オンライン市民社会に向けて: 不法行為としての誹謗中傷の自動検出技術の事例研究

氏　　名　　久田祥平

研 究 室 名　　ソーシャル・コンピューティング研究室

主指導教員名（論文博士の場合は推薦教員名）　　荒牧 英治

内 容 梗 概（1ページ目に収めること）

インターネットの普及により、オンライン上での誹謗中傷や名誉毀損が深刻な社会問題となっている。匿名性を悪用した攻撃的な書き込みが被害者に精神的苦痛を与え、社会的評価を損なうケースが増加している。対処方法は大きく二つある。一つは、プラットフォームが利用規約に基づきコンテンツモデレーションを行う方法だが、運営者の自主性やコストの問題から十分な対処が行われない場合がある。もう一つは、被害者が法的手段を用いて投稿の削除や差し止めを求める方法だが、手続きが複雑で迅速な救済が困難である。既存の研究は主にプラットフォーム内の対処を支援するもので、法的根拠に基づく技術的支援は限られている。

法的根拠に基づく技術支援に向けて、日本の民法上の不法行為としての誹謗中傷を自動検出する AI モデルを開発するために 2 つの研究に取り組んだ。

1 つ目の研究では、日本の裁判例を基に、誹謗中傷検出のための日本語データセットを構築した。名誉権や名誉感情の侵害に該当する発言を抽出し、法的判断をラベルとして付与した。このデータセットは、実際の社会問題に即し、法律の専門知識を活用して誹謗中傷の基準を明確化している点が特徴である。

2 つ目の研究では、構築したデータセットを用いて深層学習による誹謗中傷検出モデルを開発した。さらに、モデルの判断根拠をユーザに提示する説明生成機能を組み込み、説明可能性を検証した。法律専門家による評価を通じて、モデルの精度や説明の妥当性を確認し、コンテンツモデレーションの透明性と信頼性を高めることを目指した。

本研究は、法的根拠に基づく誹謗中傷の自動検出という新たなアプローチを提案し、技術と法学の融合を通じて学際的研究の発展に寄与するものである。違法性の判断を自動化し、その理由を明確に説明することで、被害者や第三者機関が迅速かつ適切に対処できる技術的支援を提供し、オンライン空間における社会規範の遵守と持続可能なデジタル社会の実現に貢献することが期待される。