日本語不法行為事件データセットの構築

山田 寛章 [1]　徳永 健伸 [1]　小原 隆太郎 [2,3]
得津 晶 [3]　竹下 啓介 [3]　角田 美穂子 [3]
[1] 東京工業大学　[2] 中村・角田・松本法律事務所　[3] 一橋大学
{yamada, take}@c.titech.ac.jp, r.ohara@ntmlo.com
{a.tokutsu, kei.takeshita, m.sumida}@r.hit-u.ac.jp

概要

本研究は日本語・日本法における法的判断予測研究のためのデータセットである，日本語不法行為事件データセット (Japanese Tort-case Dataset, JTD) を提案する．JTD は不法行為判断予測タスク及びその根拠抽出タスク向けに設計されている．根拠抽出タスクは不法行為の成否判断に際して重要な根拠となった主張を，原告または被告の主張の中から抽出するタスクである．JTD には 41 人の法律専門家によって注釈付けされた 3,477 件の民事事件判決書に基づいて構築されており，7,978 事例（事例に内包される原告・被告らの主張は 59,697 事例）が収録されている．ベースライン実験により JTD の各タスクの実現可能性を確認し，さらに不法行為判断予測・根拠抽出の両タスクを同時に学習させることで性能が改善することを示した．

1　はじめに

民事における法的な争いにおいて，非専門家の当事者が状況を整理した上で紛争解決の結果を予見することは困難であり，法律専門家への相談にも心理的・時間的・金銭的な障壁がある．紛争解決結果を予測するモデルが当事者に予測結果を提供することでこれらの障壁を解消できる．さらに，当事者間の話し合いにより合意することで紛争の解決を図る調停手続きの進行をモデルによって支援することや，モデルをオンライン裁判外紛争解決手続き (ODR) に組み込むことで，定型的な紛争の解決において省力化を行うことが可能である．このため，紛争解決に伴う法的判断予測 (Legal Judgment Prediction, LJP) を行うモデルの開発が求められている．

LJP は国外において盛んに研究されている．欧州人権裁判所のデータを用いた研究 [1, 2, 3, 4] はその代表例である．米国においても Katz ら [5] による米国最高裁判所におけるデータを用いた例，Semo ら [6] によるクラスアクション訴訟を取り扱った例が挙げられる．中国においては刑事事件向け LJP [7, 8, 9, 10, 11] の研究が盛んである．また，Chalkidis ら [12] は複数のタスクから構成される英語圏の法律分野向けのベンチマークデータセット LexGLUE を提案している．

国内においては，機械学習を用いた実験に適する規模を持つ実事件データを元にした LJP データセットは存在しない．法に関するデータを用いた Competition on Legal Information Extraction/Entailment (COLIEE) [13] が存在するものの，COLIEE では日本の司法試験短答式問題を用いている点で狙いが異なる．言語・法制度は国毎に異なるため，LJP 研究において日本独自の LJP データセットを構築することは重要である．そこで日本の民事事件判決書を用いた大規模なデータセットとして Japanese Tort-case Dataset (JTD) を構築する．

2　日本語不法行為事件データセット

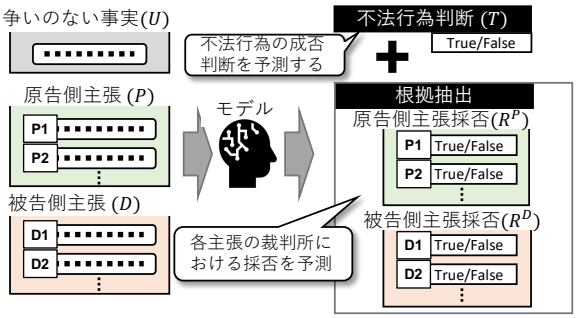
2.1　データ収集

本研究で構築するデータセットの対象は不法行為事件の判決書とする．不法行為は民事事件において重要な概念であり，日本法上は「故意又は過失によって他人の権利又は法律上保護される利益を侵害した者は、これによって生じた損害を賠償する責任を負う。」[1)] とされる．現代社会においては，インターネット上における名誉毀損・プライバシー侵害に関連して，不法行為法の果たす役割がより重要になると想定される．

判決書は，株式会社 LIC が提供する「判例秘書[2)]」

1）民法 第七百九条
2）https://www.hanreihisho.com/

図 1　不法行為判断予測タスクと根拠抽出タスク



から収集する．収集対象の判決書は，民事事件の第一審であり，名誉毀損・プライバシー侵害・信用毀損を中心とする不法行為関連事件とする．収集時には判例秘書にて提供されるキーワードベースの検索システムを利用した[3]．検索結果に混入した不法行為に関連しない事件は人手にて除去した．なお，収集対象の判決書は仮名化処理済みである．

2.2　タスク設計

不法行為事件には，不法行為を申し立てる原告側と反駁する被告側から存在する．裁判では，原告側・被告側の主張を裁判所が審理し，各主張について採用するか否かの判断が行われ，最終的な判決が導かれる．このため，各主張に対する採否の判断が最終的な不法行為判断の根拠に相当すると見なし，タスクを設計する．

図 1 に JTD に収録する 2 つのタスク，不法行為判断予測 (TP) 及び根拠抽出 (RE) の概要を示す．TP・RE 共に入力は，原告側主張のリスト P・被告側主張のリスト D・争いのない事実[4] U である．TP では不法行為の成否を示す二値 T を出力とする．RE では，P, D 中の各主張に対する採否を示す二値の系列である R^P, R^D を出力とする．図 2 に JTD を構成するデータの例を示す．図中，T_g, R_g^P, R_g^D はそれぞれ T, R^P, R^D の正解ラベルを指す．JTD には $(U, P, D, T_g, R_g^P, R_g^D)$ の組が収録される．以降，この組を事例と呼ぶ．

2.3　アノテーション

前節で定義した事例を得るため，判決書に対して人手による注釈付けを実施した．注釈付けの基準は日本語判決書に向けに設計されたものを使用した [14]．判決書への注釈付けには法に関する専門知識が必要となるため，注釈付けは法曹・法科大学院学生・法学部学生ら 41 名に依頼した．

表 1 に注釈付けの結果得られた JTD の概要を示す．3,477 件の判決書を注釈付けした結果，7,978 件の事例が得られた．全事例の 40%で不法行為が成立と注釈付けされている．また，RE の対象となる原告・被告からの主張は合計で 59,697 件得られ，そのうち 53%が原告側主張であった．全主張のうち 51%が採用と注釈付けされている．表 2 に JTD の訓練，検証，テスト用の各セットへの分配設定を示す．

3)　検索時に用いたクエリを付録 B に示す．
4)　争いのない事実は複数ある場合でも結合して一つの文字列として取り扱い，P, D のようなリストとしない．

図 2　名誉毀損に関する不法行為の訴えの事例（一部改変・縮約）

争いのない事実 (U)

インターネット上のページ「D」に「X I さん金返さないと」という書込みが，IPアドレス***.***.***.***を経由して投稿された。

原告側主張 (P)

	R_g^P
本件投稿は，一般の閲覧者の普通の注意と閲覧の仕方を基準とすると，B製作所に勤務する「X I」という人物が，特定の個人から金銭を借り入れたがその返済をしていないとの事実を摘示するものである。	True
B製作所に勤務する「X I」という姓の人物は，原告とそのいとこの2名のみである。	True
本件投稿の閲覧者のうち，原告を知っているが原告のいとこを知らない者は本件投稿の対象が原告と考えるであろう。	False
原告と原告のいとこを知る者が「X I」という記載から原告のことを思い浮かべることもあるはずである。	False
本件投稿の対象と原告との間に同定可能性はある。	False

被告側主張 (D)

	R_g^D
上記原告の主張は，いずれも争う。	False

T_g: False

3　モデル

JTD の各タスクの実現可能性の確認及び JTD のベースラインとなるスコアの確立を目的として，階層型 Transformer を元に構築したモデルを用いた実験を行う．

各主張それぞれでの内部文脈と主張間の関係の

表 1 日本語不法行為事件データセット (JTD) の概要

判決書数	3,477
判決書あたりの平均事例数	2.3
事例数	7,978
主張数	59,697
争いのない事実数	9,236

表 2 JTD における訓練・検証・テストセット

分割	文書数	事例数	主張数
検証	329	803	6,063
テスト	391	811	5,945
訓練	2,757	6,364	47,689

双方を考慮するために，二階層から成る Transformer ベースのモデルを用いる．以降，このモデルを Inter-Claim Transformer (ICT) と呼ぶ．ICT は，争いのない事実 U，原告主張の系列 P 及び被告主張の系列 D を受け取り，TP 出力 T 及び RE 出力 R^P, R^D を返す．ICT は，原告・被告の別を表現するための当事者種別表現，位置表現も考慮する．

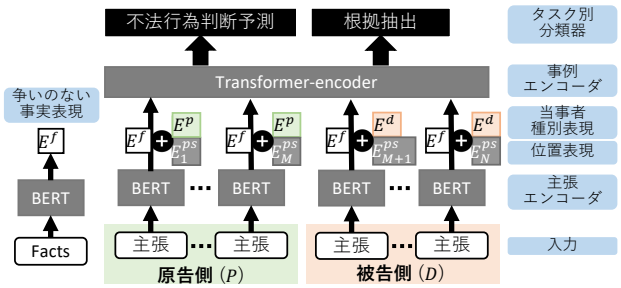
ICT には主張エンコーダと事例エンコーダが存在する．主張エンコーダには日本語 BERT を用いる．本実験では日本語 BERT として，BERT-base-Japanese (BERTja)[5] と Japanese-LegalBERT (JLBERT) [15] を用いて両者を比較する．JLBERT は日本の民事事件を用いて追加の事前学習が行われており，BERTja と比較してより高い性能が期待できる．事例エンコーダは主張間の関係を考慮する役割を担っている．事例エンコーダは主張エンコーダとは異なり，事前学習されていない Transformer エンコーダである．なお，争いのない事実については，どの主張に対しても同じ表現を伝達できるように，主張エンコーダとは独立して表現を得る機構となっている．

図 3 に ICT の内部構造の概要を示す．図中，N 個の主張があり，その内 N 個が原告側主張である．争いのない事実の表現 E^f は全主張で同じものが共有される．当事者種別表現は，各主張の主張者に応じて，原告の場合は E^p，被告の場合は E^d が割り当てられる．E^{ps}_n は n 番目の位置表現を示す．例えば，事例エンコーダに対する原告の n 番目の入力は，主張エンコーダから得られる n 番目の主張の表現に，争いのない事実 (E^f)，当事者種別表現 (E^p)，位置表現 (E^{ps}_n) を足したものになる．

TP・RE 出力は事例エンコーダの直後にある線形層を通じて得られる．ICT は TP・RE 各タスク単独で解く設定と，TP・RE を同時解くマルチタスク設定のどちらでも利用可能となっている．

5) https://huggingface.co/cl-tohoku/bert-base-japanese-v2

図 3 Inter-Claim Transformer (ICT)



以下，ICT と併せて本実験で用いるモデルを列挙する．

3.1 根拠抽出タスク (RE)

- **RE-random**: 学習セットにおけるラベルの割合を元に，ランダムに予測する手法．
- RE-BERT: 通常の BERT 単体を二値分類器として用いる手法．**RE-BERTja** の **RE-JLBERT**2 種類があり，それぞれ BERTja，JLBERT を用いている．
- RE-IST: ICT を用いる手法で，RE タスク単独で学習・推論を行う設定．**RE-ICT-BERTja** と **RE-ICT-JLBERT** の 2 種がある．

3.2 不法行為判断予測タスク (TP)

- **TP-random**: RE-random と同様の手法．
- TP-IST: ICT を用いる手法で，TP タスク単独で学習・推論を行う設定．**TP-ICT-BERTja** と **TP-ICT-JLBERT** が実験対象である．

3.3 マルチタスク

上記の各タスク単独で学習する手法に加えて，ICT を用いて TP と RE の 2 タスクを同時に学習させる実験も行う．このとき，RE タスクの損失関数 $\mathrm{Loss}_{\mathrm{RE}}$ と TP タスクの損失関数 $\mathrm{Loss}_{\mathrm{TP}}$ は式 1 のように組み合わせる．

$$\mathrm{Loss} = \alpha \mathrm{Loss}_{\mathrm{TP}} + (1-\alpha)\mathrm{Loss}_{\mathrm{RE}} \tag{1}$$

マルチタスクを採用する ICT においても，**Multi-ICT-BERTja** と **Multi-ICT-JLBERT** の 2 種類を用意する．

表 3 RE タスクの実験結果 (Accuracy 及び標準偏差)

モデル	全体	原告	被告
RE-random	0.498 (.005)	0.496 (.005)	0.501 (.007)
RE-BERT-BERTja	0.598 (.005)	0.597 (.007)	0.598 (.008)
RE-BERT-JLBERT	0.634 (.005)	0.635 (.009)	0.631 (.006)
RE-ICT-BERTja	0.637 (.012)	0.652 (.010)	0.620 (.022)
RE-ICT-JLBERT	0.663 (.008)	**0.677** (.013)	0.648 (.008)
Multi-ICT-BERTja	0.666 (.008)	0.671 (.009)	0.661 (.011)
Multi-ICT-JLBERT	**0.674** (.009)	0.675 (.007)	**0.673** (.014)

表 4 TP タスクの実験結果 Experimental results of TP (Accuracy 及び標準偏差).

モデル	マクロ平均 (σ)
TP-random	0.503 (.014)
TP-ICT-BERTja	0.649 (.023)
TP-ICT-JLBERT	0.674 (.024)
Multi-ICT-BERTja	0.680 (.007)
Multi-ICT-JLBERT	**0.683** (.020)

4 実験

4.1 実験設定

全モデルに共通して，最適化アルゴリズムには AdamW [16] を用いる．エポック数は 30 に固定し，検証用セットにおいて Accuracy スコアが最大とるエポックのチェックポイントを採用する．ICT が許容する最大の主張の数は 64, ICT の主張エンコーダの最大入力長は 512 トークンとする．ICT の主張エンコーダである BERTja または JLBERT のパラメータは固定しないで学習対象とする．ハイパーパラメタの探索では Optuna [17] を用いた．探索時には訓練セット中の 3,000 事例のみを利用して学習を行い，その評価には検証セット全てを用いた[6]．

本実験では，RE・TP 共に Accuracy を評価指標として用いる．モデルの学習と評価は，異なるシード値を用いて 5 回実施し，その平均値を結果と見なす．統計的仮説検定として Permutation test による有意水準を 5%とした両側検定を実施する．

4.2 実験結果

表 3 に根拠抽出タスク (RE) の結果を示す．表中の「全体」列は RE の対象である主張全体の結果，「原告」及び「被告」の列は，ぞれぞれ原告側・被告側主張についての結果である．

「全体」の結果によれば，ICT 系モデルは RE-BERT 系モデルから大きく性能が向上していることがわかる．これは，ICT の主張間関係を考慮する機構が期待通りに働いたためと考えられる．BERTja 採用モデルと JLBERT 採用モデル間で比較すると，JLBERT 採用モデルが常に良いスコアを示している．この点，Multi-ICT 系モデルの組を除く，全ての組で統計的に有意な差を確認している．ただし，JLBERT の事前学習時に用いられた判決書と JTD 構築に用いられた判決書の間に重複がある点を留意すべきである．

Multi-ICT 系モデルは，同じ主張エンコーダを用いる RE-ICT 系モデルと比較すると，常に Multi-ICT 系モデルが良いスコアを示している．この傾向は，特に BERTja を主張エンコーダとして採用する際に明確で，RE-ICT-BERTja と Multi-ICT-BERTja 間では統計的な有意差を確認した．「原告」「被告」それぞれに着目すると，傾向は「全体」と同様であるものの，「原告」において RE-ICT-JLBERT が最高スコアを示した点が異なっている．

表 4 に不法行為判断予測タスク (TP) の結果を示す．RE と同様，Multi-ICT 系モデルが最高性能を示している．Multi-ICT 系モデルは，同じ主張エンコーダを用いる TP-ICT 系モデルと比較すると，常により高いスコアを示している．特に，Multi-ICT-BERTja と TP-ICT-BERTja 間で統計的な有意差を確認した．

RE・TP 共に，各モデルが RE-random, TP-random を大きく超える性能を示したことから，JTD において提案した各タスクに一定の実現可能性があることを確認できた．また，RE・TP の双方で Multi-ICT 系モデルが最高性能を示した．このことは，これらのモデルが不法行為成否の判断と各主張の採否の判断に間にある相互関係を捉えて利用できていたことを示唆している．

5 結論

本研究では，日本語・日本法に基づく LJP データセットである日本語不法行為事件データセットを構築した．本データセットは，不法行為の成否判断を予測するタスクと，その根拠となりうる重要な主張を抽出するタスク向けデータを提供する．人間の法律専門家による注釈によって構築された LJP データセットとしては規模が大きく，他に類例がない．ベースライン実験では，各タスクに実現可能性があることを示し，さらに判断予測と根拠抽出の両タスクを同時に学習させることで性能が改善することを示した．

6) 付録表 7 に探索空間を示す．

謝辞

判決書データをご提供いただいた株式会社 LIC に感謝申し上げます．また，データセット構築及び注釈付け作業にご協力いただいた全ての方に感謝申し上げます．本研究は JST RISTEX JPMJRX19H3 並びに JST ACT-X JPMJAX20AM の支援を受けたものです．

参考文献

[1] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. **PeerJ Comput. Sci.**, Vol. 2, p. e93, 2016.

[2] Masha Medvedeva, Michel Vols, and Martijn Wieling. Judicial decisions of the European Court of Human Rights: looking into the crystall ball. In **Proceedings of the Conference on Empirical Legal Studies in Europe 2018**, 2018.

[3] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural legal judgment prediction in English. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4317–4323, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Josef Valvoda, Ryan Cotterell, and Simone Teufel. On the Role of Negative Precedent in Legal Outcome Prediction. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 34–48, 01 2023.

[5] Daniel Martin Katz, Michael J. Bommarito, II, and Josh Blackman. A general approach for predicting the behavior of the supreme court of the united states. **PLOS ONE**, Vol. 12, No. 4, pp. 1–18, 04 2017.

[6] Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. ClassActionPrediction: A challenging benchmark for legal judgment prediction of class action cases in the US. In **Proceedings of the Natural Legal Language Processing Workshop 2022**, pp. 31–46, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[7] Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. Learning to predict charges for criminal cases with legal basis. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 2727–2736, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[8] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. Legal judgment prediction via topological learning. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3540–3549, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[9] Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. Few-shot charge prediction with discriminative legal attributes. In **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 487–498, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[10] Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. Automatic judgment prediction via legal reading comprehension. In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, **Chinese Computational Linguistics - 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019, Proceedings**, Vol. 11856 of **Lecture Notes in Computer Science**, pp. 558–572. Springer, 2019.

[11] Nuo Xu, Pinghui Wang, Long Chen, Li Pan, Xiaoyan Wang, and Junzhou Zhao. Distinguish confusing law articles for legal judgment prediction. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3086–3095, Online, July 2020. Association for Computational Linguistics.

[12] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 4310–4330, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[13] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. Summary of the competition on legal information, extraction/entailment (coliee) 2023. In **Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law**, ICAIL '23, p. 472–480, New York, NY, USA, 2023. Association for Computing Machinery.

[14] Hiroaki Yamada, Takenobu Tokunaga, Ryutaro Ohara, Keisuke Takeshita, and Mihoko Sumida. Annotation study of Japanese judgments on tort for legal judgment prediction with rationales. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 779–790, Marseille, France, June 2022. European Language Resources Association.

[15] Keisuke Miyazaki, Hiroaki Yamada, and Takenobu Tokunaga. Cross-domain analysis on Japanese legal pretrained language models. In **Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022**, pp. 274–281, Online only, November 2022. Association for Computational Linguistics.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019**. OpenReview.net, 2019.

[17] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In **Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**, KDD '19, p. 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery.

表 5　予測結果サンプル (ID: L07530809-6654)

主張元	主張	RE 正解	RE 予測
原告側	被告 Y2 は, 平成 30 年 7 月頃から, 原告代表者に対し,「AI を利用した会計システムの開発があとわずかで完成するが資金が足りない」旨を申し向けて, 融資を依頼し, 原告代表者は, こうした要請を受けて, 平成 30 年 7 月 30 日, 本件契約を締結した上, 同日から同年 9 月 25 日までの間,3 回にわたって合計 1000 万円を交付したものであるところ, 被告らにおいて,AI 会計ソフトを開発していた事実は全く存在しなかったものである。	T	F
原告側	被告らは,AI 会計ソフトを開発していた事実はなく, 原告から借り受けた金員を被告会社の事業に用いるつもりも, これを返済する意思もなかったにもかかわらず, これがあるように装って, 原告から金員を詐取したものであって, こうした被告らの詐欺行為は, 原告に対する不法行為を構成する。	T	F
被告側	被告らが, 原告代表者に対し, 融資を依頼した事実はなく, 本件契約は, 元々, 原告代表者が事業で数億円の利益が出たことから, 被告会社の株式を 2000 万円分購入したいと申し出たことを契機とするものであり, その後, 原告側の事情で, 原告が, 被告に対して 1000 万円を貸し付けることとなったものである。	F	T
被告側	被告会社は,AI 会計ソフトの開発を行っており, しかも, 本件契約においては, 借入金の使途は「経営資金」とされるのみで,AI 会計ソフトの開発費用に限定されていたものではなく, 被告会社に本件契約に違反する点はなく, 被告らに不法行為は成立しない。	F	T

TP 正解: True, **TP 予測**: False

A　モデルによる予測例

表 5 に各タスクに対するモデル出力の例を示す.

B　判決書収集時に使用したクエリ

判決書収集の際に使用したクエリを表 6 に示す. 名誉毀損・プライバシー侵害・信用毀損などの不法行為関連事件の収集には表中のクエリ A を用いた. また，インターネット上おける不法行為に関する争いは，発信者情報開示請求事件においても扱われることから，クエリ B を用いた収集も行った.

表 6　判決書収集の際に用いた検索クエリ

種別	クエリ
クエリ A	(“名誉” OR “プライバシー” OR “信用毀損”) AND “不法行為” NOT “発信者情報開示” NOT “地位確認” NOT “無効確認” NOT “商標”
クエリ B	“発信者情報開示”

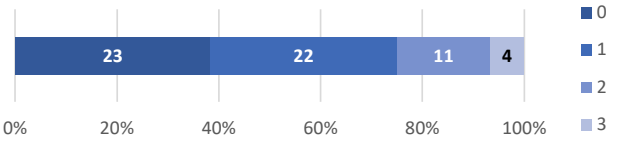
C　ハイパーパラメタ探索空間

表 7　ハイパーパラメタの候補

Parameters	Choices
Learning rate	2e-6, 4e-6, 6e-6, 8e-6
TRenc heads	2, 4, 6, 8
TRenc FF dim	64, 128, 256, 512
TRenc layers	1, 2, 3, 4
Use UF	True, False
α (if applicable)	0, 0.05, ..., 0.95

D　誤予測事例の難易度分析

図 4　人間の専門家による確信度



4 節の実験について，モデルが不法行為判断予測を誤った事例の難易度を調査した．分析には 4 人の法学専門家（本論文の著者）が参加した．分析対象のモデルは TP-ICT-BERTja, TP-ICT-JLBERT, Multi-ICT-BERTja and Multi-ICT-JLBERT とした．各モデル 5 回の出力結果は多数決により統合した．分析対象事例は，上記 4 モデルのいずれもが予測に失敗した事例 139 件から無作為に 60 件選択した．60 件は参加者に重複なく分配され，2 名が各 20 件，別の 2 名が各 10 件を担当した.

分析対象事例について，与えられている入力から法的判断予測を行う際の確信の度合いを調査した．分析参加者には，0 から 3 までの 1 点刻みでの回答を求めた．3 は専門家であれば自信を持って不法行為成否を判断できることを意味し，0 は全く不可能であること意味する．図 4 に回答の集計結果を示す．0 または 1 と回答された事例は 75%を占め，モデルが誤って予測した事例は，実際に難易度の高い事例であることが示唆された[7)].

7)　より詳細な分析:https://arxiv.org/abs/2312.00480