オンライン誹謗中傷検出に向けた裁判例データセット

久田祥平　若宮翔子　荒牧英治
奈良先端科学技術大学院大学
{s-hisada, wakamiya, aramaki}@is.naist.jp

掲載号の情報

2024 年 31 巻 4 号 p. 1598-1634
doi: https://doi.org/10.5715/jnlp.31.1598

概要

本論文では，誹謗中傷の事例を扱う性質上，不快な表現が含まれることにご注意下さい．

デジタルプラットフォーム上の誹謗中傷は、現代社会で深刻な問題となっており、多くの国や地域で法的規制やプラットフォーム事業者による対策が進められている。しかし、誹謗中傷の定義や判断基準は文化的背景や社会的通念、さらには発言者と受け手の文脈によって大きく左右されるため、その自動検出や評価には多くの困難が伴う。

そこで本研究では、日本国内の裁判例に基づいた日本語誹謗中傷検出データセットの構築を提案する[1]。具体的には、発信者情報開示請求事件や損害賠償請求事件から抽出したオンライン上の投稿について、原告が主張した権利侵害（名誉権、名誉感情、プライバシー権、私生活の平穏、営業権、その他の人格権・人格的利益）と裁判所の判断（認容・否認など）をラベルとして付与した。これにより、誹謗中傷に対する基準を、社会の問題を反映した専門家の判断としてより適切に活用できる。

しかし、法学専門家による複数アノテーターのラベル付けでも、一致率が低い部分があり、裁判例を構造化データとするアノテーションに関して多数の課題が見つかった。裁判例の判決文は必ずしも統一的・構造的に記載されていないため、内容の順序や用語の揺れ、上級審判決特有の記載形式などに起因する作業の難しさが顕在化したほか、一部のラベルに対する法的概念の解釈の違いも同定を難しくしている。

本データセットを用いた実験では、投稿が権利侵害となるかを判別する二値分類タスクと、どの権利の問題が争われているかを予測するマルチラベル分類タスクを実施した。その結果、誹謗中傷を適切に検出するには前後の投稿など文脈情報を扱うことが重要である一方、プライバシー権や私生活の平穏などサンプル数の少ない権利を分類することの難しさも浮き彫りとなった。こうした分析を通じて、裁判例を構造化データとして不法行為としての誹謗中傷を検出するタスクに伴う課題が洗い出された。今後は法的観点と現実シナリオを統合して検討することが不可欠である。

本研究は、法的観点に基づく客観性を備えたデータセットを整備することで、実社会で生起する誹謗中傷問題に即した自動検出を可能にする点に意義がある。また、自然言語処理分野における社会科学的知見の活用を促進し、オンライン上の不適切コンテンツ対策における説明可能性やアカウンタビリティの向上も期待される。こうした取り組みを通じて、より公正で安全なデジタル社会の実現を目指すことが本研究の最終的なゴールである。[1)]

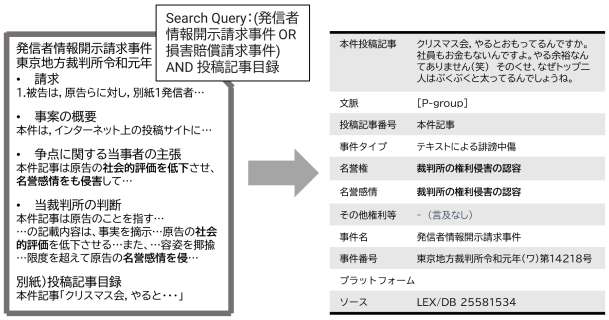

図 1　本研究の概要：民事事件の裁判例から誹謗中傷検出に向けたデータセットを作成

1）　この論文は，言語処理学会第 29 回年次大会発表「権利侵害と不快さの間：日本語人権侵害表現データセット」を拡張したものである．