大規模コーパスにおける要配慮個人情報検出の精度向上

源怜維 [1,2]　小田悠介 [2]　河原大輔 [1,2]
[1] 早稲田大学理工学術院　[2] 国立情報学研究所 大規模言語モデル研究開発センター
{ray@akane.,dkw@}waseda.jp　odashi@nii.ac.jp

概要

要配慮個人情報を含む大規模コーパスで学習した大規模言語モデル（LLM）は，出力を通じて当該情報を漏洩する危険性があり，それを検出する技術が必要である．本研究では，要配慮個人情報に関係するが非要配慮である境界的な文書を追加したデータセットを人手により構築し，検出器を学習する．その結果，再現率を維持したまま適合率が向上し，高速で実用的な検出器を構築できた．

1　はじめに

LLM の構築には大規模な事前学習コーパスが必要であり，その収集手段として Web クローリングが広く利用されているが，意図せず個人情報を収集してしまうことがある．LLM が個人情報を記憶し，出力を通じて漏洩する危険性があり [1]，特に高い機微性を有する要配慮個人情報については，漏洩時の社会的影響がより大きくなる．

個人情報の保護に関する法律（平成十五年法律第五十七号） [2] (個人情報保護法) では，要配慮個人情報を以下のように定義している．

> 「要配慮個人情報」とは、本人の人種、信条、社会的身分、病歴、犯罪の経歴、犯罪により害を被った事実その他本人に対する不当な差別、偏見その他の不利益が生じないようにその取扱いに特に配慮を要するものとして政令で定める記述等が含まれる個人情報をいう。

個人情報保護委員会は，生成 AI サービスの学習データに要配慮個人情報が含まれないように取り組み，含まれる場合は削除または適切な加工をすることを要求している [3]．このような背景から，LLM の学習データ，特に大規模事前学習コーパスを対象として，要配慮個人情報を実用的な速度で検出する技術が不可欠である．

近年公開された Common Corpus [4] や Dolma [5] のような事前学習コーパスでは，構築段階で個人情報の除去が行われているが，要配慮個人情報を含む機微性の高い個人情報を除去したことは報告されていない．また，Dolma 3 [6] では LLM を用いて個人情報に加えて，機微性の高い個人情報も除去されているが，この処理は PDF など一部のデータに限定されている．日本語では，大規模コーパス用の要配慮個人情報フィルタ[1)]が存在するが，表層的なルールに基づくフィルタであり，網羅的な検出ではなく実用的なリスク軽減を目的として設計されている．

我々は，LLM によるアノテーションによって，事前学習コーパスから抽出した各文書に要配慮個人情報の区分と対応するラベルを 1 つ付与したデータセット（自動データセット）を構築した [7]．自動データセットは，医療関係，犯罪関係，および障害の 3 区分の文書と非要配慮文書から構成される．

自動データセットで学習した検出器は，要配慮文書（要配慮個人情報を含む文書）だけでなく，カテゴリ関係文書（要配慮個人情報の区分と関係するが，要配慮個人情報を含まない文書）も検出する傾向があった．これは，複数の LLM が非要配慮と判定した文書を負例データとして採用した結果，境界的なカテゴリ関係文書が負例データに十分に含まれなかったためと考えられる．

本研究では，カテゴリ関係文書の誤検出を抑制するため，人手アノテーションによってカテゴリ関係文書を含むデータセット（人手データセット）を構築する．さらに，人手データセットを用いて学習した検出器を大規模コーパスに適用し，その挙動を分析する．

2　人手データセットの構築

人手データセットは，文書と要配慮個人情報のアノテーションから構成する．

1） https://github.com/matsuolab/jp-llm-corpus-pii-filter

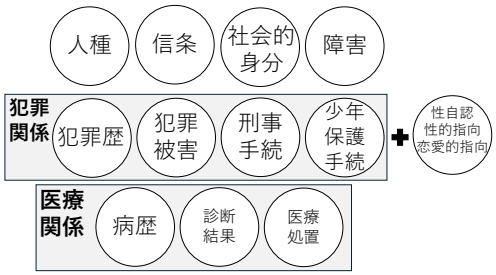

図 1　要配慮個人情報の種別

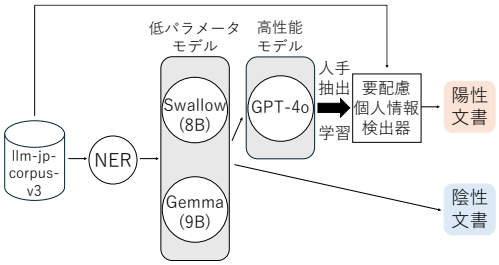

図 2　アノテーション対象文書取得フロー

2.1　定義

要配慮個人情報は個人を特定する情報と機微情報から構成される．本研究では以下のとおり，個人を特定する情報を「人物」，機微情報を「種別」および「スパン」として定義しアノテーションする．

- 人物: 要配慮個人情報の対象となる人物の情報．
- 種別: 要配慮個人情報の区分．
  個人情報保護法と個人情報の保護に関する法律施行令（平成十五年政令第五百七号）[8] で定義された区分に準拠する．なお，アノテーションの一貫性を担保するため，図 1 に示すように重複する区分は犯罪関係，医療関係として結合する．加えて，EU または英国から十分性認定に基づき提供された個人情報に，性自認・性的指向・恋愛的指向に関する情報が含まれる場合，要配慮個人情報と同様に扱う必要がある．そのため，本研究の種別では当該区分を追加する．詳細な定義を付録 B に記載する．
- スパン: 要配慮個人情報の根拠となる表現．

2.2　アノテーション対象文書

本研究ではカテゴリ関係文書の誤検出抑制を目的とするため，アノテーション対象文書にカテゴリ関係文書を含める．対象文書は，公開されている日本語 Web コーパスから取得し，要配慮個人情報を含まない可能性が高い陰性文書と，要配慮個人情報を含む可能性が高い陽性文書から構成する．

陽性文書は，我々の提案した 2 段階アノテーション [7] を用いて抽出したデータセットで学習した検出器により取得する（2.2.1 節）．この過程で取得される陽性文書には，カテゴリ関係文書が含まれる．

2.2.1　アノテーション対象文書取得手法

要配慮個人情報は個人を特定する情報が必要であるため，前処理として人名認識を適用する．

アノテーション対象文書は大規模コーパスから取得するため，精度と処理速度の両立が必要となる．そのため，陽性文書の取得にはまず，図 2 に示すように，2 種の低パラメータ LLM と高性能 LLM を組み合わせた 2 段階アノテーションを行う．3 つのモデルが要配慮個人情報を含むと判定した文書に一部人手アノテーションを行い，各種別の要配慮個人情報を含む小規模データセットを構築する．小規模データセットの規模は約 9,000 件であり，要配慮個人情報を含む文書が全体の約 10%となるよう調整した．このデータセットを用いて，要配慮個人情報の種別を出力する検出器を学習し，当該検出器を大規模コーパスに適用することで，陽性文書を取得する．

陰性文書は低パラメータモデルのいずれもが要配慮個人情報を含まないと判定した文書を取得する．

2.2.2　アノテーション対象文書取得過程

アノテーション対象文書は，Common Crawl から収集された日本語の Web コーパスである llm-jp-corpus-v3/ja/ja_cc/level0 [2] より取得した 10,000 件である．このうち，陰性文書 3,000 件と陽性文書を種別ごとに 1,000 件ずつ収集した計 7,000 件から構成する．

人名認識には固有表現認識モデルである bert-base-japanese-v3-ner-wikipedia-dataset [3] を利用し，低パラメータなモデルは Llama 3.1 Swallow 8B Instruct v0.2 [9, 10] と Gemma 2 9B IT [11] を，高性能なモデルは GPT-4o を使用した．そして，小規模データセットを gensim [12] の Doc2Vec でベクトル化し，scikit-learn [13] のサポートベクターマシン（SVM）などの機械学習モデルで検出器を構築した．

陽性文書については人手で確認し，各種別ごとに

2）https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3/-/tree/main/ja/ja_cc/level0
3）https://huggingface.co/llm-book/bert-base-japanese-v3-ner-wikipedia-dataset

表 1　要配慮個人情報スパン分布

人種	信条	社会的身分	医療関係	犯罪関係	障害	性自認・性的指向・恋愛的指向	合計
1,287	888	433	1,980	2,718	1,212	1,628	10,146

表 2　要配慮個人情報ラベルの組み合わせの分布

非要配慮	犯罪関係	医療関係	性自認等	人種	障害
5,698	964	691	557	502	474

信条	社会的身分	障害+医療関係	犯罪関係+医療関係
304	203	118	55

注:「性自認等」は，性自認・性的指向・恋愛的指向を意味する．

要配慮文書とカテゴリ関係文書の比率が 1:1 の 500 件ずつになるよう，小規模データセットの一部を追加して調整した．

なお，社会的身分に関しては前述の方法では文書を 1,000 件確保できなかった．そのため，前述の方法で取得した文書に加えて，同一の llm-jp-corpus-v3 からキーワードベースのルールにより取得した文書も陽性文書に含めた．

2.3　アノテーション

以下の条件のもと，各文書当たり人物，種別，スパンの組を最大 4 つまでアノテーションする．

1. 人物・スパンは最小限の範囲で編集せずそのまま抽出する
2. 同一人物・同一種別でも文脈が異なるスパンは複数抽出する
3. スパンには可能な限り人物を含める
4. 詳細な情報が含まれるスパンを優先する

アノテーションでは判断に迷う境界事例も抽出し，備考に記載した．要配慮個人情報の性質上，見逃しは重大なリスクにつながるため，本研究では境界事例も要配慮個人情報として扱う．

2.4　アノテーション結果の分析

表 1 に要配慮個人情報のスパン分布を示す．社会的身分はアノテーション対象文書の一部をキーワードベースのルールで取得したため，スパン数が比較的少なくなっていると考えられる．

また，要配慮個人情報の種別に「非要配慮」を加えたものを要配慮個人情報ラベルとして定義する．表 2 に，各文書に付与された要配慮個人情報ラベルの組み合わせの分布を示す．各種別は理想的には 500 件ずつ要配慮文書が得られるため，信条や社会的身分の要配慮個人情報を含む文書が比較的少ないことがわかる．また，障害と医療関係など関連性の強い種別が同じテキスト内に含まれている．

表 3　自動データセットおよび人手データセットの構成

データ	医療関係	犯罪関係	障害	非要配慮
自動	312	157	400	7,821
人手	629	830	305	3,805

表 4　新旧モデル評価（括弧内は旧モデルの結果）

モデル	クラス重み	適合率	再現率	F1 スコア
SVM	None	0.77 (0.67)	0.74 (0.97)	0.75 (0.80)
(Doc2Vec)	balanced	0.66 (0.63)	0.98 (0.98)	0.79 (0.77)
LGBM	None	0.81 (0.58)	0.72 (0.14)	0.76 (0.23)
(TF-IDF)	balanced	0.77 (0.52)	0.80 (0.15)	0.79 (0.24)

なお，低パラメータ LLM のアノテーションによって得られた陰性文書 3,000 件のうち，要配慮個人情報を含むものは 42 件（1.4%）にとどまった．

3　要配慮個人情報検出器の構築

3.1　検出器の構築手法

要配慮個人情報検出器は，入力した文書に対して要配慮個人情報ラベルを 1 つ付与する．具体的には，文書を Doc2Vec でベクトル化し，SVM で検出するモデルと，TF-IDF でベクトル化し，LightGBM（LGBM） [14] で検出するモデルを構築する．両モデルにおいて，クラス重みとして None, balanced を設定し，後者ではクラス頻度に応じて損失関数の重みを調整することで，クラス不均衡を補正する．

3.2　既存検出器との比較

2 節では，検出器がカテゴリ関係文書を誤検出するという問題を解決するために，人手データセットを構築した．そこで，本節では自動データセットで学習した検出器（旧モデル）と，人手データセットで学習した検出器（新モデル）の性能を比較する．旧モデルは人手データセットを用いて，新モデルは交差検定によって評価する．

自動データセットは医療関係・犯罪関係・障害・非要配慮しか要配慮個人情報ラベルを含まないため，手動データセットも対応するラベルのみを使用する．なお，各文書に 1 つの要配慮個人情報ラベルを付与するため，人手データセットで複数のラベルがついている場合は代表ラベルを選択する．例えば，当該文書を取得時に想定していた要配慮個人情

表 5　大規模コーパスに適用した新旧モデルの検出結果に対する人手評価（括弧内は旧モデルの結果）

モデル	クラス重み	要配慮文書	カテゴリ関係文書	無関係文書	抽出数	推定要配慮文書数
SVM	None	0.35 (0.26)	0.45 (0.53)	0.20 (0.21)	9,348 (12,045)	3,272 (3,131)
(Doc2Vec)	balanced	0.19 (0.07)	0.55 (0.34)	0.26 (0.59)	34,855 (88,529)	6,622 (6,197)
LGBM	None	0.46 (0.24)	0.49 (0.69)	0.05 (0.07)	8,046 (16,048)	3,701 (3,852)
(TF-IDF)	balanced	0.39 (0.20)	0.55 (0.69)	0.06 (0.11)	13,247 (24,813)	5,166 (4,963)

表 6　人手データセットの全種別を用いた交差検定結果

モデル	クラス重み	適合率	再現率	F1 スコア
SVM	None	0.772	0.575	0.659
(Doc2Vec)	balanced	0.614	0.947	0.745
LGBM	None	0.810	0.648	0.720
(TF-IDF)	balanced	0.757	0.767	0.762

表 7　チューニングモデルの交差検定結果

モデル	適合率	再現率	F1 スコア
SVM (balanced)	0.646	0.896	0.751
LGBM (balanced)	0.715	0.856	0.779
SVM + LGBM	0.637	0.951	0.763

報の種別が，付与されたラベルに含まれる場合は，それを優先して選ぶ．また，一部文書に重複があったため手動データセットから削除した．最終的なデータセットの構成を表 3 に示す．

表 4 に両モデルの評価結果を示す．なお，本研究において適合率と再現率は，種別を問わず要配慮個人情報を含む文書を陽性，要配慮個人情報を含まない文書を陰性として計算する．

新モデルでは SVM および LGBM ともに適合率が向上しており，これはカテゴリ関係文書を陰性文書として追加したことで，偽陽性が減少したためと考えられる．旧モデルは，人手データの取得に Doc2Vec でベクトル化した文書で学習した検出器を使用したため，Doc2Vec を使用したモデルの再現率が非常に高くなっている．一方，TF-IDF を使用したモデルでは再現率が大きく低下している．これは，Doc2Vec 由来で取得したデータセットと TF-IDF を使用したモデルの相性が悪いことが原因と考えられる．

また，各モデルの実運用時における性能を比較するため，事前学習コーパスから検出された文書を人手で評価する．要配慮文書とカテゴリ関係文書に加えて，要配慮個人情報も種別と関係性がある情報も全く含まない文書を「無関係文書」と定義する．表 5 に各モデルを 100 件ずつ評価した結果と検出数を示す．すべてのモデルで要配慮個人情報の検出率が 10%以上向上しており，抽出数と割合から計算できる推定要配慮文書数もほとんど減少していない．そのため，各検出器は再現率を保ったまま適合率が向上していることがわかる．

3.3　実運用における検出器の精度向上

前節で構築した検出器の対象は，医療関係，犯罪関係，および障害に限定されていた．本節では，人手データセットの全種別を用いて学習した検出器の精度向上を図り，実運用を想定した構成を検討する．データセットは表 2 のデータに，前節と同様の方法で代表ラベルを決定して構築する．表 6 に示すように，SVM（balanced）と，LGBM（balanced）が，再現率および F1 スコアで高い性能を示したため，両モデルに F1 スコアを指標としたハイパーパラメータチューニングを行った．

表 7 にチューニング後の交差検定結果を示す．要配慮個人情報の見逃しを防ぐため，再現率を重視し，チューニング後モデルのいずれかが要配慮個人情報と判定した場合，当該文書を要配慮個人情報として扱う論理和型の構成（SVM + LGBM）も評価した．その結果，SVM も LGBM もわずかに F1 スコアが向上した．さらに，SVM + LGBM は F1 スコアこそ低下したものの，再現率 0.95 を達成した．この検出器をコーパスに適用し，検出した文書 100 件を人手評価すると，要配慮文書の割合は 23%で，適合率もある程度担保されていた．

以上より，本検出器は，大規模コーパスに対する事前フィルタとして，実用的な処理速度と高い再現率を両立できることが示された．このようなフィルタを前段に配置することで，後段で低速だが高性能な LLM を適用することが可能であると考えられる．

4　おわりに

本研究では，要配慮個人情報検出において，カテゴリ関係文書を学習データに追加することで，再現率を維持したまま適合率を向上させることができた．実験の結果，すべての要配慮個人情報種別を検出可能な，高再現率かつ高速に動作する検出器を構築できた．今後は，本検出器と LLM を組み合わせた実用的なパイプラインを検討する予定である．

謝辞

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けた．本研究成果の一部は，データ活用社会創成プラットフォーム mdx を利用して得られた．また，産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」を支援を受けて利用した．

参考文献

[1] Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 2038–2047, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[2] デジタル庁. 個人情報の保護に関する法律（平成十五年法律第五十七号）. e-Gov 法令検索. アクセス日: 2025 年 12 月 21 日. https://laws.e-gov.go.jp/law/415AC0000000057/.

[3] 個人情報保護委員会. 生成 AI サービスの利用に関する注意喚起等について. 個人情報保護委員会. アクセス日: 2025 年 12 月 21 日. https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/.

[4] Pierre-Carl Langlais, Carlos Rosas Hinostroza, Mattia Nee, Catherine Arnett, Pavel Chizhov, Eliot Krzystof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P. Yamshchikov. Common corpus: The largest collection of ethical data for llm pre-training, 2025.

[5] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025.

[7] 源怜維, 小田悠介, 河原大輔. 大規模言語モデルの事前学習用コーパスにおける要配慮個人情報の検出. 言語処理学会第 31 回年次大会 発表論文集, pp. 2873–2878, 2025.

[8] デジタル庁. 個人情報の保護に関する法律施行令（平成十五年政令第五百七号）. e-Gov 法令検索. アクセス日: 2025 年 12 月 21 日. https://laws.e-gov.go.jp/law/415CO0000000507/.

[9] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[10] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[11] Gemma Team. Gemma. Kaggle, 2024. https://www.kaggle.com/m/3301. DOI: 10.34740/KAGGLE/M/3301.

[12] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In **Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks**, pp. 45–50, Valletta, Malta, May 2010. ELRA.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, Vol. 12, pp. 2825–2830, 2011.

[14] Microsoft. Lightgbm, 2017. version = 4.6.0.

A　倫理的配慮

本研究は，所属機関の倫理審査を通過して実施された．また，人手アノテーションでは関連法令を遵守し，適切な監督の下で実施した．

B　要配慮個人情報種別の定義

1. 人種　　人種，世系又は民族的若しくは種族的出身を広く意味する（例: アイヌ，在日朝鮮人，アジア系アメリカ人，ハーフなど）．なお，単純な国籍や「外国人」という情報は法的地位であり，それだけでは人種には含まない．また，肌の色は，人種を推知させる情報にすぎないため，人種には含まない．
2. 信条　　個人の基本的なものの見方，考え方を意味し，思想と信仰の双方を含むものである（例: ○党を支持している，○教を信仰しているなど）．政党・聖職者・宗教団体の指導者（例：僧，神父，牧師，教祖，代表など）に関する情報は，それが公的な立場の表明であり，個人の信条に言及するものでない限りは該当しない．
3. 社会的身分　　ある個人にその境遇として固着していて，一生の間，自らの力によって容易にそれから脱し得ないような地位を意味する（例: 部落出身，祖父母や親が部落に居住していたなど）．単なる職業的地位や学歴は含まない．
4. 医療関係

   a 病歴　　病気に罹患した経歴を意味するもので，特定の病歴を示した部分（例：特定の個人ががんに罹患している，統合失調症を患っているなど）が該当する．軽度で一時的な骨折や風邪などは該当しない．

   b 診断結果　　本人に対して医師などにより行われた疾病の予防および早期発見のための健康診断などの結果が該当する．

   c 医療処置　　健康診断などの結果に基づき，又は疾病，負傷その他の心身の変化を理由として，本人に対して医師などにより心身の状態の改善のための指導又は診療若しくは調剤が行われたことが該当する．
5. 犯罪関係

   a 犯罪歴　　前科，すなわち有罪の判決を受けこれが確定した事実が該当する．

   b 犯罪被害　　身体的被害，精神的被害および金銭的被害の別を問わず，犯罪の被害を受けた事実を意味する．

   c 刑事手続　　本人を被疑者又は被告人として，逮捕，捜索，差押え，勾留，公訴の提起その他の刑事事件に関する手続が行われたことが該当する．他人を被疑者とする犯罪捜査のために取調べを受けた事実や，証人として尋問を受けた事実に関する情報は，本人を被疑者又は被告人としていないことから，これには該当しない．

   d 少年保護手続　　本人を非行少年又はその疑いのある者として，調査，観護の措置，審判，保護処分その他の少年の保護事件に関する手続が行われたことが該当する．
6. 障害　　身体障害，知的障害，精神障害（発達障害を含む．）その他の心身の機能の障害があることが該当する．障害者の公知である大会（パラリンピック，デフリンピック，スペシャルオリンピックスなど）に参加する選手に関する情報は該当しない．障害者の非公式な大会，集会，企画に参加した情報は該当する（例:A さんは知的障害をもつ人達のプログラムに参加）
7. 性自認・性的指向・恋愛的指向　　性・恋愛に関する様々な特徴が該当する．いわゆる LGBTQ+ に関する情報であるが，多様な概念であるためそれに限らない．

C　要配慮個人情報の具体例

以下に，アノテーションによって抽出された要配慮個人情報スパンの具体例（匿名化済み）を示す．

1. 人種　　自身も在日コリアン二世である事務局長の山田太郎さん
2. 信条　　結婚を機に日本人からムスリムになられた山田太郎さん
3. 社会的身分　　被差別部落出身の山田太郎
4. 医療関係　　山田太郎さん、スキルス性胃がん克服から 30 年
5. 犯罪関係　　大学生の山田太郎容疑者（25）＝静岡市＝が傷害容疑で逮捕
6. 障害　　生まれつき耳が聞こえない山田太郎
7. 性自認・性的指向・恋愛的指向　　山田太郎代表取締役は、自身も LGBTQ+当事者として思い悩んだ経験から