修士論文

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

志田 宗久

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
融合科学共同専攻

令和８年３月

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。

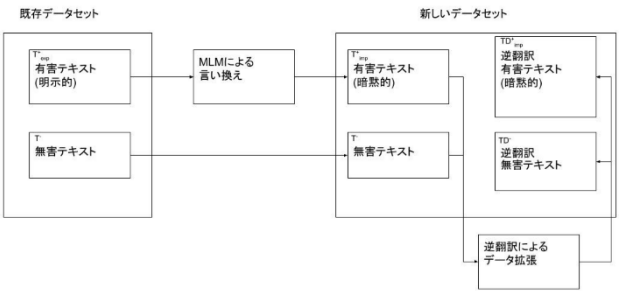

図 1 有害性判定モデルの訓練データの構築

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.

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

近年，インターネットおよびソーシャルネットワーキングサービス（SNS）の普及により，誰もが容易に情報を発信し，他者と交流することが可能となった．しかしその一方で，誹謗中傷，ヘイトスピーチ，ネットいじめといった攻撃的な書き込み（有害テキスト）の問題が深刻化している．このような背景から，健全なオンライン空間の維持を目的として，有害テキストを自動的に検出し，フィルタリングや警告を行うシステムの構築が強く求められている．

従来の有害テキスト検出技術の多くは，差別用語，侮辱語，卑語といった特定の攻撃的な単語の有無に基づく手法や，それらを特徴量として学習した機械学習モデルに依存してきた [24]．これらの手法は,「死ね」「馬鹿」といった明示的な攻撃表現（Explicit Toxicity）に対しては高い検出精度を達成している．

だが，有害な意図は必ずしも攻撃的な単語によってのみ表現されるわけではない．皮肉，嫌味，婉曲表現，ステレオタイプに基づく偏見など，一見すると無害な単語のみで構成されているにもかかわらず，文脈や常識的知識に照らし合わせることで初めて攻撃性が明らかになる表現が存在する．これらは暗黙的な有害表現（Implicit Toxicity）と呼ばれ，従来の単語ベースの手法や，明示的な有害テキストのみで学習されたモデルでは検出が極めて困難である [4]．例えば,「こんなにミスをするなんてある意味すごいですね」という発言は相手のミスを嘲笑する皮肉だが，単語自体には肯定的な意味が含まれているため，機械的な判定では「称賛」と誤分類されやすい．Wiegandra らも指摘しているが，既存の検出モデルは明示的な表現には強いが，こうした潜在的な攻撃性を見逃す傾向にある [26]．

暗黙的な有害テキストの検出が困難である最大の要因の一つは，学習データの不足である．機械学習，特に近年の深層学習モデルが高い性能を発揮するためには，大量のラベル付きデータが必要不可欠である．しかし，既存の有害テキストデータセットの大部分は明示的な有害表現によって占められており，暗黙的な有害表現を含む事例は極めて少ない．また，暗黙的な有害性は解釈が主観に依存しやすく，アノテーション（ラベル付け）のコストが高いことから，大規模なデータセットを新規に構築することは容易ではない．したがって，限られたデータ資源の中で，いかにして暗黙的な有害テキストの特徴をモデルに学習させるかが，喫緊の課題となっている．

1.2 目的

本研究の目的は，明示的な攻撃語を含まない暗黙的な有害テキストの検出性能を向上させることである．前述の通り，暗黙的な有害テキスト検出における最大の障壁は，高品質かつ大規模な学習データの欠如にある．そこで本研究では，人手による大規模なデータセット作成という高コストな手法に頼るのではなく，既存の豊富な資源である「明示的な有害テキスト」や「関連タスクのデータセット」を有効活用するアプローチをとる．

具体的には，以下の2つの主要な手法を提案・検証する．第一に，言い換え（Paraphrasing）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に対し，攻撃的な単語をマスクし，文脈を保持したまま無害または中立的な単語に置換することで，疑似的な暗黙的有害テキストを自動生成する．これにより，表層的な攻撃語に依存せずに有害な意図を汲み取るための訓練データを人工的に増強し，モデルの汎化性能を高めることを目指す．

第二に，マルチタスク学習（Multi-Task Learning）の導入である．有害性判定という単一のタスクのみならず，感情分析（Sentiment Analysis）や皮肉判定（Sarcasm Detection）といった関連性の高いタスクのデータセットを同時に用いる．暗黙的な有害表現は，否定的な感情や皮肉といった要素と密接に関連しているため，これらのタスクから得られる言語的特徴を共有することで，データの過疎性を補い，検出精度の向上を図る．

また，本研究では対象言語として英語および日本語の両方を取り扱う．有害テキスト検出の研究は英語を対象とした研究が先行しているが，日本語においてもSNS 上の誹謗中傷は深刻な社会問題である．しかし，日本語はハイコンテクストな言語であり，主語の省略や婉曲的な言い回しが多用されるため，暗黙的な有害性の検出は英語以上に困難である可能性がある．本研究では，提案手法を英語データセットで検証するだけでなく，日本語データセットに対しても適用し，言語に依存しない汎用的な有効性や，言語特有の課題を明らかにする．

最終的には，小規模な暗黙的有害テキストデータセットと大規模な明示的有害テキストデータセットを用いた評価実験により，提案手法が暗黙的有害テキストの検出精度を向上させることを実証する．

1.3 本論文の構成

本論文の構成は以下のとおりである．

第 2 章では，有害テキスト検出に関する既存研究について概観する．特に明示的有害表現および暗黙的有害表現の検出に関する先行研究を整理する．あわせて，本研究の位置づけと既存手法の課題を明確にする．

第 3 章では，本研究で提案する暗黙的有害テキスト検出手法について述べる．具体的には，明示的に有害なテキストを対象とした言い換えによるデータ拡張手法

と，複数の関連タスクを同時に学習するマルチタスク学習の枠組みについて詳細に説明する．

第 4 章では，提案手法の有効性を検証するために実施した実験および評価について述べる．まず，評価に用いた明示的・暗黙的有害テキストデータセットの概要，ならびに学習および評価手順を示す．次に，これらの設定に基づく実験結果を提示し，言い換えによるデータ拡張およびマルチタスク学習が暗黙的有害テキスト検出性能に与える影響について考察する．

第 5 章では，本研究のまとめを行うとともに，提案手法によって得られた知見を整理し，今後の課題および将来の展望について述べる．

第2章　関連研究

2.1　有害テキスト分類

Wang らは，ナイーブベイズ，Convolutional Neural Network(CNN)，Long Short-Term Memory(LSTM) といった機械学習手法を有害コメントの分類に適用した [24]．単語埋め込みを用いて分類対象のコメントを特徴ベクトルに変換し，有害か無害かのラベルが付与されたデータセットを用いて分類モデルを学習した．その結果 LSTM や CNN は有害コメントを適切に分類できる一方，ナイーブベイズは誤分類が多いことが分かった．

Bhat らは，職場における E メールを「無礼」「陰口」「攻撃的」などのカテゴリに分類する手法を提案した [4]．上記のカテゴリがラベル付けされた ToxiScope データセットを構築し，分類対象の文とその文脈を入力とする BERT モデル [9] を同データセットを用いてファインチューニングした．

Albladi らは，ソーシャルメディア上のヘイトスピーチ検出において，キーワードマッチングや従来の機械学習手法が文脈依存性の高い表現の検出に課題を抱えていることを指摘し，大規模言語モデル (Large Language Model; LLM) を用いた検出手法に関する包括的なレビューを行った [1]．彼らは，BERT や RoBERTa[22]，GPT[1] シリーズなどの LLM が，その高度な文脈理解能力によってヘイトスピーチ検出の精度を大幅に向上させている現状を整理し，特に多言語対応や低リソース言語における有効性について論じた．一方で，LLM の導入に伴う課題として，膨大な計算コスト，モデルの解釈可能性の欠如，および学習データに内在するバイアスの増幅といった倫理的・技術的問題を挙げている．これらの分析に基づき，今後の研究の方向性として，効率的なモデルの開発，説明可能性の向上，およびマルチモーダル情報の統合などが重要であると提言した．

2.2　暗黙的な有害テキストの検出

Das らは，暗黙的な攻撃的テキストの検出に向けたデータセット OffensiveLang を構築した [7]．攻撃の対象となりうる「人種」「職業」のような様々な社会的カテゴリについて，ChatGPT を用いて暗黙的にこれらを攻撃するテキストを生成させ，人手でアノテーションを行った．また，人間と ChatGPT によるアノテー

[1] https://chatgpt.com/

ションを比較し，OffensiveLang の品質が十分に高いことを示した．さらに，暗黙的な攻撃テキスト判定のベースラインとして，OffensiveLang を用いて BERT や RoBERTa などの事前学習済み言語モデルをファインチューニングし，その性能を評価した．

Wiegand らは， 暗黙的な攻撃的テキストの検出が明示的な攻撃的テキストの検出よりも難しいことを指摘した [26]．既存の攻撃的テキストのデータセットでは明示的な表現が大部分を占め，暗黙的な攻撃テキストの検出モデルを学習するには不適切であると考察した．また,「ステレオタイプ」「加害者としての描写」「非人間化」「婉曲表現」など，8 種の暗黙的攻撃表現のタイプを定義し，各タイプの定量的な分析や分類の難易度に関する分析を行った．その上で，暗黙的攻撃テキストの検出精度を高めるに，汎用モデルではなく，タイプごとに専用のデータセットと分類器を構築する「分割統治」戦略を提案した．さらに，テキスト全体ではなく句や節といったより小さな言語単位に注目することで暗黙的攻撃テキスト検出の性能が高まる可能性を指摘した．

Han らは，既存の有害テキスト検出手法が皮肉やマイクロアグレッションなどの偽装された有害テキストを十分に検出できないという問題に着目し，これに対処する手法を提案した [13]．少数のラベル付き偽装有害テキスト（プロービング例）から既存の分類器が見逃していた多数の潜在的有害発言を発見するために，影響度解析手法を導入した．勾配積や影響関数といった機械学習モデルの説明手法を用いて，誤分類に強く影響を与えた訓練データ中の事例を特定し，それらを再注釈した上でモデルを再学習することで分類器を強化した．実験では，既存モデルの偽装有害テキストの F1 スコアが 1.2%だったのに対し，提案手法では最大 51.1%まで向上した．この際，明示的な攻撃的テキストに対する検出性能は大きく損われることはなかったと報告している．

Jafari らは，暗黙的なヘイトスピーチの検出において，投稿に含まれる詳細な感情情報が有効であることを示した [16]． Latent Hatred データセットを用いて，怒りや苛立ちといった特定の感情がヘイトスピーチの投稿に多く現れる一方で，非ヘイトスピーチの投稿では承認や好奇心といった感情が多いことを分析した．この知見をもとに，単一タスク学習 (Single-Task Learning; STL) および感情を補助情報として活用するマルチタスク学習 (Multi-Task Learning; MTL) によってヘイトスピーチ検出モデルを構築した．実験の結果，感情分類とヘイトスピーチ分類を同時に行う MTL モデルは，特に “incitement” や “white grievance” といった暗黙的カテゴリの検出において高い性能を示し，その F1 スコアは最大で 6 ポイント改善された．

ElSherief らは，従来のヘイトスピーチ研究の多くが明示的なヘイト表現に焦点を当てており，婉曲的で間接的な表現を含む暗黙的なヘイトスピーチへの対応が不十分であると指摘した [11].「白人の不満」「暴力の扇動」「劣等性の言語」「皮肉」「ステレオタイプと誤情報」「脅迫と威嚇」という 6 つの暗黙的ヘイトスピーチの分類体系を定義し，これに基づいて詳細なアノテーションを行った大規模データ

セット Latent Hatred を構築した．このデータセットは，各投稿に対してカテゴリラベルだけでなく攻撃対象や発言の含意に関する自然言語記述も付与されている点に特徴がある．実験では，BERT などの事前学習済み言語モデルをファインチューニングすることで，既存の商用 API などと比較して高い検出精度が得られることを示した．また，GPT-2 を用いて暗黙的なヘイトの意図を説明生成するタスクにも取り組み，モデルが人間によるアノテーションに近い妥当な説明を生成可能であることを確認した．さらに，誤分類の分析を通じて，コード化されたシンボル，常識的推論，談話関係の理解など，暗黙的ヘイトスピーチ検出における特有の課題を明らかにした．

Hartvigsen らは，既存の有害テキスト検出システムがマイノリティへの言及を過度に有害と判定してしまうバイアスや，露骨な表現を含まない暗黙的なヘイトスピーチの検出が困難であるという課題を指摘した [14] ．彼らはこの問題に対処するため，GPT-3 を用いたデモンストレーションベースのプロンプト学習により，13 のマイノリティグループに関する有害および非有害のテキストを含む 27 万件規模のデータセット「TOXIGEN」を構築した ．特に，分類器をデコーディングのループに組み込む敵対的生成手法 ALICE (Adversarial Language Imitation with Constrained Exemplars) を提案し，人間には有害とわかるが機械には無害に見えるような，検出困難なテキストの生成に成功している．評価の結果，生成されたテキストは人間による記述と区別がつかないほど自然であり，TOXIGEN を用いて既存の検出器をファインチューニングすることで，暗黙的な有害性の検出性能が大幅に向上することを実証した．

Wei らは，既存の自動検出手法が明示的なヘイトスピーチに対しては良好な性能を示す一方で，暗黙的なヘイトスピーチの検出には苦戦していることを指摘した [25]．彼らは，暗黙的ヘイトスピーチ検出のための新たな分類体系として，6 つの符号化戦略を定義した「codetypes」を導入した．さらに，この codetypes を検出モデルに統合する手法として，LLM に直接プロンプトを与えて分類させる手法と，符号化プロセスに codetypes を埋め込んで LLM をエンコーダとして利用する 2 つのアプローチを提案した．中国語と英語のデータセットを用いた実験により，codetypes の導入が言語を問わず暗黙的ヘイトスピーチの検出精度を向上させることを示した．

Chen らは，LLM を用いた有害テキスト検出において，検閲を回避するために巧妙化された「暗黙的な有害表現」の検出が，従来の単純なバイアス検出よりも高度な推論を必要とする点に着目した [6]．彼らは，認知科学と言語学の知見に基づき，LLM に語用論的な推論プロセスを促す新たなプロンプティング手法「Pragmatic Inference Chain (PIC)」を提案した．実際のオンライン上のやり取りから収集され，人間によって高い推論負荷が必要であると検証されたデータセットを用いて評価を行った結果，PIC プロンプトを適用することで，GPT-4o や Llama-3 といった最新の LLM における暗黙的有害テキストの識別成功率が，Chain-of-Thought（CoT）などのベースラインと比較して大幅に向上することを示した．

Lee らは，暗黙的なヘイトスピーチの検出において，人間がまず攻撃対象を特定し，その後に文脈との関係を解釈するという推論プロセスに着目した [21]．彼らはこのプロセスを模倣する新たな手法「AmpleHate」を提案した．この手法では，事前学習済みの固有表現抽出（Named Entity Recognition; NER）モデルを用いて明示的なターゲットを特定するとともに，[CLS] トークンを用いて暗黙的なターゲット情報を捉える．そして，これらターゲットと文脈との間のアテンションに基づく関係性を計算し，その関係性ベクトルを最終的な文表現に直接注入することで，暗黙的なヘイト判定に重要なシグナルを増幅させた．実験の結果，AmpleHate は複数のデータセットにおいて既存の最先端手法を上回る性能を達成したことを示した．

Cabrera らは，皮肉や嫌味，当てこすりといった非直接的なヘイトスピーチの検出が依然として課題であることを指摘し，皮肉の検出学習をヘイトスピーチ検出に転移させることの有効性を検証した [5]．彼らは，Reddit の皮肉データセットなどを用いて，CNN+LSTM および BERT+BiLSTM モデルに対し，皮肉検出による事前学習を行う戦略を提案した．具体的には，皮肉データのみで学習しヘイトスピーチでテストする手法と，皮肉学習後にヘイトスピーチでファインチューニングを行う順次転移学習の手法を比較した．実験の結果，皮肉による事前学習は BERT+BiLSTM モデルの性能を改善し，ETHOS データセットでは F1 スコアが 6 ポイント向上，Implicit Hate Corpus では暗黙的サンプルの precision が 7.8 ポイント向上するなど，皮肉の言語的特徴の学習が明示的・暗黙的双方のヘイトスピーチ検出に寄与することを示した．

2.3　テキスト分類におけるデータ拡張

Beddiar らは，ヘイトスピーチやネットいじめの検出性能を向上させるために，逆翻訳と言い換えを適用したデータ拡張手法を提案した [3]．英語文を他言語に翻訳し再翻訳することで多様な攻撃的テキストを新しく生成する手法と，Transformer ベースのモデルを用いた言い換え生成手法を用いて，元の訓練データを最大 20 倍に拡張した．これらの拡張データを用いて CNN や LSTM を学習させた結果，Offensive Language Identification Dataset[27] や Twitter ヘイトスピーチデータセットにおいて F1 スコアが最大で 13 ポイント向上し，特に CNN が安定して高い性能を示した．また，少数クラスのみに対して訓練事例を拡張するよりも全てのクラスの訓練事例を拡張する方が分類性能の向上に効果的であることを示した．

Kobayashi は，テキスト分類におけるデータ拡張手法として，双方向言語モデルを用いて文脈に適した単語へ置換を行う「Contextual Augmentation（文脈的拡張）」を提案した [19]．従来のシソーラスを用いた同義語置換は，置換可能な単語数が限られるという課題があったが，提案手法では文脈情報から文脈的に置き換え可能な関係にある単語を予測・サンプリングすることで，より多様な表現への書き換えを可能にした．また，文脈のみに基づく置換では元のラベル（感情極性

など）と矛盾する単語（例：ポジティブな文脈で「素晴らしい」を「ひどい」に置換してしまう等）が生成される可能性があるため，ラベル情報をネットワークに条件として与えることでラベル整合性を保つ「ラベル条件付き言語モデル」を導入した．CNN および RNN を用いた 6 種類のテキスト分類タスクでの実験の結果，提案手法は従来の同義語ベースの拡張手法と比較して高い汎化性能を示し，特にラベル条件付きモデルが最も高い精度を達成すると報告した．

Edunov らは，機械翻訳における逆翻訳の効果について大規模な調査を行い，合成データの生成手法が最終的な翻訳性能に与える影響を体系的に分析した [10]．彼らは，ターゲット言語の単言語コーパスからソース言語の合成データを生成する際，従来のビームサーチ（MAP 推定）ではなく，モデル分布からのサンプリングやノイズ付きビームサーチを用いることが有効であることを示した．実験の結果，規則的なビームサーチに比べて，サンプリング等で生成されたデータは表現の多様性に富み，モデルにより強力な学習信号を提供することで，翻訳精度を有意に向上させることが明らかになった．さらに，数億文規模の単言語データを用いた大規模実験では当時の最高性能（SOTA）を達成し，ドメインが一致する場合には合成データが実在するパラレルデータに匹敵する効果を持つことも報告している．

2.4　日本語における有害テキスト検出

久田らは，オンライン上の誹謗中傷検出において，その定義や判断基準が文脈に依存し困難であるという課題に対し，日本の裁判例に基づいたデータセット構築を提案した [30]．彼らは，発信者情報開示請求事件や損害賠償請求事件の判決文から，原告が主張する権利侵害（名誉権，名誉感情，プライバシー権など）と，それに対する裁判所の判断（認容・否認）をラベルとして付与することで，法的観点に基づいた客観的な基準を持つデータセットを作成した．専門家によるアノテーションであっても一致率が低いなどの課題が見つかったものの，構築したデータセットを用いた実験では，投稿の文脈情報の重要性や，データ数の少ないプライバシー権などの権利侵害の分類の難しさを明らかにした．彼らは，このアプローチが実社会の課題に即した自動検出や，対策における説明責任の向上に寄与すると述べている．

荒井らは，ソーシャルメディア上の攻撃的・暴力的表現，特に社会的弱者を標的としたヘイトスピーチの自動検出に向け，日本語データセットの構築手法を提案した [29]．COVID-19 パンデミック下で懸念される排外主義的な言説の増加などの社会的背景を踏まえ，英語等と比較してリソースが不足している日本語において，ヘイトスピーチを体系的に収集・分析するための基盤整備の必要性を論じた．既存の多様な言語におけるデータセット構築の事例を概観しつつ，日本国内の状況に適したデータセット構築の試案を示した．

Zhang らは，Twitter 上の日本語のネットいじめを自動検出するために，いじめに関連する単語とその深刻度を登録した「日本語いじめ表現辞書」の構築手法を

提案した [28]．シードとなるいじめ単語を用いて収集したツイートから，SO-PMI (Semantic Orientation Using Pointwise Mutual Information) を用いて各単語のいじめ度を算出し，辞書を作成した．この辞書から得られる特徴量に加え，n-gram，Word2vec，Doc2vec などの特徴量を組み合わせ，Support Vector Machine(SVM) や Multi-layer Perceptron (MLP) など 6 種類の機械学習アルゴリズムを用いて分類性能を評価した．実験の結果，いじめ表現辞書の特徴量を導入することで多くのモデルで検出精度が向上し，最良のモデルでは F 値が 0.9 を超えたことを報告している．また，データの収集時期と辞書の構築時期のズレが性能に与える影響についても検証し，時期の一致よりも，辞書に登録されている単語数の多さが分類性能により大きな影響を与えることを明らかにした．

2.5 本研究の特徴

本研究の特徴は，暗黙的な有害テキストの検出という課題に対し，既存データセットを活用した言い換えによるデータ拡張とマルチタスク学習を組み合わせて取り組む点にある．先行研究の多くは，有害な単語や直接的な攻撃表現を含む明示的な有害テキストを主な対象としてきたが，皮肉や婉曲表現など，有害な単語を含まない暗黙的な有害表現の検出は十分に探究されていなかった．

暗黙的な有害テキストの検出が困難である要因のひとつとして，暗黙的な有害表現を十分に含む大規模データセットが少ない点が挙げられる．暗黙的な有害表現は，皮肉や婉曲表現など文脈依存性が高く，明確な判断基準を設定することが難しいため，新規にデータを収集し人手で注釈を付与する場合，作成コストが高くなるという問題がある．本研究では，この課題に対処するため，新たに暗黙的有害テキストを大量に収集・注釈付与するのではなく，既存の明示的有害テキストデータセットに対して言い換え手法を適用し，攻撃的な語を含まないが有害な意図を保持したテキストを疑似的に生成することで，暗黙的有害テキストの学習データを構築する．

さらに，有害性判定のみを単独で学習するのではなく，感情分析や皮肉検出といった関連タスクを補助タスクとして導入し，マルチタスク学習を行った点も本研究の特徴である．これにより，暗黙的有害表現と関連する感情的・語用論的な情報を共有表現として学習させ，データの過疎性による性能低下の緩和を図る．

また，本研究では暗黙的有害テキストデータセットのみを用いた評価に限定せず，明示的有害テキストデータセットも評価対象に含める．これにより，暗黙的有害表現に対する検出性能の改善を確認すると同時に，明示的な有害表現に対する検出性能が大きく損なわれていないことを検証し，提案手法が特定の表現タイプに特化しすぎることなく，一定の汎用性を維持していることを確認する．

第3章　提案手法

3.1　概要

本研究が対象とする有害テキスト検出タスクは，与えられたテキストに対し，それが有害な意味を持つか否かを分類するタスクである．特に明示的な有害表現を含まず，暗黙的に有害な意味を持つテキストを分類することに焦点を当てる．前提として，有害・無害のラベルが付与された既存のデータセットがあり，それに含まれる有害テキストは主に明示的な有害表現を含むものと仮定する．既存のデータセットをもとに暗黙的な有害テキストを含む新しいデータセット (訓練データ) を自動構築する．その概要を図 3.1 に示す．

既存データセットにおける有害テキストの集合を T^{+}_{exp}，無害テキストの集合を T^{-} とおく．T^{+}_{exp} 中のテキストに対し，それに含まれる有害な意味を持つ単語を無害な単語に言い換えることで，暗黙的に有害な意味を表すテキストの集合 T^{+}_{imp} を生成し，これと T^{-} を組み合わせて新しい訓練データを構築する．この手続きの詳細は 3.2 節にて述べる．次に，逆翻訳により T^{+}_{imp} と T^{-} の文を別の文に言い換えて，新しいラベル付きデータ TD^{+}_{imp} と TD^{-} を作成し，訓練データに追加する．このデータ拡張の手続きの詳細は 3.3 節にて述べる．最後に，構築したデータセットを用いて有害テキスト検出モデルを学習する．この詳細は 3.4 節で述べる．さらに，感情分析タスクならびに皮肉判定タスクとのマルチタスク学習を実施し，分類モデルの性能向上を図る．この手続きの詳細は 3.5 節にて述べる．

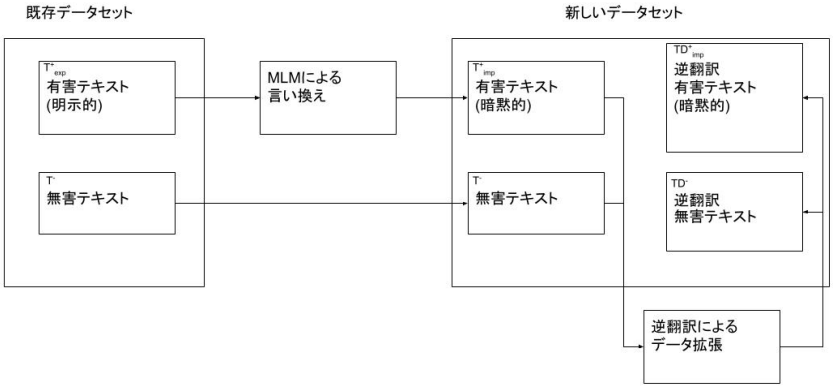

図 3.1: 有害性判定モデルの訓練データの構築

表 3.1: 有害語辞書に含まれる単語の例

種別	単語の例
一般的な有害語	fuck, shit, bitch, ass, bastard
表記ゆれ・伏せ字	f*ck, 5h1t, b!tch, @$$, a55

3.2 言い換えによる暗黙的有害テキストの生成

既存のデータセットにおける明示的な有害テキスト $s \in T^{+}_{\text{exp}}$ を $s = \{w_1, \ldots, w_n\}$ とおく．w_i はテキスト中における単語を表す．以下の手続きに従い，s を暗黙的な有害テキストに言い換える．

1. w_i の中から有害な意味を持つ単語 (以下,「有害語」と記す) を検出する．具体的には，英語の有害語辞書として profanewords[2] を使用した．profanewords に登録されている有害語の例を表 3.1 に示す．この辞書には 2,822 語が登録されており，一般的な罵倒語に加え，伏せ字やリートスピーク（文字の一部を数字や記号に置き換えたもの）などの表記ゆれにも対応している．検出された有害語を w_k とおく．

2. Bidirectional Encoder Representations from Transformers (BERT)[9] を Masked Language Model (MLM) として使用し，有害語 w_k を無害な別の単語 w'_k に置き換える．w_k を特殊トークン [MASK] に置き換えで (文 2) を生成し，事前学習済み BERT モデル bert-base-uncased[1]を用いて [MASK] に当てはまる単語を予測する．予測された単語のうち，有害語ではなくかつ予測確率が最大の単語 w'_k を選択し，[MASK] を w'_k に置き換える．

3. 新しい文 $s' = \{w_1, \ldots, w'_k, \ldots, w_n\}$ を暗黙的な有害テキストの集合 T^{+}_{imp} に加える．

上記の手続きによる，単一の有害語が含まれる場合の言い換え例を以下に示す．有害語 stupid を [MASK] に置き換え (文 2)，MLM により [MASK] を有害性の低い単語 unusual に言い換えて新しい文 3) を生成している．

1) That was a stupid thing to say.
2) That was a [MASK] thing to say.
3) That was a unusual thing to say.

一方，1 つの文に複数の有害語が含まれる場合は，それぞれを順に [MASK] に置き換えて同様の処理を繰り返すことで，全ての有害語を無害な単語に置き換える．以下に複数の有害語が含まれる場合の例を示す．文中の damn を [MASK] に置き

[1] https://huggingface.co/google-bert/bert-base-uncased

換え (文 2)，文脈を考慮して really に言い換える (文 3)．続いて，shit を [MASK] に置き換え (文 4)，同様に bad に言い換えることで文 5) を生成している．このように段階的に置換を行うことで，表面的には攻撃的ではないが，文脈によっては読み手に不快感を与えうる暗黙的な有害テキストが自動生成される．

1) This is a damn shit movie.
2) This is a [MASK] shit movie.
3) This is a really shit movie.
4) This is a really [MASK] movie.
5) This is a really bad movie.

3.3 逆翻訳によるデータ拡張

暗黙的な有害テキストのサンプル数を増やすために，逆翻訳によるデータ拡張を行う．T^{+}_{imp} ならびに T^{-} のテキストについて，機械翻訳を用いて元の文 (英語文) を別の言語に翻訳し，それを再び英語に翻訳することで新しいラベル付きデータを獲得する．これにより，文の意味を保ちつつ，異なる単語，語順，文構造で表現された別の文を訓練データに加えることができる．特に暗黙的な有害テキストについては，有害な意味を持つ別の表現を含む文を生成することで，データセットにおける暗黙的な攻撃的表現の多様性が向上し，分類モデルの汎化性能が向上することが期待される．

実装の詳細について述べる．機械翻訳システムとして Google Translate API[2] を使用し，中間言語として中国語，フランス語，ドイツ語，日本語の 4 言語を用いて，1 つの文から 4 つの拡張データを獲得する．次に，翻訳前テキスト (元データ) および逆翻訳後テキスト (拡張データ) を Sentence-BERT(all-MiniLM-L6-v2) によって文埋め込み (ベクトル) に変換する．次に，元データと拡張データの類似度をベクトル間のコサイン類似度によって測る．文間の類似度が低い場合，逆翻訳によってテキストの意味が大きく変わっていたり，もともと有害であったテキストが無害なテキストに変わっていたりする可能性がある．拡張データの品質を高めるため，コサイン類似度が 0.5 以上の場合のみ，逆翻訳によって得られたサンプルのみを候補として残し，その中からランダムに 1 つを選択して拡張データとして採用する．

表 3.2 に，本手法で生成された逆翻訳データの例を示す．ここでの入力テキスト（オリジナル）は，3.2 節の手法により明示的な有害語が置換された後の暗黙的な有害テキストである．例えば，フランス語を中間言語とした例 (ID 1) では，元の文の unusualが，逆翻訳によって strange に変化している．同様に，ドイツ語の例 (ID 2) では bad が poor に，中国語の例 (ID 3) では transgender が transsexual に，日本語の例 (ID 4) では acting が behaving に，それぞれ言い換えられている．

[2] https://cloud.google.com/translate

このように，中間言語を介することで，有害語が隠蔽された文脈を維持しつつ，多様な語彙・表現のバリエーションを持つデータを生成することが可能となる．

表 3.2: 逆翻訳によるデータ拡張の例

ID	オリジナル (英語)	中間言語	逆翻訳 (英語)
1	That was a unusual thing to say.	フランス語	That was a strange thing to say.
2	This is a really bad movie.	ドイツ語	This is a very poor film.
3	She looks like a transgender.	中国語	She looks like a transsexual.
4	Quit acting like a victim.	日本語	Stop behaving like a victim.

3.4 分類モデルの学習

構築した訓練データを用いてテキストの有害性を判定する分類モデルを学習する．以下の事前学習済み言語モデルをファインチューニングすることで有害性判定モデルを獲得する．

- **BERT**：双方向 Transformer に基づく言語モデルである．Masked Language Modeling タスクと Next Sentence Prediction タスクにより事前学習されている [9].
- **RoBERTa**：BERT を改良し，より大規模なデータとバッチサイズで学習された言語モデルである．事前学習の際には BERT とは異なり，Next Sentence Prediction タスクを廃止し，学習効率を改善している [22].
- **DistilBERT**：BERT の蒸留モデルであり，パラメータ数を削減しつつ高速化・軽量化を実現している [23].

3.5 マルチタスク学習

暗黙的な有害テキストの検出性能をさらに高めるために，マルチタスク学習を導入する．マルチタスク学習とは，複数の関連タスクのデータセットを用いて分類モデルを学習することで訓練データの過疎性の問題を解消し，かつ複数のタスクに共通する言語的特徴を学習することでモデルの汎化性能を高める手法である．

本研究では，感情分析 (Sentiment Analysis; SA) および皮肉判定 (Sarcasm Detection; SD) を補助タスクとして採用する．これらのタスクには既存のデータセットが存在するため，マルチタスク学習が可能である．

感情分析タスクは，テキストが持つ極性 (肯定，否定，中立) を判定するタスクである．有害な発言はしばしば強い負の感情 (怒り，悲しみ，嫌悪など) を伴うため，有害性判定タスクとの相関が高いと考えられる．さらに，暗黙的な有害テキス

表 3.3: SentiWordNet における単語と感情スコアの例

単語 (Synset)	品詞	Positive	Negative
excellent	形容詞	1.000	0.000
happy	形容詞	0.875	0.000
terrible	形容詞	0.000	0.875
sad	形容詞	0.125	0.750
table	名詞	0.000	0.000

トの識別能力を高めるため，感情分析タスクの訓練データとして，3.2 節で述べた MLM による文の言い換えにより暗黙的なテキストを含むデータセットを構築する．具体的には，データセット中の文に対し，感情語辞書を用いて否定的な感情語 (例. sad, angry) を検出し，それを [MASK] に置き換えて，MLM によって中立の感情極性を持つ別の単語に言い換える．感情語辞書として SentiWordNet[12] を用いる．SentiWordNet は，英語の語彙データベースである WordNet 3.0 に基づき，各同義語セット (Synset) に対して「肯定 (Positive)」および「否定 (Negative)」の感情スコアを付与したものである．スコアは 0.0 から 1.0 の実数値で与えられ，肯定スコアと否定スコアの和が 1.0 に満たない分は「客観 (Objective)」スコアとみなされる．本辞書には約 117,000 の Synset が収録されており，名詞，動詞，形容詞，副詞の各品詞に対応している．

表 3.3 に SentiWordNet における登録単語の例を示す．表に示すように，*excellent* (素晴らしい) や *happy* (幸せな) といった単語には高い肯定スコアが，*terrible* (ひどい) や *sad* (悲しい) には高い否定スコアが付与されている．一方で，*table* (テーブル) のような一般的な物体を表す単語は，肯定・否定スコア共に 0.0 となり，客観的な単語として扱われる．

皮肉判定タスクは，与えられたテキストが皮肉であるか否かを判定するタスクである．一般に皮肉とは，表層的には肯定的または中立的であっても，攻撃性や揶揄を暗に示唆するテキストである．したがって，暗黙的な有害テキストと共通の言語的特徴があると考えられ，有害性判定のマルチタスク学習の補助タスクとして適している．

マルチタスク学習の概要を図 3.2 に示す．テキストを Transformer ベースの事前学習済みモデル (BERT，RoBERTa，DistilBERT) に入力し，それから得られる文の抽象表現を全結合層 (Fully Connected Layer;FCL) に渡す．有害性判定タスクと，感情分析タスクもしくは皮肉判定タスクのそれぞれに対して個別の FCL を用意する．損失関数は各タスクのクロスエントロピーとし，各タスクのデータセットを用いて，事前学習済み言語モデルとそのタスクに対応した FCL のパラメタを更新する．また，有害性判定，感情分析，皮肉判定の 3 つのタスクのマルチタスク学習も実施する．この場合のモデル構造を図 3.3 に示す．この図に示すように，共通の事前学習済みモデルの出力に対して，それぞれのタスクに対応した

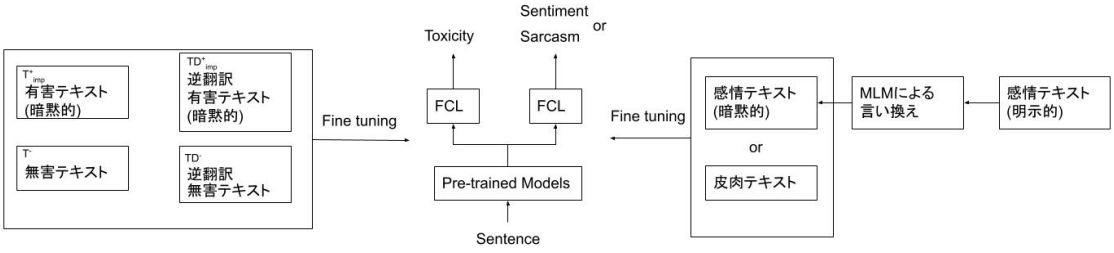

図 3.2: マルチタスク学習のアーキテクチャ

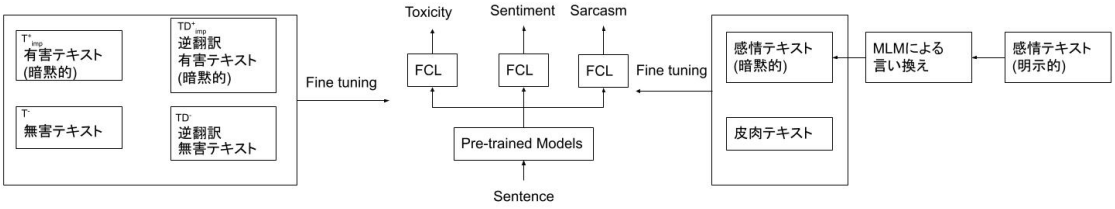

図 3.3: 3 つのマルチタスク学習のアーキテクチャ

3 つの FCL を並列に接続し，ファインチューニングを行う．

3.6 日本語テキストの有害性判定

本節では，日本語テキストを対象とした有害性判定手法について述べる．基本的には前節までの手法を日本語テキストに適用するが，異なる点もある．

英語においては，OffensiveLang [7] など，暗黙的な有害表現に焦点を当てたデータセットが存在するが，日本語においては同様の目的で構築された公開データセットは管見の限り存在しない．したがって，日本語テキストに対する評価を行うためには，まず暗黙的な有害テキストを含むデータセットを自ら構築する必要がある．そこで本研究では，大規模言語モデルを用いて日本語の暗黙的有害テキストを収集・構築し，それを用いて分類モデルの学習を行うアプローチをとる．

3.6.1 暗黙的有害テキストデータセットの構築

日本語における暗黙的有害テキストの構築にあたっては，暗黙的攻撃表現を対象とした先行研究である OffensiveLang を参考にした．OffensiveLang において設定されているカテゴリーに基づき,「人種」「宗教」「性別」などの 7 つの社会的カテゴリーを設定し，それらに含まれる「黒人」「イスラム教」「男性」などの計 38 のターゲットグループを対象としてテキスト生成を行った．

テキスト生成には大規模言語モデルである ChatGPT-4o を用いた．各ターゲットグループに対して，以下のようなプロンプトを与えることで，明示的な差別語を含まない攻撃的なテキストを生成させた．

> 「○○（ターゲット）」に対して，差別用語や暴言などの明示的に攻撃的な単語を使わずに，皮肉や偏見を含んだネガティブな発言の例文を生成してください．

この方法により，明示的な差別語や侮辱語を含まないが，特定のターゲットグループに対して有害な表現を含むテキストを収集した．生成されたテキストに対しては，3 人のアノテーターによって有害・無害のラベル付けを行った．アノテーションの品質を担保するため，3 人全員の判定が一致したもののみを採用した．以後，これを「Japanese Implicit Toxic Dataset」とし，有害性判定モデルの学習ならびに評価に用いる．

3.6.2 日本語有害テキスト検出モデルの学習

日本語においても英語実験と同様に，提案手法である「言い換えによるデータ拡張」および「マルチタスク学習」を適用する．ただし，言語資源の有無や言語的特性の違いにより，一部の設定を変更している．

まず，言い換えによる疑似データの生成については，英語と同様の手順で行う．日本語の有害語辞書として inappropriate-words-ja-master[3] を用いるとともに，既存の明示的な有害テキストデータセット Multilingual Toxicity Detection Dataset[4] に特徴的に出現する単語を tf-idf に基づいて抽出し，これらを加えたものを「有害語」として定義する．これらの有害語を日本語事前学習済み BERT モデルの MLM 機能を用いて無害な単語に置換することで，明示的な有害語を含まない疑似的な暗黙的有害テキストを生成し，学習データに追加する．

次に，マルチタスク学習については，補助タスクとして「感情分析タスク」のみを採用する．英語実験では「皮肉判定タスク」も併用したが，日本語においては皮肉に特化した大規模かつ高品質な公開データセットが不足しているため，本実験では皮肉判定とのマルチタスク学習は実施しない．感情分析タスクとのマルチタスク学習においては，英語と同様に，感情語を中立的な単語に言い換えた疑似データセットを用いた学習を行うことで，暗黙的な有害性の検出性能向上を図る．

また，比較検証のために逆翻訳によるデータ拡張も実施する．中間言語として中国語，フランス語，ドイツ語，英語の 4 言語を用い，Google Translate API によって「日本語 → 中間言語 → 日本語」への再翻訳を行うことで 4 種類の拡張データ候補を生成する．ただし，英語の場合には 4 つの拡張データ候補を全て採

[3] https://github.com/MosasoM/inappropriate-words-ja
[4] https://huggingface.co/datasets/textdetox/multilingual_toxicity_dataset

用したが，日本語の場合は生成した 4 つの拡張データのその中からランダムに 1 つを選択して元のデータに追加する．

3.7 大規模言語モデルのファインチューニング

本研究では、提案手法の有効性を検証するため、大規模言語モデル (Large Language Model; LLM) を用いた有害テキスト検出モデルの構築も行う。モデルには Llama-3-8B-Instruct [5]を採用し、パラメータ効率の良い学習手法である LoRA (Low-Rank Adaptation)[15] を用いてファインチューニングを行う。学習データには、本章で述べた提案手法によって構築・拡張されたデータセット（疑似的な暗黙的有害テキストを含む）を使用する。

学習時のプロンプトには、先行研究である OffensiveLang [7] を参考に、以下の形式を採用した。

> Is the following text offensive, hurtful, derogatory, or obscene for human?
> Answer “yes” or “no” with no explanation.

モデルは入力されたテキストに対し、有害であれば “yes”、無害であれば “no” を生成するように学習させる。また、日本語テキストに対する実験においては、上記プロンプトを日本語に翻訳した以下の形式を用いる。

> 以下のテキストは、人間にとって不快、傷つける、軽蔑的、またはわいせつなものですか？
> 説明なしで「はい」または「いいえ」で答えてください。

[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

第4章　評価

4.1　英語を対象とした暗黙的有害テキスト検出の評価

4.1.1　データセット

英語での実験に使用したデータセットを以下に示す.

Hate Speech and Offensive Language Detection (HSOLD) テキストに対して攻撃的か否かのラベルが付与されたデータセット [8]．攻撃的テキストの多くは明示的な有害表現を含む．サンプル数は 24,783 である.

OffensiveLang 暗黙的な攻撃テキストを収集したデータセット [7]．サンプル数は 8,270 である.

Sentiment and Emotion Analysis Dataset 感情分析タスクの大規模なデータセット [20]．テキストに対して Joy, Sadness, Anger, Fear, Love, Surprise のラベルが付与されている．サンプル数は 422,000 である.

Sarcastic Comments 皮肉判定のデータセット [18]．Reddit から収集したテキストに対して皮肉か否かのラベルが付与されている．サンプル数は 962,295 である.

本実験では以下に示す 3 種類のテストデータを用いる.

- $\mathcal{D}_{\text{imp}}$: 主に暗黙的な有害テキストと無害なテキストから構成されるテストデータ．OffensiveLang を用いる．ただし，OffensiveLang は全体の 80%が有害テキストである不均衡なデータセットであるため，有害テキストと無害テキストを 874 サンプルずつランダムに抽出し，これをテストデータとする.
- $\mathcal{D}_{\text{exp}}$: 主に明示的な有害テキストと無害なテキストから構成されるテストデータ．HSOLD から $\mathcal{D}_{\text{imp}}$ と同数，すなわち有害テキストと無害テキストを 874 件ずつランダムに抽出し，テストデータとする.
- $\mathcal{D}_{\text{imp+exp}}$: $\mathcal{D}_{\text{imp}}$ と $\mathcal{D}_{\text{exp}}$ を合わせたテストデータ．現実の多くの場面では明示的な有害テキストと暗示的な有害テキストは混在するため，両者を含むテストデータを作成した.

各テストデータの統計を表 4.1 に示す.

表 4.1: 英語テストデータのサンプル数

	有害	無害	合計
$\mathcal{D}_{\text{imp}}$	874	874	1,748
$\mathcal{D}_{\text{exp}}$	874	874	1,748
$\mathcal{D}_{\text{imp+exp}}$	1,748	1,748	3,496

4.1.2 実験設定

本実験では以下のモデルを比較する.「B」はベースラインモデルを,「M」は提案手法のモデルを表す.

- B_{imp}: 比較的少量の暗黙的な有害テキストを用いて学習したベースラインモデル．OffensiveLang におけるテストデータ $\mathcal{D}_{\text{imp}}$ 以外のデータから，テストデータと同様に有害テキストと無害テキストをランダムに 874 件ずつ抽出し，訓練データとする.

- B_{exp}: 比較的大規模な明示的な有害テキストのデータセット (HSOLD) を用いて学習したベースラインモデル．ただし，$\mathcal{D}_{\text{exp}}$ は訓練データとして使用しない.

- M_{imp}: 明示的有害テキストのデータから MLM による言い換えによって疑似的な暗黙的有害テキストのデータセットを生成し，これを用いて学習したモデル (3.2 節).

- $M_{\text{imp+A}}$: 上記に加え，逆翻訳によってデータ拡張した訓練データから学習したモデル (3.3 節).

- $M_{\text{imp+A+L}_1}$: 有害テキスト検出モデルの学習には $M_{\text{imp+A}}$ と同じ訓練データを用い，かつ感情分析タスクとのマルチタスク学習によって学習されたモデル.

- $M_{\text{imp+A+L}_2}$: $M_{\text{imp+A+L}_1}$ とほぼ同じだが，感情分析タスクの訓練データとして明示的な感情語を中立な単語に言い換えた暗黙的テキストのデータセットを用いたモデル.

- $M_{\text{imp+A+L}_3}$: 有害テキスト検出モデルの学習には $M_{\text{imp+A}}$ と同じ訓練データを用い，かつ皮肉分析タスクとのマルチタスク学習によって学習されたモデル.

- $M_{\text{imp+A+L}_4}$: 有害テキスト検出モデルの学習には $M_{\text{imp+A}}$ と同じ訓練データを用い，有害テキスト検出タスク，感情分析タスク，皮肉判定タスクの 3 タスクを同時に学習するマルチタスク学習によって学習されたモデル.

表 4.2: 英語実験における各モデルの訓練データのサンプル数

	有害	無害	合計
B_{imp}	874	874	1,748
B_{exp}	18,178	3,316	21,494
M_{imp}	18,178	3,316	21,494
$M_{\text{imp+A}}$	32,808	8,158	40,966

有害性判定の評価基準として，精度 (Precision; P)，再現率 (Recall; R)，F1 スコア (F1 score; F)，正解率 (Accuracy;A) を用いる．前述の 7 つのモデル (手法) と 3 つの事前学習モデル (BERT, RoBERTa, DistilBERT) の組み合わせについて，これらの指標を比較する．

4.1.3 逆翻訳によるデータ拡張の結果

3.3 項で述べた逆翻訳によるデータ拡張により訓練データがどの程度拡張されたかを評価する．各モデルの学習に用いた訓練データにおけるサンプル数を表 4.2 に示す．モデル $M_{\text{imp+A}}$ はデータ拡張後の訓練データを使用している．データ拡張を行わないモデル M_{imp} の訓練データと比べて，サンプル数をおよそ 2 倍に増やすことができた．

各モデルの訓練データについて説明を補足する．B_{imp} の訓練データは小規模であるのに対し，B_{exp} の訓練データは比較的大規模である．提案手法により自動生成した暗黙的な有害テキストを用いる M_{imp} の訓練データ数は M_{exp} と同じであり，両者を比較することで提案手法の有効性を公平に評価できる．なお，マルチタスク学習を行うモデル $M_{\text{imp+A+L}_i}$ の有害テキスト検出タスクの訓練データは $M_{\text{imp+A}}$ と同じである．

4.1.4 暗黙的有害テキストのテストデータに対する結果と考察

表 4.3: 有害テキスト検出の実験結果 (テストデータ $\mathcal{D}_{\text{imp}}$)

	DA	MTL		BERT				RoBERTa				DistilBERT			
		感情	皮肉	P	R	F	A	P	R	F	A	P	R	F	A
B_{imp}				0.56	0.62	0.59	**0.57**	0.58	0.56	0.57	**0.58**	0.58	0.63	0.60	**0.59**
B_{exp}				**0.66**	0.07	0.13	0.52	**0.68**	0.07	0.13	0.52	**0.68**	0.06	0.11	0.52
M_{imp}				0.50	0.70	0.58	0.50	0.50	0.74	0.59	0.49	0.51	0.80	0.62	0.52
$M_{\text{imp+A}}$	✓			0.49	0.68	0.57	0.49	0.50	0.67	0.57	0.50	0.52	0.82	0.63	0.52
$M_{\text{imp+A+L}_1}$	✓	O		0.50	0.83	0.62	0.50	0.50	0.69	0.58	0.50	0.50	0.85	0.63	0.50
$M_{\text{imp+A+L}_2}$	✓	P		0.51	**0.92**	**0.65**	0.51	0.50	0.76	0.61	0.51	0.50	0.92	0.65	0.51
$M_{\text{imp+A+L}_3}$	✓		O	0.51	0.91	**0.65**	0.51	0.50	**0.90**	**0.65**	0.51	0.51	0.92	0.65	0.51
$M_{\text{imp+A+L}_4}$	✓	P	O	0.51	0.76	0.61	0.51	0.49	0.82	0.62	0.49	0.52	**0.93**	**0.66**	0.52

暗黙的な有害テキストのみを含むテストデータ $\mathcal{D}_{\text{imp}}$ に対する各手法の精度，再現率，F1 スコア，正解率を表 4.3 に示す．DA の列は逆翻訳によるデータ拡張を適用するか否かを表し，MTL の列はマルチタスク学習に感情分析タスクおよび皮肉判定タスクを適用したか否かを示す.「感情」の列の「O」は元の感情分析のデータセット (Original) をそのまま学習に用いたこと,「P」は感情語をマスクして言い換えたデータセット (Paraphrased) を用いて学習したことを表す．皮肉判定タスクについては常に元のデータセット (O) をそのまま用いる．

明示的な有害表現のデータセットから学習した M_{exp} は，再現率が 0.06 から 0.07 と非常に低い．M_{exp} が明示的に有害な単語を手がかりにテキストの有害性を判定しているのに対し，テストデータにおける暗黙的な有害テキストではそのような単語が出現しないため，暗黙的有害テキストの多くが正しく検出できていないと考えられる．これに対し，明示的な有害テキストを暗黙的な有害テキストに言い換えたデータを用いる M_{imp} は，再現率が大幅に改善され，F1 スコアも M_{exp} より高くなっている．このことから，MLM による無害な単語への言い換えが暗黙的な有害テキストを検出するモデルの学習に有効であることがわかる．また，M_{imp} は小規模な暗黙的有害テキストのデータセットから学習された B_{imp} と比べて，精度は低いが再現率は高い．より大規模な訓練データを用いてモデルを学習することにより，多様な暗黙的有害テキストを正しく分類できるようになり，False Negative の誤りが減少したと考えられる．ただし，F1 スコアや正解率を比較すると，M_{imp} は B_{imp} と比べてやや劣る．ここで，訓練データとテストデータはともに OffensiveLang であるという点で B_{imp} の学習は IID (independently and identically distributed) の設定であるのに対し，M_{imp} の学習は訓練データとテストデータが異なるという点で OOD (out-of-distribution) の設定であることに注意を要する．一般に IID は OOD よりも分類モデルの性能は高くなるため，実験設定が IID であることが B_{imp} が M_{imp} を上回る要因のひとつになっていると考えられる．

次に，逆翻訳によるデータ拡張について考察する．M_{imp} と $M_{\text{imp+A}}$ を比較すると，DistilBERT ではデータ拡張ありのモデル $M_{\text{imp+A}}$ の方が F1 スコアや正解率が高いが，BERT と RoBERTa では M_{imp} の方が高い．したがって，データ拡張による効果は限定的である．とはいえ，M_{imp} と $M_{\text{imp+A}}$ の中で最も高い F1 スコア (0.63) を達成したのは，事前学習済み言語モデルとして DistilBERT を用いたデータ拡張ありのモデルである．

マルチタスク学習の効果について検証する．$M_{\text{imp+A}}$ とマルチタスク学習ありの 4 つのモデルを比較すると，いずれの事前学習済み言語モデルでも再現率が大幅に改善し，False Negative の削減に寄与していることがわかる．感情分析タスクとのマルチタスク学習について，感情分析のデータセットをそのまま用いる $M_{\text{imp+A+L}_1}$ では，単一タスクの $M_{\text{imp+A}}$ と比べていずれの指標も大きな差はなく，モデルが明示的な感情語に依存しているために暗黙的な有害テキストの特徴を十分に学習できていない可能性が示唆される．一方，感情語を中立的な単語に言い換えたデー

タセットを用いる $M_{\text{imp+A+L}_2}$ では，BERT・RoBERTa・DistilBERT のいずれにおいても再現率と F1 スコアが大きく改善した．特に BERT では再現率が 0.24 ポイント，F1 スコアが 0.08 ポイント向上した．暗黙的な有害テキストのデータセットと同様に，感情分析についても MLM を用いた言い換えによって暗黙的なテキストから構成される擬似訓練データを作成するアプローチは，分類モデルが表層的なキーワードに過度に依存することを抑制し，表層的ではない深い意味理解を促進すると考えられる．

次に，皮肉判定タスクを補助タスクとして用いた $M_{\text{imp+A+L}_3}$ では，$M_{\text{imp+A+L}_2}$ とほぼ同等の高い再現率 (BERT 0.91, RoBERTa 0.90, DistilBERT 0.92) と F1 スコア (すべて 0.65) が得られた．皮肉には暗黙的な表現が多いため，皮肉判定タスクとのマルチタスク学習が暗黙的な有害テキスト検出モデルの性能向上に貢献したと考えられる．

最後に，感情分析タスクと皮肉判定タスクを併用する $M_{\text{imp+A+L}_4}$ では，DistilBERT で再現率 0.93，F1 スコア 0.66 と全モデル中最高の結果が得られたものの，BERT や RoBERTa では $M_{\text{imp+A+L}_3}$ と比べてわずかな性能低下が見られた．以上の結果から，マルチタスク学習は補助タスクの選択や言い換え手法との組み合わせによって大きく効果を発揮する一方，事前学習済み言語モデルとの相性によっては逆に性能が下がる場合もあることがわかった．

以上の考察をまとめると，実験の結果から，MLM を用いた暗黙的有害テキストの生成，データ拡張，マルチタスク学習を併用する提案手法の有効性が示されたと言える．

4.1.5 その他のテストデータに対する結果と考察

明示的な有害テキストのテストデータ $\mathcal{D}_{\text{exp}}$ に対する実験結果を表 4.4 に示す．各指標とも暗黙的な有害テキストのテストデータ (表 4.3) に比べて高い傾向が見られることから，暗黙的な有害テキストの検出は明示的有害テキストと比べてより難しいタスクであると言える．最高の成績を得たのは明示的な有害テキストのデータセットから学習された B_{exp} である．ただし，B_{imp} を除いて各モデルの性能に大きな差は見られないことから，暗黙的な有害テキストのデータセットから学習したモデルは明示的有害テキスト検出にも有効に働くことが確認できる．モデル B_{imp} の性能は他のモデルに比べて顕著に劣るが，これは訓練データの量が他のモデルと比べて少ないためと考えられる．

明示的・暗黙的の両方の有害テキストを含むテストデータ $\mathcal{D}_{\text{imp+exp}}$ に対する実験結果を表 4.5 に示す．暗黙的もしくは明示的な有害テキストのみから学習されたベースラインモデル B_{imp} もしくは B_{exp} は，両者が混在する現実的なテストデータでは有害テキスト検出の性能が低く，その F 値は 0.59–0.69 程度に留まる．一方，提案手法のモデル M_* の F1 スコアはいずれも 0.76 を上回っており，その有効性が確認できる．データ拡張の効果，マルチタスク学習の効果については，モデル間

の差は小さくなってはいるが，4.1.4 項で述べた暗黙的有害テキストのテストデータに対する考察と概ね一致する．これは，$\mathcal{D}_{\text{imp+exp}}$ は $\mathcal{D}_{\text{imp}}$ と $\mathcal{D}_{\text{exp}}$ を合わせたテストデータであり，$\mathcal{D}_{\text{exp}}$ では提案手法のモデル間の優劣がほとんどないことに起因する．

表 4.4: 有害テキスト検出の実験結果 (テストデータ $\mathcal{D}_{\text{exp}}$)

	DA	MTL		BERT				RoBERTa				DistilBERT			
		感情	皮肉	P	R	F	A	P	R	F	A	P	R	F	A
B_{imp}				0.80	0.73	0.77	0.70	0.80	0.73	0.77	0.70	0.79	0.68	0.73	0.66
B_{exp}				**0.96**	0.91	**0.93**	**0.93**	**0.93**	0.92	**0.93**	**0.94**	**0.96**	0.92	**0.94**	**0.93**
M_{imp}				0.87	**0.92**	0.89	0.92	0.86	**0.93**	0.89	0.91	0.89	**0.93**	0.91	0.91
$M_{\text{imp+A}}$	✓			0.90	0.90	0.90	0.92	0.90	0.90	0.90	0.91	0.89	0.92	0.91	0.91
$M_{\text{imp+A+L}_1}$	✓	O		0.89	0.91	0.90	0.92	0.90	0.91	0.90	0.91	0.90	0.92	0.91	0.91
$M_{\text{imp+A+L}_2}$	✓	P		0.89	**0.92**	0.90	0.92	0.90	0.91	0.90	0.91	0.90	0.92	0.91	0.91
$M_{\text{imp+A+L}_3}$	✓		O	0.89	**0.92**	0.90	0.92	0.90	**0.93**	0.91	0.91	0.90	0.92	0.91	0.91
$M_{\text{imp+A+L}_4}$	✓	P	O	0.90	0.90	0.90	0.92	0.91	0.89	0.90	0.91	0.90	**0.93**	0.91	0.91

表 4.5: 有害テキスト検出の実験結果 (テストデータ $\mathcal{D}_{\text{imp+exp}}$)

	DA	MTL		BERT				RoBERTa				DistilBERT			
		感情	皮肉	P	R	F	A	P	R	F	A	P	R	F	A
B_{imp}				0.59	0.59	0.59	0.59	0.62	0.65	0.64	0.63	0.61	0.66	0.63	0.62
B_{exp}				**0.97**	0.54	0.69	**0.76**	**0.96**	0.54	0.69	**0.76**	**0.97**	0.53	0.68	**0.76**
M_{imp}				0.69	0.85	0.76	0.74	0.69	0.87	0.77	0.74	0.69	0.90	0.78	0.75
$M_{\text{imp+A}}$	✓			0.71	0.81	0.76	0.74	0.70	0.83	0.76	0.74	0.69	0.91	**0.79**	0.75
$M_{\text{imp+A+L}_1}$	✓	O		0.68	0.91	0.78	0.74	0.69	0.84	0.76	0.73	0.67	0.92	0.78	0.74
$M_{\text{imp+A+L}_2}$	✓	P		0.67	**0.96**	0.78	0.74	0.69	0.87	0.77	0.74	0.67	**0.96**	**0.79**	0.74
$M_{\text{imp+A+L}_3}$	✓		O	0.67	0.95	**0.79**	0.74	0.67	**0.94**	**0.78**	0.74	0.67	**0.96**	**0.79**	0.74
$M_{\text{imp+A+L}_4}$	✓	P	O	0.69	0.88	0.77	0.74	0.69	0.85	0.76	0.73	0.67	**0.96**	**0.79**	0.75

4.1.6 大規模言語モデルを用いた検出モデルの実験結果

本項では，3.7 節で述べた LLM による有害テキスト判定モデルの実験について述べる．

まず，暗黙的な有害テキスト $\mathcal{D}_{\text{imp}}$ に対する実験結果を表 4.6 に示す．明示的な有害テキストのみを学習した B_{exp} は，再現率が 0.03 と極めて低く，暗黙的な有害性をほとんど検出できていない．これは BERT 等の結果と同様の傾向であるが，より顕著な失敗が見られる．LLM であっても，適切な学習データを与えなければ暗黙的な有害性の検出は困難であることが示された．これに対し，提案手法である $M_{\text{imp+A}}$ は再現率 0.80，F1 スコア 0.63 と性能を大きく向上させている．さらに，皮肉判定タスクとのマルチタスク学習を行った $M_{\text{imp+A+L}_3}$ は，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 F1 スコアを記録した．Llama-3 は文脈理解能力が高いため，皮肉判定という関連タスクの学習

が，類似した性質を持つ暗黙的有害テキストの検出能力を効果的に強化したと考えられる．

表 4.6: Llama3 による有害テキスト検出結果（テストデータ $\mathcal{D}_{\text{imp}}$）

	DA	MTL		LLM			
		感情	皮肉	P	R	F	A
B_{imp}				0.63	0.53	0.57	**0.60**
B_{exp}				**0.98**	0.03	0.06	0.51
M_{imp}				0.52	0.60	0.56	0.52
$M_{\text{imp+A}}$	✓			0.52	0.80	0.63	0.52
$M_{\text{imp+A+L}_1}$	✓	O		0.52	0.61	0.56	0.53
$M_{\text{imp+A+L}_2}$	✓	P		0.54	0.65	0.59	0.55
$M_{\text{imp+A+L}_3}$	✓		O	0.52	**0.96**	**0.67**	0.51
$M_{\text{imp+A+L}_4}$	✓	P	O	0.53	0.80	0.63	0.53

次に，明示的な有害テキスト $\mathcal{D}_{\text{exp}}$ に対する実験結果を表 4.7 に示す．全てのモデルが高い性能を示している．提案手法のモデルは，暗黙的な有害テキストへの適応能力を獲得しつつも，明示的な有害テキストに対する高い検出能力（F1 スコア 0.93）を維持していることが確認できる．

表 4.7: Llama3 による有害テキスト検出結果（テストデータ $\mathcal{D}_{\text{exp}}$）

	DA	MTL		LLM			
		感情	皮肉	P	R	F	A
B_{imp}				0.92	0.88	0.90	0.89
B_{exp}				0.95	0.90	0.92	0.91
M_{imp}				0.93	0.92	0.92	0.92
$M_{\text{imp+A}}$	✓			0.92	0.94	0.93	0.93
$M_{\text{imp+A+L}_1}$	✓	O		0.93	0.93	0.93	0.93
$M_{\text{imp+A+L}_2}$	✓	P		0.94	0.93	0.93	0.93
$M_{\text{imp+A+L}_3}$	✓		O	0.90	0.97	0.93	0.92
$M_{\text{imp+A+L}_4}$	✓	P	O	0.92	0.93	0.93	0.93

最後に，両者が混在する $\mathcal{D}_{\text{imp+exp}}$ に対する実験結果を表 4.8 に示す．皮肉判定とのマルチタスクモデル $M_{\text{imp+A+L}_3}$ が F1 スコア 0.81 と最も高い性能を示した．

さらに，BERT，RoBERTa，DistilBERT による分類モデルならびに LLM の優劣を比較するため，手法 Mimp+A+L4 について，これらのモデルの実験結果をひとうにまとめた表を表 4.9 ，表 4.10 ， 表 4.11 に示す．

まず，暗黙的な有害テキスト $\mathcal{D}_{\text{imp}}$（表 4.9）においては，DistilBERT が最も高い再現率 (0.93) と F1 スコア (0.66) を記録した．Llama-3 も F1 スコア が 0.63 と高く，BERT や RoBERTa を上回る性能を示している．

次に，明示的な有害テキスト $\mathcal{D}_{\text{exp}}$（表 4.10）においては，すべてのモデルが高い性能を示しているが，中でも Llama-3 は F1 スコア 0.93 を達成し，最も高い性

表 4.8: Llama3 による有害テキスト検出結果（テストデータ $\mathcal{D}_{\text{imp+exp}}$）

	DA	MTL		LLM			
		感情	皮肉	P	R	F	A
B_{imp}				0.75	0.70	0.72	0.73
B_{exp}				0.92	0.55	0.69	0.70
M_{imp}				0.76	0.75	0.75	0.75
$M_{\text{imp+A}}$	✓			0.74	0.85	0.79	0.78
$M_{\text{imp+A+L}_1}$	✓	O		0.78	0.76	0.77	0.78
$M_{\text{imp+A+L}_2}$	✓	P		0.78	0.80	0.79	0.79
$M_{\text{imp+A+L}_3}$	✓		O	0.72	0.92	0.81	0.78
$M_{\text{imp+A+L}_4}$	✓	P	O	0.77	0.84	0.80	0.80

表 4.9: $M_{\text{imp+A+L}_4}$ における各言語モデルの性能比較（テストデータ $\mathcal{D}_{\text{imp}}$）

モデル	P	R	F	A
BERT	0.51	0.76	0.61	0.51
RoBERTa	0.49	0.82	0.62	0.49
DistilBERT	0.52	**0.93**	**0.66**	0.52
Llama-3 (LLM)	**0.53**	0.80	0.63	**0.53**

表 4.10: $M_{\text{imp+A+L}_4}$ における各言語モデルの性能比較（テストデータ $\mathcal{D}_{\text{exp}}$）

モデル	P	R	F	A
BERT	0.90	0.90	0.90	0.92
RoBERTa	0.91	0.89	0.90	0.91
DistilBERT	0.90	**0.93**	0.91	0.91
Llama-3 (LLM)	**0.92**	**0.93**	**0.93**	**0.93**

表 4.11: $M_{\text{imp+A+L}_4}$ における各言語モデルの性能比較（テストデータ $\mathcal{D}_{\text{imp+exp}}$）

モデル	P	R	F	A
BERT	0.69	0.88	0.77	0.74
RoBERTa	0.69	0.85	0.76	0.73
DistilBERT	0.67	**0.96**	0.79	0.75
Llama-3 (LLM)	**0.77**	0.84	**0.80**	**0.80**

能を示した．これは，大規模言語モデルが持つ豊富な事前知識と文脈理解能力が，直接的な攻撃表現の検出においても有利に働いたと考えられる．

最後に，両者が混在する現実的な設定である $\mathcal{D}_{\text{imp+exp}}$（表 4.11）においては，Llama-3 が F1 スコア 0.80，正解率 0.80 となり，全モデル中で最も高い総合性能を示した．DistilBERT は再現率 (0.96) において突出しているものの，精度 (0.67) が Llama-3 の精度 (0.77) よりも低く，False Positive が比較的多い傾向にある．対照的に Llama-3 は精度と再現率のバランスが良く，多様な有害テキストに対して安定した検出能力を発揮している．

以上の分析から，提案手法である $M_{\text{imp+A+L}_4}$ はモデルのアーキテクチャに関わらず有効に機能するが，特に実用上重要となる混合データ環境においては，大規模言語モデル（Llama-3）を用いることで最もバランスの取れた高い性能が得られることが明らかとなった．

4.2 日本語を対象とした暗黙的有害テキスト検出の評価

4.2.1 データセット

日本語テキストに対する検証のため，以下のデータセットを用いる．

Japanese Implicit Toxic Dataset 日本語における暗黙的な有害テキストを対象としたデータセット．3.6.1 項で述べたように，OffensiveLang を参考に設定した複数の社会的カテゴリーおよびターゲットグループに基づき，大規模言語モデルを用いてテキストを生成し，人手による注釈を付与した．サンプル数は 4,940 である．

multilingual_toxicity_dataset 日本語オンライン掲示板（2 ちゃんねる）から収集されたテキストを対象としたデータセット．Perspective API を用いて有害・無害のラベルが付与されており，主に明示的な有害表現を含むテキストから構成される．

WRIME: Dataset for Emotional Intensity Estimation 感情分析タスクの大規模なデータセット [17]．テキストに対して Joy, Sadness, Anticipation, Surprise, Anger のラベルが付与されている．サンプル数は 35,000 である．

日本語実験においても，英語実験と同様に暗黙的な有害テキストを中心としたテストデータ，明示的な有害テキストを中心としたテストデータ，および両者を含むテストデータの 3 種類を用いる．

- $\mathcal{D}^{\text{ja}}_{\text{imp}}$: 主に暗黙的な有害テキストと無害なテキストから構成されるテストデータ．Japanese Implicit Toxic Dataset を用い，有害テキストと無害テキストをそれぞれ 500 件ずつ抽出して構成する．

表 4.12: 日本語テストデータのサンプル数

	有害	無害	合計
$\mathcal{D}^{\text{ja}}_{\text{imp}}$	500	500	1,000
$\mathcal{D}^{\text{ja}}_{\text{exp}}$	500	500	1,000
$\mathcal{D}^{\text{ja}}_{\text{imp+exp}}$	1,000	1,000	2,000

- $\mathcal{D}^{\text{ja}}_{\text{exp}}$: 主に明示的な有害テキストと無害なテキストから構成されるテストデータ．multilingual_toxicity_dataset を用い，有害テキストと無害テキストをそれぞれ 500 件ずつ抽出して構成する．
- $\mathcal{D}^{\text{ja}}_{\text{imp+exp}}$: $\mathcal{D}^{\text{ja}}_{\text{imp}}$ と $\mathcal{D}^{\text{ja}}_{\text{exp}}$ を合わせたテストデータ．有害テキストと無害テキストをそれぞれ 1,000 件ずつ含む．

各テストデータの統計を表 4.12 に示す．

4.2.2 実験設定

本実験では以下のモデルを比較する．「B」はベースラインモデルを，「M」は提案手法のモデルを表す．

- $B^{\text{ja}}_{\text{imp}}$: 日本語の暗黙的な有害テキストデータセットを用いて学習したベースラインモデル．
- $B^{\text{ja}}_{\text{exp}}$: 日本語の明示的な有害テキストデータセットを用いて学習したベースラインモデル．
- $M^{\text{ja}}_{\text{imp}}$: 日本語の明示的有害テキストに対して言い換え手法を適用し，疑似的に暗黙的な有害テキストを生成して学習したモデル．
- $M^{\text{ja}}_{\text{imp+A}}$: $M^{\text{ja}}_{\text{imp}}$ の訓練データに対し，逆翻訳によるデータ拡張を行って学習したモデル．
- $M^{\text{ja}}_{\text{imp+L}}$: $M^{\text{ja}}_{\text{imp}}$ と同じ訓練データを用い，有害性判定タスクと感情判定タスクのマルチタスク学習によって学習したモデル．

日本語実験においては，日本語事前学習済みの BERT モデルを用いて学習および評価を行う．

表 4.13: 日本語実験における各モデルの訓練データのサンプル数

	有害	無害	合計
$B^{\mathrm{ja}}_{\mathrm{imp}}$	174	2,151	2,325
$B^{\mathrm{ja}}_{\mathrm{exp}}$	2,000	2,000	4,000
$M^{\mathrm{ja}}_{\mathrm{imp}}$	2,000	2,000	4,000
$M^{\mathrm{ja}}_{\mathrm{imp+A}}$	4,000	4,000	8,000
$M^{\mathrm{ja}}_{\mathrm{imp+L}}$	2,000	2,000	4,000

4.2.3 訓練データの詳細

日本語実験における各モデルの訓練データ構成を表 4.13 に示す．日本語実験においても，英語実験と同様に逆翻訳によるデータ拡張を適用したモデル（$M^{\mathrm{ja}}_{\mathrm{imp+A}}$）の評価を行った．しかし，後述するように日本語においては逆翻訳によるデータ拡張による暗黙的有害テキスト検出性能の向上が見られなかった．そのため，英語実験とは異なり，日本語実験におけるマルチタスク学習では，データ拡張を行わないモデル $M^{\mathrm{ja}}_{\mathrm{imp}}$ と同じ訓練データを用いて学習した．

4.2.4 暗黙的有害テキストのテストデータに対する結果と考察

まず，暗黙的有害テキストのみを含むテストデータ $\mathcal{D}^{\mathrm{ja}}_{\mathrm{imp}}$ に対する実験結果を表 4.14 に示す．明示的な有害テキストのみから学習したモデル $B^{\mathrm{ja}}_{\mathrm{exp}}$ は，再現率が 0.11 と極めて低く，多くの暗黙的有害テキストを検出できていないことが分かる．これは，モデルが特定の攻撃語や差別語といった表層的な語彙情報に強く依存しており，暗黙的に有害性が表現された文脈を十分に捉えられていないためであると考えられる．この傾向は，英語データセットに対する実験結果と一致しており，暗黙的有害テキスト検出の難しさが言語に依存しない性質を持つことを示唆している．

これに対し，暗黙的有害テキストのデータセットから学習した $B^{\mathrm{ja}}_{\mathrm{imp}}$ は F1 スコア 0.81 と高い性能を示した．この結果は，訓練データの規模が比較的小さい場合であっても，暗黙的有害性を明確に反映した高品質なデータセットを用いることで，暗黙的有害テキストの検出が可能であることを示している．ただし，この設定は訓練データとテストデータが同一分布に近い IID の条件であるため，性能が相対的に高くなっている点には注意が必要である．

一方で，日本語データセットにおいても，言い換えによって疑似的に生成した暗黙的有害テキストを用いて学習した $M^{\mathrm{ja}}_{\mathrm{imp}}$ が，明示的な有害表現のみから学習した $B^{\mathrm{ja}}_{\mathrm{exp}}$ と比較して，再現率および F1 スコアを大きく改善している点は重要である．この結果は，英語データセットに対する評価結果と同様に，明示的な攻撃語に依存した学習では暗黙的有害テキストの検出が困難である一方，言い換えによって表層的な有害語を除去しつつ有害な意図を保持した訓練データを用いることで，

表 4.14: 有害テキスト検出の実験結果（日本語 BERT，テストデータ $\mathcal{D}^{\mathrm{ja}}_{\mathrm{imp}}$）

	日本語 BERT			
	P	R	F	A
$B^{\mathrm{ja}}_{\mathrm{imp}}$	0.81	0.81	0.81	0.81
$B^{\mathrm{ja}}_{\mathrm{exp}}$	0.90	0.11	0.20	0.55
$M^{\mathrm{ja}}_{\mathrm{imp}}$	0.66	0.50	0.56	0.62
$M^{\mathrm{ja}}_{\mathrm{imp+A}}$	0.53	0.30	0.38	0.57
$M^{\mathrm{ja}}_{\mathrm{imp+L}}$	0.74	0.29	0.41	0.59

暗黙的有害性に対応した判別能力をモデルに付与できることを示している．ただし，$M^{\mathrm{ja}}_{\mathrm{imp}}$ の性能を $B^{\mathrm{ja}}_{\mathrm{imp}}$ と比較すると，依然として F1 スコアには一定の差が見られる．この差は，英語データセットにおける同様の比較と比べて大きい．日本語では，明示的な攻撃語や差別語を含まないまま，婉曲的な言い回しや文脈依存の表現によって否定的・攻撃的な意図が伝えられる例が多く，その解釈には文脈理解や語用論的知識が強く関与する．このような特徴により，表層的な語の置換を中心とした言い換え手法では，暗黙的有害性の一部は捉えられるものの，人手で収集・注釈付与された暗黙的有害テキストと同等の表現多様性を確保することが難しかったと考えられる．また，逆翻訳によるデータ拡張を行った $M^{\mathrm{ja}}_{\mathrm{imp+A}}$ の結果を確認すると，F1 スコアは 0.38 となり，データ拡張を行わない $M^{\mathrm{ja}}_{\mathrm{imp}}$（0.56）と比較して大きく性能が低下した．特に再現率が 0.50 から 0.30 へと大幅に悪化している．これは，日本語における逆翻訳の過程で，暗黙的な有害性が持つ微妙なニュアンスが失われたり，文脈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．この結果を踏まえ，日本語のマルチタスク学習モデル $M^{\mathrm{ja}}_{\mathrm{imp+L}}$ では，データ拡張を行わないデータセットを採用している．

最後に，マルチタスク学習の効果について考察する．表 4.14 に示すように，感情分析を補助タスクとして導入したモデル $M^{\mathrm{ja}}_{\mathrm{imp+L}}$ の F1 スコアは 0.41 となり，シングルタスクモデル $M^{\mathrm{ja}}_{\mathrm{imp}}$ (0.56) を下回る結果となった．詳細を見ると，精度は 0.66 から 0.74 へと向上した一方で，再現率が 0.50 から 0.29 へと大幅に低下している．英語実験においては，感情語を言い換えた疑似データを用いたマルチタスク学習が性能向上に寄与したが，日本語においては逆に性能を低下させる結果となった．この要因として，補助タスクである感情分析のデータセットを用いたモデルの学習が，主タスクである有害性判定タスクに悪影響を及ぼしたことが考えられる．特に，感情語を中立的な単語に言い換えたテキストを用いた感情分析タスクによるモデルの学習過程で，モデルが「表層的な感情語の欠如」を「無害」と結びつけるバイアスを強めてしまい，結果として感情語を伴わない暗黙的有害テキストの見落とし（False Negative）が増加したと推察される．また，英語実験で最も効果の高かった「皮肉判定タスク」を日本語ではデータセットの欠如により導入できなかったことも，マルチタスク学習の恩恵を十分に得られなかった一因であると考えられる．

表 4.15: 有害テキスト検出の実験結果（日本語 BERT，テストデータ $\mathcal{D}^{\text{ja}}_{\text{exp}}$）

	日本語 BERT			
	P	R	F	A
$B^{\text{ja}}_{\text{imp}}$	0.60	0.55	0.58	0.59
$B^{\text{ja}}_{\text{exp}}$	0.90	0.79	0.84	0.85
$M^{\text{ja}}_{\text{imp}}$	0.70	0.86	0.77	0.75
$M^{\text{ja}}_{\text{imp+A}}$	0.79	0.84	0.81	0.82
$M^{\text{ja}}_{\text{imp+L}}$	0.81	0.76	0.78	0.79

4.2.5 その他のテストデータに対する結果と考察

次に，明示的有害テキストのみを含むテストデータ $\mathcal{D}^{\text{ja}}_{\text{exp}}$ に対する実験結果を表 4.15 に示す．この条件では，大規模な明示的有害テキストデータセットから学習した $B^{\text{ja}}_{\text{exp}}$ が最も高い性能を示し，F1 スコアは 0.84 に達した．明示的有害テキストは差別語や攻撃語といった語彙的特徴が明確であるため，それらを大量に含むデータから学習したモデルが高い性能を示すことは妥当であり，この傾向は英語データセットに対する実験結果とも一致している．

一方で，$B^{\text{ja}}_{\text{imp}}$ は，F1 スコアが 0.58 と相対的に低い値に留まった．この結果は，英語データセットにおいては暗黙的有害テキストから学習したモデルが明示的有害テキストに対しても一定の検出性能を維持していたのに対し，日本語では同様の傾向が十分には再現されていないことを示している．この要因の一つとして，日本語における $B^{\text{ja}}_{\text{imp}}$ が，暗黙的な有害表現に特有の文脈的・語用論的特徴に強く適応した結果，明示的な攻撃語や差別語といった語彙的手がかりを十分に捉えられていない可能性が考えられる．すなわち，暗黙的有害性の検出に有効な特徴表現を重点的に学習する一方で，明示的有害表現に固有の語彙的特徴に対する感度が相対的に低下していることが，性能低下の一因となっていると考えられる．このことから，日本語においては，暗黙的有害テキストのみを用いた学習は，暗黙的有害性の検出には有効であるものの，明示的有害テキストに対する汎化性能という観点では一定の限界を持つことが示唆される．これに対し，言い換えによって生成した疑似的な暗黙的有害テキストを含むデータから学習した $M^{\text{ja}}_{\text{imp}}$ および $M^{\text{ja}}_{\text{imp+L}}$ は，$B^{\text{ja}}_{\text{imp}}$ を上回る性能を示し，F1 スコアはそれぞれ 0.77，0.78 に達している．このことから，暗黙的有害テキストを中心に学習したモデルであっても，言い換えによって生成された多様な訓練データを用いることで，明示的有害テキストに対する検出性能の低下をある程度抑制できることが分かる．データ拡張を行った $M^{\text{ja}}_{\text{imp+A}}$ は，F1 スコア 0.81 を記録し，$M^{\text{ja}}_{\text{imp}}$ （0.77）を上回る性能を示した．明示的な有害テキストは，特定の単語（攻撃語）の有無が判定に大きく寄与するため，逆翻訳によって多様な言い回しや語彙が生成されたことが，検出性能の向上にプラスに働いたと考えられる．

以上の結果から，日本語の明示的有害テキスト検出においては，明示的な攻撃

表 4.16: 有害テキスト検出の実験結果（日本語 BERT，テストデータ $\mathcal{D}^{\mathrm{ja}}_{\mathrm{imp+exp}}$）

	日本語 BERT			
	P	R	F	A
$B^{\mathrm{ja}}_{\mathrm{imp}}$	0.70	0.72	0.71	0.71
$B^{\mathrm{ja}}_{\mathrm{exp}}$	0.94	0.53	0.67	0.75
$M^{\mathrm{ja}}_{\mathrm{imp}}$	0.78	0.73	0.75	0.76
$M^{\mathrm{ja}}_{\mathrm{imp+A}}$	0.81	0.50	0.62	0.69
$M^{\mathrm{ja}}_{\mathrm{imp+L}}$	0.85	0.58	0.69	0.74

語を十分に含む大規模データセットによる学習が依然として有効である一方，言い換えによって生成した疑似的な暗黙的有害テキストを用いた提案手法は，暗黙的有害性への対応能力を維持しつつ，明示的有害テキストに対する性能低下を一定程度抑えるという点で，バランスの取れた学習戦略であることが確認された．

暗黙的および明示的な有害テキストが混在するテストデータ $\mathcal{D}^{\mathrm{ja}}_{\mathrm{imp+exp}}$ に対する結果を表 4.16 に示す．この現実的な設定においては，提案手法に基づくモデル $M^{\mathrm{ja}}_{\mathrm{imp}}$ が最も高い F1 スコア（0.75）を達成した．一方で，データ拡張を行った $M^{\mathrm{ja}}_{\mathrm{imp+A}}$ は F1 スコア 0.62 に留まった．これは前述の通り，明示的なテキストに対する性能は向上したものの，暗黙的なテキストに対する検出能力の低下が全体の結果に影響したためである．一方，暗黙的あるいは明示的なデータのみから学習したベースラインモデル $B^{\mathrm{ja}}_{\mathrm{imp}}$ および $B^{\mathrm{ja}}_{\mathrm{exp}}$ は，いずれもこれを下回る結果となった．このことから，日本語テキストにおいても，暗黙的有害表現と明示的有害表現が混在する状況では，暗黙的有害テキストを考慮した学習が有効であることが分かる．

以上の結果を総合すると，日本語データセットに対する実験においても，明示的な有害表現に依存したモデルは暗黙的有害テキストの検出が困難である一方，言い換えによって生成した暗黙的有害テキストを用いる提案手法は，暗黙的・明示的の両方が混在する設定において安定した性能を示すことが確認できた．これらの傾向は英語データセットに対する結果と概ね一致しており，本研究で提案した手法が特定の言語に強く依存しない，汎用的な暗黙的有害テキスト検出の枠組みである可能性を示唆している．

4.2.6 大規模言語モデルとの比較

日本語の処理能力に優れた大規模言語モデルを用いた場合の有害テキスト検出性能について確認する．モデルには，Llama 3 をベースに日本語データを継続事前学習させた Llama-3.1-Swallow-8B-Instruct-v0.1（以下，Swallow）を採用した．BERT と Swallow を用いた各テストデータに対する実験結果の比較を表 4.17，表 4.18，表 4.19 に示す．以後の考察では，まず LLM の実験結果 (表中の「Swallow」の結果) の分析について述べ，その後 BERT と LLM の比較について論じる．

表 4.17: BERT と Swallow による有害テキスト検出結果の比較（テストデータ $\mathcal{D}^{\text{ja}}_{\text{imp}}$）

	BERT				Swallow			
	P	R	F	A	P	R	F	A
$B^{\text{ja}}_{\text{imp}}$	0.81	**0.81**	**0.81**	**0.81**	0.81	**0.89**	**0.85**	**0.84**
$B^{\text{ja}}_{\text{exp}}$	**0.90**	0.11	0.20	0.55	**0.86**	0.33	0.48	0.64
$M^{\text{ja}}_{\text{imp}}$	0.66	0.50	0.56	0.62	0.84	0.43	0.57	0.67
$M^{\text{ja}}_{\text{imp+A}}$	0.53	0.30	0.38	0.57	0.79	0.39	0.52	0.65
$M^{\text{ja}}_{\text{imp+L}}$	0.74	0.29	0.41	0.59	**0.86**	0.42	0.56	0.65

LLM の考察　まず，暗黙的な有害テキスト $\mathcal{D}^{\text{ja}}_{\text{imp}}$ に対する結果を表 4.17 に示す．明示的な有害テキストのみを学習した $B^{\text{ja}}_{\text{exp}}$ は再現率が 0.33 と低く，F1 スコアも 0.48 に留まった．これに対し，提案手法である $M^{\text{ja}}_{\text{imp}}$ は再現率 0.43，F1 スコア 0.57 となり，ベースラインと比較して一定の性能向上が見られた．しかし，実際の暗黙的有害テキストを用いて学習した $B^{\text{ja}}_{\text{imp}}$（F1 スコア 0.85）と比較すると，その差は大きい．これは 4.2.4 節の BERT による実験結果（$M^{\text{ja}}_{\text{imp}}$ の F1 スコア 0.56）と同様の傾向であり，言語モデルの能力が向上しても，疑似データのみで学習するアプローチには依然として改善の余地があることを示唆している．また，データ拡張を行った $M^{\text{ja}}_{\text{imp+A}}$ の F1 スコアは 0.52 となり，$M^{\text{ja}}_{\text{imp}}$（0.57）を下回った．これは BERT を用いた実験と同様の傾向であり，日本語における逆翻訳データ拡張が暗黙的有害テキスト検出においてはノイズとなり得ることを，LLM を用いた場合でも裏付ける結果となった．一方，マルチタスク学習を行った $M^{\text{ja}}_{\text{imp+L}_1}$ の F1 スコアは 0.56 であり，$M^{\text{ja}}_{\text{imp}}$ と同等の性能を示した．

表 4.18: BERT と Swallow による有害テキスト検出結果の比較（テストデータ $\mathcal{D}^{\text{ja}}_{\text{exp}}$）

	BERT				Swallow			
	P	R	F	A	P	R	F	A
$B^{\text{ja}}_{\text{imp}}$	0.60	0.55	0.58	0.59	0.60	0.88	0.71	0.64
$B^{\text{ja}}_{\text{exp}}$	0.90	0.79	0.84	0.85	0.88	0.78	0.83	0.84
$M^{\text{ja}}_{\text{imp}}$	0.70	0.86	0.77	0.75	0.80	0.84	0.82	0.81
$M^{\text{ja}}_{\text{imp+A}}$	0.79	0.84	0.81	0.82	0.83	0.81	0.82	0.83
$M^{\text{ja}}_{\text{imp+L}}$	0.81	0.76	0.78	0.79	0.86	0.80	0.80	0.82

次に，明示的な有害テキスト $\mathcal{D}^{\text{ja}}_{\text{exp}}$ に対する結果を表 4.18 に示す．提案手法 $M^{\text{ja}}_{\text{imp}}$ は F1 スコア 0.82 を達成した．これは明示的な有害テキストで学習した $B^{\text{ja}}_{\text{exp}}$ の F1 スコア 0.83 に極めて近く，提案手法が暗黙的な有害性への適応を図りつつも，明示的な有害性の検出能力を損なっていないことを示している．$M^{\text{ja}}_{\text{imp+A}}$ および $M^{\text{ja}}_{\text{imp+L}}$ についても，F1 スコアはそれぞれ 0.82，0.80 となり，高い検出能力を維持していることが確認された．

表 4.19: BERT と Swallow による有害テキスト検出結果の比較（テストデータ $\mathcal{D}^{\mathrm{ja}}_{\mathrm{imp+exp}}$）

	BERT				Swallow			
	P	R	F	A	P	R	F	A
$B^{\mathrm{ja}}_{\mathrm{imp}}$	0.70	0.72	0.71	0.71	0.70	**0.91**	**0.79**	0.76
$B^{\mathrm{ja}}_{\mathrm{exp}}$	**0.94**	0.53	0.67	0.75	**0.90**	0.62	0.73	0.77
$M^{\mathrm{ja}}_{\mathrm{imp}}$	0.78	**0.73**	**0.75**	**0.76**	0.85	0.68	0.76	**0.78**
$M^{\mathrm{ja}}_{\mathrm{imp+A}}$	0.81	0.50	0.62	0.69	0.89	0.60	0.60	0.77
$M^{\mathrm{ja}}_{\mathrm{imp+L}}$	0.85	0.58	0.69	0.74	0.88	0.64	0.74	**0.78**

最後に，混合データ $\mathcal{D}^{\mathrm{ja}}_{\mathrm{imp+exp}}$ に対する結果を表 4.19 に示す．提案手法 $M^{\mathrm{ja}}_{\mathrm{imp}}$ が F1 スコア 0.76 を記録し，最も安定した性能を示した．データ拡張を行った $M^{\mathrm{ja}}_{\mathrm{imp+A}}$ は F1 スコア 0.60 となり性能が低下したが，マルチタスク学習モデル $M^{\mathrm{ja}}_{\mathrm{imp+L}}$ は 0.74 と，$M^{\mathrm{ja}}_{\mathrm{imp}}$ に近い性能を示した．

以上の結果から，日本語 LLM を用いた場合においても，言い換えによるデータ生成（$M^{\mathrm{ja}}_{\mathrm{imp}}$）は有効に機能することが確認された．一方で，逆翻訳によるデータ拡張は日本語においては逆効果となる場合があり，マルチタスク学習の導入についても，単独の言い換え手法を大きく上回る効果までは確認されなかった．これは，ベースとなる LLM（Swallow）が既に高い言語理解能力を有しており，単純なデータの増強や関連タスクの学習が必ずしも性能向上に直結しない可能性を示唆している．

BERT と LLM の比較　性能面においては，Swallow が多くの条件で BERT を上回るか同等の結果を示している．特にデータ拡張やマルチタスク学習を適用した際に BERT よりも高い正解率が得られている．BERT はパラメータ数が少なく計算コストが低いという実用上の利点を持つが，文脈理解や語用論的推論が不可欠となる暗黙的有害テキスト検出においては，大規模なパラメータと豊富な事前知識を持つ Swallow の方が適応力が高いことが確認された．また，BERT は訓練データの質や補助タスクの相性に敏感であり，ノイズによって性能が大きく変動する傾向が見られたのに対し，Swallow はそれらの悪影響を受けにくく，一貫して高い性能を維持できる点でも優位性がある．したがって，計算リソースが許容される環境であれば，Swallow のような日本語 LLM を採用することで，より高精度かつ安定した有害テキスト検出が可能になる．

第5章　おわりに

5.1　本論文のまとめ

本研究では，ソーシャルメディア上の深刻な問題である有害テキストの中でも，特に検出が困難な「暗黙的な有害テキスト」に着目し，その検出性能を向上させるための新たな手法を提案・検証した．従来の有害テキスト検出技術は，差別語や侮辱語といった明示的な攻撃表現を主な手がかりとしており，文脈依存性が高く，表層的には無害に見える暗黙的な有害表現に対しては効果が低かった．また，深層学習モデルの学習に不可欠な「暗黙的有害テキスト」のラベル付きデータは極めて希少であり，人手による大規模なデータセット構築もコストの観点から困難であるという課題があった．これらの課題を解決するため，本研究では，既存の豊富な言語資源を有効活用し，データの過疎性を克服する以下のアプローチを提案した．

第一に，言い換えによる疑似データの生成手法を提案した．本手法では，明示的な有害テキストに含まれる攻撃的な単語を，BERT の Masked Language Model (MLM) を利用して文脈に適した無害または中立的な単語に置換することで，疑似的な暗黙的有害テキストを自動生成した．これにより，攻撃的な語彙特徴を持たない有害テキストの訓練データを大量に確保することを可能にした．

第二に，逆翻訳によるデータ拡張とマルチタスク学習を導入した．データの多様性を高めるために複数の言語を経由する逆翻訳を行い，さらに有害性判定と親和性の高い「感情分析」および「皮肉判定」を補助タスクとしてマルチタスク学習を行うことで，モデルが表層的な単語だけでなく，文脈の背後にある否定的な感情や皮肉の意図を捉えられるようにした．

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て評価実験を行った．得られた主な知見は以下の通りである．

1. **明示的有害テキストのみによる学習の限界と提案手法の有効性**
   英語・日本語の双方において，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して極めて低い再現率を示した．これは，モデルが特定の攻撃語の有無に過度に依存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 スコアを大幅に向上させた．これにより，人手によるデータ作成を行わずとも，既存データからの変換によって暗黙的有害性の検

出能力を獲得できることが実証された．

2. **マルチタスク学習とデータ拡張の効果**
   英語データセットにおいては，逆翻訳によるデータ拡張と，皮肉判定タスクとのマルチタスク学習が最も高い性能（F1 スコア）を達成した．特に大規模言語モデル (Llama-3) を用いた実験では，皮肉判定タスクの併用が暗黙的有害テキストの検出に極めて有効であることが確認された．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考えられる．

3. **言語による特性の違い**
   日本語データセットに対する実験では，言い換えによるデータ生成の有効性が確認された一方で，逆翻訳によるデータ拡張は性能を低下させる場合があることが明らかとなった．日本語は文脈依存性が高く，機械翻訳の過程でニュアンスが変化しやすいため，データ拡張がノイズとして働いた可能性がある．また，日本語には皮肉の大規模データセットが存在しないため，感情分析のみを補助タスクとしたが，その効果は英語と比較して限定的であった．このことから，言語に応じた適切な補助タスクの選択やデータ拡張手法の検討が必要であることが示された．

本研究の貢献は，データ資源が乏しい暗黙的な有害テキストの検出というタスクに対し，既存資源の転用によって検出モデルを構築する汎用的なフレームワークを示した点にある．提案手法は特定の言語に依存しない仕組みであり，英語のみならず日本語においてもその有効性が確認されたことから，多言語の有害テキスト対策への応用が期待できる．

5.2 今後の課題

最後に今後の課題を述べる．実験では，マルチタスク学習は有害テキスト検出の性能を常に向上させるわけではなく，タスクの種類や組み合わせ，ベースとなる事前学習済み言語モデルによっては有効でないこともあった．補助タスクの選定方法や，タスク毎に FCL を結合する単純なモデルの代わりにより精緻なモデルを導入することを検討する必要がある．また，現在は有害語の言い換えによって擬似データを生成しているが，より自然で多様な暗黙的な有害テキストを生成する手法を探究することも重要である．将来的には，真に有害なテキストと皮肉や冗談など表層的には有害テキストに見えるが実際には無害な意味を持つテキストを識別したり，マルチモーダルな有害テキスト検出モデルを構築したりすることにも取り組みたい．

参考文献

[1] Aish Albladi, Minarul Islam, Amit Das, Maryam Bigonah, Zheng Zhang, and Fatemeh Jamshidi. Large language models for hate speech detection: A survey. In *IEEE Access*, pp. 20871–20892. IEEE, 2025.

[2] Zac Anger. profane-words: A list of profane English words. `https://github.com/zacanger/profane-words`, 2016. Accessed: 2025-04-19.

[3] Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, Vol. 24, , 2021. Article 100153.

[4] Meghana Moorthy Bhat, Saghar Hosseini, Ahmed Hassan Awadallah, Paul Bennett, and Weisheng Li. Say ‘YES’ to positivity: Detecting toxic language in workplace communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[5] Angelly Cabrera, Linus Lei, and Antonio Ortega. Transfer learning via lexical relatedness: A sarcasm and hate speech case study. *arXiv preprint arXiv:2508.16555*, 2025.

[6] Xi Chen and Shuo Wang. Pragmatic inference chain (PIC) improving LLMs’ reasoning of authentic implicit toxic language. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5826–5841. Association for Computational Linguistics, 2025.

[7] Amit Das, Mostafa Rahgouy, Dongji Feng, Zheng Zhang, Tathagata Bhattacharya, Nilanjana Raychawdhary, Fatemeh Jamshidi, Vinija Jain, Aman Chadha, Mary Sandage, Lauramarie Pope, Gerry Dozier, and Cheryl Seals. OffensiveLang: A community based implicit offensive language dataset. *IEEE Access*, Vol. 12, pp. 39289–39306, 2024.

[8] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2017.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 489–500, 2018.

[11] Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 345–363, 2021.

[12] Andrea Esuli and Fabrizio Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 417–422, 2006.

[13] Xiaochuang Han and Yulia Tsvetkov. Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7732–7739, 2020.

[14] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxIGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326, 2022.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[16] Amir Reza Jafari, Guanlin Li, Praboda Rajapaksha, Reza Farahbakhsh, and Noël Crespi. Fine-grained emotions influence on implicit hate speech detection. *IEEE Access*, Vol. 11, pp. 105330–105343, 2023.

[17] Koichi Kajiwara, Chu-Ren Huang, and Mamoru Komachi. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2095–2104. Association for Computational Linguistics, 2021.

[18] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[19] Sosuke Kobayashi. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 452–457, 2018.

[20] Kushagra. Sentiment and emotion analysis dataset. `https://www.kaggle.com/datasets/kushagra3204/sentiment-and-emotion-analysis-dataset`, 2022. Accessed 2025-04-19.

[21] Yejin Lee, Joonghyuk Hahn, Hyeseon Ahn, and Yo-Sub Han. AmpleHate: Amplifying the attention for versatile implicit hate detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28850–28862. Association for Computational Linguistics, 2025.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[24] Kehan Wang, Jiaxi Yang, and Hongjun Wu. A survey of toxic comment classification methods. *arXiv preprint*, 2021. arXiv:2112.07451.

[25] Lu Wei, Liangzhi Li, Tong Xiang, Xiao Liu, and Noa Garcia. Cracking the code: Enhancing implicit hate speech detection through coding classification. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pp. 112–126. Association for Computational Linguistics, 2025.

[26] Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly abusive language – what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 576–587, 2021.

[27] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1415–1420. Association for Computational Linguistics, 2019.

[28] Jianwei Zhang, Lin Li, and Shinsuke Nakajima. Constructing Japanese bullying expression dictionary for automated cyberbullying detection on Twitter. In *Vietnam Journal of Computer Science*, pp. 135–158. World Scientific, 2023.

[29] 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. 言語処理学会年次大会発表論文集 27, pp. 466–470. 言語処理学会, 2021.

[30] 久田祥平, 若宮翔子, 荒牧英治. オンライン誹謗中傷検出に向けた裁判例データセット. 自然言語処理, pp. 1598–1634. 言語処理学会, 2024.