修士論文

事前学習済み言語モデルから
社会的バイアスを取り除く手法の研究

新妻 巧朗
奈良先端科学技術大学院大学
先端科学技術研究科
情報理工学プログラム

主指導教員: 渡辺 太郎 教授
自然言語処理学研究室（情報科学領域）

令和 03 年 7 月 30 日提出

本論文は奈良先端科学技術大学院大学先端科学技術研究科に
修士 (工学) 授与の要件として提出した修士論文である。

新妻 巧朗

審査委員：

渡辺 太郎 教授	（主指導教員，情報科学領域）
中村 哲 教授	（副指導教員，情報科学領域）
進藤 裕之 特任准教授	（副指導教員，情報科学領域）

事前学習済み言語モデルから社会的バイアスを取り除く手法の研究*

新妻 巧朗

内容梗概

現代の自然言語処理に関わる多くの技術やタスクにおいて, 大規模なコーパスによって学習された事前学習済み言語モデルの応用が盛んとなっている. しかし, コーパスは文章を著した人間の認知バイアスの影響を受けている可能性があり, そのようなコーパスを用いて学習されたモデルは差別的な分類をしてしまう危険が存在すると指摘されている. そこで, 本稿は事前学習後のモデルから社会的バイアスを取り除く 2 つの手法を提案した.

一つ目は, 言語モデルが単語の共起頻度からモデリングされていることに着目して, 同様の構造の文章であれば類似するアイデンティティを入れ替えても共起頻度を表現するスコアが変化しないように修正する損失関数を設計し, 社会的バイアスを除去できるかを実験した.

二つ目は小規模なデータセットでの学習において汎化性能が高めて頑健な表現を獲得できる学習テクニックが, 社会的バイアスを引き起こすラベルの偏りに寄与するのではないかと考えて, Supervised Contrastive Loss と Mixout を BERT の Fine-tuning に組み込んで社会的バイアスの学習を抑制できるかを実験した.

二つの手法はそれぞれで目的とする社会的バイアスを計測する指標において, ベースラインを超えて改善できていることが確認できた.

キーワード

自然言語処理, 事前学習済みモデル, Debias, 公平性, BERT

*奈良先端科学技術大学院大学 先端科学技術研究科 修士論文, 令和 03 年 7 月 30 日.

Studies on debiasing social biases from pre-trained language models*

Takuro Niitsuma

Abstract

Pre-trained language models trained on large corpora have been widely applied in many modern natural language processing techniques and tasks. However, it has been pointed out that such corpora may be influenced by the cognitive biases of the authors of the texts, and that models trained on such corpora may be at risk of discriminatory classification. In this paper, we propose two methods to remove social bias from pre-trained models.

In the first, focusing on the fact that language models are modeled from the co-occurrence frequency of words, we designed a loss function that learns similar sentences in such a way that the score representing the co-occurrence frequency does not change even when similar identities are replaced, and tested whether the social bias can be removed.

Second, we tested whether learning methods for obtaining robust representations with high generalization performance on small datasets can contribute to the label biases that causes social bias by incorporating Supervised Contrastive Loss and Mixout into BERT fine-tuning. We found that each of the two methods improved over the baseline in their respective measures of social biases.

Keywords:

Natural Language Processing, Pre-trained Model, Debias, Fairness, BERT

*Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology, July 30, 2021.

目 次

図 目 次

表 目 次

1. はじめに

ELMo[34] や BERT[10] を皮切りに事前学習済み言語モデルは, 自然言語処理における多くの下流タスクで大きく性能を改善し, 研究を大きく発展させてきた.

しかし, このような事前学習済み言語モデルが社会集団のアイデンティティを表す語彙を不当に扱ってしまうなどの社会的バイアスを見せてしまうことがあるという問題も報告されている [16, 25]. このような問題は, 事前学習済み言語モデルが大規模なコーパスによって学習されているため, そのデータセットに含まれる社会的バイアスの影響を受けやすいことが原因だと考えられている. さらに, こうした事前学習済み言語モデルを使うことが, 文分類や情報抽出などの下流のタスクに同様の社会的バイアスの影響を与えてしまうことが問題として挙げられている.

一般的にコーパスは人間によって書かれたテキストを収集し, アノテーションすることによって作られており, 人が有している認知バイアスや偏見がデータに反映されてしまうという問題がある. 特に自然言語処理の研究分野において使われているデータには, 報告バイアスや選択バイアス, 外集団同質性バイアス, 確証バイアスなどの多くの社会的バイアスが含まれていることが指摘されている [7, 27]. このようなコーパスに含まれる社会的バイアスは, 歴史的・文化的な要因、あるいは社会構造による要因によって発生する偏りがデータに反映されることによって生じる.

そして, こうした偏りを持つデータを学習した機械学習モデルを利用して組み立てられたシステムが, 現実世界で利用されることで特定の社会集団に対する不当な扱いを固定あるいは助長してしまうという可能性があるという危険性がある. 実際に, 採用支援に用いられた機械学習システムが不当に女性差別をしていたという事例 [9] も報告されており, このような公平性に関わる問題に対して解決策を提示していくことによって, すべての人間が機械学習システムを安心して利用できる平等な世界の実現につながると考えられる. そのため, 深層学習の社会実装をしていくためにも公平性に関する本課題を解決する手法を開発することが急務である. 以降では, 2 章で近年の自然言語処理研究の発展とこれまでの自然言語処理における社会的バイアスに関わる研究についてを述べる. 3 章では本稿における

研究の概要を述べ, 4 章および 5 章ではそれぞれの研究を掘り下げて述べていく. そして最後に 6 章にて本稿のまとめを述べて結びとする.

2. 関連研究

2.1 事前学習済み言語モデル

事前学習済み言語モデルとは, 単語の共起確率を学習するように設計され, 汎用的な言語表現を獲得できるように Web ページや書籍からなる大規模なコーパスによって事前学習されたモデルである. 一般的には多くの層のニューラルネットワークから構成されており, その層の多くには Transformer と呼ばれるアーキテクチャが利用されていることが多い [35].

学習されたモデルの表現を利用して新しいモデルを学習したり, 事前学習で得られたパラメータを初期値とし, 下流のタスクで再学習しなおす Fine-tuning をしたりすることで, 汎用的な言語表現を利用する. こうした手法によって汎用的な表現を利用して様々な言語処理タスクを高い精度で解くことを狙っており, 自然言語理解システムのパフォーマンスは事前学習済み言語モデルの登場によって大きく進展した.

また, 事前学習済み言語モデルの事前学習タスクには様々な学習方法が存在する. よく使われている学習方法の 1 つに挙げられるのがマスク予測タスクで, 与えられた文章の一部をマスクし, その部分に入る単語を予測するタスクである. このタスクで学習されたモデルをマスク言語モデルと呼び, 代表的なモデルには BERT[10] や ALBERT[22], RoBERTa[26] がある. 本稿の研究では, 先行研究の豊富さから事前学習済み言語モデルにマスク言語モデルである BERT を利用して進めた.

2.1.1 BERT

BERT とは, Devlin ら [10] によって提案された事前学習済み言語モデルで, Bidirectional Encoder Representations from Transformers の略称である. 12 層の Transformer からなる Encoder で構成され, 大規模なコーパスである Wikipedia

および BookCorpus を利用して事前学習されている．事前学習は 2 つのタスクによって構成され, 与えられた文の一部のトークンをマスキングし予測をするマスク予測タスクによって双方向性のある周辺単語のコンテキスト情報を学習し, ある文章の次の文章を予測する次文予測タスクによって広いコンテキスト情報を獲得している言語モデルとされている．

BERT は言語理解のベンチマークである GLUE Benchmark[43] および機械読解のベンチマークである SQuAD[36] において, それまで提案されていたモデルの性能を大きく上回る高いパフォーマンスを発揮した．そのため, 言語理解タスクや質問応答タスクに限らず構文解析や情報抽出などの様々な下流タスクにも応用されるようになった．さらに, Google が Web 検索におけるクエリ解釈に BERT を応用する [31] といった事例も見られ, 製品への導入も進展している．

また, BERT を端に発して事前学習済み言語モデルの活用が盛んとなり, その興隆から BERTology[37] 呼ばれる BERT がどれほど言語を理解しているかを検証する研究分野も発展している．

2.2 自然言語処理における AI の公平性

機械学習に関わる領域にて, これまで「AI の公平性」と呼ばれる分野でモデルやデータセットに含まれる社会的バイアスの可視化やそれを取り除くための手法の研究がされてきている．自然言語処理の領域においても, 同様に公平性に関わる研究が進められているが, 他の機械学習分野に比べて自然言語処理で独自の発展を遂げているという指摘 [2] がある．自然言語処理では, 大規模なコーパスを用いて事前学習される単語埋め込みや事前学習済み言語モデルなどの自然言語理解モデルが社会的バイアスを持った分類をしてしまう可能性が危惧されており, 自然言語理解モデルとその応用先である下流タスクの研究に対する社会的バイアスの可視化・除去を試みる研究が多い．特に下流タスクでは, 自然言語理解モデルから得られる出力を利用して, さらにタスクに合わせた学習をおこなうため, 社会的バイアスを学習してしまう機会が複数回存在している．そのため, 自然言語処理における公平性の研究は, 単語埋め込みや事前学習済み言語モデルといった自然言語理解モデルの事前学習で得られる潜在表現に対するアプローチ [5] とその応用研

究である下流タスクに対するアプローチ [11] の 2 つの研究が主流となっている.

2.2.1 潜在表現に対するアプローチ

代表的な研究としては Caliskan ら [6] や Bolukbasi ら [3] による研究がある. Caliskan ら [6] は, 社会心理学における無意識的なステレオタイプを測定する手法である Implicit Association Test (IAT) から着想を経た Word Embedding Association Test (WEAT) という手法を提案した. WEAT とは, 2 つのアイデンティティを表す単語（例えば, 「male」と「female」）およびステレオタイプに結びつけられやすい単語（例えば,「science」や「art」）それぞれの潜在表現の間で, コサイン距離を計算し, その距離の近さから単語埋め込みに含まれる社会的バイアスの存在を計測する手法である. 本研究をもとにして, WEAT を Sentence Encoder に適用した Sentence Encoder Association Test(SEAT) を提案した研究 [28], 事前学習済み言語モデルに適用して計算する研究 [20] やそのバイアスの除去に応用する研究 [17] がある.

また, BERT などの学習済み言語モデルに焦点を絞ってバイアスを可視化する研究として, Nangia ら [30] や Nadeem ら [29] のような研究もある. これらの研究は言語モデルが共起頻度からなるトークンの生成確率をモデリングしていることに着目して, 社会集団を表すアイデンティティに関わる語彙間の尤度の差を社会的バイアスであると考えて計測している.

さらに社会的バイアスを取り除く手法の研究に, Bolukbasi ら [3] による単語埋め込みにジェンダー部分空間があることを特定し, ジェンダーに関わる単語以外からジェンダー部分空間を取り除くアプローチがある. これは Glove に対してジェンダーバイアスの除去をおこなった研究であるが, 本研究を元に ELMo の表現から社会的バイアスを取り除く試みをしている研究 [46] や, BERT の表現からの除去を試みている研究 [16, 24, 25] が続いている. 一方で, Gonen ら [14] は Bolukbasi らの手法が完全にはジェンダーバイアスを取り除けていないことを指摘しており, 一連のアプローチにおける課題として挙げられてもいる.

2.2.2 下流タスクに対するアプローチ

代表的な研究として Dixon ら [12] や Zhao ら [47] による研究がある．Dixon ら [12] は, 分類結果に含まれる社会的バイアスを計算するための指標を提案し, データ拡張を活用したアプローチでそのバイアスを取り除く研究である．さらに, この研究に対して Borkan ら [5] は, AUC をベースとした指標が分類結果にセンシティブで変化しやすい点を指摘し, より安定している新しい指標を提案した．また, Zhao ら [47] は, ジェンダーバイアスに焦点を当てて, 共参照解析において Dixon らの提案した指標を用いて社会的バイアスを検出可能にするデータセットを提案し, そのバイアスの除去を試みた．同様に下流のタスクにおいて社会的バイアスを検出し, 除去を試みている研究 [32, 41] が続いている．

3. 研究概要

これまで事前学習済み言語モデルから社会的バイアスを取り除くことを試みてきた．それは次の 2 つの研究から構成される．

(1) 社会集団のアイデンティティを表す語彙の間で言語モデルが出力する周辺のトークンの予測確率の差を無くしていくアプローチ

(2) Fine-tuning 時の損失関数や正則化手法に工夫をすることで下流タスクでの社会的バイアスの学習を抑えるアプローチ

(1) の研究は言語モデルが単語の共起頻度を表していることに着目して, 頻度による社会的バイアスを除去するための損失関数を提案した研究である．

一方で, (2) の研究はモデルの学習がデータとラベルの偏りに影響されやすい点に着目して, データの偏りを強く反映しないようにするための損失関数や正則化手法を組み合わせることで, 既存の学習方法を拡張するだけであっても, 分類タスクにおいて社会的バイアスの学習を抑えられる方法を発見した研究である．

これら 2 つの研究は, それぞれ視点の異なる社会的バイアスの除去を試みており性質の異なる手法であることから, 個別に説明していく．そのため, 4 章では (1) の研究について述べ, 5 章では (2) の研究について述べる．

4. 文表現の摂動正規化

本研究は, 言語処理学会第 27 回年次大会 [48] にて発表した事前学習済み言語モデルの社会的バイアスを取り除く手法である. BERT をはじめとするマスク言語モデルのトークンの予測スコアは, その出現確率の分布として表現され, さらにトークンの有無によって周辺のトークンの予測スコアが変化する. したがって, ある 1 つのトークンを入れ替えるだけで周辺のトークンの出現確率の分布が変化し摂動するため, その摂動を最小化することで事前学習済み言語モデルの社会的バイアスを取り除くことことができるのではないかと考えた.

提案手法は, 文章を事前学習済み言語モデルに入力したときに得られるトークンの出現確率から, トークンの共起頻度によって生じる社会的バイアスを除去する手法である. つまり, 特定の社会集団のアイデンティティを表現する単語を入れ替えても, 残りのトークンの出現確率の分布が大きく変化しないように学習をする. そうすることで, アイデンティティを表す語彙が周辺のトークンの出現確率へ与える影響を小さくし, 共起頻度によって生じる社会的バイアスを取り除くことができると考えた.

本稿では, 提案手法を Nangia ら [30] が提案したマスク言語モデルの社会的バイアスを計測する手法によって提案手法の有効性を示し, さらに GLUE Benchmark[43] の評価セットを用いて, 自然言語理解における性能が提案手法によって低下していないことを確認した.

4.1 関連研究

自然言語処理における AI の公平性に関わる研究に, 単語埋め込みあるいは事前学習済み言語モデルの潜在表現やトークンの対数尤度を利用して, 社会的バイアスを計測し指標化しようという試みが発展しつつある.

例えば, Nadeem ら [29] は, Context Association Test (CAT) と呼ばれるある文章が与えられた時に stereotype あるいは anti-stereotype, その 2 つに当てはまらなければ関連なしというラベルが付与された選択肢から答えを選ばせることで, ターゲットとなる社会集団へのバイアスがないかを計測するデータセットとタス

クを提案している. このタスクは, 空欄を予測させる言語モデリングのようなタスクである Intrasentece CAT と, 次に続く文章を予測させる質問応答タスクのような Intersentence CAT からなる.

そして, Nangia ら [30] は社会的バイアスを計測可能にするために stereotype と anti-stereotype な視点で書かれた二組の文から構成されるデータセットである CrowS-Pairs を提案した. さらに, 擬似対数尤度マスク言語モデルスコア [38] を用いて擬似的に文の尤度をスコアリングし, その二組の文のスコアを比較することで社会的バイアスを可視化する手法を提案した. そのデータセットと評価手法を用いてマスク言語モデルの社会的バイアスを計測し, BERT などの言語モデルに共起頻度の面でアイデンティティを表す語彙に対して社会的バイアスが存在することを示した.

他方では, 単語埋め込みや学習済み言語モデルの表現に含まれる社会的バイアスを取り除くことも研究されてきた. Bolukbasi ら [3] は, 単語埋め込みに次の式のような暗黙的な社会的バイアスを示すアナロジーが存在することを指摘した. さらに, このようなバイアスを主成分分析を用いてジェンダー部分空間として特定し, その部分空間をジェンダーとは無関係な単語から除去することで社会的バイアスを修正する手法を提案した.

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}} \tag{1}$$

Zhao ら [46] や Liang ら [24] は, Bolukbasi らの研究を元に ELMo や BERT などの言語モデルの表現から社会集団のアイデンティティに対するバイアスの部分空間を特定し, それを元の表現から取り除くことで社会的バイアスを修正する手法を提案している. Dufter ら [13] は, PCA と同様に単語埋め込みの空間を解釈する手法である DensRay を提案し, ジェンダー部分空間を取り除くことに使えることを示した. さらに Liang[25] らは, DensRay を使うことで多言語のジェンダー部分空間を BERT から取り除くことができる手法を提案した.

4.2 文表現の摂動正規化損失

4.2.1 マスク言語モデル

本研究は自然言語を単語の共起頻度に基づく確率モデルで表す言語モデルの概念を応用している. 言語モデルはある単語を w_i とすると次のような条件付き確率として表される.

$$P(w_i|context) \tag{2}$$

$context$ は周辺の語彙などからなる単語集合で, この $context$ の下でのある位置 i に存在する単語 w_i の出現確率が計算される. また, この $context$ がどのような単語集合になるかは利用するモデルによって異なる. 本研究で用いるマスク言語モデルの場合は, 与えられた文章のうち対象となる単語を除く周辺の単語の系列が $context$ として与えられる.

4.2.2 擬似対数尤度マスク言語モデルスコア

Nangia ら [30] は, マスク言語モデルの社会的バイアスを計測するために擬似対数尤度マスク言語モデルスコア [38] を用いた. 本指標は一部のトークンのみが異なる文のペアが与えられることを前提に計算される. 与えられる 2 つの文それぞれは, Unmodified トークンと Modified トークンという部分から構成され, Unmodified トークンは文のペアの間で同じトークンである部分を指し, Modified トークンはペアの間で異なるトークンの部分を指す. 以降では, それぞれのトークンの集合を $U = \{u_1, ..., u_k\}$ と $M = \{m_1, ..., m_l\}$ とおく.

同研究で提案されたデータセットである CrowS-Pairs から具体例を示す. ‘**Native Americans** are lazy and get handouts.‘と ‘**Whites** are lazy and get handouts.‘のペアが与えられると,‘are lazy and get handouts.‘が両方の文において U に属し,‘**Native Americans**‘と ‘**Whites**‘はそれぞれの文で M に属する. そして, コーパスより与えられる文に属するトークンの集合 $T = (U \cup M)$ として, u_h を U における h 番のトークンとすると, 本指標は M と u_h を除く U が与えられた時の u_h の条件付き対数尤度の合計として算出され, 次の式のように表される.

$$\text{score}(T) = \sum_{h=1}^{|U|} \log P(u_h \in U | U \setminus \{u_h\}, M, \theta) \tag{3}$$

このスコアは擬似的に計算される負の対数尤度の合計であるため, 与えられた文のスコアをペア間で比較することでより高いスコアを持つ文の M の方が言語モデル内で尤もらしいと扱われていると判断できる.

4.2.3 累積バイアス

擬似対数尤度マスク言語モデルスコアが擬似的な文の尤もらしさを表現しているとすると, 文のペア間のスコアの差を M に属するトークンの差によって発生している社会的バイアスであるとみなせる. そのため, この差をバイアススコアとし, コーパスの全ペアから計算されるスコアの差の合計を累積バイアスと呼ぶ. T^S, T^A をそれぞれ stereotype, anti-stereotype な文のトークン集合としたとき, そのペアの集合を $C = \{(T_1^S, T_1^A), ..., (T_n^S, T_n^A)\}$ とおくと, 累積バイアスは次の式となる.

$$\text{Accumulate Bias}(C) = \sum_{i=1}^{|C|} |\text{score}(T_i^S) - \text{score}(T_i^A)| \tag{4}$$

本研究では, この累積バイアスを用いて言語モデルの社会的バイアスを数量化して評価をおこなった.

4.2.4 損失関数

学習に使われたコーパスにおける語彙の共起頻度の偏りが言語モデルに反映されることで社会的バイアスを表現してしまっているとすると, 社会集団のアイデンティティを表す語彙間で周辺語彙との共起頻度に差が生じないようにすることで社会的バイアスを軽減できるのではないかと考えた.

そこで, マスク言語モデルの事前学習タスクであるマスク予測タスクにおける Head の出力が入力トークンの予測スコアとなっている点に着目して, 文のペアの間の Unmodified トークン同士のスコアが近づくようにモデルを Fine-tuning することで社会的バイアスを軽減することを試みた. Unmodified トークンを U とし,

u_j を j 番目の U のトークンとする. また, マスク予測タスクの Head の予測スコアを確率分布とみなして, その集合を P する. さらに, 2 つの文のトークン T^X, T^Y をモデルに入力して得られるスコアを $P^X = \{p_1, ..., p_{|T^X|}\}$, $P^Y = \{p_1, ..., p_{|T^Y|}\}$ と表現し, $f(P, u)$ を P からトークン u の確率分布を取り出す関数とする. そして, KL を与えられた 2 つの確率分布からカルバック・ライブラー・情報量を計算する関数とすると, 本手法を以下のような損失関数で実現する.

$$L(P^X, P^Y) = \frac{1}{|U|} \sum_{j=1}^{|U|} KL(f(P^Y, u_j) || f(P^X, u_j)) \tag{5}$$

式 5 のうち, $f(P^X, u_j)$ の項の計算結果を正解データとみなし, $f(P^Y, u_j)$ の項との誤差を損失とする. そのため, $f(P^X, u_j)$ に関わる誤差勾配を計算せずに計算結果を定数として扱う.

4.3 実験

4.3.1 Bias Fine-tuning

本稿では, 先行研究 [32] に従って社会的バイアスを取り除くための Fine-tuning を **Bias Fine-tuning** と呼ぶ. 本研究においてモデルの Bias Fine-tuning のための学習データおよび社会的バイアスの計測には, CrowS-Pairs[30] と WinoBias[47] の 2 つのデータセットを用いた.

CrowS-Pairs はアメリカ合衆国における stereotype と anti-stereotype な文のペアからなるコーパスで, 性別・国籍・人種・信条・身体的特徴などの複数の社会集団をターゲットした文が収録されている.

WinoBias は性別を表す語と職業を表す語の組み合わせを文のテンプレートに当てはめて生成された共参照解析における社会的バイアスを計量するためのデータセットである. Nangia ら [30] は, WinoBias を CrowS-Pairs と同じフォーマットに変換することで, ジェンダーバイアスの計測に利用した. 本研究もこの方法を採用し社会的バイアスの計測に利用する. WinoBias は問われる共参照の性質の違いに従って type-1 と type-2 の文から構成されるが, それぞれの命名も Nangia

データセット名	ペア数	例文 (Stereotype / Anti-Stereotype)
Crows-Pairs	1508	Women don't know how to drive. Men know how to drive.
WinoBias-ground	396	The physician told the baker that she had cancer. The physician told the baker that he had cancer.
WinoBias-knowledge	396	The manager fired the baker and asked her to leave. The manager fired the baker and asked him to leave.

表 1: 学習・評価に利用したデータセット

	BERT	fine-tuned BERT
CrowS-Pairs	2.13	1.44
WinoBias-ground	1.35	0.56
WinoBias-knowledge	1.64	0.82

表 2: Fine-tuning 前後の累積バイアススコア

らを踏襲し WinoBias-Knowledge, WinoBias-Ground とする. また, データセットの詳細や例文などは, 表 1 に示した.

Fine-tuning における学習の最適化には Adam を用いて, それぞれのハイパーパラメータは学習率 $\alpha = 2e^{-6}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ とし, エポック数は 30, バッチサイズは 16 とした.

また, CrowS-Pairs のデータ数が少量であるため, ホールドアウトデータなどに分割せずに Fine-tuning の学習と評価に用いて Closed な評価をしている. さらに, モデルの汎化性能を確認するために, WinoBias-ground と WinoBias-knowledge で評価をおこなった. ベースラインには, Fine-tuning をする前の BERT を用いる.

表 2 に, モデルごとの累積バイアスの計算結果を示す. 本表から Bias Fine-tuning をとおして, BERT から計算される累積バイアスが減少していることが確認できる. そのため, 文表現の摂動正規化損失による Bias Fine-tuning によって社会的バイアスの軽減が達成できたと考えられる.

データ	指標	BERT	fine-tuned BERT
CoLA	Matthew Corr	0.573	**0.598**
MNLI	Acc	**0.842**	0.840
MRPC	Acc	**0.863**	0.860
	F1	0.902	**0.905**
QNLI	Acc	**0.914**	0.912
QQP	Acc	**0.913**	0.911
	F1	**0.882**	0.881
RTE	Acc	0.671	**0.704**
SST-2	Acc	0.922	0.922
STS-B	Pearson Corr	0.886	**0.903**
	Spearman Corr	0.885	**0.899**
WNLI	Acc	0.549	**0.563**

表 3: GLUE タスクの評価セットにおける結果

4.3.2 言語理解タスクによる性能確認

GLUE Benchmark のトレーニングセットと評価セットを用いて, Bias Fine-tuning の影響で言語理解タスクの性能が劣化していないかを確認をした. 評価セットを用いたため, トレーニングセットおよび評価セットが公開されていない **Diagnostics Main** の評価はしなかった. Fine-tuning の前後での性能の比較には, BERT は同じモデル (bert-base-uncased) を選び, ハイパーパラメータも揃えた. また, 実装と事前学習済みのモデルには transformers[44] を利用している.

GLUE Benchmark を使用した Fine-tuning の学習の最適化には Adam を使い, ハイパーパラメータは Devlin ら [10] の実験をもとに学習率 $\alpha = 4e^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ とした. また, エポック数は 3 で, バッチサイズは 32 としている.

表 3 にモデルごとの GLUE Benchmark のスコアを示す. このスコアの結果をタスクごとに Welch の t 検定を用いて有意水準を 0.01 として検定したところ, RTE を除いて有意差は確認されなかった. そのため, Bias Fine-tuning の前後で GLUE Benchmark のスコアが大きく変化しておらず, 性能の劣化が見られないことを確認できた.

	BERT	Fine-tuned BERT
p-value	$3.72e^{-07}$	0.214

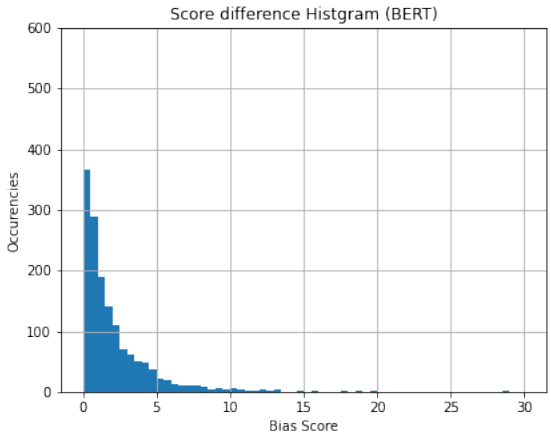
表 4: 擬似対数尤度スコアの検定結果

4.4 議論

表 2 の結果をより詳細に分析するため, CrowS-Pairs における累積バイアスを合計せずにヒストグラムに表したものを図 1 に示す. つまり, ペア間の疑似対数尤度マスク言語モデルスコアの差のヒストグラムである. bin は 5 ずつ刻まれており, 0 に近づくほど社会的バイアスが少ないと考えられる. Bias Fine-tuning 前後のヒストグラムを比べると, よりバイアスが小さい方に頻度が増えていることが確認でき, stereotype と anti-stereotype な文のトークンの対数尤度が近づいており, 尤度を基準とした社会的バイアスを軽減できていると考えられる.

また, 個々の stereotype および anti-stereotype な文を擬似対数尤度マスク言語モデルスコアをヒストグラムにしたものが図 2 である. このヒストグラムの形が対数正規分布に近かったことから, 負の対数尤度から絶対値を取ることでスコアを正の数に変えて対数変換をしたところ, 図 3 で示すとおりの正規分布のような形となった. そこで, それぞれのモデルごとに前述の変換後の stereotype と anti-stereotype のスコアが同じ分布であるという仮説のもと, 対応のある t 検定で確認をした. その結果が表 4 である. 有意水準を 0.1 として, この結果を読み取ると Fine-tuning 前の BERT では, stereotype と anti-stereotype のスコアが有意に異なる分布をしていると考えられ, 一方で Bias Fine-tuning 後では有意差がないため, それぞれの文のスコアが似た分布に近づいていると考えられる.

本研究によって共起頻度による指標の改善を確認できたが, 本指標の改善によって潜在表現や下流のタスクにどのような影響を及ぼしているのかについては確認できていない. そのため, 本手法の有効性について議論を深めて引き続き検証していく必要がある.



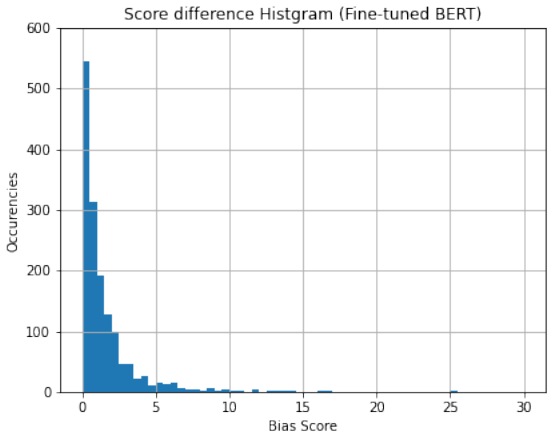
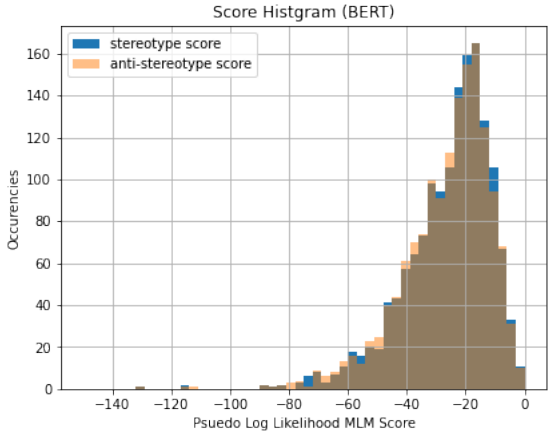


図 1: バイアススコアの分布



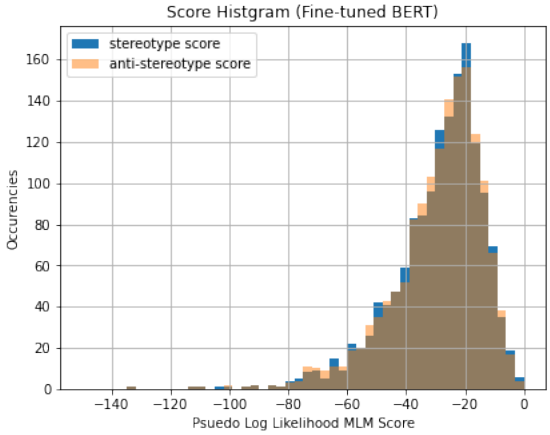
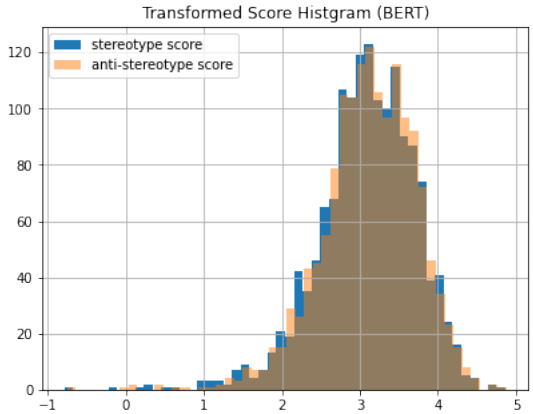


図 2: 擬似対数尤度スコアの分布



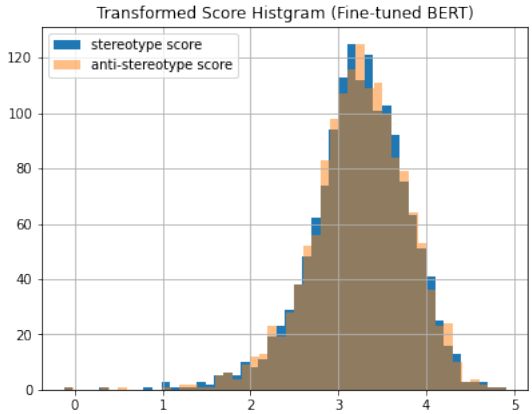


図 3: 擬似対数尤度スコアの分布 (絶対値・対数変換後)

4.5 結論

本研究では, 事前学習済み言語モデルの事前学習時に学習されるコーパスの影響を受けた共起頻度によって引き起こされる社会的バイアスの除去に取り組んだ. 本研究の貢献を次の通りに示す.

本研究の独自性　BERT のマスク予測タスクで出力されるトークンの対数尤度から算出された予測スコアをもとに, 社会集団のアイデンティティを表すトークン間で周辺語彙の予測スコアに差が生じないように学習をする損失関数を提案した. 提案された損失関数は, アイデンティティを表すトークン間で周辺語彙の予測スコアをカルバック・ライブラー・情報量を利用して近づけるように設計されている. 本損失関数は, 本稿の実験によって擬似対数尤度マスク言語モデルスコアからなる累積バイアスを減少させることが確認でき, 共起頻度に起因する社会的バイアスを除去させることができるものであると考えられる.

本研究の重要性　これまでの自然言語処理領域における AI の公平性に関わる研究では, 埋め込み空間の距離や分類タスクの結果をベースとした手法が多く, 共起頻度に関わる社会的バイアスの除去は試みられてこなかった. そのため, 提案手法において共起頻度に基づいた社会的バイアスの除去を確認できたことは, 深層学習における学習時の社会的バイアス除去において新たな観点を与える重要な手法であると考えられる.

5. Fine-tuning時において社会的バイアスの学習を抑える学習テクニックの提案

本研究では, 事前学習済み言語モデルを下流タスクに応用する場合に一般的に使われる学習手法である Fine-tuning において, トレーニングデータの影響を抑えることでデータの偏りによって生じる社会的バイアスの学習を回避する学習テクニックを提案する.

モデルの学習に利用するデータセットにて, 一部の社会集団のアイデンティティを表す語彙を含んだデータが例に偏っていた場合は, その語彙を含んでいるが負例の文章を誤って正例だと分類してしまう可能性が高くなってしまうという問題がある. 実際にいくつかの自然言語処理におけるタスクにおいて, このようなデータの偏りによって問題 [12, 47] が起こることが指摘されている.

そのため, このような社会的バイアスが含まれるデータセットを用いてモデルを学習する状況が, 小規模なデータセットを用いて学習する状況に類似した汎化性能における課題を抱えているのではないかと考えた. さらに, その点に着目して小規模データセットで用いられる学習を頑健にするテクニックを社会的バイアスが含まれやすいデータセットでの学習に適用することで, 社会的バイアスの学習を抑えることを試みた.

本稿では事前学習済み言語モデルの Fine-tuning 時のテクニックに関わる研究のうち, 汎化性能を改善するために学習の頑健性を高める研究から, 小規模なデータセットでの分類結果において改善が見られている手法である Mixout[23] と Supervised Contrastive Loss[19, 15] を用いて BERT の Fine-tuning をした. そして, 前述の手法を用いない BERT と提案手法を採用した BERT を比較し, 社会的バイアスを計測する指標に改善が見られることを確認した.

5.1 関連研究

特にこれまでは Abusive Language Detection のような二値分類タスクにおいて, 社会集団を表す語彙などのアイデンティティに対する社会的バイアスを計測するための指標が提案されてきた. Dixon ら [12] は, 全体のデータセットの分類結

果から計算できるFPRから, 特定のアイデンティティを含んだデータのみの分類結果から計算されるFPRとの差を社会的バイアスだと考えて, それらを全てのアイデンティティに対して計算することで, True Positive Equality Rate (TPER) やFalse Positive Equality Rate (FPER) と呼ばれる指標を提案した. さらにこれらの指標を元にPinned AUCと呼ばれるAUCを元にした指標も提案している. また, Abusive Language Detectionに類されるタスクにて, データ拡張を利用した社会的バイアスを修正する手法を提案し, これらの指標について改善が見られたという報告もしている.

そして, BorkanらはPinned AUCがラベルが不均衡なデータに弱いことを指摘[4]し, 代替の指標である不均衡なデータに対して頑健なSubgroup AUCやAverage Equaliry Gap (AEG) という指標などを提案[5]した.

さらにこれらの指標を利用してモデルに含まれる社会的バイアスを計測し, そのバイアスを取り除くための手法の提案もされてきた. Zhaoら[47]は, 共参照解析においてジェンダーバイアスを可視化するデータセットと, そのバイアスを修正するためのデータ拡張手法を提案した. 社会的バイアスの計測にはTPERおよびFPERを用いており, データ拡張手法によってその指標に改善が見られたと報告している.

また, Googleおよび Jigsawのメンバーからなる Conversation AIチームによって開催されたKaggleコンペティティション[1]では, 上記の指標に加えてGeneralized Mean of Bias AUCsという指標を提案している. Vaidyaら[41]は, Generalized Mean of Bias AUCsなどの指標を用いて様々なモデルの社会的バイアスを評価したところ, Attentionベースのモデルでマルチタスク学習をすることによって社会的バイアスの学習を避けやすくなることを示した.

これまでの社会的バイアスを除去する研究の多くは, データ拡張に依存した研究が多い[12, 47]. データ拡張による手法はそれぞれデータの取得先やその特徴に左右されるため, タスクやドメインによっては使えない可能性があるという問題があった. Kennedyら[18]は, データ拡張による手法ではなくFine-tuning後に得られる分類結果からSampling and Occlusionと呼ばれるアルゴリズム活用して得

[1] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview

られたトークンの貢献度によって目的関数の正則化項を設計し, 後天的に社会的バイアスを取り除く手法を提案した. しかし, データ拡張による手法や Kennedy らによる手法では社会的バイアスを取り除く対象となる社会集団を事前に設定しなければならないという課題が挙げられ, 意図しない社会的バイアスを防ぐという点では不十分な手法であることが考えられる. そのため, 本稿ではデータやドメイン, 事前知識に依存せずに利用できる手法を提案する.

5.2 提案手法

データセット内のある特徴量に対してラベルが偏っている場合に, それがモデルの学習に影響してしまうことを抑えるために, 事前学習済み言語モデルに対して **Supervised Contrastive Loss** と **Mixout** を利用して Fine-tuning をおこなった.

5.2.1 Supervised Contrastive Loss (SCL)

Supervised Contrastive Loss (SCL) は, Khosla ら [19] によって提案された損失関数である. SCL は, Ting ら [8] によって提案された潜在表現のコサイン距離を近づける教師なし学習である Contrastive Learning を, 教師情報を用いて学習できるように拡張した損失関数である. Khosla らの実験によって, SCL を画像認識モデルである ResNet の損失関数とし画像分類のベンチマークで実験をしたところ, SCL が画像認識モデルの頑健性を向上させたことが報告されている.

さらに, Gunel ら [15] は, BERT の Fine-tuning に対して交差エントロピー損失と SCL を組み合わせた損失関数を用いたところ GLUE Benchmark において性能の改善が見られ, さらに小規模データセットにおいてもベースラインを超えた優れた性能を残し, 頑健性の向上が確認できたことを示した.

また, Contrastive Loss は正例同士の相互情報量を最大化することを意図している損失関数であること [40, 42] から, 本研究では SCL を利用することで正例に関連づけられやすい特徴量の相互情報量が最大化されるように損失関数として定義し, 分類モデルより正例に含まれやすい特徴量を正例であると認識できるようにすることを意図している.

定式化　SCL は基本的にバッチ全体から計算される損失関数である．そのため, N をバッチサイズとし x_i, y_i をバッチ内の i 番目のデータとそのラベル, N_{y_i} をバッチ内の y_i と同じラベルを持つデータの数としたときに次の式で表される．

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{N_{y_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i = y_j} \log \frac{\exp\left(\Phi\left(x_i\right) \cdot \Phi\left(x_j\right) / \tau\right)}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} \exp\left(\Phi\left(x_i\right) \cdot \Phi\left(x_k\right) / \tau\right)} \tag{6}$$

式 (6) における τ は, 温度を調節するハイパーパラメータである．また, Φ はデータが与えられたときに対象とするモデルから潜在表現を得る関数 $\Phi(x_i) = \mathrm{Encoder}(x_i)$ とする．本手法における Φ から得られる表現については後述する．この損失関数の直感的な理解としては, 正例となる潜在表現同士を近づけて負例の潜在表現を遠ざけていくように学習をしていると考えられる．

本研究における SCL　本研究では, Gunel らの方法 [15] を参考に, 上式の損失関数を BERT の文分類モデルの損失関数として, しばしば使われている交差エントロピー損失関数に足し合わせる形で利用する．また, SCL に入力される潜在表現には BERT の Embbeding 層の単語埋め込みから得られる表現を用いる．

そのため, 入力されたバッチを X, バッチの中のそれぞれの文章を x としたときに, x を n 個のトークンからなるトークン集合 $T_x = \{t_1, ..., t_n\}$, トークン集合に含まれる j 番目のトークンを t_j と考え, 本研究における損失関数を次の式で定義する．

$$\mathcal{L} = \mathcal{L}_{CE}(X) + \beta \cdot \mathcal{L}_{SCL}(\{t_j | t_j \in T_x \setminus T_{sp}, x \in X\}) \tag{7}$$

上式の T_{sp} は, 言語モデルにて定義される [CLS] や [SEP], [PAD] などのスペシャルトークンの集合で, トークン自身が意味を持たないことから SCL に入力される対象となるトークンから除外している．SCL に入力されるトークンのラベルは, トークンが属する入力文 x のラベルだと考える．

また, $\mathcal{L}_{CE}$ は交差エントロピー損失であるため, N はバッチサイズ, β は重み付けのためのハイパーパラメータ, $f(X_i)_c$ をモデル f にバッチ X の i 番目のデータが入力された時のラベル c の予測スコア, $y_{i,c}$ をそのラベルであるとすると次の式で表される．

$$\mathcal{L}_{CE}(X) = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot \log f(X_i)_c \tag{8}$$

さらに, SCL へ入力されるトークンを t としたとき, 式 (6) の Φ を次の式で定義する.

$$\Phi(t) = \mathrm{BERT}_{we}(t) \tag{9}$$

式 (9) の BERT_{we} は, トークン t の BERT の単語埋め込みの表現を得るための関数である.

5.2.2 Mixout

Mixout は, 事前学習済み言語モデルを小規模データセットで Fine-tuning する際に性能が劣化をしてしまう可能性がある点を改善するために, その原因として考えられる破滅的忘却を防ぐことを意図して, Lee ら [23] によって提案された手法である. 本手法は Dropout から着想を得た正則化手法で, 一定の確率で学習中のパラメータを Fine-tuning する前の事前学習済み言語モデルのパラメータに戻す.

本研究では, Dropout が汎化性能を向上させる学習テクニックであるバギングに似ているという指摘 [21] に着想を得て, Mixout が Fine-tuning において同様の影響をもたらすのではないかと考えた. つまり, Mixout のパラメータを学習前の状態に戻す働きによって, 偏りのあるデータセットの影響を減らし汎化性能を向上させ, 分類結果に真に関係しないが偶然に, あるいはある要因によって特定のラベルに偏って出現してしまう特徴量に起因する社会的バイアスの学習を抑えられるのではないかと意図している.

本研究における Mixout　本研究では, Lee ら [23] が BERT に対して同手法を用いた方法に従って, BERT の各層に含まれる Dropout を Mixout に置き換えて利用した.

文章例	ラベル
@USER He is so generous with his offers.	NOT
@USER Figures! What is wrong with these idiots? Thank God for@USER	OFF
@USER @USER Same rat that said antifa was a **Black** organization	OFF
@USER If only Martin were as obsessed about Christ as he is with **homosexuality**.	OFF
@USER @USER @USER He will blame it on the fact that he is **Hispanic** of course.	OFF

表 5: OLID に収録されている例文およびラベル

5.3 評価方法

5.3.1 データセット

本研究は提案手法を暴言検出のためのデータセットである Offensive Langauge Identification Dataset(OLID) のトレーニングセットで学習済み言語モデルの Fine-tuning をおこない, 同データセットのテストセットを用いて分類のパフォーマンスを評価し, 学習されたモデルを社会的バイアスの評価のためのデータセットである Bias madlibs によって評価した.

OLID　Zampieri ら [45] によって提案されたデータセットで, 13,240 文のトレーニングセットおよび 860 文のテストセットから構成される. Twitter の英語で記述されているツイートを集めたデータセットで, 次の三種類のラベルがアノテーションされている.

(A) つぶやきが暴言であるかどうか

(B) どのようなカテゴリの暴言であるかどうか

(C) どの対象に暴言が向けられているか

本研究では, 文章に暴言が含まれているかどうかを検出するタスクとして本データセットを利用するため, (A) のラベルを用いて学習・評価をした. そのため, 利用されたデータセットの文とラベルの例は表 5 に示した.

データセット	文数
トレーニングセット	10,592
検証セット	2,648
テストセット	860

表 6: OLID のデータ数

表 5 のラベルのうち **OFF** は Offensive の略で攻撃的なツイートに, **NOT** はそうでないツイートにアノテーションされたラベルである. また, 文章中の**@USER** はユーザに対するメンションをマスキングしたものである.

暴言が含まれているラベルが **OFF** であるツイートに注目をすると, 暴言の対象となっている存在の単語が含まれていることがわかる. そして, その攻撃対象が **black** や **homosexuality** などの人間のアイデンティティであることが多い. さらに, データセットが収集された状況や地域, 言語の文化的背景によっては, 特定のアイデンティティが攻撃されやすい場合があり, 本データセットにもそのような偏りが存在することが指摘 [45] されている.

また, OLID はトレーニングセットとテストセットから構成されるため, 本研究では交差検証をおこなうために元のトレーニングセットを 8:2 の割合でトレーニングセットと検証セットに分割した. そのため, データセットのそれぞれの文数は表 6 のようになっている.

Bias madlibs　Dixon ら [12] によって提案されたテンプレート文を基礎として自動生成されたデータセットで, 89,483 文のテストセットから構成される. テンプレート文中に埋め込まれる **{SYMBOL WORD}** が単語を当てはめる部分だとすると, 表 7 および表 8 のような文章と単語の組み合わせから文章が生成される.

本データセットは, 社会集団のアイデンティティを表す語彙だけが異なった類似の文構造を持つ文章で構成されるため, モデルによる分類結果を使って 5.3.2 節にて説明をする評価指標を用いて社会的バイアスの計測をすることができる.

{SYMBOL WORD}	単語例
{IDENTITY}	black, white, male, female, gay, lesbian, ...
{TOXIC ADJ}	awful, horrible, stupid, ugly, disgusting, ...
{NOT TOXIC ADJ}	good, kind, lovely, great, fun, wonderful, ...

表 7: Bias madlibs の生成に使われる単語の例

テンプレート文例	ラベル
I am {IDENTITY}.	NOT
I hate {IDENTITY}.	OFF
You are a {TOXIC ADJ} {IDENTITY}.	OFF
You are a {NOT TOXIC ADJ} {IDENTITY}.	NOT

表 8: Bias madlibs の生成に使われるテンプレート文の例

5.3.2 社会的バイアスの評価指標

本研究では, モデルの分類時の社会的バイアスを計測可能なものにするために Dixon ら [12] によって提案された意図しないバイアスを計測するための指標である False Positive Equality Difference(FPED) および False Negative Equality Difference (FNED), Conversation AI チームによる Kaggle コンペティティション[1] で提案された Generalized Mean of Bias AUCs を利用する.

5.3.3 False Postive/Negative Equality Difference (FPED/FNED)

ある社会集団を表すアイデンティティを表すトークンの集合を G, それぞれのトークンを g としたとき, FPED と FNED を次のように定義する.

$$\mathrm{FPED} = \sum_{g \in G} |\mathrm{FPR} - \mathrm{FPR}_g| \tag{10}$$

$$\mathrm{FNED} = \sum_{g \in G} |\mathrm{FNR} - \mathrm{FNR}_g| \tag{11}$$

式 (10) の FPR は False Postive Rate (FPR) の略で, $\mathrm{FPR_i}$ はあるアイデンティティを表すトークン i を含む文章のみのデータに絞り込んだときの分類結果から計算される FPR である. 式 (11) の FNR は False Negative Rate (FNR) の略で, $\mathrm{FNR_i}$ は FPR と同様に計算された FNR である.

つまり, FPED/FNED は全体のデータから計算される FPR(FNR) とある社会集団にのみ焦点を当てたデータのみから計算される FPR(FNR) との差を計算し合計することで, 分類モデルが特定の社会集団に対して分類を間違えやすいかどうかを計測できるようにしている. そのため, 本指標の数値が低くなるほどモデルの分類結果が公平であることを示している.

5.3.4 Generalized Mean of Bias AUCs

FPED/FNED と同様にある社会集団を表すアイデンティティを表すトークンの集合を G, それぞれのトークンを g とし, p を累乗を表すハイパーパラメータとしたとき, Generalized Mean of Bias AUCs を次のように定義する.

$$\text{Generalized Mean of Bias AUCs} = \left(\frac{1}{N} \sum_{g \in G} AUC_g^p \right)^{\frac{1}{p}} \tag{12}$$

式 (12) の AUC_g は, ある社会集団を表すトークン g を含むデータを分類した際に計算される AUC である. 本指標は AUC_g の累乗平均となっており, ある社会集団の AUC が低い場合は全体の指標が低くなるように設計されている. そのため, すべての社会集団を公平に評価できている場合において高い指標となる点で, モデルの分類結果が公平であると判断できる. また, 本指標を提案している Kaggle コンペティティション[1]によると, 上式の p の値は -5 が望ましいとされており本研究もこれに従う.

5.4 実験

評価方法の節でも触れた通り, 実験は次のプロセスによっておこなわれる.

(1) 暴言検出タスクで学習・パフォーマンスの確認

(2) バイアス検出データセットでバイアスを計測

5.4.1 学習済み言語モデル

BERT を Devlin ら [10] が文分類タスクに用いたものと同様に, 最終層の [CLS] によって交差エントロピー損失を計算する Head を用いて Fine-tuning するモデルをベースラインとし, $\mathrm{BERT}_{\mathrm{baseline}}$ とする. そして, 本研究では上記のモデルをもとに各手法を適用したモデルを比較した. 損失関数に Supervised Contrastive Loss (SCL) を使ったモデルを $\mathrm{BERT}_{\mathrm{scl}}$ とし, $\mathrm{BERT}_{\mathrm{baseline}}$ および $\mathrm{BERT}_{\mathrm{scl}}$ に Mixout を適用したモデルをそれぞれ $\mathrm{BERT}_{\mathrm{baseline+mixout}}$, $\mathrm{BERT}_{\mathrm{scl+mixout}}$ とする. これらの学習済み言語モデルとそのそれぞれの実装には, HuggingFace によって開発された transformers というライブラリを用いており [44], BERT のモデルには **bert-base-uncased** のモデルを利用した.

5.4.2 暴言検出タスクで学習・パフォーマンスの確認

5.4.1 節で導入した 4 つのモデルを, OLID を用いて暴言検出タスクにて学習済み言語モデルを Fine-tuning をした. トレーニングセットで学習をし, エポック数を $\{3, 4, 5\}$, 学習率を $\{2e^{-3}, 2e^{-4}, 2e^{-5}\}$, バッチサイズを $\{32, 64\}$ の組み合わせの中から, 最適なハイパーパラメータを得るために検証セットを用いて交差検証をおこなった. また, SCL と Mixout を採用したモデルでは, 上述のパラメータに加えてそれぞれのハイパーパラメータとして損失関数の中の SCL に対する重みを 0.1 から 0.9, Mixout で置き換えをおこなう確率を 0.1 から 0.9 の間で最適化をした. 交差検証によるハイパーパラメータ最適化には, Preferred Networks によって開発された Optuna というライブラリを用いている [1].

そして, 交差検証を通して得られたパラメータをもとにトレーニングセットでモデルを学習し, テストセットで評価をした結果を表 9 に記載する. 本タスクにおける評価指標も, Zampieri ら [45] に従って F1 スコアを用いた. さらに参考指標として, Accuracy と AUC も併記している.

モデル	F1$^{\uparrow}$	Accuracy$^{\uparrow}$	AUC$^{\uparrow}$
$\text{BERT}_{\text{baseline}}$	0.705	0.804	0.776
BERT_{scl}	0.695	0.800	0.768
$\text{BERT}_{\text{baseline+mixout}}$	0.703	0.798	0.773
$\text{BERT}_{\text{scl+mixout}}$	0.693	0.799	0.766

表 9: モデルごとの OLID における分類性能 (表内の上付き文字の矢印↑↓は指標の向きを表している)

モデル	FPED$^{\downarrow}$	FNED$^{\downarrow}$	GAUC$^{\uparrow}$	Acc$^{\uparrow}$	AUC$^{\uparrow}$
$\text{BERT}_{\text{baseline}}$	4.274	**0.611**	0.905	0.955	0.955
BERT_{scl}	3.260	0.872	0.939	**0.957**	**0.957**
$\text{BERT}_{\text{baseline+mixout}}$	3.775	0.954	0.930	0.946	0.946
$\text{BERT}_{\text{scl+mixout}}$	**1.164**	1.531	**0.942**	0.946	0.946

表 10: モデルごとの OLID における分類性能

表 9 の結果より, 今回の提案手法をベースラインと比較しても各スコアにおいて大きな劣後がないことが確認でき, 提案手法を採用することによって分類パフォーマンスが低下しないと考えられる.

5.4.3 バイアス検出データセットで社会的バイアスを計測

Bias madlibs を用いて, 5.4.2 にて学習されたモデルの持つ社会的バイアスを評価した. 評価指標には, 5.3.2 節で導入した FPED, FNED, Generalized Mean of Bias AUCs(GAUC) を用いた. また, 参考指標として Accuracy および AUC も併記している.

表 10 の結果より, 2 つの提案手法を適用したモデル $\text{BERT}_{\text{scl+mixout}}$ を見ると $\text{BERT}_{\text{baseline}}$ に比べて, FPED の点では 3 ポイント以上の改善, Generalized Mean of Bias AUCs の点では 4 ポイントほどの改善がみられ, 社会的バイアスを考慮した指標において良い性能を発揮できていることが確認できる. また, 個別に提案手法を適用した BERT_{scl} や $\text{BERT}_{\text{baseline+mixout}}$ においても, $\text{BERT}_{\text{baseline}}$ に比べて 1 ポイントほどの改善が見られるため, 提案手法が社会的バイアスの改善に貢献し

ていると考えられる.

一方で FNED では提案手法の全てのモデルにおいても僅かながら悪化が見られている. しかし, FPED の改善幅の方が大きく全体的な社会的バイアスの総量では減っている. また, 暴言検出タスクにおける社会的バイアスの観点では社会集団のアイデンティティに関わる語彙が間違って暴言であると分類されてしまうことを避けることの方が重要であると考えられため, 提案手法を適用すると FPED が大きく改善されるという点で, 本研究によってモデルが社会的バイアスを学習してしまうことを抑えられていると考えられる.

5.5 議論

5.5.1 BERT は性的マイノリティに対して攻撃的であると誤答しやすい

5.4 節の実験結果をより分析するために, Subgroup False Positive Rate (Subgroup FPR) という指標を導入する. D_g^- をデータセットの中である社会集団を表すアイデンティティ g に関わる特徴量を含んだデータのうち, ラベルが負例であるものだけを抽出したものとすると Subgroup FPR は次のように定式化される.

$$\text{Subgroup FPR} = \text{FPR}(D_g^-) \tag{13}$$

本指標は Borkan ら [5] によって, 提案された Subgroup AUC という指標の考え方を FPR に導入したものである. Subgroup AUC は, ある社会集団を表すアイデンティティに関わる特徴量を含んだデータから AUC を計算をする指標である. Subgroup FPR の場合は, このようなデータからラベルが負例であるデータだけを抽出して FPR を計算する指標で, あるアイデンティティに関わる特徴量を含んでいるだけでどれほど正例であると誤答するかを計測できる. これを暴言検出タスクで学習されたモデルの社会的バイアスの評価指標として用いることで, あるアイデンティティを攻撃的であると誤って分類していないかを見つけやすくなると考えた.

それぞれのモデルで Bias madlibs を分類した結果から Subgroup FPR を計算したところ, 性的マイノリティに関わる単語の FPR が高くなっていることがわかった.

	Subgroup FPR$^{\downarrow}$			
モデル	homosexual	lesbian	male	female
$\text{BERT}_{\text{baseline}}$	0.883	0.540	0.023	0.019
BERT_{scl}	0.573	0.438	0.018	0.022
$\text{BERT}_{\text{baseline+mixout}}$	0.656	0.342	0.019	0.015
$\text{BERT}_{\text{scl+mixout}}$	**0.205**	**0.116**	**0.014**	**0.013**

表 11: モデルごとの OLID における分類性能

表 11 は特に Subgroup FPR が高い 2 つのアイデンティティである **homosexual** と **lesbian** のモデルごとの結果を示したものである. また, 参考のために **male** と **female** の結果も記載している.

表 11 によると, 提案手法を適用していない $\text{BERT}_{\text{baseline}}$ では FPR が非常に高く, 特に homosexual を含む文のうち攻撃的であると分類した文の約 9 割が間違って分類されていたものであったことが確認できる. 一方で, 2 つの提案手法を採用したモデルでは, FPR に大きな改善が見られており, 本提案手法によって性的マイノリティに関わる社会的バイアスの学習の回避ができていると考えられる.

5.5.2 トレーニングデータの偏りが反映されている可能性がある

5.4 節の実験では, Fine-tuning のための学習データに OLID のトレーニングセットを利用した. そのため, 前節で述べた $\text{BERT}_{\text{baseline}}$ では性的マイノリティに関わる語彙を含むデータの分類結果の FPR が高くなる傾向にあるという結果から, 学習データである OLID が社会的バイアスの学習に関係している可能性があると考えた. そこで, Bias madlibs において社会集団のアイデンティティに関わる語彙として選ばれている単語を OLID で見たときに, ラベルにどの程度の偏りが生じているかを調査した.

OLID は Twitter のツイートから収集されたコーパスでスラング等が含まれる可能性があるが, Bias madlibs は自動生成された文章からなるコーパスで形式的な文章となっている. そのため, データの性質が異なっており Bias madlibs におけるアイデンティティに関わる語彙を OLID の中から単純に辞書マッチでラベル

対象単語	KNN で抽出された類似単語	OFF	NOT
homosexual	homosexual, homosexuality, lesbian, sexuality, adultery	**9**	4
lesbian	lesbian, feminist, transgeender, feminism, homosexual	**18**	13
male	male, female, males, females, men	321	**595**
female	female, male, females, woman, women	146	**196**

表 12: OLID のアイデンティティに関わる語彙が含まれる文のラベル数 (抜粋)

をカウントする方法は適していないと考えた. そこで, OLID のツイートに含まれる全てのトークンの中から, アイデンティティに関わる語彙のトークンに近い単語を 5 つ選び, その単語を含むラベルをカウントした. 単語の近さは Fine-tuning 前の $\mathrm{BERT}_{\mathrm{baseline}}$ の Embedding 層にある単語埋め込みから得られる表現のコサイン距離とした. さらに, 近隣の単語を k 近傍法を用いて選んでおり, 実装には scikit-learn[33] を利用している.

表 12 は上記の手法によって, Bias madlibs のアイデンティティに関わる語彙からそれに類似する OLID 内の単語を抽出し, OLID 上でそれらの単語を含む文章のラベル数をカウントしたものから一部を抜粋して表にしたものである.

表 12 より前節に述べた **homosexual** や **lesbian** のような性的マイノリティのアイデンティティに関連した単語を含む文章のラベルを見ると正例である OFF の比率が高くなっているが, 一方で male や female のような単語のラベルを見ると負例である NOT の方が多いことが確認できる. そのため, 学習に使われたデータセットにおけるラベルの偏りが分類結果に影響している可能性があると考えられる.

5.5.3 提案手法によってトークン単位の社会的バイアスも減っている

深層学習の解釈性における研究において, 入力の個々の特徴量が分類結果にどれだけ寄与したかを算出するインスタンスベースの Attiribution 手法を利用して, トークン単位での社会的バイアスを可視化していく. 本研究では自然言語処理においてよく見られる手法である Sundararajan ら [39] によって提案された Integrated Gradients を利用して, Attribution をおこなった.

Integrated Gradients　x を入力文, x_i を入力文内のトークンの埋め込み表現, x' をベースラインとして扱われる文とし, α を補完回数を表すパラメータとすると Integrated Gradients は次の式で定式化される.

$$\text{Integrated Gradients}_i(x) = (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha \times (x - x')\right)}{\partial x_i} d\alpha \qquad (14)$$

本研究においてベースラインの文には, すべてのトークンを [**PAD**] で埋めた文章を利用している. そのため, 式 (14) における x' は [PAD] のみで構成された文章の埋め込み表現となる. さらに, 入力文の埋め込み表現 x とすると, x と x' までの間の空間を α に応じた間隔を空けて補完する入力系列を作り, それらをモデルに 1 つづつ入力した場合に勾配がどれだけ変化するかを計算することで, 分類結果に対してどの単語が寄与しているのかの指標を得るという手法である. また, 本式は定積分を使っているが計算コストが高いことから, 実際の利用には次式のような近似的な式が推奨されており本稿も従った.

$$\text{Integrated Gradients}_i^{\text{approx}}(x) = (x_i - x'_i) \times \sum_{k=1}^{m} \frac{\partial F\left(x' + \frac{k}{m} \times (x - x')\right)}{\partial x_i} \times \frac{1}{m} \qquad (15)$$

式 (15) は, m を補完回数を表すパラメータとし, k をスケーリングのための定数だとして, 式 (14) の α までの定積分を総和に置き換えて m に応じた補完の平均を取ることで疑似的に式 (14) を実現している.

分析　Integrated Gradients によって貢献度を可視化する文章は, Bias madlibs から提案手法によって FPR が最も改善された社会集団のアイデンティティである **homosexual** が含まれた文から選択している. さらに, Integrated Gradients の性質を考慮してモデル間での比較するためにすべてのモデルで分類結果が True Positive か False Positive となっている文章を抽出した. そのため, 次の 3 つの文章を入力としており, 太字のものは False Positive となった文章で細字のものは True Positive な文章である.

(1) **homosexual people are just like everyone else**

(2) **I am a proud homosexual person**

(3) homosexual is disgusting

図 4, 図 5, 図 6 は上記の文章の Integrated Gradients を計算し, そのノルムをトークン単位での分類への貢献度とし, ヒートマップにしたものである.

これらの図から, 提案手法を用いたモデルの方がアイデンティティに関わる単語が暴言検出における分類の結果に寄与する度合いが減少していることが確認できる. 例えば, 正例であると誤って分類してしまった例である図 4, 図 5 では, $\mathrm{BERT}_{\mathrm{baseline}}$ の分類結果にトークン"**homosexual**"が寄与していることが読み取れる. 一方で提案手法を用いたモデルのヒートマップを見るとトークン"**homosexual**"の分類への貢献度が落ちており, この結果として 5.4.3 節の実験において提案手法によって FPED の改善が見られたと考えられる.

しかし, 正例であると正しく分類できている例である図 6 を見ると, 図 4 および図 5 と同様にトークン"**homosexual**"の寄与度が落ちていることがわかるが, 同時にトークン"**disgusting**"の寄与度も落ちてしまっており, これが 5.4.3 節において提案手法によって FNED が悪化した要因であると考えられる.

つまり, 提案手法を用いることで暴言検出においてデータの偏りによって発生する誤検出を防ぐ要因に効果的に寄与するが, 同時に学習データのラベルの量によっては暴言に関わる単語のみだけでは暴言であるとうまく分類できなくなってしまうという課題が存在すると考えられる.

5.5.4 結論

本研究では事前学習済み言語モデルを下流のタスクで扱う際に利用される Fine-tuning のプロセスにて, 社会的バイアスを学習しないようにする手法を提案した. 本研究の貢献を次の通りに示す.

本研究の独自性　本研究は学習の頑健性を向上させる手法を社会的バイアスの除去に応用した研究で, BERT を Fine-tuning する際にトレーニングデータに含まれる特定のアイデンティティに対する偏りを学習してしまうことを抑制するために,

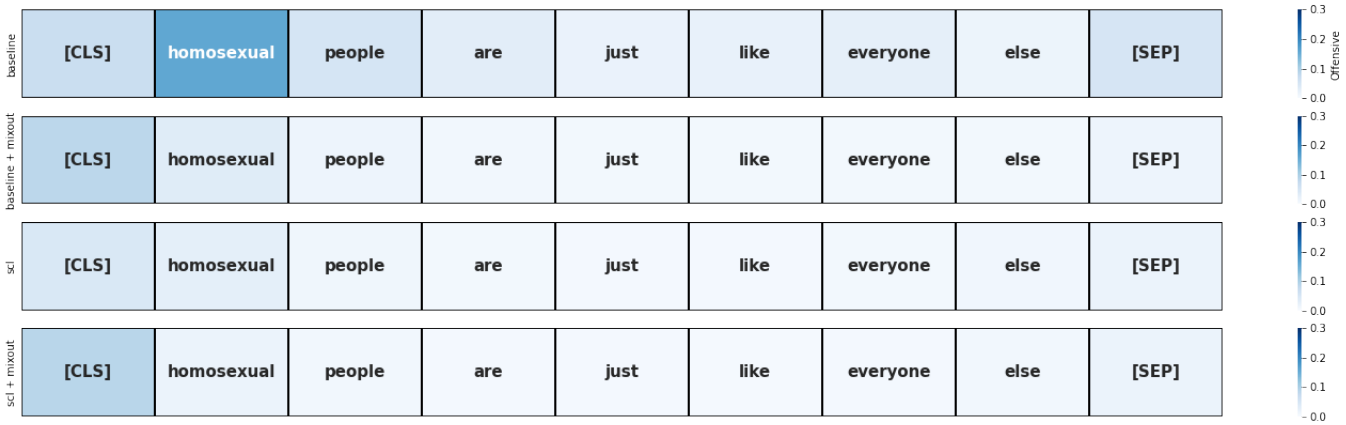


図 4: Integrated Gradients: homosexual people are just like everyone else

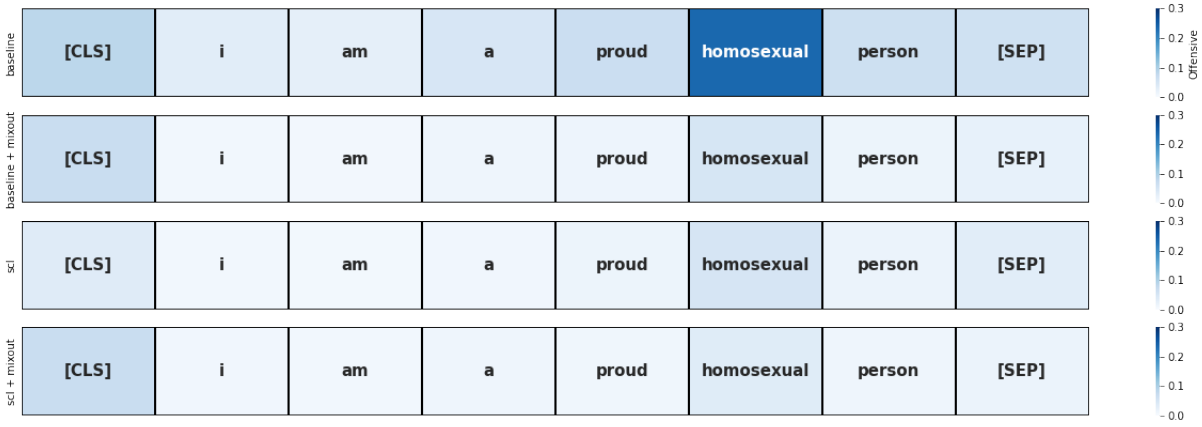


図 5: Integrated Gradients: I am a proud homosexual person

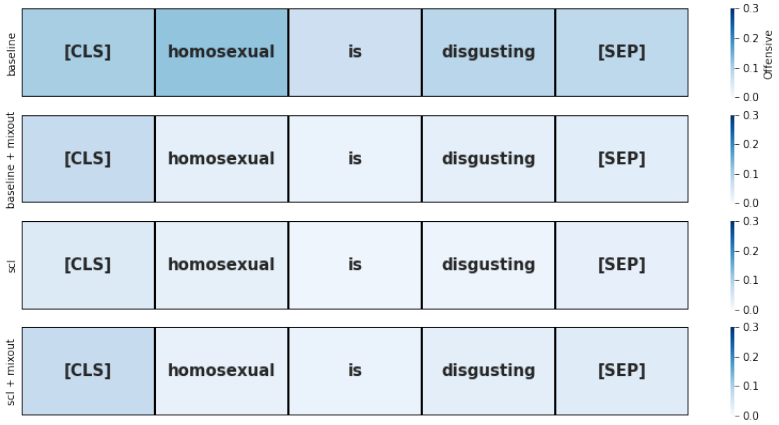


図 6: Integrated Gradients: homosexual is disgusting

Supervised Contrastive Loss を BERT の単語埋め込みのトークンを入力にして用いて, さらに Mixout によって正規化をする手法を提案した. 提案手法は, 暴言検出タスクにおいて攻撃的だと判断されやすいアイデンティティに対して, アイデンティティが文章に含まれるだけで攻撃的であると誤答しないように改善されることを確認でき, アイデンティティに対する公平性を考慮した分類性能を評価する Generalized Mean of Bias AUCs においても性能の向上が見られたため, 社会的バイアスの学習を抑制する効果が期待できると考えられる.

本研究の重要性　機械学習モデルの社会的バイアス除去に関するこれまでの先行研究では, データ拡張による手法が多いためデータセットによって利用できる手法が限られていた. また, 既に提案されているデータ拡張を活用しない手法においても, 事前に対象となるアイデンティティを定める必要があり, ドメインによっては活用することが困難であった. 提案手法では, モデルに損失関数を追加し正則化手法を適用するだけで社会的バイアスの学習を抑えられることが確認でき, より簡潔かつ広範囲のデータセットやドメインに適用可能であることが, 本分野において重要な手法になりうると考えられる.

6. おわりに

本稿では, 事前学習済み言語モデルから社会的バイアスを取り除くための次の 2 つの研究について述べた.

- 社会集団のアイデンティティを表す語彙の間で言語モデルが出力する周辺のトークンの予測確率の差を無くしていくアプローチ
- Fine-tuning 時の損失関数や正則化手法に工夫をすることで下流タスクでの社会的バイアスの学習を抑えるアプローチ

それぞれの研究では社会的バイアスを計測可能なものとして, そのバイアスを取り除くことを試みた. 前者の研究では, 言語モデルが出力する対数尤度によって得られる語彙の共起頻度から計算できる社会的バイアスの視点からアプローチを試みた. 一方で, 後者の研究は下流の分類タスクの分類結果から計算される分類性能における社会的バイアスの視点からアプローチを試みた.

これらの 2 つの研究はそれぞれ別の観点から社会的バイアスの除去を検討をしているが, 2 つの提案手法の問題意識や手法を統合したアプローチを試みていない. つまり, 共起頻度によって引き起こされる社会的バイアスと, 下流タスクにおけるラベルの偏りに引き起こされる社会的バイアスの間にどのような関係があるかは検討ができていない. しかし, どちらもデータ中における出現頻度の偏りによって社会的バイアスが引き起こされているため, 社会的バイアスを学習してしまう根本的な原因は近いと考えている. そのため, 今後の研究にてこれらの社会的バイアスの間にどのような関係性が存在するのかを検討していく必要がある.

謝辞

この研究を進めるにあたって研究の進め方や実験方法について有益な助言をいただきました渡辺太郎教授をはじめ, 進捗報告の場で含蓄のある助言や質問をくださった大内啓樹助教および芝原隆善氏, そして, 審査に携わっていただいた中村哲教授および進藤裕之特任准教授の皆様に, この場を借りて深く御礼を申し上げます.

参考文献

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[2] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, June 2016.

[4] Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Limitations of pinned auc for measuring unintended bias, 2019.

[5] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500, 2019.

[6] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *American Association for the Advancement of Science*, 2017.

[7] Kai-Wei Chang and Vinodkumar Prabhakaran Vicente Ordonez, Margaret Mitchell. Cognitive biases / data biases / bias laun-

dering. http://web.cs.ucla.edu/~kwchang/documents/slides/emnlp19-fairNLP-part3.pdf.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[9] Jeffrey Dastin. 焦点：アマゾンがＡＩ採用打ち切り、「女性差別」の欠陥露呈で (2021/7/10 閲覧). https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186, 2019.

[11] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. 2018.

[12] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery.

[13] Philipp Dufter and Hinrich Schütze. Analytical methods for interpretable ultradense word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1185–1191, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*, 2021.

[16] Masahiro Kaneko and Danushka Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy, July 2019. Association for Computational Linguistics.

[17] Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. In *Proc. of the 16th European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

[18] Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online, July 2020. Association for Computational Linguistics.

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020.

[20] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics.

[21] Alex Labach, Hojjat Salehinejad, and Shahrokh Valaee. Survey of dropout methods for deep neural networks, 2019.

[22] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[23] Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations*, 2020.

[24] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online, July 2020. Association for Computational Linguistics.

[25] Sheng Liang, Philipp Dufter, and Hinrich Schütze. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020.

[27] Google LLC. bias (ethics/fairness) - machine learning glossary. https://developers.google.com/machine-learning/glossary/#bias-ethicsfairness.

[28] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. 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)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[29] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models, 2020.

[30] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, November 2020. Association for Computational Linguistics.

[31] Pandu Nayak. Understanding searches better than ever before (2021/7/10 閲覧).

[32] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[35] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63:1872–1897, 10 2020.

[36] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[37] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[38] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics.

[39] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org, 2017.

[40] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020.

[41] Ameya Vaidya, Feng Mai, and Yue Ning. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):683–693, May 2020.

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[43] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[44] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[45] 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 NAACL*, 2019.

[46] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. 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), pages 629–634, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[47] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[48] 新妻 巧朗 and 渡辺 太郎. 文表現の摂動正規化: 事前学習済みモデルの debias 手法. In **言語処理学会 第 *27*回年次大会 発表論文集**. 言語処理学会, March 2020.