# エンコーダ型・デコーダ型言語モデルを併用した差別的発言スパン検出性能向上

小林秀太郎 [1]　荒瀬由紀 [1]
[1] 東京科学大学
kobayashi.s.08a4@m.isct.ac.jp　arase@c.titech.ac.jp

## 概要

SNS をはじめとするインターネット上の文章には差別的発言が含まれている場合がある．そうした文章が大規模言語モデルの学習データに含まれていた場合，そのモデルは差別的発言を含む文章を生成する可能性がある．そのためモデルが文章を生成した時に，そのテキストから差別的発言を検出する手法が必要である．差別的発言を検出する既存手法は多くあるが，テキストから差別的発言に該当するスパンを抽出する手法の精度は不十分である．本研究では，エンコーダ型の言語モデルである RoBERTa とデコーダ型の言語モデルである Gemma，Qwen の予測を組み合わせるアンサンブルにより，ベースモデルよりも高い F1 スコアを達成した．

注意：この文書には差別的発言の例が含まれている．また，それらの例は著者の個人的見解を示すものではない．

## 1　はじめに

本研究では差別的発言を，特定の集団や個人に対して人種や性別，宗教といったアイデンティティ要素に基づいた敵意や偏見を含む発言と定義する [1]．単なる批判や暴言と異なるのは，差別的発言は多様性や人権を否定することで社会の結束を弱体化させ，特定のコミュニティに対する人々の憎悪を肥大化させる点にある [2]．つまり，差別的発言は社会全体の安寧を脅かす可能性を持っている．

インターネット上のテキストには様々な差別的発言が含まれている [3]．そうした有害な情報を含むコーパスから学習した大規模言語モデルは，差別的発言を出力する可能性を持つ．したがって，言語モデルの出力をフィルタリングするための効果的な手法が必要である．差別的発言は文章全体ではなくその一部に含まれることが多いため，本研究ではテキストから差別的発言に該当する部分を抽出することを目的とする．このアプローチはテキスト全体が差別的であるかを検出する手法に比べて，説明可能性や実用性の面で優れている．また差別的発言の検出に関わる研究は多く存在するが，テキストから差別的発言である部分を検出する既存の手法は十分な精度を持っているとは言えない [4]．これを改善するために，本研究では複数のエンコーダ型の言語モデルとデコーダ型の言語モデルを組み合わせることで検出精度を向上させた．本研究で構築したモデルの実装は GitHub で公開している[1)]．

## 2　関連研究

### 2.1　差別的発言の検出

テキストから差別的発言を検出する手法の多くは，系列モデルによる分類が用いられる．例として，LSTM[5] を用いた分類モデルに加え，BERT[6] や Llama[7] などの Transformer[8] を用いた分類が挙げられる．また，検出対象には大きく分けて 2 種類ある．1 つは，対象のテキストが差別的発言を含むかを判別するというような，テキスト全体に対する分類である [9][10]．もう 1 つは，対象のテキストに含まれる各単語やトークンが差別的発言に含まれるかを判別するというような，テキスト内の各要素に対する分類である [11][12]．本研究は後者の検出方法を採用しており，テキスト内の各単語に対する二値分類を行った．

### 2.2　アンサンブル学習

複数のモデルを組み合わせることでより高いパフォーマンスを発揮するモデルを構築する手法をアンサンブル学習という [13]．差別的発言の検出にもアンサンブル学習を用いた研究がある．BERT や

1）https://github.com/str30342/EnsembleModelForHateSpeechDetection

RoBERTa[14] などのエンコーダ型の言語モデルを組み合わせることで，単一のモデルよりも高い性能を達成している [15][16]．また，それらにデコーダ型の言語モデルである Gemma[17] を組み合わせることで，テキスト全体に対する分類の精度を高めた研究もある [18]．本研究では，エンコーダ型の言語モデルとデコーダ型の言語モデルを組み合わせることで，テキスト内の差別的発言スパン検出の精度を高めることを目指す．

# 3 提案手法

本研究で提案するモデルは図 1 に示すように，複数のモデルから得た予測ラベルを多数決で統合することで，最終的なラベルを予測するモデルである．

## 3.1 トークン二値分類モデル

前述したとおり，入力文に含まれるそれぞれの単語が差別的発言に含まれるかの単語単位の二値分類を行った．ラベルは，0 を差別的発言に含まれない単語，1 を差別的発言に含まれる単語として付加した．

**エンコーダ型** エンコーダ型の言語モデルでは，トークン分類用分類器を付加し各トークンについて二値のラベルを予測する．得られたトークン単位のラベルは単語単位のラベルに変換する．

**デコーダ型** デコーダ型の言語モデルでは，プロンプトを用いて各単語に対するラベルを出力させた．使用したプロンプトと入出力の一例を表 1 に示す．プロンプトでは差別的発言の定義と入力の形式，タスクの説明，出力形式について説明している．入力では始めに対象の文の単語数を明示した後，対象の文に含まれる単語をリスト形式にしたものを記している．出力は JSON 形式であり，文章中の単語とその単語に対するラベルの辞書を値とするリストとなっている．このリストからラベルのみを抽出して予測ラベルのリストを作成した．

## 3.2 多数決による出力の決定

アンサンブルモデルの構成を図 1 に示す．最終的な予測ラベルを得るために，各モデルから得られた予測ラベルをもとに，各単語に付加されたラベルについて多数決をとる．このとき，デコーダ型の言語モデルは出力の形式が正しい JSON 形式となっていない場合がある．その場合は，正しい出力を行っているモデルのみで多数決をとる．また丁度半数のモデルがラベル 1 を選択した場合，優先度の高いモデルの予測を採用する．優先度は個々のベースモデルの検出性能によって決定する．こうして得られたラベルを最終的な予測ラベルとした．

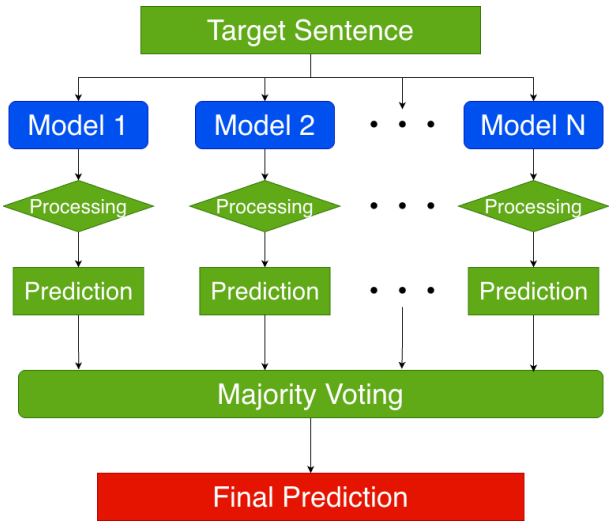


**図 1** アンサンブルモデルの構成

## 3.3 異なるサブワード単位の統合

エンコーダ型の言語モデルはそれぞれ異なるサブワード単位を持つ．例えば BERT は，「tokenization」という単語を「token」と「##ization」という 2 つのサブワードに分割する．この場合，「token」に付加されたラベルだけを「tokenization」という単語に対する予測ラベルとした．こうすることで，異なるサブワード単位であっても最終的に単語数と同じ長さの予測ラベルを得る．

# 4 実験設定

本研究では差別的発言のスパンがアノテーションされている HateNorm23[2] を用いて評価を行う [19][4]．

## 4.1 データセット

HateNorm23 には，2,421 件の訓練データと 606 件の評価データが含まれており，すべてのデータに差別的発言である部分が含まれている．本研究では訓練データからランダムに抽出した 300 件を検証データとして利用した．各データの全単語数における差別的発言に含まれる単語の割合は，訓練データにおいては 17.5%，検証データにおいては 17.1%，評価データにおいては 17.8%である．

2) https://www.kaggle.com/competitions/hatenorm23/overview

**表 1**　プロンプトと入出力の例

| | |
|---|---|
| System Prompt | You are a strict hate speech detector.<br>Hate speech is defined as the public expression of prejudice, hostility, or offensive remarks directed towards specific groups or individuals based on their identity characteristics, such as race, ethnicity, gender, or religious beliefs.<br>The input is a list of words from the sentence. Your task is to assign each word a tag of either 1 or 0. Assign 1 to words that appear to be hate speech, and 0 to all other words.<br>Format the result in JSON format, with the key "result" and the value as a list of dictionaries.<br>Output should be in the following JSON format: "result": ["text": word1, "label": Binary tag, "text": word2, "label": Binary tag, ...] |
| User Input | The number of words in the input sentence is 10.<br>input: ['what', 'with', 'this', 'racist', 'coon', 'nurse', 'a', 'murderess', 'hang', 'her'] |
| Output | {"result": [{'text': 'what', 'label': 0}, {'text': 'with', 'label': 0}, {'text': 'this', 'label': 0}, {'text': 'racist', 'label': 1}, {'text': 'coon', 'label': 1}, {'text': 'nurse', 'label': 0}, {'text': 'a', 'label': 0}, {'text': 'murderess', 'label': 0}, {'text': 'hang', 'label': 0}, {'text': 'her', 'label': 0}]} |
| Predict | [0, 0, 0, 1, 1, 0, 0, 0, 0, 0] |

**表 2**　HateNorm23 のデータ例

| id | token | label |
|---|---|---|
| 0 | Say / it / loud / , / say / it / clear / , / illegal / #immigrants / are / not / welcome / here / @user | O O O O O O O O B I O O O O O |

**表 3**　ベースモデルの実験結果

| Model | Type | Binary F1 | Soft F1 | Hard F1 |
|---|---|---|---|---|
| BERT | Encoder | 0.723 | 0.620 | 0.517 |
| RoBERTa | Encoder | **0.737** | 0.638 | 0.554 |
| hateBERT | Encoder | 0.714 | 0.615 | 0.531 |
| Llama | Decoder | 0.617 | 0.519 | 0.544 |
| Qwen | Decoder | 0.708 | 0.627 | 0.600 |
| Gemma | Decoder | 0.726 | **0.645** | **0.613** |

データセットに含まれるデータの例を表 2 に示す．データは ID と単語ごとに区切られたテキスト，それに対応する BIO ラベルを持っている．二値分類の際は B と I のタグを 1，O のタグを 0 に置き換えることで 2 値のラベルを作成した．

### 4.2　評価指標

モデルの評価には以下の 3 種類の F1 スコアを用いる．

**Binary F1**　各単語に付加された 2 値のラベルから，ラベル 1 に対して F1 スコアを計算する．どれだけ正確に差別的発言に含まれる単語を検出できているかを示す．

**Soft F1**　各単語に付加された 2 値のラベルを BIO の 3 値のラベルに変換し，ラベル B とラベル I に対する F1 スコアの micro 平均をとる [20]．どれだけ正確に差別的発言であるスパンを検出できているかを示す．

**Hard F1**　各単語に付加された 2 値のラベルから，差別的発言であるスパンの開始位置と終了位置を得る．このスパンの開始位置と終了位置が完全に一致したものを正解とし，スパン単位の F1 スコアを計算する．Soft F1 よりも厳格にスパンの検出精度を示す．

### 4.3　モデル

本研究で使用するベースモデルはすべて Hugging Face[3] に公開されているものである．エンコーダ型の言語モデルとして bert-base-uncased（BERT），roberta-large（RoBERTa），GroNLP/hateBERT（hateBERT）[21]，デコーダ型の言語モデルとして Llama-3.1-8B-Instruct（Llama）[22]，Qwen3Guard-Gen-8B（Qwen）[23]，gemma-2-9b-it（Gemma）を使用した．

訓練データセットを用いて SFT を行い，これらベースモデルのファインチューニングを行った．学習率の初期値はエンコーダ型の言語モデルで 1e-6，デコーダ型の言語モデルで 5e-5 に設定し，エポック数は 20 とした．このとき，検証データセットにおける損失関数の値が最も小さくなる段階のものを最終的なファインチューニング済みモデルとして扱った．

## 5　実験結果

### 5.1　ベースモデルの性能

実験結果を表 3 に示す．ファインチューニング済みのベースモデルにおいては，Gemma と RoBERTa が高い性能を示した．Binary F1 では

3）https://huggingface.co/

表 4　アンサンブルモデルの実験結果

| Metrics_Top-k | Models | Binary F1 | Soft F1 | Hard F1 |
|---|---|---|---|---|
| Hard_3 | Gemma, Qwen, RoBERTa | 0.744 | **0.662** | **0.618** |
| Binary_3 | RoBERTa, Gemma, BERT | 0.746 | 0.650 | 0.581 |
| Hard_4 | Gemma, Qwen, RoBERTa, Llama | 0.744 | **0.662** | **0.618** |
| Soft_4 | Gemma, RoBERTa, Qwen, BERT | 0.745 | 0.660 | 0.617 |
| Binary_4 | RoBERTa, Gemma, BERT, hateBERT | 0.744 | 0.650 | 0.590 |
| Hard_5 | Gemma, Qwen, RoBERTa, Llama, hateBERT | 0.731 | 0.644 | 0.609 |
| Soft_5 | Gemma, RoBERTa, Qwen, BERT, hateBERT | 0.738 | 0.651 | 0.611 |
| All | Gemma, Qwen, RoBERTa, Llama, hateBERT, BERT | **0.750** | 0.658 | 0.597 |

表 5　モデルの予測例

| | |
|---|---|
| True | **Muslims can rape white girls** without consequences, but if a **white man did this to a muslim girl,** he would be in jail. |
| Gemma | **Muslims can rape white girls** without consequences, but if a **white man did this to a muslim girl,** he would be in jail. |
| Qwen | **Muslims can rape white girls** without consequences, but if a white man did this to a muslim girl, he would be in jail. |
| RoBERTa | **Muslims can rape white girls** without consequences, but if a **white man did** this **to a muslim girl,** he would be in jail. |
| Hard_3 | **Muslims can rape white girls** without consequences, but if a **white man did** this **to a muslim girl,** he would be in jail. |

RoBERTa，Gemma，BERT の順に高くなった．Soft F1 では Gemma，RoBERTa，Qwen の順に，Hard F1 では Gemma，Qwen，RoBERTa の順に高くなった．

この結果からエンコーダ型の言語モデルはデコーダ型の言語モデルに比べて，単語単位で分類する能力に長けていると言える．またデコーダ型の言語モデルはエンコーダ型の言語モデルに比べて，スパンを正確に抽出する能力に長けていると言える．

### 5.2　エンコーダ LM・デコーダ LM の多数決による性能

Binary F1 と Soft F1，Hard F1 の値を元にモデルを組み合わせた．組み合わせは，上位 3 つを選択した 2 通りと，上位 4 つを選択した 3 通り，上位 5 つを選択した 2 通り，6 つすべてのモデルを選択した 1 通りの合計 8 通りである．ベースモデルでの実験結果を元に，各 F1 スコアが高い順に $k$ 個のモデルを選択して多数決を実施した．

実験結果を表 4 に示す．Binary F1 と Soft F1，Hard F1 の 3 種類全ての評価指標において，Hard_3 と Hard_4，Soft_4 の 3 種類がすべてのファインチューニング済みベースモデルの結果を上回った．また Soft F1 と Hard F1 においては，Hard_3 と Hard_4 が，他のアンサンブルモデルを合わせたすべてのモデルの結果を上回った．この 2 つの組み合わせは同じ F1 スコアを持っており，組み合わせるモデルの数が少ないほど計算資源の消費が少ないことから，スパン単位の検出において最も優れたモデルは Hard_3（Gemma+Qwen+RoBERTa）であると言える．さらに Binary F1 においては，すべてのファインチューニング済みベースモデルを組み合わせたモデルが他のアンサンブルモデルを合わせたすべてのモデルの結果を上回った．以上の結果から，複数のモデルを組み合わせることで元のモデルよりも高い性能を持たせることが可能であると言える．

Hard_3 のモデルとこれを構成するベースモデルの予測の例を表 5 に示す．True の欄の太字は差別的発言に含まれる単語であり，モデルの欄の太字は正しく差別的発言に含まれると予測できた単語である．波線部は予測ラベルが間違っていた単語である．ベースモデルで間違った予測をしている部分は「this」を除いて，Hard_3 の予測で正しくなっている．これは，多数決による決定が適切に働いた例と言える．しかし「this」については，Gemma が正しく予測しているのに対して他のモデルが間違っているため，Hard_3 の予測は間違ったものになっている．このように，間違った予測に関しても多数派に寄ってしまう点が多数決の限界であると言える．

## 6　おわりに

本研究では，複数のエンコーダ型の言語モデルとデコーダ型の言語モデルを組み合わせることで差別的発言に該当するスパンの検出性能を向上させた．実験結果よりテキストから差別的発言である部分を単語単位で抽出するタスクにおいて，エンコーダ型の言語モデルとデコーダ型の言語モデルを組み合わせることが有効であることが示された．

今後の展望としては，入力文の特徴や各モデルの特性を考慮するなど，ラベル選択の方法を改善したアンサンブル手法を検討したい．また，本研究では一種類のデータセットのみを用いて評価を行ったため，より多くの事例を含むデータセットや差別的発言を含まない事例を含むデータセットを用いた検証を行うことが挙げられる．さらに，使用したデータセットに含まれる差別的発言は明示的なものであるため，本研究で構築したモデルは明示的な差別的発言を検出することに特化している．そのため今回使用したデータセットに加えて，暗黙的な差別的発言が含まれるデータセットを用いた学習と評価を行うことで，多様な差別的発言に対応できる実用的なモデルを構築することが考えられる．

## 謝辞

本研究は，JST 経済安全保障重要技術育成プログラム JPMJKP24C3 の支援を受けたものである．また本研究は，東京科学大学のスーパーコンピュータ TSUBAME4.0 を利用して実施した．

## 参考文献

[1] Jingjie Zeng, Liang Yang, Zekun Wang, Yuanyuan Sun, and Hongfei Lin. Sheep's skin, wolf's deeds: Are LLMs ready for metaphorical implicit hate speech? In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 16657–16677, Vienna, Austria, July 2025. Association for Computational Linguistics.

[2] António Guterres. Remarks at the launch of the united nations strategy and plan of action on hate speech. Speech by the United Nations Secretary-General, June 2019. United Nations Headquarters, New York, 18 June 2019.

[3] Daniel Hickey, Daniel M. T. Fessler, Kristina Lerman, and Keith Burghardt. X under musk's leadership: Substantial hate and no reduction in inauthentic activity. PLOS ONE, Vol. 20, No. 2, pp. 1–24, 02 2025.

[4] Shrey Satapara, Sarah Masud, Hiren Madhu, Md. Aflah Khan, Md. Shad Akhtar, Tanmoy Chakraborty, Sandip Modha, and Thomas Mandl. Overview of the hasoc subtracks at fire 2023: Detection of hate spans and conversational hate-speech. In Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '23, p. 10–12, New York, NY, USA, 2024. Association for Computing Machinery.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Computation, Vol. 9, No. 8, pp. 1735–1780, 11 1997.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, Vol. abs/1810.04805, , 2018.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[9] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In Proceedings of the NAACL Student Research Workshop, pp. 88–93, San Diego, California, June 2016. Association for Computational Linguistics.

[10] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[11] John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. SemEval-2021 task 5: Toxic spans detection. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 59–69, Online, August 2021. Association for Computational Linguistics.

[12] Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. Automated hate speech detection and span extraction in underground hacking and extremist forums. Natural Language Engineering, Vol. 29, No. 5, pp. 1247–1274, 2023.

[13] Thomas G. Dietterich. Ensemble methods in machine learning. In Multiple Classifier Systems, pp. 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[14] 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. CoRR, Vol. abs/1907.11692, , 2019.

[15] Shahriar Farhan Karim, Anower Sha Shajalal Kashmary, and Hasan Murad. CUET_Blitz_Aces@LT-EDI-2025: Leveraging transformer ensembles and majority voting for hate speech detection. In Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pp. 133–139, Naples, Italy, September 2025. Unior Press.

[16] Ganesh Sundhar S, Durai Singh K, Gnanasabesan G, Hari Krishnan N, and Mc Dhanush. Wise@LT-EDI-2025: Combining classical and neural representations with multi-scale ensemble learning for code-mixed hate speech detection. In Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion, pp. 54–62, Naples, Italy, September 2025. Unior Press.

[17] Gemma Team. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118, 2024.

[18] Jebish Purbey, Siddartha Pullakhandam, Kanwal Mehreen, Muhammad Arham, Drishti Sharma, Ashay Srivastava, and Ram Mohan Rao Kadiyala. 1-800-SHARED-TASKS@NLU of Devanagari script languages 2025: Detection of language, hate speech, and targets using LLMs. In Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pp. 223–235, Abu Dhabi, UAE, January 2025. International Committee on Computational Linguistics.

[19] Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. Proactively reducing the hate intensity of online posts via hate speech normalization. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, p. 3524–3534, New York, NY, USA, 2022. Association for Computing Machinery.

[20] Yan Hu, Vipina K Keloth, Kalpana Raja, Yong Chen, and Hua Xu. Towards precise pico extraction from abstracts of randomized controlled trials using a section-specific learning approach. Bioinformatics, Vol. 39, No. 9, p. btad542, 09 2023.

[21] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), pp. 17–25, Online, August 2021. Association for Computational Linguistics.

[22] Llama Team. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[23] Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3guard technical report. arXiv preprint arXiv:2510.14276, 2025.