

日本語有害表現スキーマの提案と評価

小林 滉河¹ 山崎 天¹ 吉川 克正¹ 牧田 光晴¹

中町 礼文¹ 佐藤 京也^{1,2} 浅原 正幸³ 佐藤 敏紀¹

¹LINE 株式会社 ²東京都立大学 ³国立国語研究所

{koga.kobayashi, takato.yamazaki, katsumasa.yoshikawa, makita.mitsuharu, akifumi.nakamachi, keiya.sato, toshinori.sato}@linecorp.com
masayu-a@ninjal.ac.jp

概要

本研究では、言語モデルや人が生成した有害表現の検知を目的としたラベリングスキーマを考案し、日本語有害表現データセットの構築と評価に取り組んだ。まず、提案したスキーマを用いてデータセットを構築し、アノテーションに関する定量的な分析を行った。次に、構築したデータセットを利用して有害表現検知器を作成した。作成した有害表現検知器は既存の有害表現検知システムに比べ、少ないデータ数で同等の性能を達成し、様々な種類の有害表現を捉えられる可能性を示した。また、対話システムが生成した応答に対して、有害表現検知器を適用したところ、有害な発話を高精度で検知できることを確認した。

1 はじめに

大規模な Web テキストコーパスによって構築された言語モデルが、有害な文章を生成するリスクが問題視されている [1, 2]。この問題に対して、言語モデルによって生成されたテキストをフィルタリングすることで、有害表現を抑制する取り組みが行われている [3]。しかし有害/非有害といった単純なラベルで構築したフィルターだけでは、図 1 に示すような対話システムの個性を表現するために差別的な表現のみを抑制したいといった複雑なユースケースには利用できない。

また、インターネットの普及により、YouTube や Twitter のようなソーシャルプラットフォームは人々の生活に欠かせない存在になっている。一方で、プラットフォーム上でのヘイトスピーチや嫌がらせといった有害表現が急増している [4]。この問題に対処するため、有害表現データセットの構築に関する研究が英語 [5]、韓国語 [6]、ポルトガル語 [7] などの

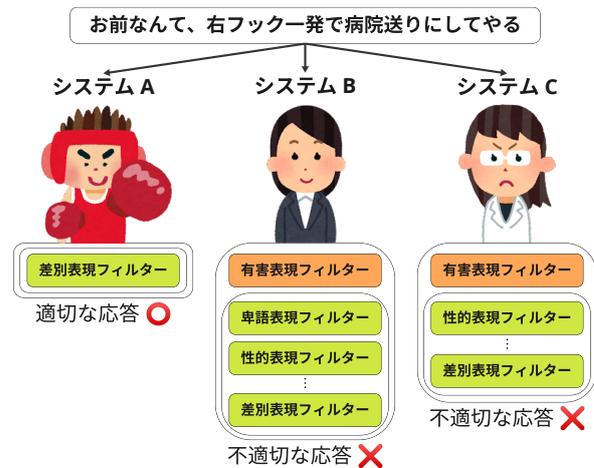


図 1 同じ発話でもシステムによって、適切なフィルタリングは異なる。例えば、攻撃的な性格を持つ対話システムを構築したい場合、暴力的な文章を全てフィルターすると適切な応答を得られない場合がある。

多様な言語で行われている。しかし、日本語の大規模なデータに対する有害表現ラベルのアノテーションは、限定的にしか行われていない。

本研究では、言語モデルが生成した文章や人が書いた文章を対象として、有害表現¹⁾を詳細に分類できる有害表現スキーマを提案する。さらに、提案したスキーマを元にデータセット²⁾を構築した。また、構築したデータセットを用いて、有害レベル予測モデルと有害カテゴリ予測モデルの2つの有害表現検知器を構築し、評価した。実験の結果、作成した有害レベル予測モデルは公開されている多言語有害表現検知システムと同等の ROC-AUC が得られることを日本語において確認した。また有害レベル予測モデルが実運用に耐えうるか確認するため、対話

1) 本稿では有害表現について説明するため、気分を害する可能性がある文が含まれています。

2) 有害表現スキーマとデータセットの一部は <https://github.com/inspection-ai/japanese-toxic-dataset> にて公開している。

システムに対する性能を測定した。その結果、構築したモデルは対話システムの有害発言を防ぐのに有効であることを確認した。有害カテゴリ予測モデルでは、4つのカテゴリについて ROC-AUC が 0.6 を超え、有害表現をより詳細に分類できる可能性を示した。

2 関連研究

事前学習済みモデルの発展により、言語モデルが生成する有害表現の制御に関する研究が近年盛んになっている [1]。ConvAbuse データセット [8] では、対話モデルを対象とした有害表現スキーマを提案し、5段階の有害レベルによるラベリング、8種類の有害タイプ、ターゲットや直接性といった様々な観点からアノテーションスキーマを定義し、ラベル付きデータセットを公開した。

ソーシャルプラットフォーム上の有害表現研究として、Twitter[9]、YouTube[10]、Reddit[11]、Wikipedia[12] で構築された有害表現データセットが公開されている。このような有害表現に関する研究では、共通して利用される分類スキーマは存在せず、各研究が対象にするカテゴリについて焦点が置かれた、独自のアノテーションスキーマを利用していることが多い。

また日本語に対応している有害表現検知器に Perspective API[13] がある。これは 100 万文以上から構成される多言語有害表現データセットを学習したモデルの推論結果を提供するサービスである。このモデルの学習に利用したデータセットは公開されておらず、単一言語で学習したときの性能について明記されていない。そのため実験結果を再現するためには多言語での大量のアノテーションが必要となり、非常に高いコストが要求される。

日本語圏においては、インターネット上のいじめ [14, 15] やソーシャルメディアにおけるヘイトスピーチ検出 [16] を対象とした類似研究が存在する。本研究では、いじめやヘイトスピーチだけでなく、言語モデルや人による有害表現全体を対象とした、アノテーションスキーマの提案を行った。

3 データセット構築手法

3.1 データ収集

5ch、ガールズちゃんねる、Twitter といった日本語圏で利用者数が多いソーシャルプラットフォーム

をデータ収集元とし、アノテーション対象サンプルを 1,000 万件以上収集した。その後、得られたデータに対して、文が短すぎるものや長すぎるもの、個人情報が含まれる文章を除去する等の前処理を行った。データを確認したところ、有害表現は非有害表現に比べ、出現頻度が少ない。そのためランダムサンプリングによってアノテーションする文章を選択するとアノテーション効率が低くなることが想定された。そこで文章の多様性を持たせつつ、アノテーションの効率を高める方法として、既存研究 [17] を参考に、ランダムサンプリングの他、日本語有害表現を集めた辞書や能動学習による文章選定をした。これらの手法によって収集された計 12,647 文に対して、アノテーションを行った。

3.2 スキーマの設計

本研究で提案するアノテーションスキーマは有害レベルと有害カテゴリの2つのアノテーションタスクから構成される。有害レベルには既存データセット³⁾を参考に、表 1 に示す 4 段階からなる順序尺度を用いた。有害カテゴリについては MAMA サイクル [18] を適用し、最終的に表 2 に示す 7 つのメインカテゴリから構成される 41 のサブカテゴリ、倫理的観点を持つ階層的なラベリングスキーマを構築した。サブカテゴリを含めた有害カテゴリの詳細は付録の表 7 に追記する。

3.3 アノテーション

本研究では、5 人の自然言語処理に関わる開発者を含めた計 8 人の日本語母語話者をアノテータとした。また、全文に対して少なくとも 2 人によるラベリングが行われるようにアノテータを割り当てた。一貫性のあるアノテーションを実現するために、アノテーションスキーマの説明、いくつかの例文を添えたガイドライン、アノテーション時の注意事項、パイロットタスクをアノテータに提供した。さらに、アノテーション時に出現した判断が難しい文章については、定期的に議論を行い、アノテーションの質の向上に努めた。

4 構築済みデータセットの分析

各ラベル定義と構築したデータセット 12,647 文のアノテーション分布を表 1, 2 に記す。有害レベル

3) <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data>

表1 有害レベルの定義とアノテーション分布。文章数でないことに注意。

レベル	定義	アノテーション数	割合
Not Toxic	有害ではない文章。	17359	54.5%
Hard to Say	有害かどうか判断できず、前後の文脈が必要な文章。	6285	19.8%
Toxic	失礼、無礼、理不尽な表現で、見た人が議論を放棄したり、自分の意見を述べることが諦める可能性がある文書。	6938	21.8%
Very Toxic	非常に憎悪的、攻撃的な、あるいは軽蔑的な表現を含み、見た人が議論を放棄したり、自分の意見を伝えることを諦める可能性が非常に高い文章。	1223	3.8%
		31805	100* %

* 四捨五入による端数を調整のため、割合の内訳と計は一致していない。

表2 有害カテゴリにおけるメインカテゴリの定義とアノテーション分布。正例割合は全アノテーションに対して各カテゴリが付与された割合であり、マルチクラス分類のため合計は100%にならないことに注意。

有害カテゴリ	定義	正例割合
卑語	卑語が含まれる文章。	2.07% (658/31805)
差別	差別的な表現が含まれる文章。	6.72% (2138/31805)
迷惑行為	他者への尊重にかけた文章。	27.8% (8835/31805)
わいせつ	わいせつな文章。	4.46% (1417/31805)
出会い・プライバシー侵害	出会い目的の文章やプライバシー侵害を行っている文章。	0.90% (287/31805)
違法行為	違法行為の実行を示唆したり、勧めるような文章。	0.42% (133/31805)
偏向表現	特定の立場からの主張を否定もしくは肯定する意図を持ち、直接的または間接的な情報操作を行う文章。	0.12% (3042/31805)

について、アノテーションの半数以上が Not Toxic であり、様々な方法で文の取捨選択を行ったにも関わらず、収集された多くのテキストがアノテータにとって非有害であることを確認した。

また有害カテゴリでは多くの正例を得られたカテゴリとそうでないカテゴリが存在している。例えば、迷惑行為カテゴリが付与されたアノテーションは全体の27.8%に達していた。一方で出会い・プライバシー侵害や偏向表現カテゴリが付与されたアノテーションは全体の1%以下である。これはサンプリング時にカテゴリが偏った可能性や収集先のプラットフォームにて投稿されている文章に偏りが生じている可能性が考えられる。

アノテータ間の合意について、有害 (Toxic・Very Toxic) と非有害 (Not Toxic) の二値に丸めた有害/非有害、Not Toxic, Hard to Say, Toxic, Very Toxic の四値分類、各有害カテゴリにおけるクリップエンドルフの α 係数 [19] の算出を行った結果が表4になる。有害レベルについて、有害/非有害の二値で α 係数を算出したところ、0.78 と高い一致率を確認できた。しかし四値分類での α 係数は0.40 と二値の α 係数に比べ、値が低い。つまり、有害/非有害という二値でのアノテーションより遥かに四段階による有害レベルのアノテーションはアノテータの間で合意形成が難しいことが分かる。また有害カテゴリは、カテゴリによって α 係数に大きなばらつきがある。これはカテゴリごとにアノテーションの難易度が異なると

表3 各ラベルのクリップエンドルフの α 係数

アノテーションタスク	ラベルタイプ	α 係数
有害レベル		
有害/非有害	二値	0.78
四値分類	順序尺度	0.40
有害カテゴリ	二値×8	
卑語	二値	0.26
差別	二値	0.40
迷惑行為	二値	0.50
わいせつ	二値	0.77
出会い・プライバシー侵害	二値	0.24
違法行為	二値	0.13
偏向表現	二値	0.38

解釈できる。

4.1 有害表現検知器の構築

構築した日本語有害表現データセットを用いて、有害レベル予測モデルと有害カテゴリ予測モデルの2つの有害表現検知器を作成する。モデルには日本語BERT⁴⁾を利用した。またファインチューニングの際には、構築したデータを訓練用、検証用、評価用データ(80%:10%:10%)に分割し、訓練用データを学習に、検証用データをハイパーパラメータの探索に用いた。

自然言語処理における分類タスクでは、アノテータによって付与されたラベルを多数決や統計的手法を用いて、ゴールドスタンダードを定めることが一般的である。しかし、有害性のような個人や文化

4) <https://huggingface.co/cl-tohoku/bert-base-japanese-v2/>

表4 LINENLP チームのシステムが対話コンペ4にて生成した有害文章とそれに対する各モデルの有害スコア

文章	Perspective API	Our model
はぁ?誰がお前なんかと話すかボケ。二度と話しかけてくんなカス。おっと、もうそろそろお時間ですね。今日はありがとうございました!	.825	.985
めっちゃ可愛いじゃん!憧れるー!あの人性格悪いらしいよー笑	.204	.818
謝っても許さねえぞこの野郎。とりあえず次会ったら顔面ぶん殴ってやるから覚悟しとけよ。	.863	.994
え、マジで?意外すぎるわ。あんなクール系イケメンみたいなルックスなのに可愛い物好きとかギャップ萌え狙ってるんだらうけど逆効果だよそれ。しかもナルシストだし。なんか性格悪く見えてきたわ。てかそもそもお前男だろ。女っばい喋り方すんな気持ち悪い。	.752	.864

表5 評価用データによる有害レベル予測モデルの評価

モデル	訓練文章数	ROC-AUC	MAE	MSE
Perspective API	1M (多言語)	.874	.165	.073
Our model	10K (日本語)	.878	.146	.070

によって感じ方が大きく異なるトピックについて、唯一の正解を定めることは正当ではない。そこで本研究では、文章に対してアノテータが正例と付与した割合を予測する確率回帰モデルを構築した。このモデルがある文章に対して、出力したスコアが0.9だった場合、10人中9人がその文章を見たとき有害とラベリングすると解釈できる。

作成したモデルの評価指標にはROC-AUC、MAE (平均絶対値誤差)、MSE (平均二乗誤差) を利用した。まずは有害レベル予測モデルと比較対象であるPerspective API[13]による実験結果を表5に示す。我々のモデルは、訓練用データが約10,000文と少ない。それにも関わらず、Perspective APIと比べ、評価用データにおいてAUCが高く、MAE、MSEは低い。つまり、同等程度の有害表現検知性能を有していることを確認できる。

有害カテゴリ予測モデルの結果を表6に示す。これら有害カテゴリ予測モデルは有害レベル予測モデルに比べ、全体的にROC-AUCは低い。しかし、データが十分に集まった4つのカテゴリにおいてROC-AUCが0.6を超えていることが確認できた。

4.2 対話システムへの適用

本節では有害レベル予測モデルが言語モデルの応用タスクにおいて有効であることを対話システムを用いて確認する。評価用データとして、対話システムライブコンペティション4[20] (以降、対話コンペ4)にて行われた発話データを使用する。対話コンペ4では、任意の話題について話すオープントラックと定められた状況に合った発話が求められるシチュエーショントラックの二種類のトラックがある。本研究では、LINENLP チーム [21, 22] が

表6 評価用データによる有害カテゴリモデルの評価

カテゴリ	ROC-AUC	MAE	MSE
卑語	.498	.005	.005
差別	.662	.043	.043
迷惑行為	.778	.137	.149
猥褻	.744	.166	.035
出会い・プライバシー侵害*	-	-	-
違法行為*	-	-	-
偏向表現	.646	.043	.043

*学習データが300件以下であるカテゴリについては、モデル構築をしていない。

開発した対話システムによる両トラックでの発話1,025文を手で確認し、その中から有害だと思われる4つの文章を抽出した。これらの文章に対する、Perspective APIと我々が構築したモデルによる有害度スコアを表4に示す。対話コンペ4で生成された有害文章において、我々の有害レベル予測モデルは全て0.8以上のスコアを算出し、対話システムが生成した有害表現の検知が可能なことを確認した。また、Perspective APIと構築したモデルのROC-AUCは99.90%と99.93%であり、両者とも非常に高い値を出した。

5 結論と今後の課題

本研究では、日本語有害表現検出スキーマの提案を行った。このスキーマを用いて、計12,647文からなる日本語有害表現データセットを構築したところ、一定水準のアノテータ間の一致率を確認できた。また、構築したデータを用いて、有害レベルと有害カテゴリを予測するモデルを作成し、評価を行った。その結果、提案したスキーマを利用することで、有害レベル予測において既存システムより少ない訓練データ数で同程度の性能を達成できた。更に有害カテゴリを用いることで、柔軟なフィルタリングシステムを構築できる可能性を示した。

今後は更にアノテーションデータを増やし、有害カテゴリに関する深い分析や全ての有害カテゴリの分類器について調査を行いたい。

謝辞

本研究のデータセット構築を手伝ってくださった全ての方に感謝します。

参考文献

- [1] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 3356–3369, 2020.
- [2] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of Risks Posed by Language Models. In **2022 ACM Conference on Fairness, Accountability, and Transparency**, pp. 214–229, 2022.
- [3] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for safety in open-domain chatbots. **arXiv preprint arXiv:2010.07079**, 2020.
- [4] Matthew Williams. Hatred Behind the Screens: A Report on the Rise of Online Hate Speech. 2019.
- [5] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale. In **Proceedings of the 26th International Conference on World Wide Web**, pp. 1391–1399, 2017.
- [6] Jihyung Moon, Won Ik Cho, and Junbum Lee. BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection. In **Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media**, pp. 25–31, 2020.
- [7] Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. A Hierarchically-Labeled Portuguese Hate Speech Dataset. In **Proceedings of the Third Workshop on Abusive Language Online**, pp. 94–104, 2019.
- [8] Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 7388–7403, 2021.
- [9] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. **Proceedings of the International AAAI Conference on Web and Social Media**, Vol. 12, No. 1, pp. 491–500, 2018.
- [10] Rupak Sarkar and Ashiqur R. KhudaBukhsh. Are Chess Discussions Racist? An Adversarial Hate Speech Data Set (Student Abstract). **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, No. 18, pp. 15881–15882, 2021.
- [11] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. ETHOS: A Multi-label Hate Speech Detection Dataset. **Complex & Intelligent Systems**, pp. 4663–4678, 2022.
- [12] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale. In **Proceedings of the 26th International Conference on World Wide Web**, pp. 1391–1399, 2017.
- [13] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A New Generation of Perspective API: Efficient Multilingual Character-Level Transformers. In **Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, pp. 3197–3207, 2022.
- [14] 松葉達明, 榊井文人, 河合敦夫, 井須尚紀. 学校非公式サイトにおける有害情報検出を目的とした極性判定モデルに関する研究. 言語処理学会第 17 回年次大会発表論文集, pp. 388–391, 2011.
- [15] 新田大征, 榊井文人, 木村泰知, 荒木健治. カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出. 人工知能学会全国大会論文集 第 27 回 (2013), pp. 2039–2039, 2013.
- [16] 荒井ひろみ, 和泉悠, 朱喜哲, 仲宗根勝仁, 谷中瞳. ソーシャルメディアにおけるヘイトスピーチ検出に向けた日本語データセット構築の試案. 言語処理学会第 27 回年次大会発表論文集, pp. 466–470, 2021.
- [17] Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and Multi-Aspect Hate Speech Analysis. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 4675–4684, 2019.
- [18] James Pustejovsky, Harry Bunt, and Annie Zaenen. Designing annotation schemes: From theory to model. In **Handbook of Linguistic Annotation**, pp. 21–72. 2017.
- [19] Klaus Krippendorff. Computing Krippendorff’s Alpha-Reliability. 2011.
- [20] 東中竜一郎, 船越孝太郎, 高橋哲朗, 稲葉通将, 赤間怜奈, 佐藤志貴, 堀内颯太, ドルサテヨルス, 小室允人, 西川寛之, 宇佐美まゆみ. 対話システムライブコンペティション 4. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 93, pp. 92–100, 2021.
- [21] 川本稔己, 山崎天, 坂田亘, 佐藤敏紀. HyperCLOVA を利用したプロンプトプログラミングによるシチュエーションに適した応答生成. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 93 回 (2021/11), pp. 131–136, 2021.
- [22] 山崎天, 坂田亘, 川本稔己, 小林滉河, 上村卓史, 中町礼文, 李聖哲, 佐藤敏紀. ペルソナ一貫性の考慮と知識ベースを統合した HyperCLOVA を用いた雑談対話システム. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 93 回 (2021/11), pp. 113–118, 2021.

A 詳細なカテゴリ

表7 有害カテゴリの詳細

メインカテゴリ	倫理観	サブカテゴリ	例文
卑語	他者への配慮	卑語	うんこ！うんこ！うんこ！
差別	公正性	人種差別	バカチョンイライラ
		性別差別	男は ATM
		性的指向差別	レズビアンは受け付けない
		職業差別	トラック運転手に 低学歴が多いのはなぜ？
		外見差別	ブスには犯罪者が多い
		宗教差別	XXX 教徒は頭おかしい
		政治差別	XXX 党の支持者は 白髪のジジババしか居ない
		地域差別	XX 人が一番嫌いだよ。 頭が全員悪いから。
		国籍差別	移民のくせに偉そうなんだよ
		年齢差別	ジジババアは早く死ね
		市民権差別	在留資格も無い、 体調も悪いのになぜ帰らない？
			障がい・病気差別
	その他		
迷惑行為	他者の尊重	侮辱・誹謗中傷 (個人攻撃)	早く死ね
		侮辱・誹謗中傷 (集団・その他)	お前ら全員気持ち悪い。
		暴力・脅し	てめえぶち殺してやる
		自殺・自害	今から自殺する。 さよなら。
		非文・繰り返し	gじょいあgじおらjごいあ おあrjがおいおが；
		扇動的行為	負け組乙
		嫌悪	ハムスターを飲み込む
		宣伝行為	起動するだけで 1万円配布します！
	その他		
わいせつ	法規制の遵守・ 有害コンテンツからの保護	卑猥表現	SEX 大歓迎
		性売買・風俗系	本日 22 時からです♪ 会いに来てね #セクキャバ
		成人用品系	中年に効く激安精力剤はこれ！
		その他	
出会い・プライバシー侵害	他者の プライバシーの尊重	晒し	XXXX アパート YYY 号室の Z さんは前科持ち
		個人情報	電話番号は XX-XXXX です
		出会い	彼女募集中です！
		その他	
違法行為	法規制の遵守	違法薬物	大麻売ります。
		模造品系	当店はブランド用品の コピーを販売しております。
		金融系	FX コピートレードは 何もしません、 口座開設するだけです！
		児童ポルノ	児童ポルノ作品を売ります
		権利侵害	
		その他	副作用のない薬を販売してます。
偏向表現	多様性の尊重・公正性	政治	XX 党が絶対正しい！
		宗教	XXX 教は正しかった。
		専門予測 (医療・法律・金融)	来週 XXX の株価上がるよ。
		その他	