ヘイトスピーチ検出における GPT-4 による擬似ラベル付与の手法と評価

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

概要

本研究では，GPT-4 を利用した擬似ラベル付与手法を提案し，その有効性を実証した．初期の F1 スコアが 0.51 であったが，本手法の適用により 0.68 まで向上し，従来の機械学習によるラベリングを上回る精度を達成した．正解率に関しては 94.90%を達成している．これにより，GPT-4 の応用範囲が拡がり，データセットの品質向上に寄与する可能性が示唆された．

1 はじめに

本研究は，SNS の普及に伴い問題となっているヘイトスピーチの検出に焦点を当てる．オンライン上での誹謗中傷や差別的な発言は増加の一途を辿り，これらを自動的に識別する技術の需要が高まっている．しかし，すべてのネガティブなテキストを単純に排除するわけにはいかず，言論の自由を尊重しつつ適切なバランスを見つける必要がある．

1．1 日本語データセットの課題

特に日本語においては，ヘイトスピーチの識別に用いるデータセットが不足しており，存在するものの精度も低い．この課題を解決するためには，大量のテキストデータに対する正確なアノテーションが求められるが，これは膨大な人的労力と時間を要する作業である．

1．2 GPT-4 の活用

本研究では，GPT-4 を活用した擬似ラベル付与の手法を提案し，この問題に対処する．GPT-4 を用いることで，アノテーションの時間的および金銭的コストを削減しつつ，精度の高いデータセットの作成を目指す．この手法は，ヘイトスピーチ検出に留まらず，広範囲なデータアノテーションタスクに応用可能である．

1．3 本研究の意義

このように, GPT-4 を用いた擬似ラベルの付与は，効率的かつ高精度なデータセット構築を実現し，オンライン上のヘイトスピーチ検出技術の進展に寄与するだけでなく，人工知能の適用範囲を広げることが期待される．

2 関連研究

2．1 GPT によるアノテーション

先行研究 [1]での GPT-3 の適用は，テキストアノテーションにかかる時間的および金銭的コストを削減する可能性を示している．特に，GPT-3 でラベリングしたデータと人間によるラベリングを組み合わせることで，どちらか一方のデータのみを用いた場合と比較して，より精度の高いモデルの構築が可能であることが示されている．このアプローチは，ヘイトスピーチ検出といった特定のタスクにおいても有効性を持つことが考えられる．

本研究では，GPT-4 を利用した擬似ラベルの付与手法を提案し，日本語での有効性とヘイトスピーチ検出タスクにおける有効性を検証する．

2．2 日本語ヘイトスピーチ

ヘイトスピーチの検出には，その定義の曖昧さという課題が伴う．荒井ら[2]の研究では，ヘイトスピーチをどのように定義し，それに基づいたデータセットを作成するかが議論されている．

本研究では, これらの定義付けの難しさを認識し, GPT-4 を用いた品質の良いデータセットの作成を試みる．この過程で，先行研究の定義と方法論を参考にしつつ, より進化した GPT モデルがもたらす新たな可能性について検討する．

3 提案手法

本章ではヘイトスピーチ検出用の日本語データセットを GPT-4 で生成する手法を提案する. GPT 以外にも大規模言語生成モデルが存在するが，今回は自然言語処理能力が高く，拡張性の高い GPT-4 を採用している.

プロンプトやパラメータなどが変数として与えられるが，本研究では下記 5 つの手法を提案する.

3. 1 single

single は，各テキストに対して一回のみラベリングを行う手法である．本アプローチは単純かつ迅速であり，大量のデータに対する初期フィルタリングに適している．また，logprobs パラメータを True に設定することにより，得られたデータの特性を把握し，後続の手法での活用を容易にする.

本手法により，ラベリングされたデータの基本的な特徴や傾向を初期段階で理解することができる．データの効果的な処理戦略を立案し，後続の分析やアプローチの基盤を築くことも可能となる.

3. 2 majority

majority は，テキストに対して GPT-4 を用いて 3 回のラベリングを行い，その結果に基づいて多数決で最終ラベルを決定する手法である．本手法は Nishika2chJP データセットの作成時にも同様に採用されている.

複数回ラベリングを行うことにより，単一回のラベリングに比べてより精度の高いラベル付けが可能になると期待される．各ラウンドのラベリングで得られる異なる視点や解釈により，最終的な判断において，よりバランスの取れた結果が得られると考えられる．この多数決方式は，特に複雑なテキストや微妙なニュアンスを含むデータに対して，より信頼性の高いアノテーション獲得を目指す際に採用する.

3. 3 0_priority

0_priority は，3 回のラベリング全てでヘイトスピーチと判定された場合のみ，該当テキストをヘイトスピーチとする手法である．偽陽性（ヘイトスピーチでないものをヘイトスピーチと誤判定する現象）の減少を目指して設計されているものである.

本手法により，あるテキストが基準を満たした場合にのみヘイトスピーチと判定されるため，偽陽性の発生確率が低下する．今回は 3 回のラベリング全てでヘイトスピーチと判定された場合を基準として設けているが，５回などに判定回数を増やした場合には適切な基準を設定する必要がある.

この基準によって，実際にヘイトスピーチであるにもかかわらず見逃される（偽陰性）リスクが高まるという側面も存在する．そのため，本手法の基準は慎重に設定する必要がある.

3. 4 miss

miss は，トレーニングデータにおいて誤ったラベリングがなされたケースをプロンプトに含め，それによってモデルが過去のミスから学習する手法である. 本アプローチは, モデルが以前の誤りを反映し，より精確なラベリングを行うことを目指す．誤ったラベリングの事例を分析し，それをモデルの学習プロセスに取り入れることで，類似の状況における将来のラベリングの正確性の向上が期待される.

3. 5 logprob

logprob は, 2023 年 11 月に発表された logprobs（出力トークンのログ確率）を活用する手法である．本アプローチでは, logprob が設定した特定の値に達するまで最大 5 回のリクエストを行い，その基準を満たさない場合はテキストを「normal」と判定する.

手法のキーとなるのは, logprob の閾値設定である. logprob は, モデルが出力したトークンの確からしさを数値化したもので，この数値が高いほどモデルはそのトークンの出力に自信を持っていると解釈できる．手法の適用にあたっては，max_tokens=1, temperature=0, top_p=1 というパラメータも同時に設定している.

4 外部データセット

4. 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[3]で作成されたデータセット（以下，Nishika2chJP とする）を活用した．このデータセットは，国連および日本法務省のヘイトスピーチの定義に基づき，3 人のアノテータによってラベリングされている．ラベリング対象は，おーぷん 2 ちゃんねる対話コーパス[6]で，最終的なラベル付けは多数決で行われた．データセット全体の件数は 5,256 件で，各ラベルの数は表 1 のとおり.

表 1　Nishika2chJP の各ラベルの数

	normal	hate
各ラベルのデータ数	4,950	306

5 実験

荒井ら[2]の研究に基づきプロンプトを作成し，ラベリングする実験を行う．GPT-4 と記載されているのは，gpt-4-1106-preview モデルであり，特に断りのない場合，パラメータはデフォルト値である．プロンプトの詳細は Appendix A に記載した．

5．1 実験手順

本研究の実験手順は以下のとおりである．

1. Nishika2chJP のデータをトレーニングデータとテストデータに分ける
2. GPT_single にてベースラインの設定を行う
3. 2 の結果から logprob の値を決める
4. 各手法にてラベリングを行う
5. 各種法を組み合わせてラベリングを行う

5．2 評価方法

ラベリングの精度評価は Nishika2chJP のうちランダムに抽出した 4,000 件をトレーニングデータ，1,000 件をテストデータと分割した上で，Accuracy, Precision, Recall, F-measure を評価指標とする．分割後の train, test データの分布は表 3 のとおりである．

表 3　Nishika2chJP の各ラベルの数

	normal	hate
Nishika2chJP_train	3,756	244
Nishika2chJP_test	939	61

5．3 実験結果

single で実験した際の logprob を表 4，5 のとおりまとめた．予測が正しい場合は 80%以上が logplob >= -0.01 になることがわかったため，本実験では logprob の値が-0.01 以上になるまで最大 5 回ラベリングを行い，5 回のラベリングでも logprob の値が -0.01 未満の場合は normal とした．

表 4
Nishika2chJP の予測が正しい場合の logprob 割合

	count	ratio
TRUE_logprob >= -0.01	739	81.93%
TRUE_logprob < -0.01	163	18.07%

表 5
Nishika2chJP の予測が誤りの場合の logprob 割合

	count	ratio
FALSE_logprob >= -0.01	62	63.27%
FALSE_logprob < -0.01	36	36.73%

実験結果は表 6 のとおりである．nomal を優先する 0_priority と，確率を計算する logprob の f1 スコアが高かったため組み合わせたところ，最も良い結果となった．また，single 手法にて最初に設定したベースライン 0.507 から，最終的には 0.683 まで f1 スコアを向上させることができた．

表 6 Nishika2chJP のラベリング結果

	Acc.	Pre.	Rec.	F1
single(base)	89.70％	0.358	0.869	**0.507**
majority	89.80%	0.361	0.869	0.510
0_priority	92.10%	0.426	0.852	0.568
miss	75.40%	0.180	0.852	0.297
logprob	94.20%	0.514	0.918	0.659
Logprob &majority	94.40%	0.523	0.918	0.667
logprob &0_priority	94.90%	0.550	0.902	**0.683**
機械学習（参考）	95.50%	0.605	0.754	**0.672**

この 0.683 という f1 スコアは Nishika2chJP のトレーニングデータを用いて Bert を事前学習モデルとして機械学習した結果，0.672 よりも高い値であり，機械学習より生成モデルでラベリングを行うほうがラベリングの精度が高いことがわかった．

5．4 考察

本研究で行った実験では，ヘイトスピーチ検出において，単一ラベリングよりも複数回ラベリングを行う majority や 0_priority 手法の方が効果的であることが確認された．single の Precision が低いことは，

偽陽性の多さを示しており，特にヘイトスピーチでない内容をヘイトスピーチと誤判定するケースが多いことを意味している．この結果は，複数回のラベリングにより，精緻なデータ分析が可能になることを示唆しています．

一方で，miss 手法では，過去の誤りから学習させる試みがモデルを混乱させ，精度を著しく低下させた．これは，適切な量のトレーニングデータの選択が重要であることを示している．また，logprob と 0_priority の組み合わせにより，精度が向上することが観察された．これは，logprob のパラメータ調整がモデルの精度向上に貢献し，0_priority の厳格な判定基準が偽陽性を減少させたことによるものと考えられる．これらの結果は，さらなるパラメータチューニングやラベリング回数の調整によって，モデルの精度をさらに高めることが可能であることを示唆している．

6 おわりに

本研究では，二値分類タスクにおける生成モデル，特に GPT-4 を用いた分類手法を提案した．一方のカテゴリを優先する手法やパラメータチューニングを行うことで，既存の機械学習手法よりも高い精度を達成することができた．特にヘイトスピーチ検出のような複雑で微妙な判断が求められるタスクにおいて，GPT-4 の高度な機能を活用することで，より正確で信頼性のある分類を実現した点は，本研究の重要な成果といえる．

今後の研究においては，さらなるパラメータチューニングやラベリング回数の増加を通じて，分類精度の向上を目指す．また，本研究で得られた知見は，ヘイトスピーチ検出に限らず，他の多岐にわたる二値分類タスクにも応用可能であると考えられる．生成モデルの能力を生かし，様々な分野での新たな可能性を探求することが，次なるステップとして期待される．

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

[2] 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.

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

[4] Achiam, Josh, et al. "GPT-4 Technical Report." ArXiv abs/2303.08774, 2023.

[5] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, Jianfeng Gao. “Instruction Tuning with GPT-4. ”CORR abs/2304.03277, 2023.

[6]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.

[7] Microsoft. “Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine”, ArXiv abs/2311.16452, 2023.

[7] Wang, Yaqing, et al. "Adaptive self-training for few-shot neural sequence labeling." ArXiv abs/2010.03680, 2020.

A ヘイトスピーチ検出のためのプロンプト

以下の設定にて行った.

ただし,<input_text>という変数にラベリングを行うテキストを入力すること.

{"role": "system", "content": "あなたは発言がヘイトスピーチかどうか判定を行うアノテーターです. 返答は 0 か 1 のみで行います. 数字は 1 つしか答えません. "},
{"role": "user", "content": """
以下のアノテーション基準に則り, アノテーションを行ってください.
また, 制約条件は必ず守ってください.

＃アノテーション基準 :
STEP①
まずは発言が誰を対象としたものか判定します.

A1. 容易に変更できないヒトの属性に基づくグループ（例 : 人種, 国籍, 民族, 信条, 年齢, 性別, 性的指向, 性自認, 出身, 障害など）
A2. A1 に当てはまらないその他のグループ（例 : 職ミュージシャン, クリエイター, 性風俗従事者, 生活保護受給者, 引きこもり, ヴィーガンなど）
A3. 個人
A4. 対象がはっきりしないもの

このとき, A1 または A2 を対象にする攻撃性のある発言文章はヘイトスピーチとします. A1 または A2 を対象とするものの攻撃性の無いものはヘイトスピーチとしません.

STEP②
次に発言の内容を判定します.

B1. コミュニティや地域から排除を告知したり扇動するような内容（例 : ○○は出ていけ」）
B2. 生命を脅かす, 精神的・身体的に危害を加えるような内容（例 : 「○○は死んだほうが良い」）
B3. 口汚くののしる, 侮蔑する（例 : 「○○はゴキブリ以下！」「○○は劣等民族」など）
B4. 不確かな根拠にもとづいた情報を言いふらす, 拡散する（例 : 「○○は凶悪犯罪を起こすので危険」）
B5. その他の内容

このとき, 発言の対象が A1 または A2 であり, B1・B2 ・B3 ・B4 のいずれかの要素を含む発言はヘイトスピーチとします.

STEP③
最後にヘイトスピーチの判定を行います.
ヘイトスピーチの場合は 1, ヘイトスピーチでない場合は 0 とします.

＃ヘイトスピーチではない例 :
以下はヘイトスピーチではないとします.
1. 「個人的な攻撃」（例 : 「お前はクソか？」）
※攻撃の対象が A1 か A2 ではない場合はヘイトではない. 攻撃性の有無だけでは判断しないこと.
2. 「趣味嗜好に基づくグループ」（例 : 「喫煙者」「撮り鉄」「ギャンブル」など）
※「容易に変更できないヒトの属性に基づくグループ」ではないのでヘイトスピーチではないとします
3. 「政党, 政治的な主義」（例 : 「○○党」「共産主義」「保守派」など）
※「容易に変更できないヒトの属性に基づくグループ」ではないのでヘイトスピーチではないとします

「パヨク」「ネトウヨ」などの政治的な思想集団に対する蔑称を使った攻撃もヘイトスピーチではないとします.
ヘイトスピーチでないのにヘイトスピーチと判定してしまうことが多いため, 判定に自信がない場合は 0 としてください.

＃制約条件 :
・返答は 0 か 1 のみ
・関係ない発言は行わず, ヘイトスピーチ判定のみ回答する
・数字以外発言しない
・ヘイトスピーチの判定はアノテーション基準に則る
・ヘイトスピーチではない例も参考にする

＃発言内容 :
""" + <input_text>}