日本語ヘイトスピーチ検出における疑似ラベルを用いた精度向上効果の検証

鍛原 大成　　倉嶋 将矢　　櫻井 義尚
明治大学 総合数理学部
ev201083@meiji.ac.jp　　ev201090@meiji.ac.jp　　sakuraiy@meiji.ac.jp

概要

本論文では，疑似ラベルを付与したデータセットを学習させることで，人間がラベリングせずに高精度のヘイトスピーチ検出モデルが構築可能であることを示す．この方法により，日本語ヘイトスピーチのラベル付きデータセットの作成にかかるコスト削減が期待できる．検証では，人間がラベリングしたデータセットと疑似ラベルを付与したデータセットを用いて学習した時の精度を比較した．その結果，人間がラベリングしたデータセットの精度よりは劣るが，LLM により疑似ラベルを付与したデータセットの F 値が 0.5，正解率が 0.9 を達成した．そのため，LLM により疑似ラベルを付与したデータセットが精度向上に有用であることが分かった．

1 はじめに

Twitter(現 X)などの SNS は，コミュニケーションツールとして広く普及しており，多くの人が利用している．一方で，SNS 上での投稿が誹謗中傷や炎上などに繋がり，事件に発展するなど，大きな社会問題となっている．しかし，誹謗中傷や炎上を引き起こすヘイトスピーチは膨大であるため，手動で判別するのは難しい．そこで，機械学習を使ってヘイトスピーチを自動的に検出することで，ユーザーを保護し，安心安全な SNS を実現できる．

現状，日本語におけるヘイトスピーチ検出は，英語に比べて精度が劣っており [2][3]，実用性は乏しい．その原因として，日本語ヘイトスピーチのラベル付きデータセットの不足 [1] がある．一方で，ラベル付きデータセットを用意するには高コストで時間も要してしまうため，困難である．

そこで，本研究では疑似ラベルを付与したデータセットを作成し，それを学習させることで，高精度なヘイトスピーチ検出モデルの構築が可能であることを示す．この方法により，人間がラベリングする際のコストや時間を削減することが期待できる．疑似ラベルの付与は 2 種類の方法で行った．1 つ目は，ラベル付きの英語の大規模データセットを機械翻訳する方法である．2 つ目は，ラベルなしの日本語データセットを，大規模言語モデル(LLM)を用いてラベリングする方法である．

2 関連研究

2.1 日本語ヘイトスピーチ検出

機械学習を用いた SNS 上の不適切な文章の検出を行う研究は広く行われている．しかし，不適切な文章の対象は研究によって異なっている．関連研究では，攻撃的な文章 [3][4] や，煽り [5]，誹謗中傷 [6]，炎上する文章 [7] などを対象としていた．また，不適切な文章の検出では，BERT の分類器が最も高精度であることが，松本ら [5] の研究により分かっているため，本研究でも，BERT を使用している．

ヘイトスピーチの定義は曖昧であるため，適切なラベリングが難しい．そこで，荒井ら [8] の研究では，日本語ヘイトスピーチ検出用のデータセットの構築手法を考案した．具体的には，検索語リストを使ったデータの収集方法や，ラベリングに関するガイドラインを設計した．本研究では，日本語ヘイトスピーチのデータセット構築及びラベリングにおいて，荒井ら [8] の研究を参考にしている．

日本語ヘイトスピーチ検出では，ラベル付き学習データの不足が課題であるため，近年は，この課題に対する研究が多く行われている．相川ら [2] の研究では，英語の Wikipedia の議論ページとコメントに対して攻撃的であるかのラベルを付与した 2 万件のデータを機械翻訳し，そのデータを学習した時の精度検証を行った．その結果，高い検出精度は得

られなかったが，日本語のラベル付き学習データがなくても検出可能であることが示されていた．また，及川ら [3] の研究では，能動学習手法を用いて学習データの選択を行うことで，学習データの大幅な削減が可能である事を示した．本研究では，疑似ラベルデータセットの作成により，データ不足に対処する．

2.2 疑似ラベルの有用性

疑似ラベルの付与により，安価で高速に大規模データセットの構築が可能になる．実際に，吉越ら [10] は，英語の自然言語推論データセットの機械翻訳により日本語のデータセットを構築した.そして，このデータセットを BERT により学習することで，93%の精度で自然言語推論タスクを解くことが可能であることを示した．また，Shuohang ら [11] は，NLU や NLG のタスクにおいて，人間がラベリングしたデータと GPT-3 でラベリングしたデータの精度とコストを比較した．その結果，GPT-3 でラベリングすることで，精度を下げずに，人間がラベリングするよりも 50%以上コストが低くなることを示した.

以上のことから，自然言語処理では，これらの疑似ラベルの付与方法は，すでに有用であることが示されている.そのため, 本研究では, 機械翻訳と LLM を用いて疑似ラベルを付与する方法を採用した.

3 外部データセット

3.1 日本語ヘイトスピーチデータセット

本研究では, Nishika のヘイトスピーチ検出コンペで作成されたデータセット [9] (以下，Nishika2chJP とする)を使用した．Nishika2chJP では，Nishika がおーぷん 2 ちゃんねる対話コーパス [12] に対してラベリングを行っていた．ラベリングの基準は，国連及び法務省のヘイトスピーチの定義と，荒井ら[8]の研究を参考に設定していた．この基準を基に，3 人のアノテータでラベリングを行い，最終的に多数決でヘイトスピーチかどうかのラベルを決定した．全体のデータ数は 5256 件で，各ラベルの数は表 1 の通りである．

表 1 Nishika2chJP の各ラベルの数

	normal	hate
各ラベルのデータ数	4950	306

3.2 英語ヘイトスピーチデータセット

本研究では，機械翻訳に使用する英語のデータセットとして， t-davidson, hate-speech-and-offensive-language データセット(以下, t-davidsonEN とする)を使用した. t-davidsonEN は, Davidson ら [13] がヘイトスピーチ検出精度の向上のために作成したデータセットである．Davidson らは，Twitter API を用いてユーザーがヘイトスピーチと認定した単語やフレーズを含むツイートを取得し, 各ツイートに対して「ヘイトスピーチである」，「ヘイトスピーチではないが攻撃的である」，「どちらでもない」の 3 カテゴリにラベリングした. ラベリングは 3 人以上で行い，多数決によりラベルを決定した．全体のデータ数は 24783 件で，各ラベルの数は表 2 の通りである．

表 2 t-davidsonEN の各ラベルの数

	hate	offensive	neither
各ラベルのデータ数	1430	19190	4163

4 提案手法

本章では，疑似ラベルを付与したデータセットの構築方法を説明する．

4.1 機械翻訳によるデータセットの構築

英語のデータセット(t-davidsonEN)を日本語に翻訳することで，機械翻訳データセットを構築した．

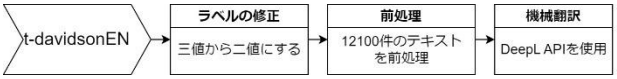

図 1 機械翻訳の流れ

機械翻訳の流れは図 1 の通りである．まず，今回のタスクは二値分類であるので，t-davidsonEN のラベルを三値から二値に直した．次に，t-davidsonEN のテキストを前処理した．最後に，前処理済みのテキストの内，12100 件を機械翻訳した．機械翻訳には，DeepL の API を使用した．機械翻訳したデータセットの各ラベルの数は表 3 の通りで，ラベルの分布を Nishika2chJP と同じになるように調整した．

表 3 機械翻訳データセットの各ラベルの数

	normal	hate
各ラベルのデータ数	11318	782

4.2 LLM によるデータセットの構築

疑似ラベルを付与するデータセットとして，2ch と Twitter のデータセットを用意した．

2ch のデータセットは，おーぷん 2 ちゃんねる対話コーパス [12] から 12500 件を使用した．

一方で，Twitter のデータセットは，Twitter API を用いて以下のキーワードで検索を行い，10000 件取得した．ただし，リツイートは自身の発言でないため，除外している．ここで，!付きは差別語句，＊付きは中立語句，@付きは排外主義である．

- 「白人主義!」「肌＊」「在日!」「韓国*」「シナ!」「中国＊」「半日@」「売国@」「黒人!」「障害者!」「気違い!」「きちがい!」「めくら!」「くろんぼ!」「外人!」

次に，疑似ラベルの付与では，GPT-4 を用いて Few-Shot で行った．プロンプトには，ヘイトスピーチのラベリング基準と，具体例を記載した．ラベリング基準は，Nishika2chJP のラベリングに使用したアノテーションドキュメントを参考にした．

Twitter，2ch のデータに対して GPT-4 でラベリングしたものをそれぞれ GPT4 (Twitter)，GPT4 (2ch) とする．それぞれの各ラベルの数は表 4，表 5 の通りである．

表 4 GPT4(Twitter)の各ラベルの数

	normal	hate
各ラベルのデータ数	7125	2875

表 5 GPT4(2ch)の各ラベルの数

	normal	hate
各ラベルのデータ数	11354	1146

5 検証方法

本章では，本研究で行った実験の流れと，ヘイトスピーチ検出モデルの構築方法を述べる．

5.1 実験の流れ

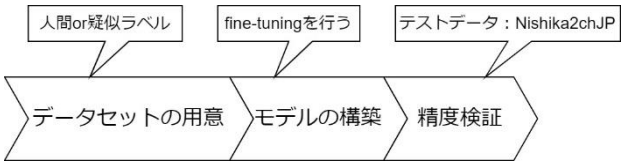

図 2 精度検証の流れ

本研究で行う実験では，疑似ラベルを付与したデータセットを用いたヘイトスピーチ検出の有用性を検証するために，図 2 の流れで精度検証を行った．まず，精度検証を行うデータセットを用意した．次に，そのデータセットを用いて fine-tuning を行い，ヘイトスピーチ検出モデルを構築した．最後に，テストデータに対する精度を検証した．テストデータには Nishika2chJP の内，1000 件を使用した．テストデータの各ラベルの数は表 6 に示した．

表 6 テストデータの各ラベルの数

	normal	hate
各ラベルのデータ数	939	61

5.2 ヘイトスピーチ検出モデルの構築

本実験では，LUKE [14] という事前学習済みの自然言語処理モデルに対して fine-tuning を行うことで，ヘイトスピーチ検出モデルを構築した．LUKE は，Studio Ousia の研究者が開発した，単語とエンティティ(固有表現)を事前学習した RoBERTa ベースの言語モデルである．LUKE の事前学習ではランダムにマスクされた単語とエンティティを予測するように学習されている．また，LUKE は日本語自然言語理解のベンチマークにおいて，Tohoku BERT や Waseda RoBERTa をはじめとする既存の言語モデルをこえる性能を獲得している．

6 実験 1:各データセットの精度比較

6.1 実験概要

実験 1 では，人間がラベリングしたデータセットと疑似ラベルを付与したデータセットを学習させた時の精度を検証した．さらに，それぞれの精度を比較し，疑似ラベルを付与したデータセットを用いたヘイトスピーチ検出の有用性を検討した．人間がラベリングしたデータセットには，Nishika2chJP を使用した．このデータセットを Base と表記する．一方で，疑似ラベルを付与したデータセットには，機械翻訳データセット・GPT4 (Twitter)・GPT4 (2ch)を使用した．疑似ラベルを付与したデータセットは，ラベルの割合を維持したままデータ数を段階的に増やして学習させて，精度を検証した．

6.2 実験結果

表 7 実験 1 の精度結果①

	Base	機械翻訳		
件数	4000	4000	8000	12000
正解率	0.955	0.921	0.917	0.896
F 値	0.672	0.288	0.357	0.381
適合率	0.605	0.320	0.338	0.299
再現率	0.754	0.262	0.377	0.525

表 8 実験 1 の精度結果②

	GPT4 (Twitter)			GPT4 (2ch)		
件数	4000	8000	10000	4000	8000	12000
正解率	0.897	0.865	0.884	0.898	0.889	0.906
F 値	0.443	0.440	0.468	0.386	0.448	0.510
適合率	0.331	0.294	0.325	0.305	0.321	0.374
再現率	0.672	0.869	0.836	0.475	0.738	0.803

6.3 考察

表 7，8 から，疑似ラベルを付与したデータセットの精度は，学習件数を増やしても Base に比べて大きく劣ることが分かった．しかし，GPT4 ラベル(2ch)に関しては，12000 件のデータを学習させることで，正解率が 0.9，F 値が 0.5 を超えることができた．また，データ数を増やして学習させることで，全体的に精度が向上したため，さらに学習件数を増やすことで，人間がラベリングして学習した時の精度に近づくことが期待できる．以上のことから，テストデータと同一プラットフォームの文章に対して GPT-4 でラベリングすることで，人間によるラベリングをせずに，高精度のヘイトスピーチ検出モデルの構築が可能であることが示せた．

7 実験 2:組み合わせによる精度検証

7.1 実験概要

実験 2 では，人間がラベリングしたデータセットに疑似ラベルを付与したデータセットを組み合わせて学習させた時の検出精度の向上効果を検証した．具体的には，Base のデータセットに対して疑似ラベルを付与したデータセットのデータ数を段階的に増やして加えた時の検出精度を検証した．

7.2 実験結果

表 9 実験 2 の精度結果①

	Base	Base+機械翻訳		
件数	4000	8000	12000	16000
正解率	0.955	0.953	0.948	0.945
F 値	0.672	0.630	0.629	0.587
適合率	0.605	0.606	0.557	0.542
再現率	0.754	0.656	0.721	0.639

表 10 実験 2 の精度結果②

	Base+GPT4 (Twitter)			Base+GPT4 (2ch)		
件数	8000	12000	14000	8000	12000	16000
正解率	0.943	0.948	0.942	0.952	0.944	0.940
F 値	0.596	0.629	0.603	0.647	0.627	0.600
適合率	0.525	0.557	0.518	0.587	0.528	0.506
再現率	0.689	0.721	0.721	0.721	0.771	0.738

7.3 考察

表 9，10 から，人間がラベリングしたデータに疑似ラベルを付与したデータを組み合わせて学習させると，精度が悪くなっている．さらに，疑似ラベルを付与したデータを増やすほど，精度が悪くなっている．そのため，ヘイトスピーチ検出において，人間がラベリングしたデータがある場合，新たに疑似ラベルを付与したデータを作成してデータ数を増やした精度向上は，効果的でないことが分かった．この要因は，疑似ラベルの不正確さが考えられる．疑似ラベルには，誤分類しているデータも多く含まれており，そのデータも学習させている．そのため，疑似ラベルを付与したデータを増やすほど，精度が悪くなっていると考えられる．

8 おわりに

本研究では，日本語のヘイトスピーチ検出におけるラベル付きデータセット不足を解消するために，疑似ラベルを付与したデータセットを作成し，それを学習させてヘイトスピーチ検出モデルを構築した．その結果，実験 1 より，人間によるラベリングを学習させた時の精度よりは劣るが，F 値が 0.5，正解率が 0.9 を達成できた．そのため，人間によるラベリングをせずに高精度のモデルの構築が可能であることが分かった．一方で，実験 2 より，人間がラベリングしたデータセットに疑似ラベルを付与したデータセットを加えて学習させると，精度は悪化してしまった．また，疑似ラベルを付与したデータセットを学習させた時の精度も人間がラベリングしたデータセットに比べて大きく劣る．これらは，疑似ラベルの精度の低さが要因であると考えられる．そのため，今後は，LLM に入力するプロンプトの工夫などにより，疑似ラベルの分類精度の精度を向上させることを目指す．

謝辞

本研究は JSPS 科研費 20K11960 の助成を受けたものです.

参考文献

[1] Md Saroar Jahan and Mourad Oussalah, 2023. A systematic review of hate speech automatic detection using natural language processing. Neurocomputing 546: 126232.

[2] 相川和希, 河合新, 延原肇. “不適切なテキストコンテンツの検出法 —Doc2Vec を橋梁とした多言語対応—”. 情報処理学会第 81 回全国大会, 2019, 2U-06.

[3] 及川正樹, 井上潮. “BERT と能動学習を用いた攻撃的なツイートの分類”. DEIM Forum, 2023, 1b-2-1.

[4] 吉田基信, 松本和幸, 吉田稔, 北研二. “BERT を用いた SNS 上における攻撃的文章訂正システム”. 情報処理学会全国大会講演論文集, 2022, p.725-726.

[5] 松本典久, 上野史, 太田学. “BERT を利用した煽りツイート検出の一手法”. DEIM2021, I14-2, 2021.

[6] 石坂達也, 山本和英. “Web 上の誹謗中傷を表す文の自動検出”. 言語処理学会第 17 回年次大会, E1-6, 2011.

[7] 大西真輝, 澤井裕一郎, 駒井雅之, 酒井一樹, 進藤裕之. “ツイート炎上抑制のための包括的システムの構築”. 人工知能学会全国大会論文集, 2015, vol. 29, p. 1-4.

[8] 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, 2021, vol.17, p466-470.

[9] Nishika 株式会社. “ヘイトスピーチ検出”. Nishika. https://competition.nishika.com/competitions/hate/summary#description

[10] 吉越卓見, 河原大輔, 黒橋禎夫. “機械翻訳を用いた自然言語推論データセットの多言語化”. 第 244 回自然言語処理研究会, 2020, vol.6, p. 1-8.

[11] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. “Want To Reduce Labeling Cost? GPT-3 Can Help”. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[12] 稲葉通将. “おーぷん 2 ちゃんねる対話コーパスを用いた用例ベース対話システム”. 第 87 回言語・音声理解と対話処理研究会(第 10 回対話システムシンポジウム), 人工知能学会研究会資料 SIG-SLUD-B902-33, 2019, p129-132.

[13] Davidson, Thomas and Warmsley, Dana and Macy, Michael and Weber, Ingmar. 2017. “Automated Hate Speech Detection and the Problem of Offensive Language”. Proceedings of the 11th International AAAI Conference on Web and Social Media: ICWSM '17, pages 512-515

[14] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6442–6454, Online. Association for Computational Linguistics.