of companies, although they are similar to the privacy policies in the previous section. Cookie policies or cookie notifications are used to clarify the handling of cookie data, given that several regulations such as GDPR define cookies as personal information. Moreover, security policies are used to inform users of the security measures implemented by companies. In this study, we analyzed security policies related to Japanese laws, in addition to privacy policies.

III. Proposed Method

This paper proposes an analysis method for privacy policies and legal compliance. The analysis pipeline of the proposed method is shown in Fig 1. First, we collected and extracted privacy policies and contents by crawling URLs. Thereafter, the proposed method classified these contents into multiple categories per line using CNN and word matching. Finally, the classification results, i.e., categories, were compared with the legal requirements to verify privacy compliance. In the compliance analysis, we used pre-defined logical expressions. In addition to privacy policies, we applied the proposed method to security policies and evaluated the increase in the compliance rate.

A. Extraction of Policy Content

We collected privacy and security policies using web crawling. The downloaded policies were HTML documents that contained unnecessary components such as headers and footers. Therefore, we removed these noise components using `HtmlToPlaintext` [8]. In particular, `HtmlToPlaintext` can parse HTML data and extract the policy contents of plain text using natural language processing (NLP) and heuristics. We updated `HtmlToPlaintext` to analyze Japanese content, as the original version only supports English content.

B. Classification of High-level and Low-level Categories

1) Word Embeddings: The policy contents were converted into vectors to train and build subsequent CNN classifiers. We used `MeCab` [9] to divide the privacy policy sentences into their parts of speech, and `fastText` [10] to convert each word into a vector. In particular, `fastText` learns the similarity of words by NLP and converts similar words into a nearby vector space, which allows for the absorption of orthographical variants. We trained and developed the `fastText` model using over 10,000 privacy policies of Japanese corporate websites collected by web crawling in advance.

2) CNN Classification: We developed and used CNN models to classify policy vectors into high-level and low-level categories. The high-level and low-level categories are shown in Fig. 2. We employed categories based on the OPP-115 dataset introduced by Wilson et al [3]. In addition, we updated these categories to comply with the APPI, as detailed in the following section. The architecture of the proposed CNN models is shown in Fig. 3. The input data of the classifiers were privacy policy sentences, and the output data were the probabilities of each category for each line. The word embedding layer converted policy sentences into vectors using

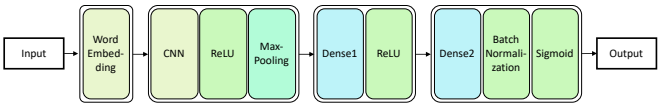

Fig. 3. Architecture of CNN classifier.

the `fastText` model, as described in the previous section. Thereafter, these vectors passed through a convolution layer using a filter with a size equal to the number of output classes, and a rectified linear unit (ReLU) activation function was applied. Thereafter, a max-pooling layer with a kernel size (*ks*) was loaded with the maximum elements from each vector, combined the elements, and generated a single vector. Subsequently, the vector passed through two dense (fully-connected) layers. Finally, we applied batch normalization with a batch size of *bs* to accelerate the learning process. Moreover, a sigmoid function converted the vector into probabilities for the possible output classes. We built one classifier for high-level categories and multiple classifiers for low-level categories for each high-level category. If the probability was higher than 50%, we adopted this category as a result.

3) Update Category Scheme: We added "Data Sharing" and "Trustee" to the original category scheme [3] and removed "Do Not Track." This was based on the following facts: (1) Article 23 (5) of the APPI defines a special requirement when personal data are provided to and jointly utilized by a specified person. (2) Article 23 (5) in the APPI states that entrusting the handling of personal information is different from provision to a third party, and it defines different requirements for each case. and (3) The APPI does not refer to "Do Not Track," and Safari no longer supports the "Do Not Track" feature [11], [12]. Moreover, we added 35 low-level categories and removed one low-level category based on the law.

4) Word Matching: Figure 2 includes several categories related to foreign laws, instead of Japanese laws. These categories may not appear in the Japanese privacy policies. In this case, the accuracy of the CNN classifiers was low due to the lack of training data. Therefore, we adopted word-matching for these categories. We manually selected words that frequently appeared in laws and privacy policies, and developed word lists for matching in advance.

C. Value Detection

For several legal requirements, we identified whether specific words (i.e., values in Fig. 2) were written in privacy policies. For example, the APPI requires companies to describe "Supervision over a Trustee" and "Employee Training" as "Security Measure" in "Data Security." We detected these specific sentences by word matching, as mentioned in the previous section. To develop word lists, the proposed method computed the importance of a word for each low-level category using TF-IDF on the training data. The application of word matching to privacy policies generated many false positives, we applied it to sentences with classification results for each low-level category.

D. Compliance Analysis

The proposed method verified legal compliance using the classification results. The APPI determines the requirement(s) under the specific condition(s), e.g., "A personal information handling business operator shall, in case of having acquired personal information except in cases where a utilization purpose has been disclosed in advance to the public, promptly inform a principal of, or disclose to the public, the utilization purpose (Article 18 (1))." Therefore, we represented the provisions as logical expressions with the categories and values in the form of "if sentences about Category A are described in a privacy policy, it should also write about Category B" in advance. Several examples of manually defined logical expressions are listed in Table I. It should be noted that $L = \{l_i\}$ refers to a set of categories/values described in a privacy policy, i.e., classification results, and l_i represents the following types: "high-level category," "high-level category_low-level category," and "high-level category_low-level category_value." For example, the proposed method checks whether a privacy policy contains sentences about "1st-party Collection_purpose" when the policy contains sentences about "1st-party Collection" to verify compliance with Article 18 (1).

IV. Datasets and Classification Models

A. Datasets

1) Training Data: We manually collected privacy policies that cover a variety of contents from 64 Japanese companies listed in Hoovers D&B [13] to create the training data. These companies were part of 11 industries: retail (RET), financial (FIN), wholesale (WHO), telecommunications (TEL), public (PUB), construction (CON), transportation (TRA), manufacturing (MAN), medical (MED), and energy (ENE), among others. Based on these privacy policies, three researchers labeled each sentence with high-level and low-level categories and values, as shown in Fig. 2. When more than two researchers annotated the same label as the same sentence, we adopted the label as training data. The number of labels with high-level categories as the training data is listed in Table II. We adopted 3,099 labels as training data for the CNN models from a total of 5,166 labels. In the same manner for low-level categories, we adopted 2,777 labels. In addition to these high-level and low-level categories, researchers annotated phrases and words with values. We adopted these values labeled by a minimum of one researcher as training data due the slight differences in the ranges of phrases and words among them.

It should be noted that there was a bias in the number of categories and values in the training data, and several categories and values had no training data. For example, we could not find sentences labeled with "3rd-party Collection_Action." Therefore, we excluded these categories and values from the compliance analysis in this study, given that the proposed method could not classify them.

2) Word Lists: The numbers of privacy sentences labeled with high-level categories such as "Data Retention," "Choice / Control," and "Specific Audience" were low, namely, 7, 4, and

TABLE I
EXAMPLES OF LOGICAL EXPRESSIONS FOR COMPLIANCE ANALYSIS.

Provision	Condition	Requirement
Article 18 (1)	1st-party Collection $\in L$	1st-party Collection_purpose $\in L$
Article 23 (2)	(3rd-party Collection $\in L$) $\vee$ (3rd-party Collection_action_receive from 1st-party $\in L$)	(1st-party Collection_purpose_third-party provision $\in L$)$\wedge$ (3rd-party Collection_information type $\in L$)$\wedge$ (3rd-party Collection_action $\in L$)$\wedge$ (Access, Edit, and Deletion_Access Rights_cease third-party provision $\in L$)$\wedge$ (Access, Edit, and Deletion_Procedures_cease third-party provision $\in L$)
Article 36 (3)	1st-party Collection_identifiability_aggregated or anonymized $\in L$	1st-party Collection_infromation type $\in L$

TABLE II
NUMBER OF LABELS WITH HIGH-LEVEL CATEGORIES IN TRAINING DATA.

High-level Category	# of training data
1st-party Collection	952
3rd-party Collection	270
Data Sharing	595
Trustee	49
Access, Edit, and Deletion	311
Policy Change	47
Data Security	312
Data Retention	7
Choice/Control	4
Specific Audience	5
Other	547
Total	3,099

TABLE III
STATISTICS OF POLICIES IN TEST DATA.

	Privacy Policy			Security Policy		
Industry	# of companies	# of policies	Avg of words	# of companies	# of policies	Avg of words
RET	180	196	1,151	3	3	433
FIN	194	261	1,490	20	22	639
WHO	102	110	1,152	13	13	460
TEL	185	255	1,610	50	65	721
PUB	16	16	1,520	6	6	405
CON	104	120	1,123	5	5	543
TRA	53	64	984	5	5	336
MAN	286	351	1,094	19	23	447
MED	73	86	970	3	3	286
ENE	10	11	2,041	2	2	360
Other	101	119	1,526	14	17	881

5, respectively, as shown in Table II. This is because the APPI does not strictly require companies to disclose information regarding these categories. In particular, effort is required to delete unnecessary personal data and to set data retention periods. In addition, there are no rules regarding personal data of children in the APPI. The proposed method detected sentences with high-level and low-level categories that contained a small number of training data by word matching instead of CNN classification. We manually developed word lists for these high-level categories, e.g., the list for "Data Retention" includes "save," "store," and "period." In addition to the high-level categories, we manually developed word lists for the low-level categories of "Data Sharing_Responsibility" and "Access, Edit, and Deletion_Reason," e.g., "responsible" and "management." Moreover, we used word matching for the value detection. To develop word lists, we calculated TF-IDF values for words with each value label and adopted words with TF-IDF scores larger than 0.2, and which were unique within the low-level category.

3) Test Data: We used Hoovers D&B [13] to obtain URLs and policy contents of major Japanese corporate websites as the test data. First, we crawled 3,728 URLs and explored various policies on the websites by following links with URLs or link texts including policy-related words. Thereafter, to filter-out noise content, the proposed method checked the titles of web pages based on two criteria; (1) whether the title includes policy-related words, e.g.,"policy," "guideline," or "statement;" and (2) whether the title includes "privacy" or "security." We collected 2,643 privacy policies and 641 security policies from the crawling and filtering processes. Thereafter, the proposed method removed unnecessary components (e.g., headers and footers) and extracted policy contents from the collected policy data using `HtmltoPlaintext`. We excluded policies with fewer than 150 words, such as those consisting of only links. In addition, we excluded company security policies with no privacy policies to analyze privacy compliance. After the data cleaning process, the test data consisted of 1,589 privacy policies and 164 security policies. The statistics of the test data are presented in Table III. The number of policies exceeded the number of companies in several industries. This indicates that these companies have several policies described in Section II-D in addition to privacy policies. In the compliance analysis, we evaluated these policies comprehensively. Table III reveals that the average number of words (i.e., the length of privacy policies) in the energy industry was the highest (longest) and security policies were pervasive in the telecommunications industry.

B. Classification Models

We developed CNN classifiers using the training data and searched for optimal hyperparameters, i.e., the kernel size *ks* and the batch size *bs*, for each classifier by comparing the evaluation metrics. First, we developed a CNN model for high-level categories and compared the accuracies of models with different kernel sizes of *ks* = 2, 3, 4, 5, and combinations of 2, 3, 4, and 5 at different epoch numbers. Thereafter, we compared the accuracies of models with different batch sizes (*bs* = 11, 22, 44, and 88) at different epoch numbers. As a result, the optimal hyperparameters were *ks* = 5 and *bs* = 44 at epoch number 50. It should be noted that the F-score of the CNN model with these hyperparameters was 0.80.

TABLE IV
PERCENTAGE OF PRIVACY POLICIES WITH HIGH-LEVEL CATEGORIES. GREEN ITEMS ARE OVER 80%, AND RED ITEMS ARE UNDER 20%.

High-level Category	RET	FIN	WHO	TEL	PUB	CON	TRA	MAN	MED	ENE	Other	All
1st-party Collection	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.0**	**98.1**	**99.7**	**100.0**	**100.0**	**100.0**	**99.8**
3rd-party Collection	**87.8**	**89.2**	**88.2**	**89.7**	**93.8**	**85.6**	**88.7**	**88.5**	**93.2**	**100.0**	**93.1**	**89.2**
Data Sharing	**81.7**	**91.8**	**87.3**	**91.9**	**87.5**	**82.7**	73.6	78.3	69.9	**90.0**	**91.1**	**84.3**
Trustee	37.2	47.9	31.4	51.9	50.0	36.5	26.4	28.0	24.7	10.0	41.6	37.5
Access, Edit, and Deletion	**92.2**	**96.4**	**92.2**	**94.1**	**100.0**	**90.4**	79.2	**90.9**	**89.0**	**100.0**	**94.1**	**92.3**
Policy Change	57.2	57.2	56.9	70.3	75.0	47.1	34.0	53.1	56.2	50.0	63.4	57.0
Data Security	**97.8**	**99.5**	**98.0**	**97.3**	**100.0**	**91.3**	**100.0**	**97.2**	89.0	**100.0**	**99.0**	**97.1**
Data Retention	4.4	10.8	4.9	15.7	6.3	6.7	9.4	10.8	9.6	0.0	12.9	9.7
Choice/Control	26.7	22.2	26.5	45.4	37.5	22.1	17.0	25.5	20.5	60.0	27.7	27.8
Specific Audience	7.8	4.6	2.0	9.7	12.5	1.9	1.9	10.1	2.7	0.0	5.0	6.4
Other	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.9**

In the same manner, we searched for the optimal hyper-parameters of the CNN models for low-level categories. We adopted the kernel size $ks = 5$, given that the input data of the high-level and low-level category classifiers were the same. We developed CNN models for low-level categories and compared the accuracies of the CNN models with different batch sizes at different epoch numbers. As a result, a batch size twice the number of output classes was optimal at epoch number 150. The average F-score of the CNN models was 0.84.

V. PRIVACY POLICY ANALYSIS

This section presents an analysis of the classification results of privacy policies using the proposed method, and an evaluation of privacy compliance using the results to answer **RQ 1** and **RQ 2**.

A. Classification Results of High-level Categories

The percentages of privacy policies that contained each high-level category are shown in Table IV. Over 90% of the privacy policies contained descriptions of high-level categories: "1st-party Collection," "Access, Edit, and Deletion," and "Data Security." Moreover, descriptions of high-level categories such as "Data Retention" and "Specific Audience" were expressed in under 10% privacy policies. The highest rate of descriptions of "Data Retention" practices was 15.7% in the telecommunications industry. Given that the guidelines for the telecommunications industry contain rules regarding data deletion, companies are encouraged to define and describe data retention periods on their privacy policies in this industry when compared with other industries.

To analyze the number of descriptions of each high-level category, we counted the number of words labeled with high-level categories. As shown in Fig. 4, privacy policies in the energy industry had more descriptions of "Data Sharing" than other policies. These companies should share customer data to provide their services. In particular, the Gas Business Act [14] requires a gas retailer to notify the gas service providers of the investigation results for gas appliances when they obtain consent from the owner(s) of the appliances. Therefore, they elaborated on "Data Sharing" with the relevant law references in their privacy policies.

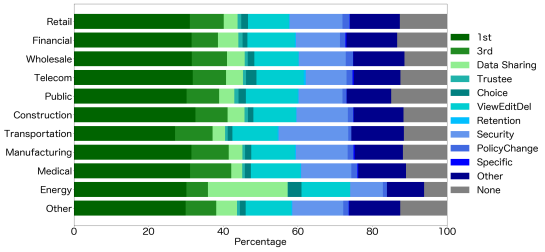

Fig. 4. Percentages of descriptions of high-level categories in privacy policies.

B. Classification Results of Low-level Categories

Table V presents low-level categories included in over 80% of privacy policies in one or more industries. The high average percentages across all industries were found in low-level categories such as "1st-party Collection_Info Type," "1st-party Collection_Purpose," "Access, Edit, and Deletion_Access Right," "Data Security_Security Measure," and "Other_Contact Info." Privacy policies of energy companies had descriptions with various low-level categories, especially low-level categories of "Data Sharing." As mentioned in the previous section, the business characteristics and laws in the energy industry have an influence on the results.

C. Detection Results of Values

To investigate the diversity and details of descriptions in the high-level and low-level categories, we evaluated value labels for each low-level category. We divided the total number of detected values by the number of privacy policies that contained the values for each low-level category. This was carried out to calculate the average number of values. Table VI presents the results of the low-level categories with more than five values. We observed more than 9 values of "Data Security_Security Measure" in the privacy policies of all industries. Therefore, the low-level category of "Data Security_Security Measure" had a wide range of values, and the frequently detected values were "Data access limitation" and "not identify a principal of anonymously processed information."

D. Analysis Results of Privacy Compliance

We applied the logical expressions shown in Table I to the classification results and calculated the compliance rates

TABLE V
PERCENTAGE OF PRIVACY POLICIES WITH LOW-LEVEL CATEGORIES. GREEN ITEMS ARE OVER 80%.

High-level Category	Low-level Category	RET	FIN	WHO	TEL	PUB	CON	TRA	MAN	MED	ENE	Other	All
1st-party	Info Type	76.7	**86.6**	79.4	**87.6**	**87.5**	75.0	71.7	78.0	**80.8**	**90.0**	**82.2**	**80.8**
1st-party	Purpose	**80.0**	**87.6**	**88.2**	**96.2**	**87.5**	**83.7**	73.6	**82.9**	79.5	**90.0**	**88.1**	**85.5**
3rd-party	Vague	67.2	74.2	69.6	68.1	50.0	59.6	71.7	72.0	**80.8**	70.0	72.3	70.2
3rd-party	Does Not	71.7	75.3	75.5	72.4	56.3	61.5	77.4	76.9	**83.6**	70.0	72.3	73.7
Data Sharing	Info Type	48.9	61.3	51.0	**81.1**	62.5	45.2	54.7	44.1	45.2	**90.0**	66.3	56.0
Data Sharing	Purpose	29.4	49.5	34.3	50.8	43.8	40.4	18.9	29.4	27.4	**80.0**	46.5	38.0
Data Sharing	Entity	59.4	70.1	66.7	79.5	75.0	65.4	52.8	59.8	50.7	**80.0**	75.2	65.8
Access, Edit, and Deletion	Access Right	79.4	**90.7**	**82.4**	**80.5**	**100.0**	76.0	67.9	**81.8**	**83.6**	**100.0**	**83.2**	**82.2**
Data Security	Security Measure	**93.3**	**97.4**	**95.1**	**93.0**	**100.0**	**89.4**	**94.3**	**93.4**	**86.3**	**100.0**	**97.0**	**93.8**
Other	Introduction	76.7	**82.5**	75.5	75.7	75.0	75.0	73.6	76.9	67.1	**80.0**	**82.2**	77.0
Other	Contact Info	78.3	**92.3**	**90.2**	**90.8**	**93.8**	**84.6**	67.9	**81.1**	74.0	**80.0**	**87.1**	**84.4**

TABLE VI
AVERAGE NUMBER OF DETECTED VALUES. GREEN ITEMS ARE VALUES OF FIVE OR MORE.

High-level Category	Low-level Category	# of value types	RET	FIN	WHO	TEL	PUB	CON	TRA	MAN	MED	ENE	Other
1st-party	Info Type	16	3.5	3.3	2.7	**5.3**	1.9	2.2	1.9	3.2	2.8	4.2	3.4
1st-party	Purpose	10	4.9	**5.6**	4.7	**7.1**	**5.4**	**5.4**	**6.4**	4.5	4.4	**8.2**	**5.5**
Access, Edit	Access Right	6	3.7	**5.9**	3.8	**6.9**	4.9	3.9	4.3	4.2	4.0	**5.9**	**5.2**
Data Security	Security Measure	12	**11.5**	**13.0**	**9.7**	**14.4**	**11.2**	**10.8**	**11.5**	**10.6**	**9.8**	**10.7**	**12.8**

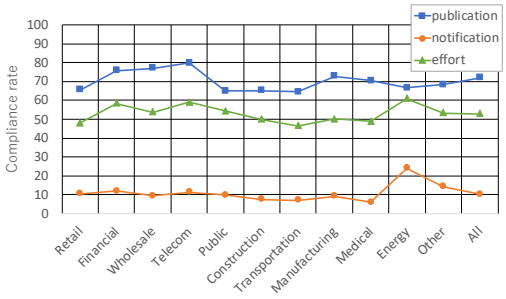

Fig. 5. Analysis results of privacy compliance for each industry.

for each provision. We grouped provisions into three types: "publication," "notification," and "effort." The "publication" provision requires companies to disclose information, the "notification" provision requires to inform users of privacy practices in privacy policies or other methods (e.g., emails and direct messages), and the "effort" provision requires internal corporate behaviors. Figure 5 presents the compliance rate of each provision type for each industry. The average compliance rate of "publication" was 71.9%, whereas those of "notification" and "effort" were 10.2% and 52.8%, respectively.

The compliance rates of "publication" in the financial, wholesale, and telecommunication industries were higher than 75%; however, that of the transportation industry was lower than 65%. It should be noted 79.2% of privacy policies matched the condition of Article 18 (1), which requires companies to disclose the purpose of personal information collection. Moreover, 85.2% of these policies met the requirement (i.e., the compliance rate of this provision); and 15.5% of privacy policies matched the condition of Article23 (2), which requires companies to disclose information about data provision to third parties. However, the compliance rate of this article was 0%; and 97% of companies did not disclose the types of personal information provided to third parties. Furthermore, 91% did not disclose any methods to stop the data provision to third parties.

With respect to the compliance rates of "notification," the energy industry yielded the highest percentage due to business characteristics. However, the overall compliance rate was under 20%, whereas the privacy policies met the conditions of more than half of the provisions. These privacy policies were lacking in the descriptions of "Policy Change," "Data Sharing," and "Other_Company Name."

There were 13 "effort" provisions, where more than half of the privacy policies met the conditions. The compliance rates of more than 70% accounted for 5 out of 13 provisions. These privacy policies were sufficiently described for "1st-party Collection_Purpose," "Data Security_Security Measure," and "Access, Edit, Deletion_Procedures_Edit," and "Other_Contact Info." Moreover, five provisions had the compliance rates of less than 50%. These privacy policies were insufficiently described for "Data Security_Security Measure_Privacy Training," "Data Security_Security Measure_Assurance about Accuracy," "Access, Edit, and Deletion_Access Rights_Utilization Cease," "Access, Edit, and Deletion_Reason," and "Data Retention_Retention Period."

VI. SECURITY POLICY ANALYSIS

In this section, we analyze the classification results of security policies using the proposed method, and evaluate the contributions to privacy compliance to answer **RQ 3**.

A. Classification Results

We calculated the percentages of security policies with descriptions of each high-level category. Only the "Data Secu-

TABLE VII
NUMBER OF POLICIES WITH VALUES IN "SECURITY MEASURE" OF "DATA SECURITY."

Value	Privacy	Security	Only Security
Supervision over a trustee	704	28	17
Data access limitation	1,202	57	4
Secure data transfer	182	8	8
Employee training	520	72	40
Assurance about accuracy	676	34	12
Not identify anonymous information	1,193	136	8
Secure data storage	1,083	85	11
Organizational structure or program	1,100	116	18

TABLE VIII
NUMBER OF COMPANIES IN LEGAL COMPLIANCE BY MULTIPLE POLICIES.

Article	Only Privacy	Only Security	Combination
Article 36 (6)	40	0	40
Article 39	40	0	40
Article 19 (1)	59	0	61
Article 20 (1)	962	124	967
Article 21 (1)	394	64	427
Article 22 (1)	379	16	394
Article 36 (1)	40	0	40
Article 36 (2)	40	0	40

rity" category was described in 100% of security policies. In the low-level category, we found that "Data Security_Security Measure" had the most frequent descriptions. In particular, 99% of security policies had descriptions of this category. The proposed method detected eight value types in the low-level category of "Data Security_Security Measure" in security policies. As shown in Table VII, we found that organizational security measures were frequently expressed in security policies, and several values were expressed only in security policies and not in privacy policies.

B. Contribution to Privacy Compliance

We applied the logical expressions of only security-related articles listed in Table VIII to the classification results of multiple policies. Table VIII presents the number of companies complying with privacy policies, security policies, and a combination of both. We confirmed that the number of compliances in the combination of both policies was higher than only privacy policies in several articles. Moreover, the compliance rate increased for the 46 companies. In Article 19 (1), we found that the compliance rate with respect to the combination increased, although the compliance rate with only security policies was zero. After a detailed investigation, we found that one requirement was complied with in a privacy policy and the other in a security policy. Therefore, security policies detail the security measures implemented by companies, and contribute to privacy compliance with respect to corporate security.

VII. DISCUSSION

A. Privacy Policies and Guidelines

It should be noted that "1st-party Collection_purpose," "Data Security_Security Measure," "Contact Information," and "Access, Edit, and Deletion_procedures" are recommended for inclusion in privacy policies in both the financial and medical guidelines [5], [6]. Among them, "1st-party Collection_purpose," "Data Security_Security Measure," and "Contact Information" were expressed adequately in various privacy policies of all industries. The compliance rate of the medical industry was under 50%, even with respect to industry-specific guidelines [6]. This can be attributed to the discrepancy between the companies covered by the guidelines and the company data used in this study. The company data in the financial industry included various types of companies, such as funds, banks, securities, insurance, and commodities. Moreover, 95% of the companies were covered by the financial guidelines [5]. The majority of company data in the medical industry was corporate websites related to biotechnology, medical devices, and pharmaceuticals. The medical guidelines mainly target companies that directly handle personal data such as patient data, namely, companies that provide healthcare and nursing care services. Therefore, the description rates of the abovementioned recommendation items were low due to the discrepancy between the companies covered by the guidelines and the company data used in this study.

B. Privacy Policies and Business

The contents of privacy policies are influenced by business characteristics and business laws. For example, energy companies share customer data to provide their own services, given that the Gas Business Act [14] requires a gas retailer to notify the gas service providers of the investigation results for gas appliances. Hence, the privacy policies of the energy companies contain adequate contents about data sharing, as shown in Fig. 4 and Table IV.

C. Privacy Policies and Security Policies

We comprehensively evaluated the company attitudes of privacy by analyzing multiple policies when companies published various policies such as privacy policies, privacy guidelines, privacy statements, and security policies. In this study, we identified corporate security efforts for privacy more accurately by analyzing security policies in addition to privacy policies. Moreover, it is important to identify the policy target and range, i.e., what the policy is defined for. For example, there are privacy policies for *website* users and privacy policies for *service* users. We cannot conduct an accurate analysis of privacy compliance if we analyze multiple policies for different targets and integrate these results. The application and evaluation of a method to automatically identify policy targets will be carried in future research. Moreover, we evaluated security policies in addition to privacy policies based on Japanese law. For foreign websites, the proposed method can evaluate privacy compliance more accurately by analyzing policies related to the laws of each country such as cookie policies and cookie notifications, in addition to privacy policies.

D. Limitations

In this study, we used the privacy policies of 64 Japanese corporate websites as the training data. The quantity of the

training data was insufficient, although the quality was enhanced by multiple annotators. To achieve an accurate evaluation of privacy compliance, we will develop more training data and improve the accuracy of the developed models in future research. Moreover, we adopted word matching to detect labels for which there was minimal data, to develop a CNN model. Given that the detection accuracy is dependent on the word list, updating the word list is within the scope of future research. In addition, the APPI is reviewed every three years [2]. Therefore, we should update the logical expressions that correspond to the law revision.

VIII. Related Work

A. Content Analysis of Privacy Policy

Various researchers proposed methods for analyzing privacy policies and providing policy-related datasets. Wilson et al. [3] analyzed 115 English-language privacy policies and developed a category scheme (the basis of the scheme in Fig. 2) for privacy policies. They created and published a dataset of annotated privacy policies, which is referred to as the OPP-115 Corpus. Zimmeck et al. [15] analyzed privacy policies by combining crowdsourcing and machine-learning classification. Harkous et al. [16] developed the Polisis tool to automatically annotate privacy policies by CNN-based classifiers using OPP-115 corpus, and visualized the results. Sarne et al. [17] proposed a framework for the topic extraction of privacy policies using unsupervised learning techniques. They analyzed the changes in the topics of interest in privacy policies using the framework. Other methods automatically identified opt-out choices in privacy policies using machine learning and the OPP-115 corpus [18], [19].

All existing methods mentioned above analyzed privacy policies written in **English**. The proposed method analyzed privacy policies written in **Japanese**, although we utilized several existing methods in the proposed method. In addition, we adopted a hybrid classification approach using a CNN (deep learning) and word matching (static analysis) in the proposed method while updating the category scheme for classification in accordance with Japanese laws.

B. Compliance Analysis of Privacy Policy

Degeling et al. [20] studied the differences between privacy policies before and after the enforcement of the GDPR. They revealed that 15.7% of websites added new privacy policies, and 72.6% with existing privacy policies updated them close to the date of enforcement of the GDPR. In other studies, legal compliance was analyzed by classifying privacy policy contents and comparing the results and the legal requirements [1], [7]. Nejad et al. [21] automatically mapped sentences in privacy policies with relevant GDPR articles using semantic text-matching techniques. Reyes et al. [22] analyzed mobile app compliance with the Children's Online Privacy Protection Act (COPPA) in the United States. They reported that the majority of popular free apps for children potentially violated the COPPA.

In this study, we designed new logical expressions for Japanese legal requirements in the proposed method and evaluated the compliance of Japanese privacy policies. In addition, we evaluated privacy compliance by analyzing multiple policies such as privacy policies and security policies.

IX. Conclusions

We analyzed both the privacy and security policies of Japanese companies and evaluated the compliance rates of Japanese laws. As a result, we identified the over- and understatements in these policies and the impact of guidelines and business characteristics on the policies for each industry. Moreover, we found that security policies complemented privacy policies by detailing the practices involving the provision of data to third parties and security measures. We therefore suggest the analysis of multiple policies in addition to privacy policies to check corporate privacy practices.

References

[1] T. Linden *et al.*, "The privacy policy landscape after the GDPR," *Proc. Priv. Enhancing Technol.*, 2020.

[2] Personal Information Protection Commission, "Act on the Protection of Personal Information "The Every-Three-Year Review" Outline of the System Reform." https://www.ppc.go.jp/en/aboutus/roles/international/cooperation/20200124/.

[3] S. Wilson *et al.*, "The creation and analysis of a website privacy policy corpus," in *ACL*, 2016.

[4] Personal Information Protection Commission, "Act on the Protection of Personal Information." https://www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf.

[5] Financial Service Agency, "Guidelines for Protection of Personal Information in the Finance Sector." http://www.japaneselawtranslation.go.jp/common/data/notice/052908.html.

[6] Ministry of Health, Labour and Welfare, "Guidelines for the appropriate handling of personal information by medical and nursing care providers." https://www.ppc.go.jp/files/pdf/01_iryoukaigo_guidance3.pdf. (in Japanese).

[7] S. Liu *et al.*, "Have you been properly notified? automatic compliance analysis of privacy policy text with GDPR article 13," in *WWW*, 2021.

[8] B. Andow, "HTML Privacy Policy to Plaintext Converter." https://github.com/benandow/HtmlToPlaintext.

[9] T. Kudo, "MeCab." https://taku910.github.io/mecab/.

[10] Facebook Inc., "fastText." https://fasttext.cc/.

[11] "DNT." https://developer.mozilla.org/ja/docs/Web/HTTP/Headers/DNT.

[12] C. Business, "What apple killing its do not track feature means for online privacy." https://edition.cnn.com/2019/02/13/tech/apple-do-not-track-feature/.

[13] Dun & Bradstreet, Inc., "D&B Hoovers." https://www.dnb.com/products/marketing-sales/dnb-hoovers.html.

[14] T. Ministry of Economy and Industry, "Gas business act." http://www.japaneselawtranslation.go.jp/law/detail/?id=3331&vm=04&re=02.

[15] S. Zimmeck and S. M. Bellovin, "Privee: An architecture for automatically analyzing web privacy policies," in *USENIX Security*, 2014.

[16] H. Harkous *et al.*, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *USENIX Security*, 2018.

[17] D. Sarne *et al.*, "Unsupervised topic extraction from privacy policies," in *WWW*, 2019.

[18] K. M. Sathyendra *et al.*, "Identifying the provision of choices in privacy policy text," in *EMNLP*, 2017.

[19] V. B. Kumar *et al.*, "Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text," in *WWW*, 2020.

[20] M. Degeling *et al.*, "We value your privacy ... now take some cookies: Measuring the gdpr's impact on web privacy," in *NDSS*, 2019.

[21] N. M. Nejad *et al.*, "Knight: Mapping privacy policies to GDPR," in *EKAW*, 2018.

[22] I. Reyes *et al.*, ""won't somebody think of the children?" examining COPPA compliance at scale," *Proc. Priv. Enhancing Technol.*, 2018.