＜修士論文＞

敵対的学習に基づくドメイン不変表現の獲得とそれを用いたアスペクトベース感情分析

滋賀大学大学院
データサイエンス研究科
データサイエンス専攻

修了年度　：2023
学籍番号　：6022139
氏 名　　：水谷 宏太
指導教員　：南條 浩輝
提出年月日　：2024年1月10日

目 次

1　はじめに

1.1　研究背景

近年，Instagram や X（旧 Twitter）などの SNS の普及に伴い，テキストデータを対象とする感情分析の重要性が増している．テキストデータを対象とする感情分析おける一般的なタスクは，対象とするテキストを読み取り，ポジティブな内容か，ネガティブな内容か，あるいはニュートラルな内容かを判定するタスクである [1]．張ら (2022)[1] によると，感情分析の対象となるテキストには主に三つの粒度のタスクがある．具体的には，粒度の大きい順に，ドキュメント単位，文単位，アスペクト単位，の感情判定タスクである．アスペクトとは，ある対象の 1 つの側面であり [1]，アスペクトを対象とする感情分析のことをアスペクトベース感情分析 (Aspect Based Sentiment Analysis: 以下，ABSA) という．ABSA では，テキストを文レベルではなく単語レベルにタグ付けをして感情分析を行う．例えば樊ら (2022)[2] では “*The food is delicious, but the service is too bad.*” というレビュー文を例を用いて説明がされている．文単位で感情極性を判定するタスクでは，例で挙げたような文章中にポジティブとネガティブの両方を含む文を分析しにくい．一方，ABSA では，この文における話者が “*the food*” に対して “*delicious*” と評価していることから，“*food*” に対して肯定的な感情を持っており，“*the service*” に対し “*too bad*” と評価していることから，“*the service*” に対して否定的な感情を持っているといったことを分析できる [2]．上記の例のように，ABSA はドキュメントレベルの感情分析や文レベルの感情分析と比べてより細かい粒度で感情分析を行えるため，多くの場面で実用化が期待されている．しかし，ABSA はその対象の粒度の細かさから，データセットのアノテーションコストが従来の文を対象とした感情分析のためのデータセットよりも高いというデメリットがある [3]．また，ABSA 研究用の一般的なデータセットが収録するテキストドメインは，文レベルの感情分析用のデータセットよりも限定的であり，データセットのサイズも小さい．加えて，ABSA のためのデータセット構築は人間の手によるアノテーションが必要となるため，新たにデータセットを作成するのは容易ではない．このような背景から，ABSA を行いたいケースにおいても，分析対象のテキストと同じドメインのデータセットがないため，高精度な ABSA が行えないという問題がある．

1.2 研究概要

現状の ABSA における問題は，分析対象としたいドメイン（ターゲットドメイン）のラベル付きデータが不足していることにより，既にある十分な教師ラベルを持つドメイン（ソースドメイン）のデータを用いて学習してもターゲットドメインとソースドメインの分布の違いにより，高い精度で分類が行えないという点である．そこで本研究では，敵対的不変表現学習という手法を用いることで上記の課題の解決を試みる．敵対的不変表現学習とは，Ganin et al.(2016)[4] で提案された，分析対象としたいドメイン（ターゲットドメイン）のデータと学習用のドメイン（ソースドメイン）のデータの分布が異なる場合にドメインに関する情報を持たない表現を獲得することでソースドメインに対する分析精度の低下を防ぐアプローチである．本研究では，分析対象と異なるドメインのデータセットで ABSA モデルの学習を行った場合に，敵対的不変表現学習を取り入れたモデルとそうでないモデルを比較することにより，ABSA における敵対的不変表現学習の有効性を検証する．

1.3 本論文の構成

本論文の今後の構成は，以下の章からなる．第 2 章では，ABSA についてのより詳細な内容や関連する研究について説明する．第 3 章では，提案手法の概要と用いる手法，モデル構成について説明する．第 4 章では，各実験の内容について説明する．第 5 章では，各実験の結果とそれに対する考察を示す．第 6 章では本論文の結論と考察，展望を述べる．

2　関連研究

2.1　ABSA について

本章では，ABSA についてより詳細な内容とその先行研究を紹介する．1 章でも述べた通り，ABSA はドキュメントレベルの感情分析や文レベルの感情分析と比べてより細かい粒度の単位を対象に感情分析を行う手法である．ABSA は様々なタスクによって構成されている [5]．それぞれのタスクは ABSA に関連する主に四つの要素を特定することを目的としている．その四つの要素とは，アスペクト用語 (Aspect Term)，アスペクトカテゴリ (Aspect Category)，意見用語 (Opinion Term)，感情極性 (Sentiment Polarity) である [6]．アスペクト用語とは，与えられたテキスト内において明示的に表れる意見の対象のことである [5]．例えば，三浦ら (2020)[7] では，「みかんが美味しい」という文章を使って説明がなされている．「みかんが美味しい」という文章が与えられた時，アスペクト用語は “みかん” となる．アスペクトカテゴリは，エンティティとアトリビュートから構成され，エンティティがアスペクト用語が何のカテゴリに属するかを意味し，アトリビュートがその対象のどの属性について言及されているかを示す [7]．例えば，先の例で示した「みかんが美味しい」という文で考えると，アスペクト用語のみかんは “食べ物” に属し，文中ではみかんの “質” に関して言及しているため，アスペクトカテゴリは “食べ物#質” となる [7]．なお，アスペクトカテゴリを特定する際は事前に候補となるカテゴリを定義しておく必要がある [5]．意見用語は，テキスト内でアスペクト用語に対する感情を表している表現である [5]．「みかんが美味しい」という文では，話者がみかんに対して “美味しい” と表現しているので，“美味しい” が意見用語となる．感情極性はアスペクト用語に対する意見用語の感情極性を示し，通常は “positive”，“negative”，“neutral” のいずれかに属する [5]．「みかんが美味しい」という文では “美味しい” という意見用語が肯定的な意見を表しているため，“positive” となる．なお，研究によって各要素が異なる呼び方で用いられている場合があるが，一般的に ABSA ではいずれかの要素を一つ，または複数の異なる要素を特定するタスクによって構成されている．

英語分野における ABSA の初期の研究では，ルールベースでのアプローチや bag-of-n-gram を特徴量とした機械学習によるアプローチなど，特徴量エンジニアリングに依存した古典的な機械学習手法を用いていた [8]．その後，RNN (Recursive Neural Networks) などのより高度なニューラルネットワークベースの手法を用いることによって，以前より

高い精度で ABSA が行えるようになった [9]．近年では他の自然言語処理のタスクと同様，BERT(Bidirectional Encoder Representations from Transformers)[10] をはじめとする事前学習済みモデルを活用した手法についても注目されている．Sun et al.(2019)[11] では，ABSA を BERT モデルが得意とする QA タスクや NLI タスクのような文ペア分類問題の形式に変換する方法が提案されている．

日本語分野における ABSA の研究では，英語分野におけるニューラルネットワークベースの ABSA の手法を日本語文へ適用する手法の提案 [12] や，最近では英語分野の ABSA と同様に事前学習済み言語モデルを利用してモデルを構築する手法も提案されている [7]．

2.2 ABSA における課題

ABSA を行うことによってテキストからより詳細な情報を得ることが期待されているが，依然として文レベルの感情分析と比較して ABSA に関する研究は少ない．その要因の一つとして考えられるのが 1.1 節でも述べたデータセットの少なさである．例として，ABSA と一般的な文レベルの感情分析，それぞれの分析における代表的なデータセットについて紹介する．ABSA の代表的なデータセットとして SemEval-2014 Task 4 データセット [13] がある．これは，レストランに関するレビュー文とラップトップ（ノート PC）に関するレビュー文によって構成されており，合計で 7485 文のデータセットサイズとなっている．一方で，文レベルの感情分析で用いられる代表的なデータセットとして Multilingual Amazon Reviews Corpus [14] がある．これは，テキストの分類を目的として提供されているデータセットであり，英語，日本語，ドイツ語，フランス語，中国語，スペイン語のレビュー文が含まれている．各言語のデータセットには 210,000 件のレビューが収録されており，様々なカテゴリの製品に対するレビュー文とそれに対する評価を表す星の数などのデータがまとめられている．このように，感情分析でよく用いられるデータセットと比較すると，ABSA を目的として提供されているデータセットはそのサイズも小さく，収録しているテキストのドメインの範囲も狭い．また，この傾向は日本語のデータセットの場合も同様である．

日本語における ABSA を目的とした代表的なデータセットに TIS 株式会社が公開している chABSA-dataset[1] がある．chABSA-dataset は上場企業の決算報告書をドメインとしており，日本語分野を対象とした ABSA に関する研究では最も用いられているデータセッ

[1] https://github.com/chakki-works/chABSA-dataset

トの一つである．しかし，chABSA-dataset は日本語の ABSA データセットとしてはサイズの大きいデータセットではあるが，収録されている文は 6,119 文と先述した Multilingual Amazon Reviews Corpus の日本語のレビュー文だけでも 210,000 件あることを考えるとやはり少ない．以上で示したように，ABSA を目的としたデータセットは限られた範囲のドメインのデータセットしか存在せず，また新たなデータセット構築にも多大なコストがかかるという問題があることがわかる．以上の問題を解消するために，本研究では，敵対的不変表現学習という学習手法を用いることにより，限られた範囲のドメインのデータセットで学習したモデルを用いて未知のドメインのデータセットに対する分類性能を向上させることを目標とした．

3　提案手法

3.1　概要

先述した課題に対して，本研究では，事前学習済み自然言語処理モデルである BERT モデルを用いたマルチタスクモデルに対し敵対的不変表現学習を行い，学習とはドメインの異なるデータセットに対する分類精度を検証する．以下では BERT モデル，敵対的不変表現学習，本研究のフレームワークについてそれぞれ詳しく説明する．

3.2　BERT

BERT とは，Devlin et al.(2018)[10] で提案された自然言語処理モデルである．BERT のモデル構造には Vaswani et al.(2017) で提案された Transformer モデル [15] のエンコーダ部分に基づくアーキテクチャが採用されている．BERT の特徴として，自然言語処理のタスクに応用する際に少ない訓練データで様々なタスクに応用することができる点が挙げられる．BERT を用いる際は，既に大規模なデータセットで事前学習したモデルを用いて，それを特定のタスクに合わせて少量のラベル付きデータで fine-turning するというのが一般的である．

BERT の事前学習には単語穴埋めと次文予測の 2 種類があり，単語穴埋めは [MASK] したトークンを予測するタスクで，文脈を考慮した単語埋め込みを得るために行われる [16]．

次文予測タスクは，特殊トークン [SEP] により連結された二つのテキストを入力してその二つが元の文書で連続していたのか，別々の文書からランダムに選ばれたものかを推定するタスクで，先頭に付加された特殊トークン [CLS] に対応する最終層の埋め込み表現に二値分類のための全結合層を連結することによって行われる [16]．これらの事前学習により，BERT は入力文全体の埋め込み表現（ベクトル）や与えられた 2 文の関係を表す埋め込み表現を学習することができる．この学習された [CLS] トークンに対応する出力は，文の意味や 2 文間の関係を捉えるベクトルとして感情分析などのタスクに有効に活用することができる．

本研究で行う ABSA においては，アスペクト用語とそれを含む文章をを事前学習済み BERT に入力して [CLS] トークンに対応する出力を得ることで，文中のアスペクト用語に関する埋め込み表現を抽出し，これを感情極性分類に用いる．なお，本研究では事前学習

済みモデルとして，日本語のウィキペディアデータで事前学習された東北大 BERT モデル[2] を利用した.

3.3 敵対的不変表現学習

敵対的不変表現学習とは，ソースドメインのデータとターゲットドメインのデータの分布が異なる場合に用いられる敵対的学習手法の一つである．教師あり学習では通常，学習データとテストデータが同一の分布に従うという仮定のもとで行われるが，この仮定が成り立たない場合に機械学習モデルのテストデータに対する予測精度は著しく低下する [17]. このような状況はドメインシフトと呼ばれ，このドメインシフトを想定した研究領域の一つにドメイン汎化と呼ばれるものがある．ドメイン汎化では，複数のソースドメインのラベル付き訓練データを利用することにより，ソースドメインと異なるターゲットドメインから得られるテストデータに対する予測を行う．ドメイン汎化は手書き文字認識 [18] や加速度センサーを用いた行動認識 [19] など様々な分野において応用されている技術である.

本研究では, 敵対的不変表現学習の手法として Ganin et al.(2016)[4] が提案した Domain Adversarial Neural Networks（以下 DANN）を用いる．DANN は，ドメインに依存しない特徴量，すなわちドメインによって不変な特徴量をもとにソースドメインのデータを分類するよう学習することによってターゲットドメインに対する予測精度を下げないようにする学習手法である．Ganin et al.(2016) で提案されている DANN のアーキテクチャを図 1 に示す．Ganin et al．(2016) では，異なるドメインの画像分類問題に対して DANN を提案してその有効性を示している．図 1 によると，特徴抽出を行うネットワーク（緑色）を出たあと，クラス分類を行うネットワーク（青色）とドメイン分類を行うネットワーク（赤色）に分岐しているが，ドメイン分類のネットワークに入力する際，Gradient Reversal Layer（GRL，以下勾配反転層）と呼ばれる層を通っている．この層は順伝播方向の計算では何も行わないが，逆伝播計算を行う際に前の層へ伝わる勾配の符号を反転させるという機能を持つ．これによって，図 1 のモデルが学習を進めるごとに特徴抽出を行うネットワークでは，クラス分類はできるもののドメイン分類は難しい特徴量（図 1 中の f）を抽出するようにパラメータが更新される. 一方，ドメイン分類を行うネットワークではこの f を用いてもドメイン分類ができるようパラメータが更新される．すなわち，特徴抽出とドメイン分類のネットワーク間で敵対的な学習が行われる．これを繰り返すことで，特徴抽

[2] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

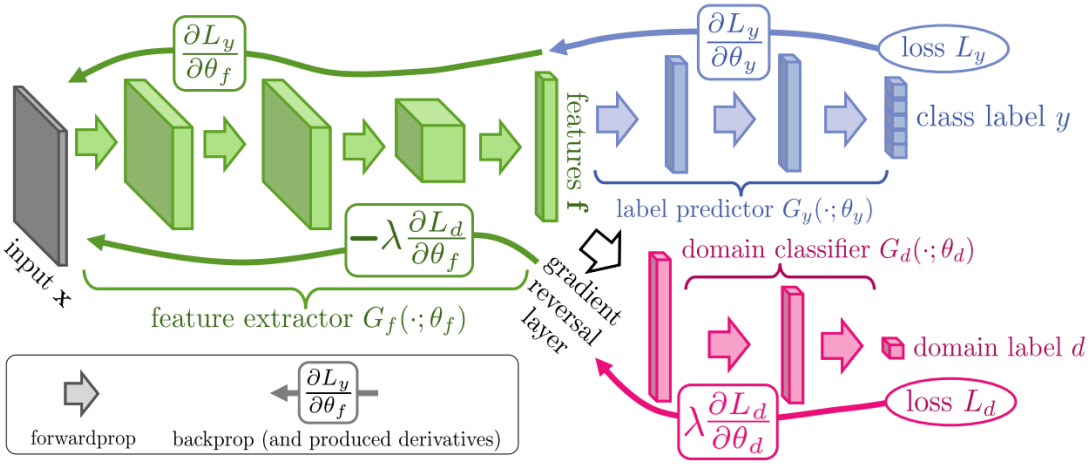


図 1: DANN のアーキテクチャ（[4] の Figure 1から引用）

出ネットワークでは，ドメイン分類ネットワークがどうやってもドメイン分類できないような特徴量，すなわち「ドメイン不変な特徴量」を獲得することを目指す仕組みである．

DANN の特徴として，標準的な誤差逆伝播法を用いて学習をするアーキテクチャのほとんどのモデルに対して DANN を適用できるという汎用性の高さがある．本研究では，テキストのアスペクトカテゴリ（エンティティ#アトリビュート）とアスペクトの極性 (positive/negative/neutral) を分類するマルチタスクモデルに対して DANN を適用することで，異なるドメイン（本研究では学習に使わないアスペクトカテゴリを指す）のデータに対する感情極性の分類精度がどのように変化するかを標準的なマルチタスクモデルや感情極性分類のみを行うモデルと比較する．

3.4 本研究のフレームワーク

本研究で提案するモデルの概要を図 2 に示す．まず，分析の対象となるテキストを事前学習済み BERT モデルへ入力する．その際，入力となるテキストは，分析対象となるアスペクト用語とそれを含む文章である．本研究では，BERT モデルに対してアスペクト用語と文章を入力する二文入力の形式で分析を行う．図 3 に入力の例を示す．この例では，アスペクト用語が「日本経済」，文章が「日本経済の見通しは明るい」であり，これらを [SEP] トークンで区切って BERT に入力する．先頭の [CLS] トークンに対応する最終層の埋め込み表現を極性分類ネットワークとカテゴリ分類ネットワークに入力することでアスペクトの感情極性分類，アスペクトカテゴリ分類をそれぞれ行う．感情極性分類のネットワー

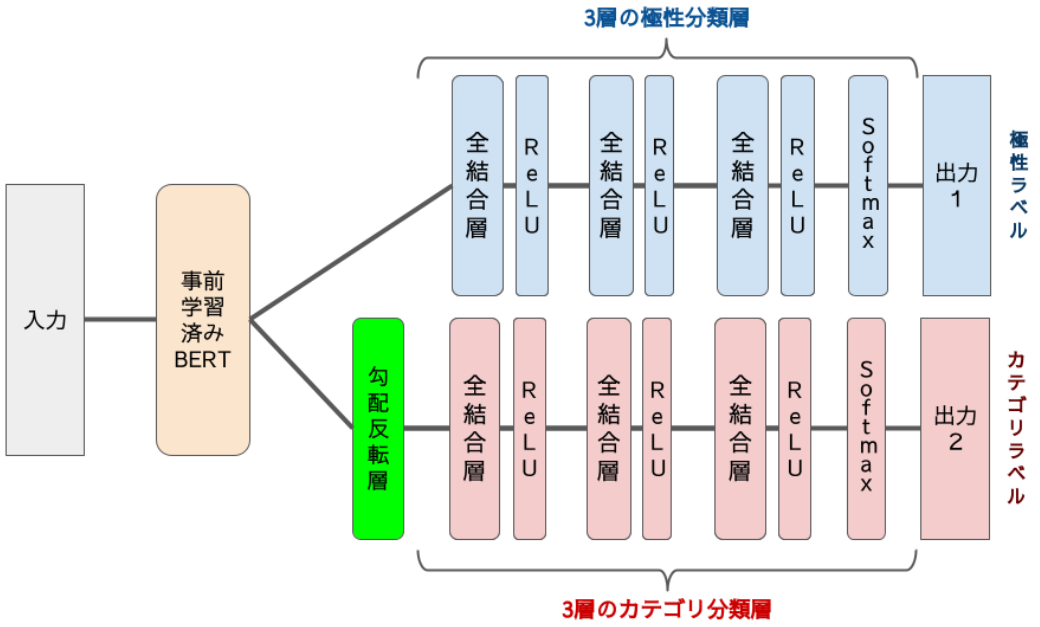


図 2: モデルの構造

図 3: 入力の例

クでは入力されたターゲットが文中でどのように表現されているかについて，“positive”，“negative”，“neutral” のいずれかの感情極性を予測をする．カテゴリ分類のネットワークでは，ターゲットが文中でどのような観点で言及されているかを与えられたアスペクトカテゴリの候補をもとに予測する．その際に，カテゴリ分類のネットワークに入力される前に 3.3 節で説明した勾配反転層を通る．それによりモデルの学習が進むにつれて，この

ネットワーク中で特徴量抽出器として働いている BERT は，与えらえたアスペクト用語と文から，感情極性分類は可能だがカテゴリ分類はできないような特徴を取り出せるようになる．その結果，アスペクトカテゴリに依らない，すなわちドメイン不変な特徴量をもとに感情極性分類が行われることが期待できる．このモデルがドメインの異なるデータセットに対しても高精度に分類が行えるか実験により検証する．

4　実験

本章では，先述した課題に対して本研究で行なった実験について説明する．

4.1　実験 1．マルチタスクモデルによる敵対的不変表現学習

4.1.1　chABSA-dataset

実験 1 では，2.2 で説明した chABSA-dataset を用いて分析を行った．改めて chABSA-dataset について説明すると，chABSA-dataset は 2016 年度の上場企業の有価証券報告書をベースに作成されたデータセットである．各文に対して “positive”, “negative”, “nertral” の感情極性のラベルだけでなく，各ターゲットのアスペクトに対して感情極性のラベルが付与されている．chABSA-dataset のアスペクトカテゴリーは，ターゲットの単語が何のカテゴリに属するかを意味するエンティティが 4 つ (company，business，product，NULL)，その対象のどの属性について言及されているか示すアトリビュートが 6 つ (general，sales，profit，amount，price，cost) であり，それらを組み合わせた 23 種類のアスペクトカテゴリー（company#price が dataset に存在しないため 23 種類）とそれに market#general を合わせた合計 24 個である．各アスペクトカテゴリーはエンティティとアトリビュートを組み合わせて company#general，product#amount といった形になっている．chABSA-dataset では，各文に対してターゲットフレーズとそのアスペクトカテゴリ，極性が付与されている．例えば，「当社グループの主力事業が属するインターネット広告市場は，当年度においても広告市場全体の伸びを上回る成長が続きました」という文に対して，ターゲットフレーズは「インターネット広告市場」，アスペクトカテゴリーは “market#general”，感情極性が “positive” というように情報が付与されている．

本研究では，chABSA-dataset の中の感情情報が含まれているアスペクトカテゴリ 7,723 個のうち，ターゲット・文のペアで複数のアスペクトカテゴリ・感情極性が付与されているサンプルを排除した 6,849 個のサンプルを用いる． これは本研究では，ターゲット・文のペアを入力とする際に，アスペクトカテゴリと感情極性をそれぞれ一つずつ予測するためである．排除したサンプルの例を挙げる . ターゲットが「食品製造販売部門」，文が「食品製造販売部門においては、取引先の見直し等により、売上高は昨年を下回りましたが、原料価格が低下したことと、販売経費の削減により、所定の利益を確保することができました」であるサンプルの場合，文中では「食品製造販売部門」の売上高に関する言及と利益に関す

表 1: 各データセットに含まれるアスペクトカテゴリー

トレーニングデータ	“NULL#amount”, “NULL#cost”, “NULL#general”, “NULL#price”, “NULL#profit”, “NULL#sales”, “company#amount”, “company#cost”, “company#general”, “company#profit”, “company#sales”, “market#general”
バリデーションデータ	“product#amount”, “product#cost”, “product#general”, “product#price”, “product#profit”, “product#sales”
テストデータ	“business#amount”, “business#cost”, “business#general”, “business#price”, “business#profit”, “business#sales”

る言及がある. このうち，売上高に関しては「昨年を下回りましたが」とあるためネガティブに，利益に関しては「所定の利益を確保することができました」とあるため，ポジティブに述べられていることがわかる. このような場合にデータセットではアスペクトカテゴリが “business#sales” の場合は “negative”，“business#profit” の場合は “positive” とラベリングしてあるが，今回はターゲットと文のペアに対してアスペクトカテゴリと感情極性をそれぞれ一つずつ予測するため，このようなサンプルは分析対象から排除している.

6849 個のデータセットに対して，実験 1 ではアスペクトカテゴリごとに学習・検証・テストデータを分割する．具体的には表 1 に記載した．なお，各データセットのサンプルサイズはトレーニングデータが 5,150，バリデーションデータが 1,130，テストデータが 566 である.

4.1.2 実験 1 の内容

実験 1 では 3 つモデルの精度を比較する．3 つのモデルとは，カテゴリ分類と感情極性分類を行うマルチタスクモデルに勾配反転層を加えたモデル（以下 GRL+Multi），カテゴリ分類と感情極性分類を行う通常のマルチタスクモデル（以下 Multi），感情極性分類のみを行うモデル（以下 Single）である．各モデルはトレーニングデータで学習し，デフォルトのエポック数を 30 回と設定して各エポックごとにバリデーションデータに対する Accuracy を測る．その際に最も Accuracy が高かったモデルのエポック数を採用し，そのモデルのテストデータに対する精度をそのモデルの最終的な精度とする．データセットごとに異なるアスペクトカテゴリのテキストが割り当てられることによって，学習に用いていないアスペクトカテゴリのテキストに対する感情極性分類の精度を比較することができる．すなわち，擬似的に未知のドメインのテキストデータに対して感情極性分類を行

い，その精度を比較するのが実験 1 である．

4.2 実験 2．異なるドメインのテキストデータに対する精度比較

実験 2 では，実験 1 で学習したモデルで他のドメインのデータセットに対して感情極性分類を行い，その精度を比較する．用いるデータセットは X のテキストデータを使用したデータセット（以下 Twitter データセット）である．以下ではまず，データセットのテキストの収集方法とアノテーションの方法について説明する．

4.2.1 Twitter データセット

Twitter データセットの作成方法を述べる．まず X で 6 種類のターゲットフレーズに対して感情的な意見を述べている文章を “positive”，“negative”，“neutral” がそれぞれ均等になるようそれぞれ 25 文ずつ合計 150 文選ぶ．次にその文章を 6 種類のターゲットフレーズごとに 3 つのグループに分ける．その 3 つのグループに対してそれぞれアノテーターを 2 人割り当てる．グループごとに割り当てられた 2 人のアノテーターと執筆者を含む 3 人でアノテーションを行った．アノテーションの方法として Google form のアンケートに回答してもらう形でアノテーションを行った．図 4 は実際にアノテーションに用いたアンケートの 1 ページ目である．まずアンケートの 1 ページ目でアンケートの説明を行った．なお，このアンケートの説明については以降のページの最初の項で再度確認できるように設計した．次に，どのようにアノテーションを行うかを示す具体例を提示した．それが図 5 である．各アノテーターにアノテーションしてもらう前にアンケートの説明と具体例を提示することにより，アノテーションの方針についての認識が揃うようにした．その後，各アノテーターに割り当てられたテキストについて回答してもらった．図 6 は設問の例である．下記のような形式でアノテーターに回答してもらった．各設問では，実際に X でポストされた意見文におけるターゲットフレーズと，それを含むテキストを提示し，ターゲットフレーズがそれを含むテキストにおいてどのように言及されているかを読んでもらい，該当すると思う感情極性を回答してもらった．各アノテーターに回答してもらった後，分析用のデータセットに採用するラベルの選定を行った．具体的な方法として，まず，3 人のアノテーターが揃った回答に関してはそのままそのラベルを採用した．逆に，3 人のアノテーターの回答が全く一致しなかったターゲットとテキストのペアに関してはデータセットから排除した．また，アノテーター 3 人のうち，2 人の回答が一致した場合は回答

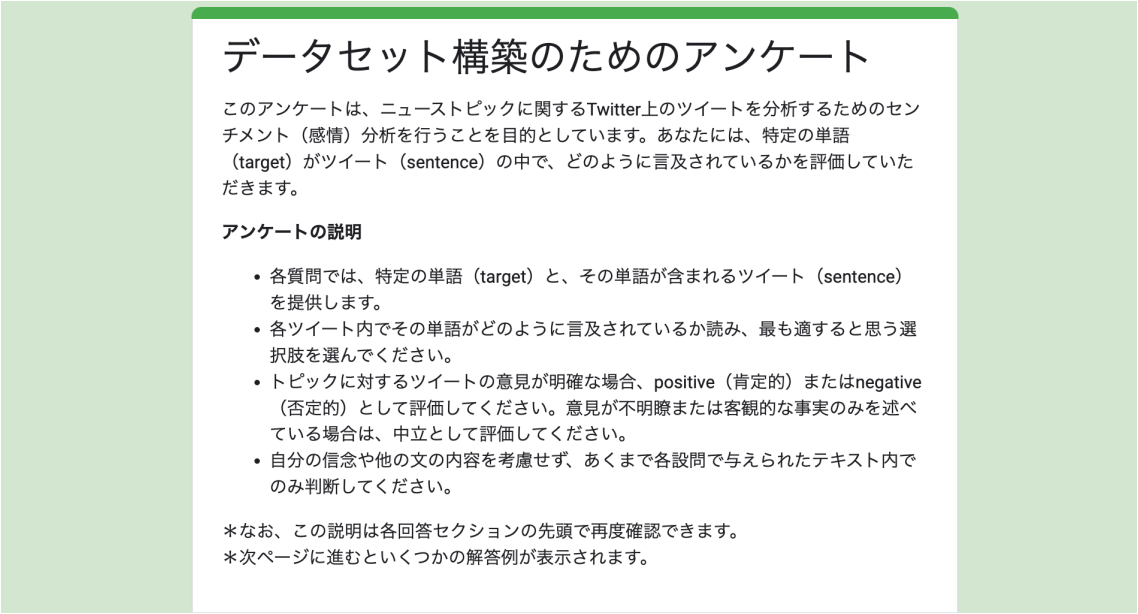


図 4: アンケート 1

図 5: アンケート 2

が多かった方のラベルを採用した．ただし，アノテーター 3 人のうち，2 人の回答が一致した場合に，“positive” に回答した人が 2 人で “negative” に回答した人が 1 人といったケースまたはその逆のケース（“positive” に回答した人が 1 人で “negative” に回答した人が 2 人）の場合はそのターゲットとテキストのペアに関してもデータセットから排除し

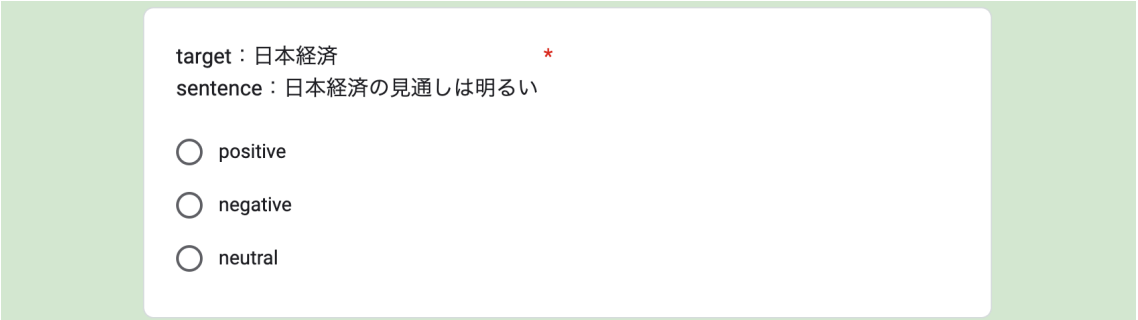


図 6: アンケートの設問例
※実際のデータではない

表 2: Twitter データセット

target	total	positive	negative	neutral
マイナンバーカード	23	9	7	9
テレワーク	25	9	8	8
ライドシェア	24	9	8	8
インボイス制度	23	7	10	8
ベーシックインカム	25	7	8	10
選択的夫婦別姓	25	9	8	8
total	145	51	46	48

た．その結果，Twitter データセットとして採用した各ターゲットごと，感情極性ごとのテキスト数は表 2 の通りである．この Twitter データセットを用いて実験を行った．

4.2.2　実験 2 の内容

実験 2 では，作成した Twitter データセットに対して，実験 1 で学習したモデルをそのまま用いて感情極性分類を行う．すなわち実験 2 は，追加学習を行わないことによって，全く未知なドメインのデータセットに対する各モデルの感情極性分類の精度を比較することを目的とした．

4.3　実験 3．Twitter データセットで追加学習を行った場合の精度比較

実験 3 では，対象としたいドメインにおける学習データが少ない場合における各モデルの感情極性分類の精度を比較することを目的とする．それにあたって，実験 1 で学習したモデルに対して Twitter データセットで追加学習を行ったモデルでクロスバリデーション

を行い，精度を比較した．追加学習として，Twitter データセットをターゲットフレーズごとに 6 分割し，トレーニングデータ：バリデーションデータ：テストデータを 4:1:1 の比率で割り当てた．全てのデータが一度はテストデータとして利用されるように割り当てのプロセスを 6 回繰り返し，学習ごとのテストデータに対する精度の平均値を取ることによってクロスバリデーションを実施した．なお，追加学習の際のエポック数は 10 回をデフォルトとし，各モデルの最初の学習の際に各エポックごとにバリデーションデータに対する Accuracy を測る．その Accuracy が最も高かったエポック数を採用し，残りの 5 回の学習もそのエポック数で行った．

5　結果

本章では，4 章で説明した各実験の結果と結果を受けた考察を述べる．

5.1　実験 1．chABSA-dataset での敵対的不変表現学習の効果

実験 1 では 4.1.2 節で述べた方法で，感情極性分類を行った．なお，3 つのモデルのハイパーパラメータを表 3 に示す．

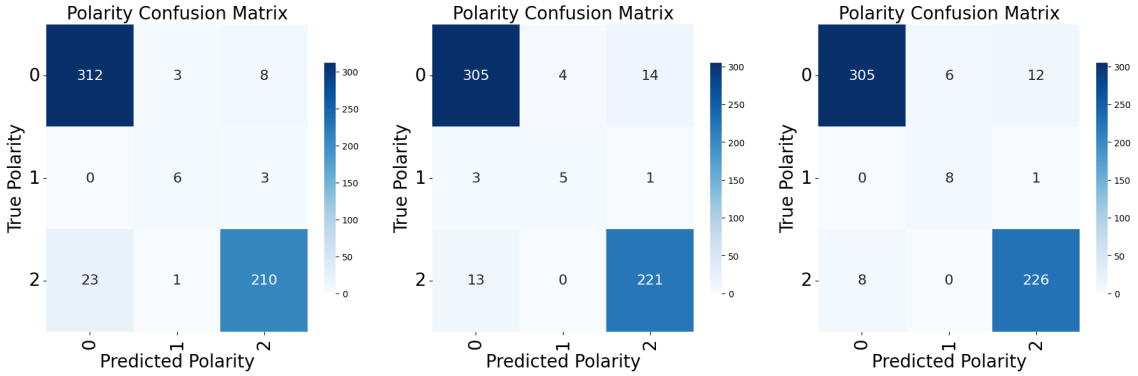
実験 1 のテストデータに対する各モデルの感情極性分類の精度を表 4 に示す．また,3 つのモデルのテストデータに対する感情極性分類の結果の Confusion Matrix を可視化したものが図 7a,7b,7c である．　表 4 の通り，感情極性分類のみを行ったモデルが Accuracy，Macro-F1 score ともに最も精度が高いという結果になった．実験 1 では，chABSA-dataset

表 3: モデルのハイパーパラメータ

	最適化アルゴリズム	学習率	バッチサイズ	エポック数
GRL+Multi	RMSProp	1e-05	16	26
Multi	RMSProp	1e-05	16	26
Single	RMSProp	1e-05	16	23

表 4: 実験 1 の結果

	Accuracy	Macro-F1
GRL+Multi	0.933	0.834
Multi	0.938	0.814
Single	0.952	0.870



(a) GRL+Multi　(b) Multi　(c) Single

図 7: 実験 1 の Confusion Matrix

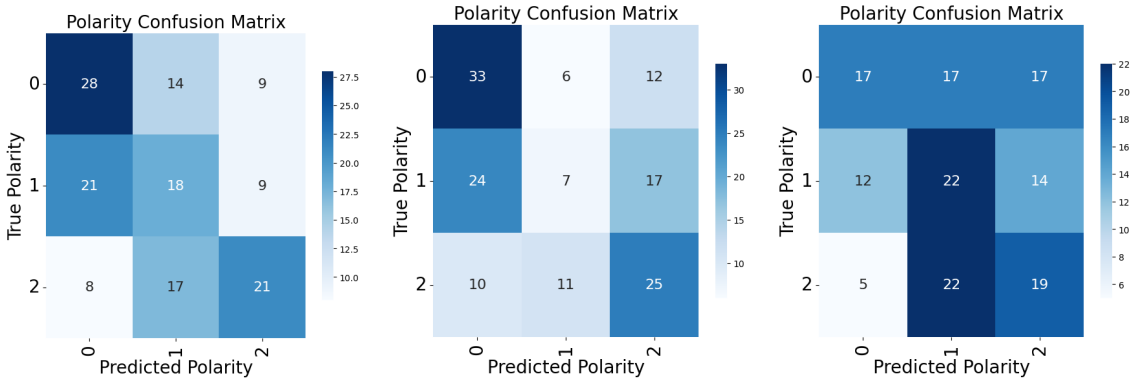
のうち，トレーニングデータ，バリデーションデータ，テストデータに分ける際にアスペクトカテゴリーが重複しないようした．これにより，学習したデータと異なるアスペクトカテゴリーのデータに対する分類精度を比較し，擬似的に未知のドメインのデータに対する分類精度を検証するというのが実験 1 の目的である．しかし結果を見ると，通常のマルチタスクモデルや感情極性分類のみを行ったモデルにおいて，アスペクトカテゴリの違いによる分類精度の低下がそこまで見られなかった．これはトレーニングデータとテストデータがどちらも有価証券報告書を元に作られた同一データセットに含まれるテキストであることから意見を表す表現が似ていたために，ドメイン不変の特徴量を用いなくても高精度で感情極性分類ができたのではないかと考えられる．

5.2 実験 2 の結果

実験 2 では，実験 1 で学習したモデルを追加学習せずに Twitter データセットに対して感情極性分類を行った．その結果を表 5 に示す．また，3 つのモデルの Twitter データセットに対する感情極性分類の結果の Confusion Matrix を可視化したものが以下の図 8a,8b,8c である．

表 5: 実験 2 の結果

	Accuracy	Macro-F1
GRL+Multi	0.462	0.461
Multi	0.448	0.418
Single	0.400	0.400



(a) GRL+Multi (b) Multi (c) Single

図 8: 実験 2 の Confusion Matrix

表 5 の通り，GRL+Multi モデルが Accuracy，Macro-F1 score ともに最も精度が高いという結果になった．実験 1 ではトレーニングデータとテストデータは同一のデータセットに含まれるテキストであったが，実験 2 ではトレーニングデータと全く異なる文章を用いて各モデルの分類精度を比較した結果，GRL+Multi モデルが最も高い精度であり，逆に実験 1 で最も精度が高かった Single モデルが精度が一番低かった．これは，実験 1 では意見を表す表現が似ていることから高精度に分類できたものの，実験 2 で用いた異なるドメインのデータに対しては学習時のデータでの意見を表す表現に依存してしまい，上手く分類が行えなかったと考えられる．この結果により，トレーニングデータと異なるドメインのデータに対しての感情極性分類において敵対的不変表現学習の有効性が示唆された．

5.3 実験 3 の結果

実験 3 では，実験 1 で学習したモデルに対して Twitter データセットで追加学習を行い，クロスバリデーションを行うことによって各モデルの精度を比較した．追加学習の際のハイパーパラメータを表 6 に示す． 実験 3 の各モデルの感情極性分類の精度を表 7 に示す．また，3 つのモデルの Twitter データセットに対する感情極性分類の結果の Confusion Matrix を可視化したものが以下の図 9a,9b,9c である．

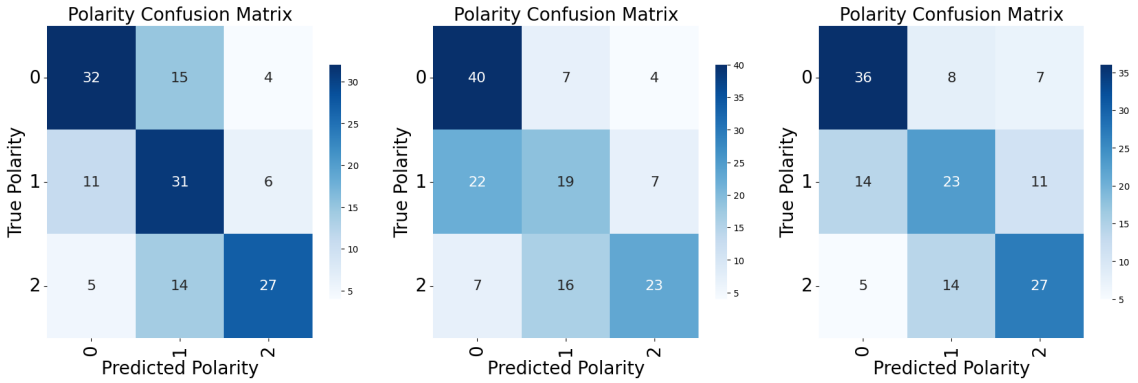
表 7 の通り，実験 2 と同様 GRL+Multi モデルが Accuracy，Macro-F1 score ともに最も精度が高いという結果になった．実験 2 の結果は敵対的不変表現学習の有効性が示唆されたものの，Accuracy が 0.400 から 0.462 と決して高いとは言えない数値であった．

表 6: モデルのハイパーパラメータ

	最適化アルゴリズム	学習率	バッチサイズ	エポック数
GRL+Multi	RMSProp	1e-05	8	9
Multi	RMSProp	1e-05	8	8
Single	RMSProp	1e-05	8	8

表 7: 実験 3 の結果

	Accuracy	Macro-F1
GRL+Multi	0.625	0.615
Multi	0.569	0.529
Single	0.597	0.576



(a) GRL+Multi (b) Multi (c) Single

図 9: 実験 3 の Confusion Matrix

そこで，対象としたいドメインにおける学習データが少ない場合における各モデルの感情極性分類の精度を比較することを目的として実験 3 を行った．結果としていずれの 3 つのモデルも実験 2 よりも精度は向上したが，各モデル間を比較すると，実験 2 と同様に GRL+Multi モデルが最も精度が高かった．この実験 3 により，対象としたいドメインにおける学習データが少ない場合において敵対的不変表現学習が有効であることが示唆される結果になった．

6　おわりに

本研究では，ABSA 研究用の一般的なデータセットが不足しているという課題に対して，敵対的不変表現学習による未知のドメインのデータセットに対する ABSA を検討した．この章では実験の結果を受けた結論と課題，今後の展望について述べる．

6.1　結論

本研究では，分析対象としたいドメイン（ターゲットドメイン）のラベル付きデータが不足しているという課題に対して，既存のデータセットを用いた未知のドメインのデータを対象とする ABSA における敵対的不変表現学習の有効性を検討した．結果として，学習時と異なるドメインのデータセットを対象とする場合や対象としたいドメインの学習データが少ない場合において，敵対的不変表現学習が有効であることが示唆された．

6.2　課題と今後の展望

この節では、本研究で行った実験の結果を受けた課題や今後の展望について述べる。

まず，4.1.1 節で述べた通り，一つのアスペクト用語と文章のペアに対して複数のアスペクトカテゴリと感情極性が割り当てられているサンプルに関しては今回は分析対象から除いたが，こういったケースの場合も分類できるようにすることは今後の課題といえる．

また，BERT 自体の持つ課題が挙げられる．文章による意見の表現には，「賛成だ」「反対だ」と言ったような直接的な表現の他に，例えば否定的な意見を主張する際に肯定的な表現を用いる皮肉的な表現などの修辞法といった表現がある．このような場合は実際に示したい意見とは反対の表現の単語を用いることがあるので，上手く分類できない可能性がある．実際に，感情分類において，皮肉的な表現が用いられている文章を計算機が正しく理解するのは難しいことが指摘されており，BERT を用いて皮肉文を検出するための手法についての先行研究も行われている [20]．同様に，主張したい意見とは反対の内容を疑問の形で述べることで断定を強調する反語表現も修辞法の一種であり，これも主張したい意見とは逆の表現が文章中に含まれるために正しく理解することは難しいと推察される．例えば，「増税に一体どんなメリットがあるのですか？」という文章が存在したとする．この場合，文章の意図としては「増税」に対して否定的な意思を表していることが読み取れるが，文章中に否定的な表現は用いられておらず，むしろ「メリット」という肯定的な表現

があることがわかる．実際，今回の実験 3 において，これと同じような構文のテストデータに対して，３つのモデル全てが誤った予測をした．従ってこのような問題に対しては今回行っていない対策を講じることによってこれらの誤りを防いだり，別の事前学習済みモデルを利用することによってこれらの問題が解消される可能性がある．

加えて，今回敵対的不変表現学習として DANN を用いたが，将来的には他の先進的な手法を採用することで、さらなる精度向上の可能性が考えられる。以上の点に関して今後の研究課題とする．

謝辞

本研究を進めるにあたり多くの方々からのご協力・ご指導を頂きました．指導教員である滋賀大学の南條浩輝教授には多大なご指導とご支援を頂きましたこと，心より深く感謝申し上げます．また，本研究を進行するにあたり，多大なる助言とご支援いただいた南條ゼミおよび市川ゼミの皆様に深く感謝いたします．ありがとうございました．

参考文献

[1] 張懿陽, ラファウ・ジェプカ, 荒木健治. BERT モデルと補助文自動生成に基づいた日本語アスペクトベース感情分析の精度向上. ことば工学研究会 : 人工知能学会第 2 種研究会ことば工学研究会資料, Vol. 70, pp. 83–89, 12 2022.

[2] 樊惠, 杉本徹. RGCN と Attention を用いたアスペクトベースの感情分析. 第 84 回全国大会講演論文集, Vol. 2022, No. 1, pp. 681–682, 02 2022.

[3] Yiming Zhang, Min Zhang, Sai Wu, and Junbo Zhao. Towards Unifying the Label Space for Aspect- and Sentence-based Sentiment Analysis. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 20–30, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of machine learning research*, Vol. 17, No. 59, pp. 1–35, 2016.

[5] Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[6] Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. Aspect Sentiment Quad Prediction as Paraphrase Generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9209–9219, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] 三浦義栄, 赤井龍一, 渥美雅保. 文中の複数アスペクトのセンチメント分析のための自己注意ニューラルネットワーク. 人工知能学会全国大会論文集 第 34 回 (2020), 3Rin441, pp. 1–4. 一般社団法人 人工知能学会, 2020.

[8] Joachim Wagner, Piyush Arora, Santiago Cortes, Utsab Barman, Dasha Bogdanova, Jennifer Foster, and Lamia Tounsi. Dcu: Aspect-based Polarity Classification for SemEval Task 4. 2014.

[9] Thien Hai Nguyen and Kiyoaki Shirai. PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2509–2514, 2015.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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. 380–385, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[12] 赤井龍一, 渥美雅保. 自己注意機構を利用したアスペクトベースの感情分析の日本語文への適用. 人工知能学会全国大会論文集 第 33 回 (2019), 3Rin213, pp. 1–2. 一般社団法人 人工知能学会, 2019.

[13] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 27–35, Dublin, Ireland, 2014. Association for Computational Linguistics.

[14] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The Multilingual Amazon Reviews Corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in neural information processing systems*, Vol. 30, , 2017.

[16] 岡崎直観, 荒瀬由紀, 鈴木潤, 鶴岡慶雅, 宮尾祐介. IT Text 自然言語処理の基礎. オーム社, 2022.

[17] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528, 2011.

[18] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing Across Domains via Cross-Gradient Training. *arXiv preprint arXiv:1804.10745*, 2018.

[19] Sarah Erfani, Mahsa Baktashmotlagh, Masud Moshtaghi, Xuan Nguyen, Christopher Leckie, James Bailey, and Rao Kotagiri. Robust Domain Generalisation by Enforcing Distribution Invariance. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp. 1455–1461. AAAI Press, 2016.

[20] 畑玲音, 森野穣, 松下光範. 皮肉文検出のための皮肉状況の検出. 人工知能学会全国大会論文集 第 37 回 (2023), 3M1GS1003, pp. 1–4. 一般社団法人 人工知能学会, 2023.