言い換えによるデータ拡張に基づく暗黙的な有害テキストの自動検出
Implicit Toxic Text Detection Based on Data Augmentation by Paraphrase

北陸先端科学技術大学院大学　2450006
氏名 志田宗久
主任研究指導教員氏名　白井清昭

1. はじめに

インターネットおよびソーシャルメディアの普及に伴い、誹謗中傷やヘイトスピーチなどの攻撃的な書き込み(有害テキスト)の拡散が深刻な社会問題となっている。従来の検出技術は、差別用語や侮辱語などの明示的な攻撃表現を含むテキストに対しては高い精度を達成している。しかし、皮肉や婉曲表現、ステレオタイプに基づく攻撃など、攻撃的な単語を明示的に含まない暗黙的な有害テキストの検出は依然として困難である[1]。暗黙的な有害表現は文脈依存性が高く、表層的な語彙のみでは無害なテキストとの区別がつきにくい。さらに、深層学習モデルの学習に不可欠な暗黙的有害表現を含む大規模なラベル付きデータセットが不足していることが研究の進展を妨げている。人手によるデータ作成はコストが高いため、既存資源の有効活用が求められる。本研究では、このデータの過疎性を解決するため、既存の「明示的な有害テキスト」データセットから「暗黙的な有害テキスト」を疑似的に生成する手法を提案する。自動構築したデータセットから暗黙的な有害表現特有の特徴を学習することで、人手によるデータセット構築のコストをかけずに暗黙的有害テキストの検出性能を向上させる。また、逆翻訳によるデータ拡張や、感情分析や皮肉判定といった関連タスクとのマルチタスク学習を導入し、検出精度のさらなる向上を図る。

2. 研究方法

本研究では、暗黙的な有害テキストの検出性能向上のため、以下の 3 つのアプローチを組み合わせた手法を提案する。第一に、「言い換えによるデータ拡張」である。既存のデータセットに含まれる明示的な有害テキストから、辞書等を用いて有害語を特定し、BERT[2]の Masked Language Model (MLM)を用いて、文脈を保ちつつ無害または中立的な単語に言い換える。これにより、表層的には攻撃的な単語を含まないが、有害な意図や不快な意味合いを保持した疑似的な暗黙的有害テキストを自動的に生成し、訓練データとして利用する。第二に、「逆翻訳によるデータ拡張」[3]である。生成した疑似データに対し、他言語（中国語、フランス語、ドイツ語、日本語）への翻訳と英語への再翻訳を行うことで、意味を保ちつつ表現の多様性を持たせたテキストを生成し、訓練データの量を増強する。これによりデータセットの語彙・表現のバリエーションを拡充し、汎化性能の向上を図る。これら一連のデータセット構築プロセスを図 1 に示す。第三に、「マルチタスク学習」[4]の導入である。有害性判定という主タスクに加え、関連性の高い「感情分析」および「皮肉判定」を補助タスクとして、これらのタスクのデータセットを用いてひとつの分類モデルを学習する。これにより、暗黙的な有害テキストに共通する言語的特徴や、文脈に潜む否定的な感情、皮肉的なニュアンスをモデルに捉えさせることを狙いとする。提案手法の有効性を検証するため、英語および日本語のデータセットを用いた評価実験を行った。分類モデルには、BERT、RoBERTa、DistilBERT などの事前学習済み言語モデルに加え、Llama-3 や Swallow といった大規模言語モデル（LLM）をファインチューニングしたモデルも検証した。

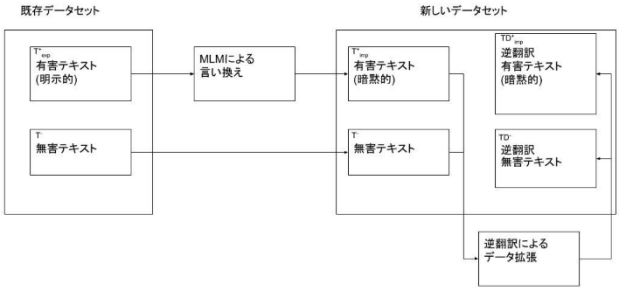

図 1 有害性判定モデルの訓練データの構築

3. 結果と考察

英語データセット[5]を用いた実験の結果、明示的な有害テキストのみを学習したベースラインモデルは、暗黙的な有害テキストに対する再現率が 0.06～0.07、F1 が 0.11~0.13 と著しく低いことが確認された。これは、モデルが表層的な攻撃語の有無に強く依存しており、それらを含まない暗黙的な有害性を看過していることを示唆している。一方、提案手法である言い換えによる疑似データを用いたモデルは、再現率および F1 スコアにおいて大幅な改善を示した。特に、DistilBERT を用いたモデルでは再現率が 0.80、F1 が 0.62 まで向上した。攻撃的な単語をマスクし、文脈に基づいた推論をモデルに強いることで、暗黙的な有害性の特徴を効果

的に学習できたと考えられる。マルチタスク学習に関しては、特に皮肉判定を補助タスクとした場合に性能向上が顕著であった。大規模言語モデルである Llama-3 を用いた実験では、皮肉判定を併用することで再現率 0.96、F1 スコア 0.67 を達成し、本研究における全実験の中で最も高い性能を示した。これは、暗黙的な有害表現の多くが皮肉や反語の構造を持つため、タスク間で共有される言語的特徴が有効に機能したためと推察される。次に、日本語データセットを用いた実験結果について述べる。日本語においても、明示的有害テキストのみで学習した場合の再現率は 0.11 と低かったが、提案手法を用いることで、BERT モデルでの F1 スコアは 0.20 から 0.56 へと向上した。しかし、日本語実験においては、英語実験とは異なる傾向も見られた。まず、逆翻訳によるデータ拡張が日本語では性能を低下させる結果となった。これは、日本語がハイコンテクストな言語であり、機械翻訳の過程で文脈の機微やニュアンスが失われたり、文が不自然になったりしたことで、訓練データにノイズが多く混入したためと考えられる。また、感情分析とのマルチタスク学習においても、日本語では性能向上が見られなかった。これは、感情語を中立的な語に言い換えたデータを用いたことで、モデルが「感情語がない＝無害」というバイアスを強めてしまい、感情語を伴わない暗黙的有害テキストの見落とし（False Negative）が増加したためであると分析される。大規模言語モデル（Swallow）を用いたモデルでは、BERT よりも高い F1 スコア（0.57）が得られた。これは、LLM が持つ豊富な事前知識と文脈理解能力が、語用論的推論を必要とする暗黙的有害テキストの検出において有効的に働いたことを示している。最後に、明示的な有害テキストと暗黙的な有害テキストが混在する現実的なデータセットを用いた評価においては、提案手法が安定した性能を示した。特に英語の Llama-3 を用いたモデルは F1 スコア 0.80、正解率 0.80 を達成しており、明示的な有害性への検出能力を維持しつつ、暗黙的な有害性へも適応できるバランスの取れた手法であることが実証された。

4. まとめ

本研究では、検出が困難な暗黙的な有害テキストに対し、既存の明示的な有害テキストデータセットを活用した言い換えによるデータ拡張手法を提案した。 英語および日本語における評価実験の結果、提案手法はデータの過疎性を緩和し、暗黙的有害テキストの検出性能を大幅に向上させる有効なアプローチであることが確認された。特に、言い換えによる疑似データの生成は言語を問わず有効であり、コストのかかる人手によるデータ作成を行わずとも、既存資源の転用によって有害テキスト検出モデルを構築できる点に大きな意義がある。 一方で、逆翻訳やマルチタスク学習の効果には言語による差異が見られ、特に日本語のような文脈依存性の高い言語においては、データ拡張の手法や補助タスクの選定に慎重な検討が必要であることが明らかとなった。今後は、より自然で多様な言い換え生成手法の探求や、マルチモーダル情報を用いた検出モデルへの拡張などが課題として挙げられる。

参考文献

[1] M. Wiegand, J. Ruppenhofer, and E. Eder, “Implicitly Abusive Language – What Does It Actually Look Like and Why Are We Not Getting There?,” Proc. NAACL, pp. 576–587, Jun. 2021.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proc. NAACL-HLT, vol. 1, pp. 4171–4186, Jun. 2019.

[3] D. R. Beddiar, M. S. Jahan, and M. Oussalah, “Data Expansion using Back Translation and Paraphrasing for Hate Speech Detection,” arXiv preprint arXiv:2106.04681, 2021.

[4] A. R. Jafari et al., “Fine-Grained Emotions Influence on Implicit Hate Speech Detection,” IEEE Access, vol. 11, pp. 105330–105341, 2023.

[5] A. Das et al., “OffensiveLang: A Community Based Implicit Offensive Language Dataset,” IEEE Access, vol. 12, pp. 39289–39306, 2024.

発表論文・口頭発表

志田 宗久, 白井 清昭, “言い換えによるデータ拡張に基づく暗黙的な有害テキストの自動検出,” 情報処理学会 第 264 回自然言語処理研究発表会, Vol.2025-NL-264, 2025 年 7 月.