情報処理学会論文誌 Vol.62 No.1 332–345 (Jan. 2021)

日本語のプライバシポリシに対する完全性を考慮したリスク要約手法の評価

中村 徹[1,a)] ウェルデルファエル B. テスファイ[2] バネッサ ブラカモンテ[3]
清本 晋作[3] 鈴木 信雄[4,1]

受付日 2020年3月31日, 採録日 2020年10月6日

概要：本研究では，自然言語処理と機械学習を用いて，プライバシポリシの要約を行うことにより，より理解しやすいプライバシポリシの実現を目指す．本論文ではまず，日本語のプライバシポリシを収集し，これにラベル付与を行い，評価するコーパスを作成する．ラベル付与にあたり，プライバシリスクに関する項目だけでなく，プライバシポリシの完全性に関する項目を設定する．プライバシポリシの完全性とは，プライバシポリシがトピックを網羅していることを示す性質である．このコーパスに対し，プライバシポリシのテキスト部の特徴抽出および機械学習を行い，学習モデルの精度評価を行う．本論文では，特徴抽出アルゴリズムとして，Bag-of-Words（以下，BoW），TF-IDF，Doc2Vec，学習アルゴリズムとして，サポートベクタマシンとランダムフォレストを採用した場合のラベル推測精度について評価を行う．評価の結果，本コーパスに対しては，BoW を用いた場合と TF-IDF を用いた場合については，両者間に有意な差は確認できなかったが，BoW または TF-IDF とランダムフォレストを組み合わせた場合が最も高精度にラベル推測を行うことができることが明らかになった．

キーワード：プライバシ保護，プライバシポリシ，機械学習，自然言語処理

Evaluation of Risk Summarization for Privacy Policies in Japanese with Considering Completeness

Toru Nakamura[1,a)] Welderufael B. Tesfay[2] Vanessa Bracamonte[3]
Shinsaku Kiyomoto[3] Nobuo Suzuki[4,1]

Received: March 31, 2020, Accepted: October 6, 2020

Abstract: The purpose of this study is to realize more understandable privacy policies by privacy policy summarization. We first made the corpus for evaluation by collecting Japanese privacy policies and labeling to the policies. As labels of corpus, we set not only items related to privacy risk, but also those related to completeness of privacy policy, that is the property to cover necessary topics for privacy policies. Next, we made prediction models by feature extraction algorithms for text and machine learning algorithm with this corpus. We also evaluated the accuracy if we used these prediction models. In this paper, Bag-of-Words (BoW), TF-IDF, and Doc2Vec were used as feature extraction algorithms for text and support vector machine and random forest were used as machine learning algorithms. As the result of evaluation, we obtained the fact that the combination of BoW or TF-IDF and random forest achieved the best F1 value for predicting labels of privacy policies. There was no significant difference between the case with BoW and that with TF-IDF in this evaluation.

Keywords: privacy protection, privacy policy, machine learning, natural language processing

[1] 株式会社国際電気通信基礎技術研究所（ATR）
Advanced Telecommunications Research Institute International (ATR), Sorakku-gun, Kyoto 619–0237, Japan
[2] ヨハンヴォルフガングゲーテ大学
Goethe University Frankfurt, Theodor-W.-Adorno-Platz 1, 60323 Frankfurt am Main, Germany
[3] 株式会社 KDDI 総合研究所
KDDI Research, Fujimino, Saitama 356–0003, Japan
[4] 近畿大学
Kindai University, Iizuka, Fukuoka 820–8555, Japan
[a)] tr-nakamura@atr.jp

1. はじめに

1.1 研究の背景

IT の発展により，多量のデータが収集可能となり，またこれを有効に活用することが可能になってきた．特に，個人に関するデータであるパーソナルデータの有効利用が，産業界を中心として期待されている．一方で，活用されるパーソナルデータの対象となっている個人にとっては，自身のパーソナルデータが適切に取り扱われているか等，プライバシに関する懸念もある．パーソナルデータの利用にあたり，OECD8 原則等により，データの利用目的を明確にし，またデータ収集の実施方針等を明示することが推奨されている．通知や公表の手段として，Web ページへのプライバシポリシの公開が一般的である．利用者がプライバシポリシを確認することにより，自身のパーソナルデータがどのように活用されるかを確認し，サービスを利用するかを判断することにより，個人の権利侵害を防止することが期待できる．

一方で，プライバシポリシの有効性については批判も多い．たとえば，Costante らは，プライバシポリシは非常に長く理解し難いため，人々はプライバシポリシを読まないと批判した [1]．また，McDonald らは，プライバシポリシを読むために費やさなければならない時間を見積もり，膨大な時間の浪費であると批判した [2]．

EU において，2018 年 5 月より，一般データ保護規則（GDPR）が施行されている．GDPR 第 12 条第 1 項において，管理者は，データ主体に対し，簡潔で，透明性があり，理解しやすく，容易にアクセスできる方式により，明確かつ平易な文言を用いて，情報提供を行うための措置を講じる必要があると述べられている．このことから，今後はより理解しやすい形式でプライバシポリシを表示することが求められる可能性が高い．

利用者に対してプライバシポリシを理解しやすくする取り組みとして，P3P Expandable Grid [3] や Privacy Label [4] 等の，プライバシポリシを要約した重要情報を表形式で表示する方法が提案されている．**図 1** に，Privacy Label の例を示す．これらの研究では，プライバシポリシに記載すべき情報のうち，利用する情報の種類や利用目的等の重要情報を表形式で表示する．しかしながら表形式への対応は，各事業者が行わなければならないため，実用化には課題がある．実際，P3P Expandable Grid は，P3P [5] に準拠した機械可読なプライバシポリシを事業者が準備することを前提としているが，P3P に対応する事業者は限られており，十分に普及しなかった．

上記の問題に関して，自然言語処理と機械学習を用いて，半自動的にプライバシポリシをユーザにとって理解しやすい形式に変換する研究が存在する．既存研究は，プライバシポリシの完全性を評価する研究 [1], [6] と，プライバシポリシからリスクを評価する研究 [7], [8], [9] に分類できる．プライバシポリシの完全性とは，プライバシポリシがトピックを網羅していることを示す性質である．プライバシポリシの完全性を評価する研究は，プライバシポリシの品質を評価するためには有用であるが，プライバシポリシから要約された重要情報を抽出することを目的としていない．一方で，プライバシポリシからリスクを評価する研究は，プライバシポリシから要約された重要情報を抽出，またはその重要情報についてのリスク度を推測することを目的としている．これを用いることで，たとえば重要情報やそのリスク度を表形式で表示し，理解しやすい形式で利用者に情報提供することが可能となる．

図 1 Privacy Label の例 [4]
Fig. 1 Example of privacy label [4].

これまでに，日本語のプライバシポリシを用いて，ポリシ要約技術を評価した研究は，著者らの知る限りない．本論文では，日本語のプライバシに対する有効性について調査する．また，プライバシポリシのリスクを評価する研究について，完全性を考慮している研究がない．完全性が欠如している，すなわちリスクを判断するのに十分な情報が記載されていないプライバシポリシについては，適切にリスク評価することはできない．たとえば，収集するパーソナルデータの種類について明記されていない場合，プライバシポリシから，メールアドレスの扱いに関するリスクについて評価することはできない．

1.2 本論文の貢献

本論文では，既存研究の課題に対し，下記のアプローチで解決を目指す．

日本語のプライバシポリシに対する要約技術の評価

既存のアプローチが日本語のプライバシポリシに対しても有効であることを示す．日本語のプライバシポリシを用いて要約技術を評価をするにあたり，日本語のコーパスが必要となる．しかしながら，オープンアクセス可能な日本語のプライバシポリシに関するコーパスは著者らの知る限り存在しない．そこで，日本語のプライバシポリシを収集し，これに重要情報に関するラベル付与を行い，評価するコーパスを作成する．ラベルの付与にあたり，日本国内と諸外国では法制度が異なるため，記載すべき情報も異なる．本論文では，ラベル付与を行う項目について，国内の状況を考慮して設定する．上記のように作成したコーパスに対し，プライバシポリシのテキスト部の特徴抽出および機械学習を行い，学習モデルの精度評価を行う．特徴抽出アルゴリズムおよび学習アルゴリズムは複数の手法の組合せを評価し，その効果を観察する．本論文では，特徴抽出アルゴリズムとして，Bag-of-Words（以下，BoW），TF-IDF，Doc2Vec，学習アルゴリズムとして，サポートベクタマシン（以下，SVM）とランダムフォレストを採用した場合のラベル推測精度について，F1 値により評価を行う．

その結果，作成したコーパスに対し最も効果のあったアルゴリズムの組合せを特定した．精度については，同じコーパスを使用していないため直接比較することはできないが，英語のコーパスを使用した既存研究で達成されていたものと同程度の精度を達成した．

完全性を考慮したリスク評価

プライバシポリシからリスクを評価する場合，本来は完全性を評価したうえで，十分高い完全性を持つプライバシポリシに対してのみリスク評価を実施するようにすべきである．そこで本研究では，プライバシポリシのラベル付与に関して，完全性に関する項目とプライバシリスクに関する項目，双方についてラベルを付与することとする．これにより，提供するプライバシポリシの品質が低く，リスク評価が困難な事業者についても，プライバシポリシの品質が低いことに対して注意喚起することができる．また，GDPR のガイドライン等に即してラベル項目を設定した場合には，ガイドラインに準拠しているプライバシポリシであるかどうか判定することができる．

1.3 本論文の構成

本論文の構成は以下のとおりである．2 章で，本研究で扱うプライバシポリシについて述べる．3 章で，関連する既存研究について述べる．4 章で，本論文で評価を行う，テキストの特徴抽出アルゴリズムと，ラベル推測に用いる学習アルゴリズムについて述べる．5 章で，本論文で扱うプライバシポリシのラベル項目について述べる．6 章で，評価のために構築したコーパスについて述べる．7 章で，プライバシポリシのラベル推測の評価方法および評価結果について述べる．8 章で，評価結果についての考察について述べる．9 章で，本論文をまとめる．

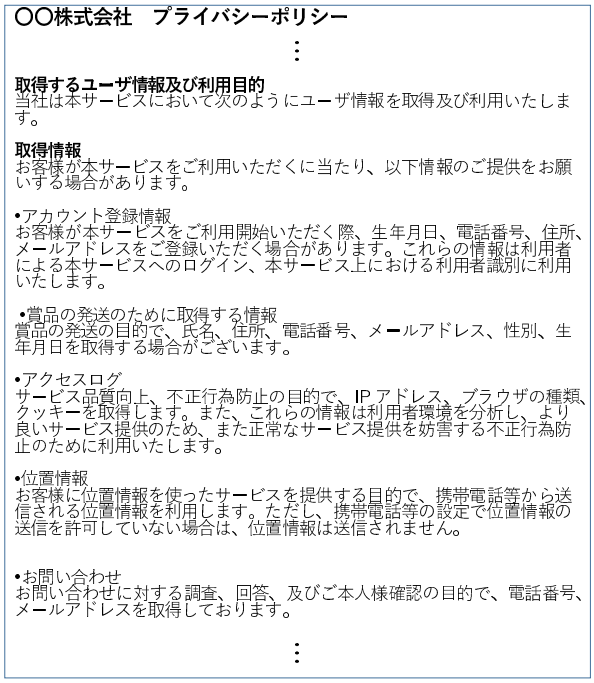
2. プライバシポリシ

2.1 概念

McDonald らによれば，プライバシポリシは，企業が自身の（プライバシに対する）プラクティスを開示するメカニズムであり，OECD（経済協力開発機構）が定めた “Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data” [10] が基盤となっている [2]．このガイドラインでは，以下の 8 つの原則を定めている．

(1) **収集制限の原則**
(2) **データ内容の原則**
(3) **目的明確化の原則**
(4) **利用制限の原則**
(5) **安全保護の原則**
(6) **公開の原則**
(7) **個人参加の原則**
(8) **責任の原則**

わが国の個人情報保護法においても，第 15 条（利用目的の特定）や第 18 条（取得に際しての利用目的の通知等）等により，データの利用目的を明確にし，通知，公表を行うことが義務付けられており，これを実現する手段の 1 つとしてプライバシポリシを公表していると考えられる．消費者基本法第 2 条（基本理念）においても，消費者保護の観点から，合理的な選択の機会を確保するべきであると記述されている．プライバシポリシには，消費者がプライバシリスクを評価し，合理的な選択をするに十分な情報が記載されているべきである．また，EU においても，2018 年 5 月より，一般データ保護規則（GDPR）が施行されている．GDPR 第 12 条第 1 項において，「管理者は，データ主体に対し，簡潔で，透明性があり，理解しやすく，容易にアクセスできる方式により，明確かつ平易な文言を用いて，取扱いに関する第 13 条及び第 14 条に定める情報並びに第 15 条から第 22 条及び第 34 条に定める連絡を提供するために，特に，子どもに対して格別に対処する情報提供のために，適切な措置を講じる．その情報は，書面により，又は適切であるときは電子的な手段を含めその他の方法により，提供される．データ主体から求められたときは，当該データ主体の身元が他の手段によって証明されることを条件として，その情報を口頭で提供できる．」と述べられている．このことから，今後はより理解しやすい形式でプライバシポリシを表示することが求められる可能性が高い．


〇〇株式会社　プライバシーポリシー

⋮

取得するユーザ情報及び利用目的
当社は本サービスにおいて次のようにユーザ情報を取得及び利用いたします。

取得情報
お客様が本サービスをご利用いただくに当たり、以下情報のご提供をお願いする場合があります。

•アカウント登録情報
お客様が本サービスをご利用開始いただく際、生年月日、電話番号、住所、メールアドレスをご登録いただく場合があります。これらの情報は利用者による本サービスへのログイン、本サービス上における利用者識別に利用いたします。

•賞品の発送のために取得する情報
賞品の発送の目的で、氏名、住所、電話番号、メールアドレス、性別、生年月日を取得する場合がございます。

•アクセスログ
サービス品質向上、不正行為防止の目的で、IP アドレス、ブラウザの種類、クッキーを取得します。また、これらの情報は利用者環境を分析し、より良いサービス提供のため、また正常なサービス提供を妨害する不正行為防止のために利用いたします。

•位置情報
お客様に位置情報を使ったサービスを提供する目的で、携帯電話等から送信される位置情報を利用します。ただし、携帯電話等の設定で位置情報の送信を許可していない場合は、位置情報は送信されません。

•お問い合わせ
お問い合わせに対する調査、回答、及びご本人様確認の目的で、電話番号、メールアドレスを取得しております。

⋮

図 2　プライバシポリシの例（一部）
Fig. 2　Example of privacy policy.

2.2　プライバシポリシの例

プライバシポリシの例を**図 2** に示す．ここには，「取得情報」として，「アカウント登録情報」，「商品の発送のために取得する情報」，「アクセスログ」，「位置情報」，「お問い合わせ」といったカテゴリの情報が示されており，たとえば「アカウント登録情報」として，「生年月日」，「電話番号」，「住所」，「メールアドレス」が具体的に明記されている．

本論文においては，プライバシポリシを，(1) トピックの有無，(2) 具体的な記述，の 2 つの観点から評価することを想定する．1 つ目のトピックの有無は，プライバシポリシに記載されるべき事項について記述が存在するかどうかという観点での評価である．たとえば，プライバシポリシに，「収集するデータ項目」のトピックについて記載があるべきであると考えたとき，先に例示したプライバシポリシにはこのトピックについて記載があるため，このプライバシポリシは「収集するデータ項目」のトピックを持つといえる．もう 1 つの具体的な記述は，プライバシポリシがトピックを持つとき，そのトピックにどのような記述が存在するかという観点での評価である．たとえば，先に例示したプライバシポリシには，「メールアドレス」を収集することを記述しているが，「クレジットカード情報」を収集することは記載していない．「収集するデータ項目」のトピックを持ち，「クレジットカード情報」について具体的な記述がないことから，この会社は「クレジットカード情報」を収集しないことが期待できる．

表 1　Guntamukkala らのトピック
Table 1　Topics in Guntamukkala's paper.

カテゴリ	概要
1. アクセス	パーソナルデータを修正する手段
2. 選択	第三者提供等のオプトインまたはオプトアウトに関する議論
3. 収集	収集するパーソナルデータおよび収集手段
4. クッキー	クッキーや履歴を追跡するためのそのほかの技術
5. 利用目的	パーソナルデータの利用目的
6. 保管	パーソナルデータの保管の目的，期間，理由
7. セキュリティ	採用しているセキュリティ技術や Web サイトのセキュリティプラクティス
8. 共有	パーソナルデータの外部機関への共有

3.　既存研究

3.1　プライバシポリシの完全性

Costante らは，自然言語処理と機械学習を用いて，プライバシポリシの完全性を評価する手法を提案した [1]．ここで完全性とは，プライバシポリシがトピックを網羅していることを示す性質であり，具体的な記述について評価するものではないことに注意したい．Costante らは，特徴抽出アルゴリズムとして，TF-IDF を用い，学習アルゴリズムとして k-最近傍法，SVM，LSVM，決定木を用い，パラグラフについての完全性推測を評価した．Guntamukkala らもまた同様に，プライバシポリシの完全性推測を提案および評価した [6]．Guntamukkala らは，特徴抽出アルゴリズムとして，TF-IDF を用い，学習アルゴリズムとして k-最近傍法，LSVM，ランダムフォレストを用い，プライバシポリシ全体についての完全性推測を評価した．たとえば，Guntamukkala らによる評価では，**表 1** に示す 8 つのカテゴリをトピックとして評価する．先述のとおり，これらの研究はトピックを網羅しているかどうかを推測することを目的としており，プライバシポリシ本文から重要情報を抽出することを目的としておらず，重要情報の表形式での表示に直接用いることはできない．

3.2　プライバシポリシのリスク評価

Zaeem らは，自然言語処理と機械学習を用いて，プライバシリスクをプライバシポリシから自動的に評価する手法を提案した [7]．Zaeem らは，特徴抽出アルゴリズムとして，TF-IDF，およびそれに加えてさらに目視でフィルタリングする手法を用いた．学習アルゴリズムとして LSVM，k-最近傍法，ランダムフォレストを用い，プライバシポリシ全体についてのリスク推測について評価した．Zaeem らは，**表 2** に示す 10 項目を，プライバシリスクの評価指標

表 2 Zaeem らのプライバシリスク項目
Table 2 Items of privacy risk in Zaeem's paper.

1. メールアドレス
2. クレジットカード番号
3. 社会保障番号
4. 広告への利用
5. 位置情報
6. 子供の個人情報収集
7. 司法機関への情報共有
8. ポリシーの変更
9. データコントロール
10. データ集積

として定義している．たとえば 1，2，3，5 番目の項目については，それぞれ緑，黄，赤の三段階でリスクの評価を行い，緑は提供を要求しない，黄は本来のサービス利用のためにのみ使用する，赤は第三者と共有する，と記載されていることを表す．

Tesfay らは，GDPR に基づいたプライバシポリシの自動的なリスク評価手法および評価結果を表示する GUI を提案した [8]．Tesfay らは特徴抽出アルゴリズムとして TF-IDF，学習アルゴリズムとしてナイーブベイズ，SVM，決定木，ランダムフォレストを用いて評価した．Harkous ら [9] は，Online Privacy policies（OPP-115）データセット [11] のラベル推測を行う手法を提案した．プライバシポリシの特徴量抽出において，fastText と呼ばれる Document Embedding を用い，学習アルゴリズムには CNNs（Convolutional Neural Networks）を用いた．

これらの手法は，評価したラベル項目をプライバシポリシのラベル項目と見なし，表形式のような理解しやすい表示形式に適用することができる．しかしながら，日本語のプライバシポリシに対して評価を行った既存研究は，著者らの知る限りない．また，ラベル付与にあたり，品質の低いプライバシポリシについては，リスクを判断するのに十分な情報が記載されておらず，ラベルが適切に付与できない問題があるが，このようなポリシのリスク度の扱いが明確ではない．

3.3 日本語のプライバシポリシでの評価

上記の既存研究は英語のコーパスを用いて評価されており，日本語のコーパスで評価している研究は少ない．Fukushima らは特徴抽出アルゴリズムとして Bag-of-Words を用い，学習アルゴリズムとしてナイーブベイズを用いて，プライバシポリシに含まれる文が言及する内容のカテゴリ（データ種別，利用目的，そのほか）を推測する手法を評価した [12]．Fukushima らが文単位で，その文が言及しているカテゴリを推測することを目的にしているのに対し，本論文ではポリシ全体に対してトピックの有無の推測，およびトピックを持つ場合に具体的な記述の有無を推測することを目的にしている点が異なる．また，文献 [12] ではリスクの評価を行っていない点も異なる．リスクに関する記述は複数の文にまたがって記載されていることが多く，文単位のリスク評価ではポリシ全体についての妥当な結果が得られないことが懸念される．そのため本論文では文単位ではなくポリシ単位での評価を行うこととした．

4. 要素技術

本章では，本論文で評価を行うテキストの特徴抽出アルゴリズムと，ラベル推測に用いる学習アルゴリズムについて述べる．

4.1 特徴抽出アルゴリズム

プライバシポリシのような文書を入力として機械学習を行うためには，文書の特徴を抽出し，ベクトルに変換する必要がある．よく知られた方式として，BoW や TF-IDF があるが，順序の情報を保持できないことや，単語間の意味や文法上の関係を考慮できない問題点がある．また，特徴ベクトルが，次元数が辞書として用いる単語数の疎行列となるため，巨大なコーパスを扱うことは困難である．上記の問題点を解決する手法として，Doc2Vec 等の Document Embedding がある．Document Embedding とは，単語や文書を，意味や文法を扱える形式で実数上のベクトル空間に写像する手法である．本章では以下，BoW，TF-IDF，および Doc2Vec について概要を述べる．

4.1.1 Bag-of-Words（BoW）

BoW は最もよく知られたテキストの特徴抽出アルゴリズムである．BoW では，変換後のベクトルの要素として，変換するテキストに出現する単語の出現頻度を使用する．BoW ではまず，すべてのテキストに出現する単語を辞書に登録する．辞書に登録された単語数を N としたとき，変換後のベクトル長は N となる．ベクトルの要素はそれぞれ辞書に登録された単語に対応する．あるテキストをベクトルに変換する際には，そのテキストに出現する単語に該当するベクトルの要素を，その単語の出現頻度とする．BoW による特徴抽出の例として，以下の 3 テキストに対して BoW を適用した場合を示す．

- txt1：私は数学が得意です．スポーツも得意です．
- txt2：私は数学が得意です．スポーツは苦手です．
- txt3：私は国語の教師です．スポーツは趣味です．

ここでは，辞書に名詞のみ登録することとする．このとき，辞書に登録される単語は，「私」，「数学」，「得意」，「スポーツ」，「苦手」，「国語」，「教師」，「趣味」の 8 単語であり，変換後のベクトル長は 8 となる．**表 3** に，各テキストの変換後のベクトルを示す．

BoW を適用することで，$\mathrm{txt1} = (1, 1, 2, 1, 0, 0, 0, 0)$ のように，テキストをベクトルで表現することができる．

表 3 BoW による特徴抽出の例
Table 3 Example of feature extraction with BoW.

	txt1	txt2	txt3
私	1	1	1
数学	1	1	0
得意	2	1	0
スポーツ	1	1	1
苦手	0	1	0
国語	0	0	1
教師	0	0	1
趣味	0	0	1

4.1.2 TF-IDF

TF-IDF は，BoW 同様，テキストに出現する単語の出現頻度に基づく特徴抽出アルゴリズムである．TF-IDF では，多くのテキストに出現する単語は重要度が低く，少数のテキストにしか出現しない単語は重要度が高くなるように，IDF（Inverse Document Frequency）という指標を用いている．まず，TF（Term Frequancy）について，テキスト t_j における単語 w_i の出現頻度を $n_{i,j}$ とすると，テキスト t_j における単語 w_i の $tf_{i,j}$ は，

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

と表される．ここで，$\sum_k n_{k,j}$ は，テキスト t_j に含まれるすべての単語の出現回数である．単語 w_i における idf_i は，T をすべてのテキスト集合とすると，

$$idf_i = \frac{|T|}{|\{t : t \ni w_i\}|}$$

と表される[*1]．ここで，$|\{t : t \ni w_i\}|$ は，単語 w_i を含むテキストの数である．テキスト t_j における単語 w_i の $tfidf_{i,j}$ は，

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i$$

と表され，これを変換後のベクトルの要素として用いる．

表 3 の txt2 を例とすると，「私」という単語が 1 回出現し，txt2 に含まれる単語の総数が 5 であることから，$tf = 1/5$ である．idf は，「私」という単語は 3 つのテキストすべてに出現するため，$3/3 = 1$ となるが，「苦手」という単語は txt2 内のみ出現するため，$3/1 = 3$ となる．このように，IDF が出現頻度に応じた一種の重みとなっていることが分かる．以上から，txt2 における「私」の tfidf は $1/5 \cdot 1 = 1/5$ である．同様にその他の単語についても変換を行うことができ，このとき txt2 $= (1/5, 3/10, 3/10, 1/5, 3, 0, 0, 0)$ となる．

4.1.3 Doc2Vec

BoW や TF-IDF の問題点として，順序の情報を保持できないことや，単語間の意味や文法上の関係を考慮できない点がある．表 3 の txt2 を例とすると，BoW でベクトル変換した場合，「数学が得意でスポーツが苦手」なのか「スポーツが得意で数学が苦手」なのかといった情報は失われている．n 個の単語の組合せをベクトルの要素とする n-gram という概念が使われることもあるが，ベクトル空間が増大するため，巨大なコーパスを扱うことは困難である．また，特徴ベクトルが，次元数が辞書として用いる単語数の疎行列となるため，このこともまた巨大なコーパスを扱うことを困難にする．

上記の問題点を解決する手法として，Doc2Vec 等の Document Embedding がある．Doc2Vec [13] はニューラルネットワークベースの Document Embedding アルゴリズムであり，特徴として，任意の次元数のベクトルに変換できる点，似た文脈で使用された単語は，距離の近いベクトルで表現できる点がある．Doc2Vec は PV-DM（Distributed Memory Model of Paragraph Vectors）と PV-DBOW（Distributed Bag of Words version of Paragraph Vector）の 2 つの手法をサポートしている．PV-DM は Word2Vec と呼ばれる Word Embedding アルゴリズムをベースとしている．一般に，PV-DBOW はメモリ効率性に優れるが，精度は PV-DM に劣る．本論文では計算効率については考慮しないため，高精度が期待できる PV-DM のみ評価を行う．

4.2 学習アルゴリズム

本論文で扱うプライバシポリシのラベルは，あるプライバシに関する事項について，その記載があるかどうかを表すものである．そこで，本論文で扱う学習アルゴリズムは，2 値分類を行う学習アルゴリズムとなる．本論文では，サポートベクタマシン（SVM）およびランダムフォレストについて評価を行う．

4.2.1 サポートベクタマシン（SVM）

SVM は基本的には，2 値分類を行う線形分類器である．SVM の特徴は，データ群を 2 クラスに分類する境界となる超平面（境界面）について，マージンが最大になる境界面を探索する点にある．ここでマージンとは，境界面に最も近い点から境界面までの距離を指す．SVM は基本的には，線形分離可能な場合にしか適用できないが，ソフトマージンやカーネル法の導入により，境界面からデータが逸脱する場合や，境界面が非線形の場合についても適用可能となった [14]．ソフトマージンを導入する場合には，境界面から逸脱したデータをどの程度許容するかを示すパラメータ（コストと呼ぶ）が指定される．カーネルについては，RBF（Radial Basis Function）カーネル（ガウシアンカーネル）がよく用いられる．RBF カーネルは，以下の式で表される．

$$K(x, x') = \exp(-\gamma \parallel x - x' \parallel^2)$$

[*1] 実際には対数をとることが多く，本論文の実験においても対数をとる．ここでは説明を簡単にするため，対数をとらずに説明する．

パラメータγは境界面の複雑さを表し，小さいほど単純な境界面となり，大きいほど複雑な境界面となる．

4.2.2 ランダムフォレスト

ランダムフォレストは，決定木に基づくアンサンブル学習アルゴリズムの一種である [15]．アンサンブル学習とは，複数の学習器を組み合わせることにより高い推測精度を達成する学習アルゴリズムである．ランダムフォレストにおいては，学習データの一部をランダムに抽出することを繰り返して複数のサブセットを生成し，各サブセットごとに決定木を生成する．未知のデータに対してクラスを推測する際には，すべての決定木における推測結果を算出し，最後に投票を行うことによって最終的な推測結果を決定する．

5. プライバシポリシのラベル項目

本論文では，**表 4** で示す項目についてラベリングを実施する．子項目，孫項目については，それぞれ自身の親項目，子項目が有だった場合のみラベリングする．親項目については，以下の 12 項目を設定し，これが本論文におけるトピックに該当する．子項目，孫項目は，具体的な記述に該当する．

- **D. データ項目記述**：収集するデータ項目について具体的な記述の有無を示す．「個人情報」のように具体的なデータ項目が特定できない場合は記述なしとする．子項目として，E メールアドレス，クレジットカード情報，位置情報，Web 閲覧操作履歴の収集の有無を示す項目を持つ．
- **P. 利用目的記述**：データ収集の目的について具体的な記述の有無を示す．子項目として，広告配信への利用の有無を示す項目を持つ．
- **T. 第三者提供に関する記述**：収集したデータの第三者提供について，具体的な記述の有無を示す．子項目として，第三者提供するデータ項目に関する記述の有無と，利用目的に関する記述の有無を示す項目を持つ．これらは完全性の評価に相当する．さらに，孫項目として，データ項目については E メールアドレス，クレジットカード情報，位置情報，Web 閲覧操作履歴の収集の有無を示す項目を，利用目的については広告配信への利用の有無を示す項目を持つ．
- **S. 共同利用記述**：収集したデータの共同利用についての記述の有無を示す．
- **O. 業務委託記述**：収集したデータの業務委託についての記述の有無を示す．
- **An. 匿名加工情報記述**：匿名加工したデータ利用についての記述の有無を示す．
- **Ch. ポリシ変更時の記述**：事業者がポリシを変更した場合の同意の扱いについての記述の有無を示す．子項目として，変更された時点で自動的に適応されることがないとの記述の有無を示す項目を持つ．
- **Pe. 開示修正削除の記述**：個人情報の開示修正削除について記述の有無を示す．子項目として，開示，修正，

表 4 プライバシポリシのラベル項目
Table 4 Labels of privacy policies.

親項目	子項目	孫項目
D. データ項目記述	D-E. E メールアドレス収集	
	D-Cr. クレジットカード情報収集	
	D-Lo. 位置情報収集	
	D-W. Web 閲覧操作履歴収集	
P. 利用目的記述	P-Ad. 広告配信	
T. 第三者提供に関する記述	T-D. 第三者提供するデータ項目記述	T-D-E. E メールアドレス
		T-D-Cr. クレジットカード情報
		T-D-Lo. 位置情報
		T-D-W. Web 閲覧操作履歴
	T-P. 第三者提供する目的記述	T-P-Ad. 広告配信
S. 共同利用記述		
O. 業務委託記述		
An. 匿名加工情報記述		
Ch. ポリシ変更時の記述	Ch-N. 変更した時点で自動的に適用されない	
Pe. 開示修正削除の記述	Pe-Di. 開示可能	
	Pe-M. 修正可能	
	Pe-De. 削除可能	
Q. 問い合わせ窓口記述		
Se. セキュリティ関連記述		
Co. 自動収集・外部収集記述		
I. 収集機能無効化記述		

表 5 Guntamukkala らが設定した項目 [6] との対応
Table 5 Correspondence table with items in Guntamukkala's paper [6].

文献 [6] におけるカテゴリ	対応する親項目
1. アクセス	Q. 問い合わせ窓口記述
2. 選択	T. 第三者提供に関する記述
3. 収集	D. データ項目記述 Co. 自動収集・外部収集記述 I. 収集機能無効化記述
4. クッキー	(D-W. Web 閲覧操作履歴収集)
5. 利用目的	P. 利用目的記述
6. 保管	
7. セキュリティ	Se. セキュリティ関連記述
8. 共有	T. 第三者提供に関する記述 S. 共同利用記述 O. 業務委託記述

削除が可能である記述を示す項目を持つ．

- **Q. 問い合わせ窓口記述**：問い合わせを行うための窓口や連絡先について記述の有無を示す．
- **Se. セキュリティ関連記述**：個人情報を保護するセキュリティ技術に関する記述の有無を示す．
- **Co. 自動収集・外部収集記述**：個人情報の自動収集や外部収集に関する記述の有無を示す．
- **I. 収集機能無効化記述**：個人情報の収集機能を無効化する手続きについて記述の有無を示す．

項目の選定は，Guntamukkala ら [6] が用いた項目と，Zaeem ら [7] が用いた項目をベースに行った．**表 5** は，Guntamukkala ら [6] が用いた項目と，本論文でラベリングを実施する項目の対応を示している．「6. 保管」に対応する項目は，予備調査において，記述されているポリシがなかったため，本論文では設定しなかった．「4. クッキー」に対応する項目は，親項目ではなく，子項目である「D-W. Web 閲覧操作履歴収集」となる．「Pe. 開示修正削除の記述」およびその子項目，「D. データ項目記述」，「P. 利用目的」の子項目は，Zaeem ら [7] が用いた項目を用いている．

「An. 匿名加工情報記述」については，日本の個人情報保護法特有の事情によるものであり，文献 [6] では扱っていないが，本論文ではプライバシポリシに記載すべき内容であると考え，ラベリングを行う項目として設定した．「Ch. ポリシ変更時の記述」については，近年，ポリシ変更時にオプトアウトにより自動的に変更後のポリシが適用されることで利用者に不利益を与える懸念があるとして議論を呼んでいる[*2]．したがって，「ポリシ変更時の記述」があるか否かはプライバシポリシの重要な特徴と考え，参考情報としてこの項目を追加した．実際には，ユーザが懸念する項目はその他にも多く考えられるが，ここではすべての項目を網羅することを目的とせず，手法の妥当性を確認することのみを目的とし，一部の項目について評価を行う．

[*2] Evernote における一例：https://www.rbbtoday.com/article/2016/12/16/148016.html

6. コーパス

本論文で扱うコーパスはプライバシポリシのテキスト部と，プライバシポリシに付与されたラベル部から構成される．

6.1 プライバシポリシ収集

プライバシポリシは，Android アプリ，iOS アプリ，Web サービス，国内企業の 4 カテゴリから収集した．プライバシポリシの収集は，2017 年 7 月 31 日に実施した．アプリやサービス，企業の選択は以下のように実施した．

- Android アプリ：Google Play の「無料トップ Android アプリ」ランキングおよび「新着有料トップ Android アプリ」ランキングから上位のアプリを選択．
- iOS アプリ：Apple Store の「無料 App」ランキングおよび「有料 App」ランキングから上位のアプリを選択．
- Web サービス：Alexa Top の日本版を用いて，アクセス数が上位 200 のサービスを選択．
- 国内企業：日経 225 を構成する企業を選択．

プライバシポリシの探索は，検索エンジンおよびクローリングを用いてプライバシポリシが記載されている Web ページの候補を推測し，URL と HTML ファイルを自動的に収集した．後に作業者がブラウザを用いて目視による確認を行う．対応する URL が示す Web ページの確認を行い，真にプライバシポリシが記述されている Web ページを表すテキストのみコーパスとして扱う．クローリング時に文字化け等の不適切なデータが取得された場合にはこれを除外した．その結果，計 329 のプライバシポリシが取得できた．

6.2 ラベリング

ラベリングは，作業者によって目視で行った．作業者はまず，ブラウザを用いて収集した URL が示す Web ページを閲覧し，対象のプライバシポリシが含まれているかどうかを確認した．対象のプライバシポリシが含まれていた場合には，そのプライバシポリシに対してラベリングを行った．日本語のプライバシポリシが存在しない場合については除外した．

表 6 に，Android，iOS，Web サービス，国内企業から収集したプライバシポリシ数の内訳を示す．異なるカテゴリで同一のプライバシポリシが出現した場合には，同一のプライバシポリシを重複してカウントすることを避けるため，先のカテゴリにのみカウントすることとした．たとえば Android アプリと iOS アプリで同一のプライバシポリシを参照していた場合は，Android アプリにのみカウントする．収集したポリシのラベルの内訳を**表 7** に示す．

表 7 収集したプライバシポリシのラベル内訳
Table 7 Breakdown of the number of collected privacy policies.

ラベル	D	D-E	D-Cr	D-Lo	D-W	P	P-Ad	T	T-D	T-D-E	T-D-Cr	T-D-Lo
記述有	218	134	54	57	158	230	187	114	67	5	8	7
記述無	111	84	164	161	60	99	43	215	47	62	59	60
ラベル	T-D-W	T-P	T-P-Ad	S	O	An	Ch	Ch-N	Pe	Pe-Di	Pe-M	Pe-De
記述有	55	108	66	141	245	24	148	15	278	267	263	213
記述無	12	6	42	188	84	305	181	133	51	11	15	65
ラベル	Q	Se	Co	I								
記述有	263	273	93	120								
記述無	66	56	236	209								

表 6 収集したプライバシポリシ数
Table 6 The number of collected privacy policies.

カテゴリ	ポリシ数
Android	52
iOS	35
Web サービス	97
国内企業	145
合計	329

6.3 テキスト部の前処理

まず，収集した HTML ファイルからコメントタグや script タグ，CSS/スタイルシート部分を除去し，プライバシポリシが記述されていることが期待されるテキスト部のみ抽出した．次に，テキスト部から，正規表現を用いた英数記号や特殊文字の除去，および SlothLib [16] によるストップワードの除去を行う．最後に，python の日本語形態素解析ライブラリである janome [17] を用いて，形態素解析を行う．ここでは，名詞，動詞，形容詞（形容動詞も含まれる），助動詞の基本形（base_form）のみ抽出し，それ以外の品詞については除去することとする．

本コーパスは，9,723 種類の単語，約 128 万語から構成される．

7. ラベル推測手法の評価

本論文で扱うラベル推測手法の評価手順の概要を以下に示す．

(1) 特徴量抽出アルゴリズムを用いて，コーパスのテキスト部を特徴ベクトルに変換する．
(2) コーパスの一部を教師データとし，学習モデルを作成する．
(3) コーパスの残りを評価データとし，精度として F1 値を算出する．

本論文では，プライバシポリシの特徴量抽出アルゴリズムとして，BoW，TF-IDF，Doc2Vec を用いる．学習アルゴリズムには，SVM とランダムフォレストを用いる．BoW，TF-IDF，SVM，およびランダムフォレストの実装には scikit-learn [18] を用いた．Doc2Vec の実装には gensim [19] を用いた．

7.1 特徴量抽出アルゴリズムのパラメータ

まず，各特徴量抽出アルゴリズムのパラメータチューニングについて述べる．BoW および TF-IDF については，max-df と min-df についてパラメータチューニングを行う．max-df，min-df は単語の出現頻度に関するパラメータであり，max-df は，全文書に対し，ある単語の出現する文書が指定した引数以上の割合（整数値の場合は文書数）であった場合に，その単語を削除する．同様に，min-df は，ある単語の出現する文書が指定した引数以下の割合であった場合に，その単語を削除する．その他のパラメータについてはデフォルトとした．

Doc2Vec については，vec_size と window についてパラメータチューニングを行う．vec_size は，出力するベクトル長を決定するパラメータである．window は同じ文脈として考慮する前後の単語数を決定するパラメータである．また，max_epochs = 30, min_count = 1, min_alpha=0.00025 に固定し，そのほかのパラメータについてはデフォルトとした．

7.2 学習アルゴリズムのパラメータ

SVM ではまず kernel パラメータを選択する．kernel は linear と rbf があり，linear の場合は線形カーネル，rbf の場合は RBF カーネルを用いる．linear の場合はコスト C についてパラメータチューニングを行う．rbf の場合はコスト C に加え，gamma についてチューニングを行う．その他のパラメータについてはデフォルトとした．

ランダムフォレストでは，max_features，max_depth，min_sample_leaf についてパラメータチューニングを行う．max_features は考慮する特徴の数を指定するパラメータである．max_depth は決定木の深さの最大値を指定するパラメータである．min_sample_leaf は葉を構成するのに必要な最小限のサンプル数を指定するパラメータである．アンセンブル学習のために生成する決定木の数を示す n_estimators は 100 に固定した．そのほかのパラメータはデフォルトとした．

表 8 グリッドサーチのパラメータ
Table 8 Parameters for grid search.

SVM	'kernel':linear	'C': [0.1, 1, 10, 100, 1000]
	'kernel':rbf	'C': [0.1, 1, 10, 100,1000]
		'gamma':[0.0001, 0.001, 0.01, 0.1]
ランダムフォレスト		'max_features':[1, 'auto', None]
		'max_depth':[1, 3, 5, None]
		'min_samples_leaf':[1, 2, 4]

表 9 min-df，max-df を変化させた場合の SVM の F1 値
Table 9 F1-values of SVM in changing min-df and max-df.

BoW+SVM

max/min	1	3	5	7
0.7	0.804	0.804	0.805	0.801
0.8	0.8	0.803	0.804	0.805
0.9	0.802	0.8	0.797	0.797
1	0.803	0.802	0.802	0.803

TF-IDF+SVM

max/min	1	3	5	7
0.7	0.802	0.801	0.8	0.799
0.8	0.807	0.806	0.801	0.799
0.9	0.801	0.796	0.798	0.793
1	0.793	0.796	0.796	0.784

7.3 BoW および TF-IDF を用いた場合の評価

まず max-df と min-df を変化させた場合の，BoW および TF-IDF の性能について評価を行った．max-df は {0.7, 0.8, 0.9, 1.0} で変化させ，min-df は {1, 3, 5, 7} で変化させた．上記のパラメータにて文書をベクトル化し，これを入力として，SVM およびランダムフォレストを用いて学習および評価を行った．このとき，それぞれ**表 8** に示すパラメータを用いてグリッドサーチを実施し，最も良い精度を達成したパラメータを用いて生成されたモデルを決定した．精度については，既存研究と同様に，項目ごとに評価を行うこととした．学習データと評価データの比率は 3 : 1 とし，ランダムに 10 回データ分割を行い，すべての項目について F1 値を算出し，これを評価指標とした．ただし，記述有のラベルが少数の項目については，評価データの正解値および推測値がすべて negative となり，F1 値が 0 となる場合がある．本評価ではこのようなケースは除外することとした．**表 9** に SVM の結果を，**表 10** にランダムフォレストの結果を示す．SVM の結果から，BoW の場合は，max-df = 0.8，min-df = 7 の場合で最も精度が高く，0.805 となった．TF-IDF の場合は，max-df = 0.8，min-df = 1 の場合で最も精度が高く，0.807 となった．

ランダムフォレストの結果からは，BoW の場合は，max-df = 1.0，min-df = 5 の場合で最も精度が高く，0.861 となった．TF-IDF の場合は，同様に max-df = 1.0，min-df = 3 の場合で最も精度が高く，0.861 となった．

表 10 min-df, max-df を変化させた場合のランダムフォレストの F1 値
Table 10 F1-values of Random Forest in changing min-df and max-df.

BoW+RF

max/min	1	3	5	7
0.7	0.855	0.854	0.854	0.853
0.8	0.86	0.86	0.858	0.86
0.9	0.86	0.859	0.86	0.857
1	0.859	0.86	0.861	0.855

TF-IDF+RF

max/min	1	3	5	7
0.7	0.855	0.855	0.855	0.851
0.8	0.86	0.861	0.859	0.857
0.9	0.859	0.859	0.86	0.857
1	0.859	0.861	0.851	0.855

表 11 vec_size, window を変化させた場合の SVM およびランダムフォレストの F1 値
Table 11 F1-values of SVM and Random Forest in changing vec size and window.

Doc2Vec+SVM

window/vec_size	50	100	200
1	0.783	0.792	0.8
3	0.804	0.796	0.795
5	0.796	0.797	0.792

Doc2Vec+RF

window/vec_size	50	100	200
1	0.793	0.794	0.79
3	0.806	0.802	0.802
5	0.806	0.801	0.8

SVM，ランダムフォレストともに，BoW と TF-IDF での差は小さい．そこで，以後 Doc2Vec との比較は BoW の場合のみ行う．

7.4 Doc2Vec を用いた場合の評価

次に，vec_size と window を変化させた場合の，Doc2Vec の性能について評価を行った．vec_size は {50, 100, 200} で変化させ，window は {1, 3, 5} で変化させた．上記のパラメータにて文書をベクトル化し，これを入力として，SVM およびランダムフォレストを用いて学習および評価を

表 12 項目別の精度比較
Table 12 Comparison of accuracy in each label.

項目名	BoW+SVM	BoW+RF	Doc2Vec+SVM	Doc2Vec+RF
D	0.834	0.865	0.842	0.807
D-E	0.747	0.817	0.768	0.733
D-Cr	0.677	0.807	0.675	0.676
D-Lo	0.682	0.822	0.704	0.729
D-W	0.876	0.944	0.867	0.881
P	0.855	0.87	0.843	0.82
P-Ad	0.903	0.904	0.9	0.902
T	0.674	0.728	0.682	0.629
T-D	0.666	0.711	0.717	0.655
T-D-E	N/A	N/A	N/A	N/A
T-D-Cr	0.667	N/A	N/A	N/A
T-D-Lo	0.533	N/A	N/A	N/A
T-D-W	0.912	0.913	0.92	0.912
T-P	0.966	0.971	0.97	0.971
T-P-Ad	0.782	0.808	0.789	0.78
S	0.764	0.934	0.693	0.719
O	0.855	0.966	0.839	0.847
An	0.499	0.554	0.326	N/A
Ch	0.728	0.781	0.737	0.715
Ch-N	N/A	N/A	0.417	N/A
Pe	0.9	0.945	0.902	0.9
Pe-Di	0.977	0.973	0.976	0.976
Pe-M	0.97	0.97	0.972	0.969
Pe-De	0.863	0.931	0.869	0.864
Q	0.881	0.883	0.881	0.884
Se	0.924	0.928	0.927	0.918
Co	0.709	0.773	0.709	0.63
I	0.684	0.808	0.673	0.624
F1 の平均	0.805	0.861	0.804	0.806

行った．グリッドサーチに用いるパラメータや，学習データの分割，評価指標については BoW, TF-IDF での評価と同様とした．**表 11** に Doc2Vec を特徴量抽出アルゴリズムとして使用した場合の，SVM およびランダムフォレストの結果を示す．この結果から，SVM については最も精度が良いのは window=3，vec_size=50 のときで F1 値が 0.804 となり，またランダムフォレストについては window=3, vec_size=50 のときで F1 値が 0.806 となった．以上から，Doc2Vec を使用した場合，SVM については BoW，TF-IDF と同等の性能を示し，ランダムフォレストについては精度が劣化する結果となった．

8. 考察

8.1 項目別の精度比較

表 12 に，BoW でベクトル化した場合と，Doc2Vec でベクトル化した場合の各ラベル項目の推測を行った場合の F1 値を示す．この結果から，他と比較して特に BoW とランダムフォレストを用いた場合に高い精度となった項目は，D-E，D-Cr，D-Lo，D-W，T，S，O，Co，I 等である．

8.2 各手法の比較

まず，各手法間の統計的な有意差を確認するため，(1) BoW-SVM と Doc2Vec-SVM，(2) BoW-ランダムフォレストと Doc2Vec-ランダムフォレスト，(3) BoW-SVM と BoW-ランダムフォレスト，(4) Doc2Vec-SVM と Doc2Vec-ランダムフォレスト，の 4 通りの組合せに対し，各 10 回ずつの試行における全項目の F1 値の平均について，t 検定を行った．上記のように 4 回検定を繰り返す多重比較となるため，Bonferroni 法による補正を行い，有意水準は $\alpha/N = 0.05/4 = 0.0125$ とする．BoW と SVM を用いた場合と Doc2Vec と SVM を用いた場合の結果は $p = 0.839$ となり，有意差は確認できなかった．BoW とランダムフォレストを用いた場合と Doc2Vec とランダムフォレストを用いた場合の結果は $p = 4.44e^{-5}$ となり，有意差が確認できた．以上から，学習アルゴリズムとして SVM を用いる場合には，特徴量抽出アルゴリズムとして BoW を用いても Doc2Vec を用いても差は見られない．一方で，ランダムフォレストを用いた場合には，特徴量抽出アルゴリズムとして BoW を用いた場合のほうが精度が高いといえる．

表 13 S および Q における重要度が上位である 10 語
Table 13 Top 10 important words in S and Q.

S		Q	
単語	重要度	単語	重要度
共同	0.0647	ます	0.0142
グループ	0.0153	情報	0.0123
個人	0.0091	連絡	0.0105
範囲	0.009	お願い	0.0096
取扱い	0.0081	住所	0.009
目的	0.0078	個人	0.0085
いたす	0.0078	窓口	0.0084
する	0.0075	する	0.0074
利用	0.0074	目的	0.0071
実施	0.0069	利用	0.0071

また，BoW と SVM を用いた場合と BoW とランダムフォレストを用いた場合の結果は $p = 2.00e^{-5}$ となり，有意差が確認できた．Doc2Vec と SVM を用いた場合と Doc2Vec とランダムフォレストを用いた場合の結果は $p = 0.68$ となり，有意差は確認できなかった．上記の結果と 7 章から，BoW を用いた場合には，ランダムフォレストを用いた場合のほうが精度が高く，Doc2Vec を用いた場合には，精度の差がなかった．

本実験結果（表 12）から，BoW とランダムフォレストの組合せが，そのほかの組合せより高い精度を達成していることが分かる．以下に，このような傾向となる要因について分析する．まず，他の組合せと比較して特に高い精度となった項目について，「単語の重要度」との関連について調査した．例として**表 13** に，項目「S. 共同利用記述」および項目「Q. 問い合わせ窓口記述」について，max-df = 1.0，min-df = 3 で BoW によって特徴量抽出を行い，ランダムフォレストでモデル生成したときに得られる単語の重要度をそれぞれ上位 10 語に限定して示す．BoW とランダムフォレストの組合せで他の組合せよりも高精度となる項目 S では，最も重要度の高い単語である「共同」の重要度が，2 番目に重要度の高い「グループ」の 4 倍超であるのに対し，どのモデルでも同程度の精度となっている項目 Q では，著しく重要度の高い単語は存在しなかった．上記の傾向と精度の関係について他の項目についても確認するため，項目全体を，BoW とランダムフォレストの組合せが他の組合せと比べて，顕著に高精度である項目（A 群），やや高精度である項目（B 群），同程度である項目（C 群），の 3 つに分類し，各項目の上位 5 番目までの単語の重要度の傾向を**図 3** に示した．また，各群に属する項目を**表 14** に示す．図 3 から，BoW とランダムフォレストの組合せにより高い精度となった項目では，少数の単語の重要度が他の単語に対し著しく高い傾向にあることが分かる．

上記の特徴から，SVM とランダムフォレストを比較した場合，重要単語の存在が分類に効果的となる二分木ベース

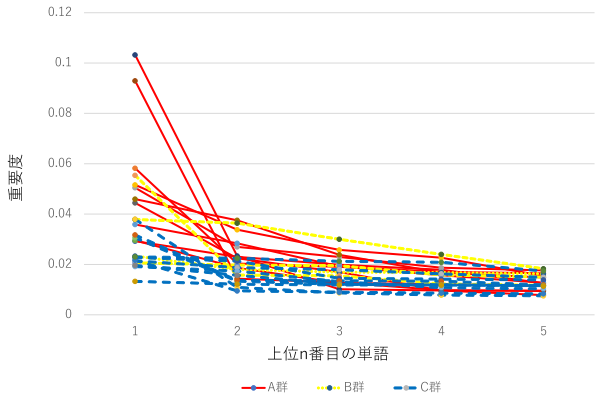


図 3 各項目の上位 5 番目までの単語の重要度
Fig. 3 Top 5 importance of words in each label.

表 14 各群に属する項目
Table 14 Labels belonging to each group.

A 群（顕著に高い）	D-E, D-Cr, D-Lo, D-W, T, T-P-Ad, S, O, P-De, I
B 群（やや高い）	D, An, Ch, Pe
C 群（同程度）	P, P-Ad, T-D, T-D-W, T-P, Pe-Di, Pe-M, Q, Se, Co

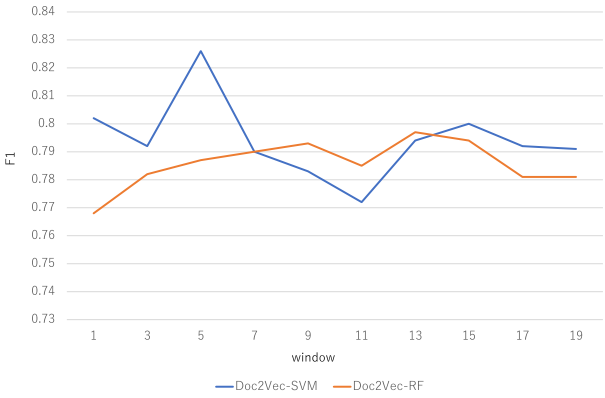


図 4 window と F1 の平均
Fig. 4 Relationship between window and average of F1-value.

のランダムフォレストの方が精度が高くなったものと推測される．一方，BoW と Doc2Vec を比較した場合，Doc2Vec では，コンテキストを考慮した単語の分散表現により単語自身の価値が弱まり，ランダムフォレストでの精度が劣化したと推測される．

また，Doc2Vec が十分な効果を上げられなかった理由として，本研究で扱うコーパスにおいては，順序関係やコンテキストを考慮することによる利得が十分でなかったものと推測される．実際，Doc2Vec の window を変化させた場合の，F1 の平均の変動を調査した**図 4** の結果から，window を変化させても，F1 の変化は小さく，順序関係やコンテキストが項目の推測に与える影響が小さいことが分かる．

8.3 日本語のプライバシポリシへの適用可能性

Zaeem ら [7] の研究では，3 値での推測において，一致

率の平均は 0.60 であった．Tesfay ら [8] の研究では，カテゴリレベルでの推測精度は，3 値での推測において，平均で 0.74 であった．Harkous ら [9] の研究では，カテゴリレベルでの推測精度は，2 値での推測において，F1 値で平均 0.84 であった．それに対し，本論文で評価した結果は，TF-IDF とランダムフォレストを組み合わせた場合で，2 値での推測において，F1 値で平均 0.861 であった．既存手法とはデータセットや推測するラベルの値域が異なるため，直接的な比較はできないが，既存手法と同等以上の精度での推測が実現できており，日本語のプライバシポリシに対しても適用可能性があるといえる．

8.4 本研究の限界

本論文では，329 のプライバシポリシを用いて評価を行った．BoW を用いる場合，特徴ベクトルの次元数が登場する単語数になることを考慮すると，使用するポリシ数が十分大きいとはいえない．しかしながら，人手によりポリシにラベルを付与する作業は労力がかかるため，ポリシ数を増やすことは容易なことではない．実際，たとえば OPP-115 データセット [11] に含まれるプライバシポリシ数は 115 であるため，本論文で扱うポリシ数が既存手法と比較して少ないわけではない．評価するラベル項目については，Harkous ら [9] は，OPP-115 データセットを用いて，約 70 の項目について評価を行っているが，本論文では収集した項目数が 29，評価を行った項目数は 21 と少ない．今後より実際的な評価を行うことを考えたとき，コーパスの拡充は必要である．

9. おわりに

本論文では，日本語のプライバシポリシに対して，プライバシポリシの重要情報を抽出する手法の評価を行い，有効性を明らかにした．日本語のポリシを用いて評価するにあたり，日本語のコーパスが存在しないため，日本語のプライバシポリシを収集し，これにラベル付与を行い，評価するコーパスを作成した．このとき，品質の低いプライバシポリシについては，リスクを判断するのに十分な情報が記載されておらず，ラベルが適切に付与できない問題があることが分かった．そこで本論文では，完全性評価，リスク評価いずれも実施可能なように，双方についてラベルを付与することとした．

特徴量抽出アルゴリズムとして, BoW, TF-IDF, Doc2Vec を用いた場合でラベル推測を行い，比較評価を行った．その結果，本論文で扱うコーパスに対しては，TF-IDF とランダムフォレストを用いた場合に最も精度の高い推測ができることが明らかになり，その精度は約 0.864 となった．この結果は，既存研究と同等以上の精度であるため，本研究で採用した解決方針は，日本語のポリシに対しても，既存研究で達成したと報告されている結果と同程度の効果は期待できるといえる．

参考文献

[1] Costante, E., Sun, Y., Petković, M. and den Hartog, J.: A machine learning solution to assess privacy policy completeness: (short paper), *Proc. 2012 ACM Workshop on Privacy in the Electronic Society, WPES '12*, pp.91–96, ACM (2012).

[2] McDonald, A.M. and Cranor, L.F.: The cost of reading privacy policies, *A Journal of Law and Policy for the Information Society*, Vol.4 (2008).

[3] Reeder, R.W., Kelley, P.G., McDonald, A.M. and Cranor, L.F.: A user study of the expandable grid applied to p3p privacy policy visualization, *Proc. 7th ACM Workshop on Privacy in the Electronic Society, WPES '08*, pp.45–54, ACM (2008).

[4] Kelley, P.G., Bresee, J., Cranor, L.F. and Reeder, R.W.: A "nutrition label" for privacy, *Proc. 5th Symposium on Usable Privacy and Security, SOUPS '09*, pp.4:1–4:12, ACM (2009).

[5] W3C: Platform for privacy preferences (P3P) project, available from ⟨https://www.w3.org/P3P/⟩.

[6] Guntamukkala, N., Dara, R. and Grewal, G.: A machine-learning based approach for measuring the completeness of online privacy policies, *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference*, pp.289–294 (2015).

[7] Zaeem, R.N., German, R.L., Barber, K.S. and German, R.L.: Privacycheck: Automatic summarization of privacy policies using data mining (2016).

[8] Tesfay, W.B., Hofmann, P., Nakamura, T., Kiyomoto, S. and Serna, J.: Privacyguide: Towards an implementation of the eu gdpr on internet privacy policy evaluation, *the 4th ACM International Workshop on Security and Privacy Analytics (IWSPA2018)* (2018).

[9] Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K.G. and Aberer, K.: Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning, *27th USENIX Security Symposium (USENIX Security 18)*, pp.531–548 (2018).

[10] OECD.org: OECD guidelines on the protection of privacy and transborder flows of personal data, available from ⟨http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm⟩.

[11] Wilson, S., Schaub, F., Dara, A.A., Liu, F., Cherivirala, S., Leon, P.G., Andersen, M.S., Zimmeck, S., Sathyendra, K.M., Russell, N.C., Norton, T.B., Hovy, E., Reidenberg, J. and Sadeh, N.: The Creation and Analysis of a Website Privacy Policy Corpus, *Proc. 54th Annual Meeting of the Association for Computational Linguistics*, pp.1330–1340 (2016).

[12] Fukushima, K., Nakamura, T., Ikeda, D. and Kiyomoto, S.: Challenges in classifying privacy policies by machine learning with word-based features, *Proc. 2nd International Conference on Cryptography, Security and Privacy (ICCSP 2018)* (2018).

[13] Le, Q. and Mikolov, T.: Distributed representations of sentences and documents, *Proc. 31st International Conference on International Conference on Machine Learning - Volume 32, ICML'14*, pp.II-1188–II-1196, JMLR.org (2014).

[14] 中川裕志：情報工学機械学習，丸善出版 (2015).

[15] Breiman, L.: Random forests (2001), available from

⟨https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf⟩.
[16] Slothlib: Slothlib wiki, available from ⟨http://www.dl.kuis.kyoto-u.ac.jp/slothlib/⟩.
[17] janome: Janome v0.3 documentation (ja), available from ⟨http://mocobeta.github.io/janome/⟩.
[18] scikit learn: scikit-learn, available from ⟨https://scikit-learn.org/stable/⟩.
[19] gensim, available from ⟨https://radimrehurek.com/gensim/⟩.

中村 徹 （正会員）

2011 年九州大学大学院システム情報科学府情報工学専攻博士後期課程修了．博士（工学）．同年 KDDI（株）入社，同年（株）KDDI 研究所（現（株）KDDI 総合研究所），2018 年（株）国際電気通信基礎技術研究所．2020 年より（株）KDDI 総合研究所情報セキュリティグループに所属．プライバシ保護技術やトラストに関する研究に従事．2016 年コンピュータセキュリティシンポジウム SPT 論文賞受賞．電子情報通信学会会員．

ウェルデルファエル B. テスファイ

2013 年ルレオ工科大学コンピュータ科学科修士課程修了．同年ヨハンヴォルフガングゲーテ大学助手．2019 年ヨハンヴォルフガングゲーテ大学で博士号（コンピュータ科学）取得．現在はヨハンヴォルフガングゲーテ大学主任研究員．プライバシ保護技術，ユーザブルプライバシ，機械学習応用の研究に従事．2018 年 Best demo award of The Web Conference 2018 受賞．

バネッサ ブラカモンテ

2016 年総合研究大学院大学情報学専攻博士後期課程修了．博士（情報学）．2018 年より（株）KDDI 総合研究所情報セキュリティグループに所属．技術受容性，ユーザブルセキュリティとプライバシ，トラスト，AI 認知に関する研究に従事．電子情報通信学会，IEEE，各会員．

清本 晋作

2000 年筑波大学工学研究科物質工学専攻博士前期課程修了．同年 KDD（株）入社．現在，（株）KDDI 総合研究所情報セキュリティグループグループリーダー．ストリーム暗号，暗号プロトコル，プライバシ保護技術等の研究に従事．2004 年電子情報通信学会学術奨励賞，2016 年電子情報通信学会業績賞，各受賞．日本物理学会会員，電子情報通信学会各会員．2008～2009 年，London 大学 Royal Holloway 校客員研究員．博士（工学）．

鈴木 信雄 （正会員）

2007 年筑波大学大学院博士後期課程修了．博士（システムズ・マネジメント）．1982 年国際電信電話株式会社（現 KDDI）入社，2008 年トヨタ自動車株式会社 IT・ITS 企画部，2011 年（株）KDDI 研究所スマートワイヤレスグループグループリーダー．2016 年（株）国際電気通信基礎技術研究所広帯域ワイヤレス研究室室長．2019 年より近畿大学産業理工学部教授．モバイルネットワーク，IoT セキュリティ，データマイニング等の研究に従事．電子情報通信学会，電気学会各会員．