卒 業 研 究 報 告 書

題　　目

皮肉データセットを用いた深層学習による
皮肉推定手法

研 究 グ ル ー プ　　第 1 研究グループ

指 導 教 員　　森 直樹 教授

令和　3　年（　　2021　年 ）度卒業

（No.　　1181201087　　）　　多田 瑞葵

大阪府立大学工学域電気電子系学類情報工学課程

皮肉データセットを用いた深層学習による皮肉推定手法

第 1 グループ　多田 瑞葵

1. はじめに

皮肉表現とは，文章の文字通りの意味と書き手の意図とが異なる修辞表現である．皮肉表現の理解は，人間にとっては凡そ容易である一方で，機械にとっては未だ困難である．そのため機械による皮肉推定は，文章の理解や感情分析等の分野でも課題となっている．

本研究では皮肉表現にドメイン依存性が存在するという仮説のもと，SARC データセット [1] を用いて文章の話題ごとの皮肉推定に取り組んだ．また文脈を考慮できるモデルとして深層言語処理モデル BERT [2] を使用した．

2. BERT

Bidirectional Encoder Representations from Transformers (BERT) は，Transformer による双方向のエンコーダーを用いた言語モデルである．文章を文頭と文末の双方向から学習することによって文脈を理解することを可能としている．様々な事前学習済みモデルが発表されており，ファインチューニングすることで各タスクに応用可能である．本研究では，Wikipedia の英語記事で事前学習されたモデルである bert-base-uncased[1] を使用した．

BERT は文の入力に対して，入力された文および文に含まれる各単語に対応する分散表現を出力する．また特殊トークンが用意されており，文頭の“[CLS]”トークンに対応する分散表現を入力された文全体の分散表現とみなして分類問題に使用することができる．

3. SARC データセット

本研究では，自己注釈付き Reddit コーパス（Self-Annotated Reddit Corpus; SARC）を使用した．このデータセットは Reddit[2] に投稿されたコメントから構築されており，皮肉ラベルが付与されている．各コメントには，投稿者の識別子，subreddit と呼ばれる投稿トピック，ユーザ評価，投稿日時，親投稿の情報も付与されている．親投稿とは，最初の投稿から各コメントに至るまでの一連の投稿であり，全てのコメントは 1 つ以上の親投稿を保持している．図 1 にデータの例を示す．本研究では，皮肉推定の対象となるコメントと，その直接の投稿元である 1 つの親投稿のコメントを実験に使用した．

本研究では SARC 2.0 main の均衡データセットから subreddit ごとのデータセットを作成し，テストデータが 1,500 件以上であった 4 つの subreddit（politics, AskReddit, worldnews, pcmasterrace）を対象に皮肉推定をした．また main からデータをランダムサンプリングした random データセットを作成し，比較対象とした．

4. 実験

データセットの各コメントが皮肉か非皮肉かの二値分類をした．まず，コメントを BERT に入力し，出力された分散表現から“[CLS]”トークンに対応する分散表現を取得した．そして取得した分散表現を線形層に入力し，皮肉か非皮肉かを推定した．その際，訓練データを用いて BERT をファインチューニングし，テストデータに対する精度を確認した．

[1] https://huggingface.co/bert-base-uncased

[2] https://www.redditinc.com

subreddit : AskReddit	
親投稿	What's something that's completely legal, but that pisses you off when you see someone doing it?
	People reclining their seats fully on aeroplanes.
非皮肉	There is a special hell for these people.
皮肉	yes people who spell airplanes wrong are just horrible

図 1: データセットのデータ例

表 1: 実験結果

subreddit	Accuracy	Precision	Recall	F1 score
politics	**0.730**	**0.722**	**0.748**	**0.735**
AskReddit	0.619	0.630	0.580	0.604
worldnews	**0.690**	0.655	**0.803**	**0.721**
pcmasterrace	0.648	0.635	**0.698**	**0.665**
random	0.653	0.664	0.618	0.640

5. 結果と考察

表 1 にテストデータの実験結果を示す．太字の項目は random の評価値を上回ったことを示している．politics, worldnews, pcmasterrace の F1 値は random を上回り，politics ではどの評価値においても random を上回る結果となった．このことから，ドメインごとに皮肉を学習，推定することで，皮肉推定の精度が向上する場合があると考えられる．一方で，AskReddit ではどの評価値においても random を下回る結果となった．このことから，AskReddit における皮肉推定が困難であり，データセット全体の推定の際に精度を下げる要因の一つとなることが考えられる．

6. まとめと今後の課題

本研究では，BERT を用いた皮肉推定に取り組み，皮肉表現のドメイン依存性について確認した．皮肉表現にドメイン依存性が存在するという仮説のもと，subreddit ごとのデータセットを作成して皮肉推定をし，精度を比較した．結果として，subreddit ごとに皮肉推定することで精度が向上したものと低下したものがあり，ドメインを考慮して皮肉推定する手法の有効性と，皮肉推定の困難なドメインの存在を示すこととなった．

今後の課題として，ドメインごとの弱学習器を用いたアンサンブル学習によって皮肉推定に取り組むことや，皮肉推定に使用する文脈情報やメタデータを拡大することが挙げられる．

参考文献

[1] M. Khodak, N. Saunshi, and K. Vodrahalli. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

目 次

1 はじめに **1**

2 関連研究 **2**

2.1 皮肉表現の心理的・言語的研究 2

2.2 皮肉推定のためのデータセット 2

2.3 SARC データセットを用いた皮肉推定 3

3 要素技術 **5**

3.1 BERT 5

3.2 Optuna 8

3.3 SARC データセット 8

4 実験 **11**

4.1 実験 1 11

4.2 結果と考察 13

4.3 実験 2 15

4.4 結果と考察 15

4.5 誤識別したデータの分析 17

4.6 t-SNE を用いた次元圧縮による可視化 22

5 まとめと今後の課題 **25**

謝辞 **26**

参考文献 **27**

図 目 次

3.1 BERT のファインチューニングの概要図（文献 [10] Figure 1 より引用） . 7
3.2 Reddit に投稿されたコメントの例（文献 [7] Figure 1 より引用） 9
3.3 SARC データセットのデータの例 10
4.1 実験モデルの概要 12
4.2 t-SNE による可視化 (politics) 23
4.3 t-SNE による可視化 (AskReddit) 23
4.4 t-SNE による可視化 (worldnews) 23
4.5 t-SNE による可視化 (pcmasterrace) 24
4.6 t-SNE による可視化 (random) 24

表 目 次

3.1 subreddit ごとのデータ数の分布 10
4.1 各データセットのデータ数 12
4.2 実験パラメータ 13
4.3 実験結果（実験 1） 14
4.4 実験結果（実験 2） 15
4.5 正解したモデルの数とラベルの内訳 16
4.6 全てのモデルで誤識別したデータの内訳 16
4.7 誤識別したデータの例（politics） 18
4.8 誤識別したデータの例（AskReddit） 19
4.9 誤識別したデータの例（worldnews） 20
4.10 誤識別したデータの例（pcmasterrace） 21

1　はじめに

計算機による文章の理解において，様々な修辞表現の理解が重要な課題となっている．修辞表現の一つである皮肉表現は，文章の文字通りの意味と書き手の意図とが異なる言語表現である．人間の場合，ある文章が皮肉であるか否かを推定することは凡そ容易である．一方で，文章を文字情報に基づいて処理する場合，皮肉推定は困難である．そのため，計算機による皮肉推定は，文章の理解や創作，感情分析等の分野でも解決すべき課題となっている．

皮肉の表現の幅は広く，全ての皮肉表現を一括りに扱って推定することは良策ではない．そこで本研究では，皮肉推定の前段階として，文章の話題ごとの皮肉推定に取り組んだ．ソーシャルメディアサイト Reddit には，subreddit と呼ばれる投稿トピックがある．この subreddit のそれぞれについて皮肉推定をし，その結果を分析することにより，文章の話題ごとの皮肉推定の有効性を検証した．また，モデルには深層言語モデル BERT を使用することで，文脈を考慮した皮肉推定に取り組んだ．

本稿では，まず第 2 章で皮肉推定に関する関連研究を紹介する．続いて第 3 章で本研究で使用した要素技術とデータセットについて記述する．第 4 章では本研究で取り組んだ実験について記述する．

2 関連研究

本章では皮肉表現の心理的・言語的側面の研究と，計算機による皮肉推定に関連する研究を紹介する．

2.1 皮肉表現の心理的・言語的研究

人間にとって，ある発話や文章が皮肉であることを理解するのは容易であるが，その現象を説明することは難しい [1]．例えば，相手がその文章によって伝えたいことを理解する際に，その文章に皮肉表現が含まれていることを認識する必要はないことから，人間は無意識に皮肉を理解できると言える．さらに話し手や書き手が皮肉を意図していなくても，受け取る側によって皮肉であると解釈されることがある．これらは皮肉表現の心理的側面に焦点を当てた説明である．言語的側面からは，皮肉表現は語用論的原則の違反によって説明されることがある [2]．これは皮肉表現が文章の文字通りの意味と書き手の意図とが異なる言語表現であるという説明を含んでいる．以上のように皮肉表現は多角的に研究されているが，全ての皮肉表現を説明しきる理論は未だ存在しない [3]．

2.2 皮肉推定のためのデータセット

皮肉推定の研究に利用されているデータセットにはソーシャルネットワーキングサービスに投稿された文章を収集したものが多い．特に Twitter [1] に投稿された文章（ツイート）を収集し，皮肉推定に使用している研究は多く見られる．Riloff らは皮肉を明示するハッシュタグ“#sarcasm”，“#sarcastic”を手がかりに皮肉データを収集した [4]．一方で Ghosh & Veale はハッシュタグだけを手がかりとするとタグ付けされていないデータを収集できないことを指摘している [5]．彼らは Twitterbot を用いて，皮肉と思われるツイートに対して投稿者に直接確認を取ることで皮肉データを収集した．

[1] https://about.twitter.com/en

News Headlines Dataset For Sarcasm Detection [6] は，ウェブニュースサイト TheOnion [2] と HuffPost [3] からヘッドラインを収集したデータセットである．TheOnion は時事問題に対して皮肉的なヘッドラインを付けることで有名である．そのため TheOnion から収集したヘッドラインデータに皮肉ラベルを，HuffPost から収集したデータに非皮肉ラベルを付与している．ヘッドラインのみのデータであるため各データに文脈が存在せず，1 つの文章で理解可能な皮肉表現となっている．

2.3 SARC データセットを用いた皮肉推定

本研究では SARC データセット [7] を用いて皮肉推定をした．このデータセットの特徴として Twitter データセットやヘッドラインデータセットと比べて大規模であることが挙げられる．またメタデータも充実しており，近年の皮肉表現に関する研究に大きく貢献しているデータセットである．本節ではこのデータセットを用いた関連研究を紹介し，SARC データセットについては後の章で詳述する．

2018 年に，ソーシャルメディアサイトへの投稿文章に対する皮肉推定のモデルとして CASCADE [8] が提案され，state-of-the-art を達成した．このモデルは文章の情報に加えて，投稿者と投稿トピックを元にした情報を特徴量として利用する．具体的には，Convolutional Neural Network (CNN) を用いて取得した文章の分散表現 $\overrightarrow{c}_{i,j}$，投稿者の過去の投稿文章から取得した分散表現 $\overrightarrow{u}_i$，投稿トピックにおける投稿文章から取得した分散表現 $\overrightarrow{t}_j$ を連結して分類に用いている．結果として，文章の情報と投稿者や投稿トピックを元にした情報を併用することで，皮肉推定の精度が向上することを示した．

文章の情報のみを使用した研究もされている．2019 年に Pelser & Murrell [9] は，投稿者などの情報はプライバシー設定やデータ欠損により常に利用可能とは限らないことを指摘し，56 層の深層ネットワークによる文章のみを利用した推定手法を提案した．結果として，上述した CASCADE の性能を上回ることはできなかったものの，文章以外の情報を利用した他の既存手法に

[2] https://www.theonion.com
[3] https://www.huffpost.com

匹敵する性能を示した.

本研究では文章情報からの皮肉推定に重きを置くため投稿者の情報は使用しない. また subreddit の情報を使用することで話題ごとの皮肉推定の有効性を確認する.

3 要素技術

本章では，本研究で使用した要素技術とデータセットについて記述する．

3.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) [10] は，2018 年 10 月に Jacob Devlin らが発表した Transformer [11] による双方向のエンコーダーを用いた言語モデルである．公開当時に様々な自然言語処理タスクで state-of-the-art の性能を示した．文章を文頭と文末の双方向から学習することによって文脈の理解を可能としている．

BERT は大規模コーパスを用いて事前学習することでモデルの性能を向上させている．事前学習には，[MASK] トークンに当てはまる単語を予測する Masked Language Modeling と，入力された 2 文が意味的に連続するかを予測する Next Sentence Prediction の 2 つのタスクが用いられる．

BERT は様々な事前学習済みモデルが発表されており，ファインチューニングすることで各タスクに応用可能である．図 3.1 はその概略図を表している．本研究では，英語の事前学習済みモデルである bert-base-uncased [4] を使用した．このモデルは本のデータセットである BookCorpus [12] と Wikipedia [5] の英語記事を用いて事前学習している．BookCorpus は約 8 億語，Wikipedia は約 2 億 5 千万語を含んでいる．またこのモデルは英語の大文字と小文字の区別をしない．モデルの層数は 12 層，隠れ層の次元数は 768 次元，最大入力長は 512 トークンである．

BERT に文を入力する際は，まず tokenizer を用いて文をトークンに分割し，ID に変換する．本研究では BERT モデルと同じ bert-base-uncased の tokenizer を使用した．この tokenizer は WordPiece [13] モデルを用いて，文を単語よりも細かいサブワードに分割する．これによって未知語を分解し，削減することができる．BERT には特殊トークンが用意されており，入力の際は変換した ID 列の先頭に [CLS] トークンを，各文の末尾に該当する位置に [SEP] トークンを挿入する．入力に対して BERT は各トークンに対応する分

[4] https://huggingface.co/bert-base-uncased
[5] https://en.wikipedia.org/wiki/Main_Page

散表現を出力する．このとき [CLS] トークンに対応する分散表現は入力された文全体の特徴を捉えており，分類問題に使用することができる．

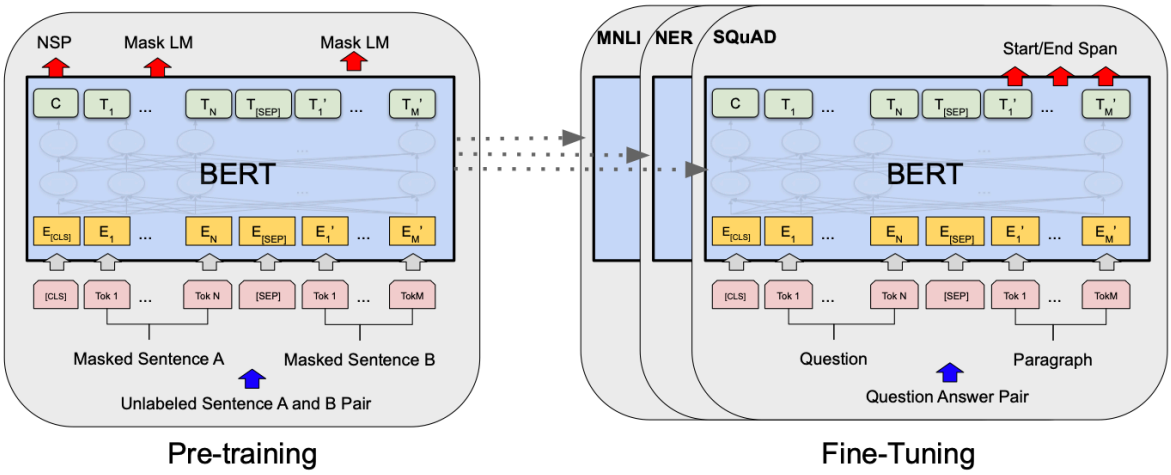


図 3.1: BERT のファインチューニングの概要図（文献 [10] Figure 1 より引用）

3.2 Optuna

Optuna [14] は，オープンソースのハイパーパラメータ自動最適化フレームワークである．Tree-structured Parzen Estimator というベイズ最適化アルゴリズムを用いて，過去の試行に基づいて有望そうな領域を推定し再度試行する．これを繰り返すことで最適なハイパーパラメータの値を自動的に発見する．使用する際は目的関数を定め，その値がより大きくまたは小さくなるように推定を進める．Optuna の主な特徴として，Define-and-Run スタイルの API，学習曲線を用いた試行の枝刈り，並列分散最適化が挙げられる．

3.3 SARC データセット

本研究では，皮肉の研究や皮肉検出システムの学習･評価のための大規模コーパスである，自己注釈付き Reddit コーパス（Self-Annotated Reddit Corpus; SARC）[7] を使用した．このデータセットはソーシャルメディアサイト Reddit に投稿されたコメントから構築されている．図 3.2 に Reddit に投稿されたコメントの例を示す．図の上のコメントは投稿元であり，下のコメントはその投稿元に反応したものである．

図 3.3 に SARC データセットに含まれるコメントの例を示す．各コメントには文章だけではなく，投稿者の識別子，subreddit と呼ばれる投稿トピック，ユーザ評価（good/bad），投稿日時，親投稿の情報も付与されている．親投稿とは，最初の投稿から各コメントに至るまでの一連の投稿であり，全てのコメントは 1 つ以上の親投稿を保持している．SARC データセットはこのようなやりとりのうち，末端のコメントのみに皮肉か非皮肉かのラベルが付与されている．

Reddit へ投稿する際，ユーザは自分のコメントが皮肉を意図していることを明示する記号“/s”を使用する．このことを利用し，投稿コメントに“/s”を含むことを必要条件としてデータに皮肉ラベルが付けられている．すなわち SARC データセットは，投稿者が皮肉を意図しているかどうかを判断基準としてラベル付けしていると言える．ただしこの記号は，データセットに含まれる文章からは除去されている．

またこのデータセットには，全ての subreddit からデータを抽出した main データセットと，politics の subreddit からデータを抽出した pol データセットがある．それぞれに対して均衡データセット（balanced）と不均衡データセット（unbalanced）が用意されており，全部で 4 つのデータセットが用意されている．均衡データセットには，同じ親投稿に対して 2 つのコメントが存在する．そして 2 つのコメントの内訳は，皮肉データと非皮肉データが 1 つずつである．本研究では SARC 2.0 main balanced [6] データセットを使用した．このデータセットに含まれる訓練データは 257,082 件，テストデータは 64,666 件である．

次に，使用したデータセットの subreddit の情報について述べる．訓練データに含まれる subreddit は 4,902 種類，テストデータに含まれる subreddit は 2,666 種類，訓練データとテストデータに共通する subreddit は 2,194 種類であった．表 3.1 に subreddit ごとのデータ数を示す．データ数が最も多い subredit は politics で，訓練データ 13,668 件，テストデータ 3,406 件であった．またデータ数が最も少ない subreddit は複数あり，データ数は 2 件であった．

図 3.2: Reddit に投稿されたコメントの例（文献 [7] Figure 1 より引用）

[6] https://nlp.cs.princeton.edu/SARC/2.0/main/

subreddit : AskReddit	
親投稿	What's something that's completely legal, but that pisses you off when you see someone doing it?
	People reclining their seats fully on aeroplanes.
非皮肉	There is a special hell for these people.
皮肉	yes people who spell airplanes wrong are just horrible

図 3.3: SARC データセットのデータの例

表 3.1: subreddit ごとのデータ数の分布

データ (件)			subreddit (種類)	
(以上)	-	(未満)	訓練	テスト
	-	100	4,575	2,567
100	-	500	248	77
500	-	1,000	40	13
1,000	-	2,000	17	6
2,000	-	3,000	6	1
3,000	-	4,000	7	2
4,000	-	5,000	4	0
5,000	-	6,000	1	0
6,000	-	7,000	0	0
7,000	-	8,000	1	0
8,000	-	9,000	0	0
9,000	-	10,000	1	0
10,000	-		2	0
合計			4,902	2,666

4 実験

本章では，本研究で取り組んだ実験について記述する．

4.1 実験 1

最初に，文章の話題ごとに皮肉か非皮肉かを推定する実験に取り組んだ．モデルには BERT を使用し，データセットには SARC 2.0 main balanced データセットを使用した．

まず subreddit ごとにデータを抽出し，新たに実験用データセットを構築した．その際にテストデータが 1,500 件以上取得できたものを選択し，politics [7]，AskReddit [8]，worldnews [9]，pcmasterrace [10] の 4 つの subreddit のデータセットを実験に使用した．各 subreddit はそれぞれ，politics は米国の時事ニュースや政治的なニュースを，AskReddit はユーザ同士の質問や回答のやりとりを，worldnews は米国以外の時事ニュースや政治的なニュースを，pcmasterrace はコンピュータやゲームに関する話題を扱っている．比較のため，全ての subreddit からランダムにデータをサンプリングした random データセットを構築し，こちらも実験に使用した．このとき同じ親投稿を持つ皮肉データと非皮肉データのペアは崩さないように抽出した．random データセットに含まれる subreddit を確認したところ，訓練データは 1,257 種類，テストデータは 569 種類の subreddit から構成されていた．表 4.1 に各データセットのデータ数を示す．なお各データセットには皮肉データと非皮肉データは同数含まれている．

図 4.1 にモデルの概要を示す．まず入力の先頭に [CLS] トークンを，各コメントの末尾に [SEP] トークンを付けて BERT に入力し，分散表現を取得した．分散表現のうち，文頭の [CLS] トークンに該当する分散表現を線形層に入力して，皮肉か非皮肉かで二値分類した．訓練データでモデルを学習し，テストデータでその性能を評価した．評価指標には Accuracy（正解率），

[7] https://www.reddit.com/r/politics/
[8] https://www.reddit.com/r/AskReddit/
[9] https://www.reddit.com/r/worldnews/
[10] https://www.reddit.com/r/pcmasterrace/

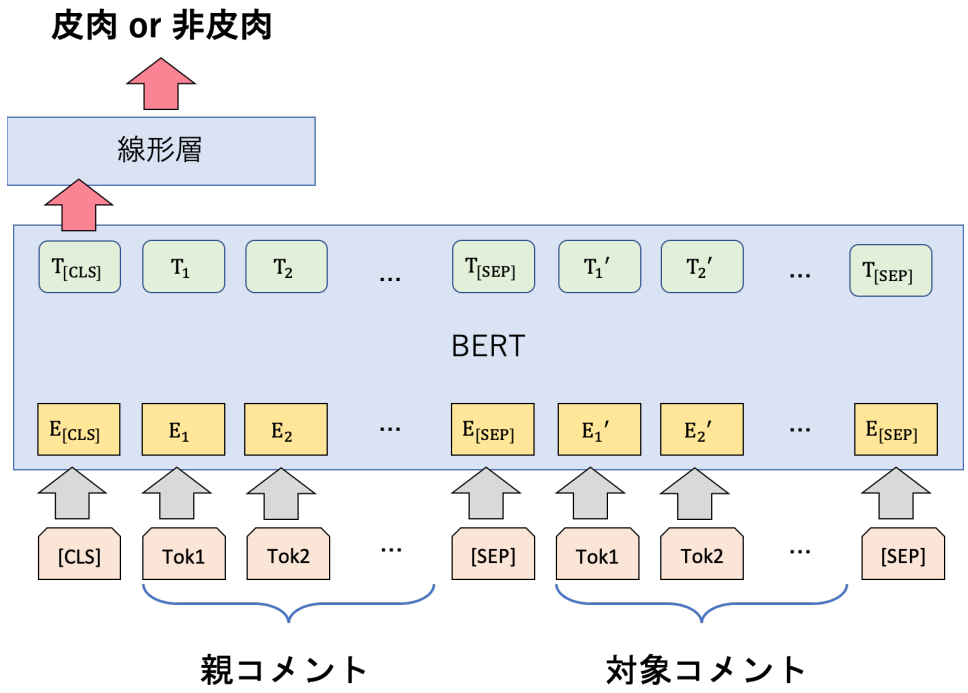


図 4.1: 実験モデルの概要

Precision（適合率），Recall（再現率），F1 値を使用した．

表 4.2 に実験のパラメータを示す．BERT と線形層の学習率は予備実験によって値の範囲を定め，Optuna によって探索した．目的関数にはテスト時の Accuracy を設定し，この値を最大化するように探索した．学習時は BERT の全層をファインチューニングし，線形層も重みを更新した．

表 4.1: 各データセットのデータ数

データセット	訓練 (件)	テスト (件)
politics	13,668	3,406
AskReddit	11,660	3,006
worldnews	9,444	2,246
pcmasterrace	7,400	1,772
random	14,000	3,500

表 4.2: 実験パラメータ

BERT	
モデル	bert-base-uncased
入力層次元数	512
出力層次元数	768
学習率（探索範囲）	1×10^{-6} 〜 1×10^{-4}
線形層	
入力層次元数	768
出力層次元数	2
学習率（探索範囲）	1×10^{-6} 〜 1×10^{-4}
学習	
エポック数	20
バッチサイズ	16
損失関数	Cross Entropy Loss
最適化関数	Adam
	$\begin{pmatrix} \text{learning rate} = 0.001 \\ \beta_1 = 0.9,\ \beta_2 = 0.999 \end{pmatrix}$

4.2　結果と考察

表 4.3 に実験結果を示す．太字の項目はベースラインである random データセットでの評価値を上回ったことを表している．表より random データセットでの評価値を上回った項目が多いことが分かる．このことは，subreddit ごとに皮肉推定をすることが，推定精度向上に有効に働いていると考えられる．また politics データセットでは全ての評価指標で random データセットでの皮肉推定精度を上回った．一方で AskReddit データセットでは全ての評価指標で random データセットでの皮肉推定精度を下回った．このことから，politics データセットの皮肉表現の特徴を上手く捉えることができ，正しく推定できていると考えられる．反対に，AskReddit データセットは何らかの理由で皮肉推定が困難であったと考えられる．その理由としては以下のことが

表 4.3: 実験結果（実験 1）

dataset	Accuracy	Precision	Recall	F1 score
politics	**0.730**	**0.722**	**0.748**	**0.735**
AskReddit	0.619	0.630	0.580	0.604
worldnews	**0.690**	0.655	**0.803**	**0.721**
pcmasterrace	0.648	0.635	**0.698**	**0.665**
random	0.653	0.664	0.618	0.640

考えられる．

- 皮肉文章と非皮肉文章が類似している
- 含まれる話題が多く，学習によって捉えきれない
- 皮肉・非皮肉ラベル付けが適切でない
- 訓練データとテストデータの違いが大きい

また worldnews データセットでは Recall の値が高くなった．このことは，データセットに含まれる皮肉データに対して皮肉であると正しく予測したものが多いことを表している．反対に，Precision の値は random データセットでの評価値を下回り，このことは，皮肉であると予測したデータのうち真のラベルが皮肉であったものが少なかったことを表している．すなわち，真のラベルが非皮肉であるものに対して皮肉であると予測したものが多いことを意味している．

4.3 実験 2

次に，politics，AskReddit，worldnews，pcmasterrace のデータセットに対して，学習時とテスト時で異なる subreddit のデータセットを使用して皮肉推定をした．実験の手順は実験 1 と同様である．

4.4 結果と考察

表 4.4 に実験結果を示す．各項目の数値は Accuracy の値である．表中の太字の項目は，各テストデータに対して最も高い評価値であることを表している．結果として，politics，AskReddit，pcmasterrace の 3 つのデータセットでは，訓練データとテストデータに同じ subreddit のデータセットを使用した際に最も評価値が高くなった．これは subreddit ごとの皮肉推定の有効性を裏付ける結果である．一方で，worldnews データセットでは，politics データセットで学習したモデルを用いて識別した場合に最も評価値が高くなった．このことから，politics と worldnews のデータセットに含まれるデータが類似していることが考えられる．また表 4.3 にあるように，実験 1 で，worldnews データセットでの学習では Recall が高く Precision が低い傾向が見られたことも踏まえて，politics データセットでの学習の方が安定していることも原因の 1 つではないかと考えられる．

この結果について，各データが 5 つのモデルのうちいくつで正解したかを調べた．表 4.5 に結果を示す．各項目の数値はデータ数である．なお正解したモデルの数が 0 とは，5 つ全てのモデルで誤識別したことを表している．

表 4.4: 実験結果（実験 2）

テストデータ	訓練データ			
	politics	AskReddit	worldnews	pcmasterrace
politics	**0.730**	0.640	0.690	0.637
AskReddit	0.586	**0.619**	0.591	0.591
worldnews	**0.723**	0.652	0.690	0.646
pcmasterrace	0.613	0.583	0.598	**0.648**

表 4.5: 正解したモデルの数とラベルの内訳

ラベル	正解したモデルの数						合計
	5	4	3	2	1	0	
皮肉	2,255	1,220	966	906	905	713	6,965
非皮肉	2,087	1,722	1,206	792	610	548	6,965
合計	4,342	2,942	2,172	1,698	1,515	1,261	13,930

表 4.6: 全てのモデルで誤識別したデータの内訳

データの所属	ラベル			
	皮肉		非皮肉	
politics	102	(6.0%)	160	(9.4%)
AskReddit	260	(17.3%)	82	(5.5%)
worldnews	56	(5.0%)	115	(10.2%)
pcmasterrace	107	(12.1%)	60	(6.8%)
random	188	(10.7%)	131	(7.5%)

表 4.5 より，5 つ全てのモデルで正解したデータが最も多く，また多くのデータは 1 つ以上のモデルで正解できていることが分かる．しかし全てのモデルで誤識別したデータも存在し，その件数は皮肉データ 713 件，非皮肉データ 548 件であった．これはそれぞれ皮肉データの 10.2%，非皮肉データの 7.8% にあたる．

次に，5 つ全てのモデルで誤識別したデータの内訳を調べた．表 4.6 に結果を示す．各項目の数値はデータ数で，括弧内は各データセットの皮肉・非皮肉データ数における割合である．AskReddit データセットに含まれる皮肉データは 1,503 件であり，そのうち全てのモデルで誤識別したデータは 260 件で 17.3% にあたる．これは他よりも高い割合であり，AskReddit データセットに含まれる皮肉データを正しく識別することが困難であったことを表している．このことは表 4.3 にあるように，実験 1 での AskReddit データセットでの評価値が低かったことと一致する．

4.5 誤識別したデータの分析

実験 2 で全てのモデルで誤識別したデータについて，各 subreddit ごとに内容を確認し定性的に評価した．表 4.7 から表 4.10 に誤識別したデータの例を示す．例 5，6，9，10，13 は特に皮肉らしいと感じられるが，モデルにとっては非皮肉であると識別された例である．例 5，10 は自分の考えを述べている文章であるが，投稿者は文章の文字通りの考えは持っていないと判断できる．例 6，13 は投稿者が嫌味や揶揄を意図していることが明らかであるが，その判断には文章の深い理解が求められる．例 9 はバベルの塔がどのような物語であるかを知っていて初めて皮肉であると理解できる．これらの例は文章の文字通りの情報からでは特に皮肉推定が困難であると考えられる．そのため皮肉推定精度の向上のためには，このような皮肉表現に対しても有効な推定手法を取り入れることが必要である．

また例 8，11 は，人間には皮肉であると判断できるように思われるが，非皮肉ラベルが付与されているデータである．例 8 は，死刑を廃止してはいけないという文章によって，レゴを床に置きっぱなしにすることは危険なことであり投稿者にとって許し難いことであることを伝えていると解釈できる．例 11 は，北朝鮮ではインターネットへの接続が制限されていることを，DDOS 攻撃を受けたコンピュータの台数に言及することで揶揄していると解釈できる．この 2 つの例は，投稿された文章の文字通りの意味と投稿者が伝えたいことが異なっており，皮肉表現の一例であると言えるため，SARC データセットにおけるラベル付けの誤りによる誤識別ではないかと考えられる．SARC データセットは，皮肉を明示する記号“/s”がコメント内に含まれていることを必要条件として皮肉ラベルを付けている．そのため投稿者が皮肉を意図していても，この記号が含まれていないために非皮肉ラベルが付けられているデータが存在する．このように誤識別したデータの中には，人間にとっては皮肉であるように感じるが，データには非皮肉ラベルが付与されているものが少なくなかった．このことは SARC データセットを利用する点での課題であるが，本研究においてモデルの実際の皮肉推定性能が算出された数値よりも高いことが期待できる．

表 4.7: 誤識別したデータの例（politics）

例 1	非皮肉と誤識別
親投稿	Obama To Visit A Mosque For The First Time As President
	オバマは大統領として初めてモスクを訪問する
皮肉	...except for the secret one in the basement of the White House of course.
	... もちろん，ホワイトハウスの地下にある秘密のモスクを除いて.
例 2	非皮肉と誤識別
親投稿	Bill Nye: Louisiana floods due to climate change
	ビル・ナイ「ルイジアナ州の洪水は気候変動が原因」
皮肉	Says who?
	誰が言ったの？
例 3	皮肉と誤識別
親投稿	Hillary Clinton has 1 Year to Live, says Medical School Professor
	ヒラリー・クリントンの余命は 1 年と医学部教授が発言
非皮肉	He’s not a medical professor because such people wouldn’t speak ill of Hillary.
	そういう人はヒラリーの悪口を言わないから医学部教授じゃないんだよね.
例 4	皮肉と誤識別
親投稿	Shaun King: Clinton should quit presidential race over email scandal
	ショーン・キング「クリントンはメールスキャンダルで大統領選を辞めるべき」
非皮肉	I’m sure she will get right on that.
	彼女はきっとすぐにそうするでしょう.

表 4.8: 誤識別したデータの例（AskReddit）

例 5	非皮肉と誤識別
親投稿	Reddit, what is your biggest achievement on Reddit?
	Reddit であなたの最大の功績は何ですか？
皮肉	I survived a month in this scary place.
	この怖い場所で1ヶ月生き延びたことです.
例 6	非皮肉と誤識別
親投稿	Police officers who don't use their turn signals.
	ウインカーを使わない警察官.
皮肉	In their defense, they're probably on their cell phone and typing on their laptop while steering the car with their knee, so they don't really have a hand free for the turn signal.
	彼らの言い分としては，おそらく携帯電話を使ったり，膝で車を操りながらラップトップを打ったりしているので，ウインカーを出す手が空いていないのでしょう.
例 7	皮肉と誤識別
親投稿	Instead of arguing about inequality, why don't we just kill the top 5% of the wealthiest and redistribute what they have to everybody as an annual check?
	不平等について議論する代わりに上位 5% の富裕層を殺して，彼らが持っているものを毎年の小切手としてみんなに再分配してはどうでしょう？
非皮肉	Learn to statistics, there will always be a top 5% .
	統計学によれば，上位 5% は常に存在します.
例 8	皮肉と誤識別
親投稿	I used to leave all my lego just laying around the house floor.
	レゴを床に置きっぱなしにしてた.
非皮肉	This is why we shouldn't abolish the death penalty.
	だから死刑を廃止しちゃダメなんだ.

表 4.9: 誤識別したデータの例（worldnews）

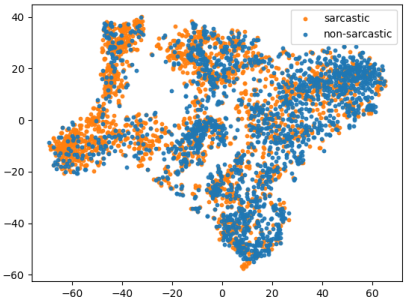
例 9	非皮肉と誤識別
親投稿	Magnetic Wormhole Created in Lab - “This device can transmit the magnetic field from one point in space to another point, through a path that is magnetically invisible,” said study co-author Jordi Prat-Camps, a doctoral candidate in physics at the Autonomous University of Barcelona in Spain.
	磁気ワームホールが研究室で作られる「この装置は空間のある点から別の点へ，磁気的に見えない経路で磁場を伝達できる」と，スペイン・バルセロナ自治大学の物理学博士候補である研究共著者ジョルディ・プラット・カンプス氏は述べた.
皮肉	Is this the new tower of babel?
	これは新たなバベルの塔か？
例 10	非皮肉と誤識別
親投稿	World Bank wants water privatized, despite risks
	世界銀行，リスクはあっても水の民営化を希望
皮肉	In that case, I would like to privatize air.
	それなら，空気の民営化を希望します.
例 11	皮肉と誤識別
親投稿	North Korea’s internet is offline; massive DDOS attack presumed
	北朝鮮のインターネットがオフラインに．大規模な DDOS 攻撃と推定される
非皮肉	All three North Korean computers connected to the internet are down!
	インターネットに接続されている北朝鮮のコンピューター 3 台全てがダウン！

表 4.10: 誤識別したデータの例（pcmasterrace）

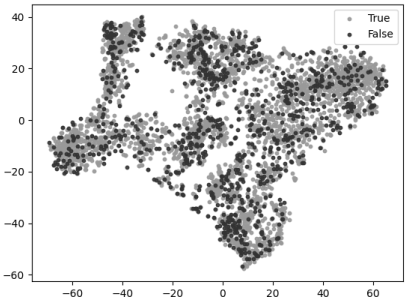
例 12	非皮肉と誤識別
親投稿	We’ve come a looooooong way baby... How to send an“E Mail”in 1984
	私たちは長い道のりを歩んできたんだ... 1984 年当時の「E メール」の送り方
皮肉	We still send email... haven’t came that far :P
	今でも E メールを送る... そんなに遠くへは来ていない :P
例 13	非皮肉と誤識別
親投稿	In 3+ years my PC has been shutdown a total of maybe 150-200 times.
	3 年以上の間，私の PC は合計で 150-200 回シャットダウンされたかもしれません.
皮肉	Hmm you use your computer around once a week then.
	ふーん，じゃあ週に一回くらいは PC を使ってるんだね.
例 14	皮肉と誤識別
親投稿	Because he can.
	だって彼はできるんだもの.
非皮肉	Yeah because he can!
	そうだ，彼はできるんだ！
例 15	皮肉と誤識別
親投稿	I suppose console players are drunk all the time
	ゲーム機プレイヤーはいつも酔っぱらっているのだろう
非皮肉	But the average age of console gamers is below the legal drinking age.
	しかし，ゲーム機ゲーマーの平均年齢は法定飲酒年齢を下回っているのである.

4.6　t-SNE を用いた次元圧縮による可視化

t-SNE [15] によって次元圧縮し可視化することでデータの分布を確認する．各データセットの訓練データによりファインチューニングした BERT に，テストデータを入力し出力を得た．得られた分散表現のうち [CLS] の分散表現 768 次元を t-SNE によって 2 次元に圧縮した．図 4.2 から図 4.6 に結果を示す．各図の (a) は橙色が皮肉データを，青色が非皮肉データを表している．(b) は灰色が皮肉・非皮肉を正しく識別したデータを，黒色が誤って識別したデータを表している．各図 (a) より，皮肉データと非皮肉データの分布に大きな差はないことから，BERT が皮肉表現の特徴を学習できたために皮肉推定が可能となったと考えられる．また各図 (b) より，皮肉データと非皮肉データが混在している箇所の皮肉推定を誤っており，その範囲は全体にわたっていることが分かる．このことから，皮肉表現は多様であり，一律の基準による皮肉推定は適切ではないと考えられる．

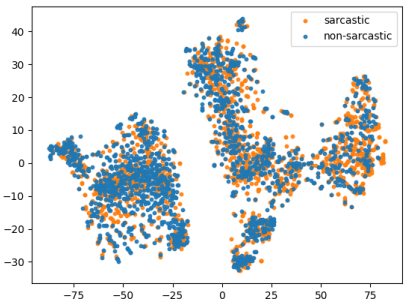


(a) 皮肉・非皮肉

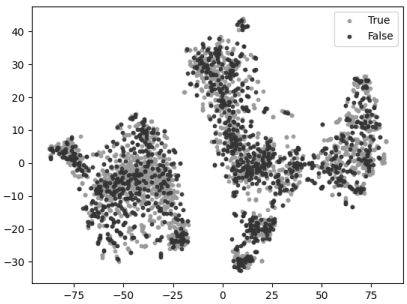


(b) 正解・不正解

図 4.2: t-SNE による可視化 (politics)

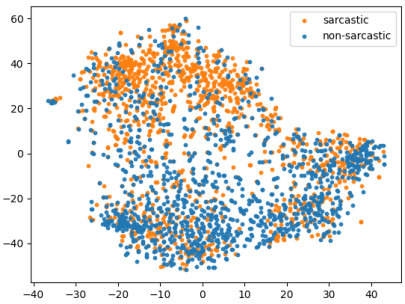


(a) 皮肉・非皮肉

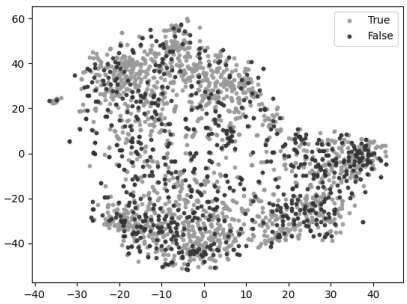


(b) 正解・不正解

図 4.3: t-SNE による可視化 (AskReddit)

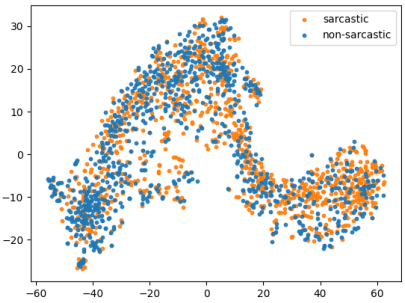


(a) 皮肉・非皮肉

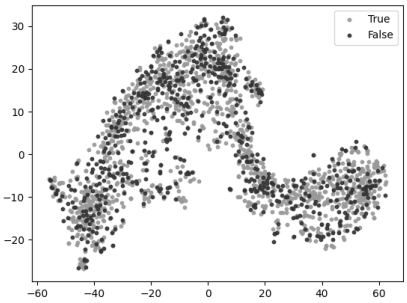


(b) 正解・不正解

図 4.4: t-SNE による可視化 (worldnews)

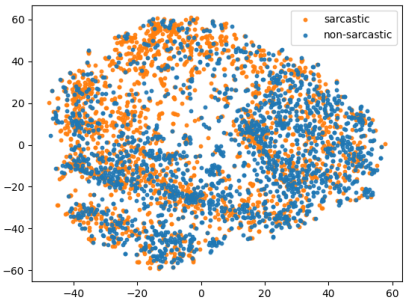


(a) 皮肉・非皮肉

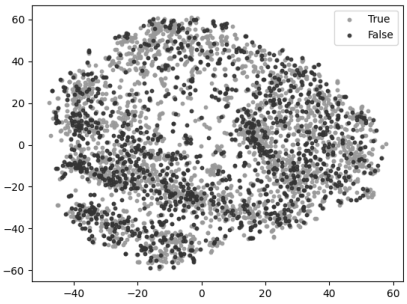


(b) 正解・不正解

図 4.5: t-SNE による可視化 (pcmasterrace)



(a) 皮肉・非皮肉



(b) 正解・不正解

図 4.6: t-SNE による可視化 (random)

5　まとめと今後の課題

本研究では皮肉推定の前段階として，文章の話題ごとの皮肉推定の有効性を検証した．Reddit における投稿トピック subreddit のそれぞれについて皮肉推定した結果，皮肉表現を一括りに扱って推定するよりも推定精度が向上することが確認できた．

今後の課題として，本研究での成果を踏まえ，各話題ごとに学習した皮肉推定モデルを組み合わせたアンサンブル学習によって，データセット全体に対する皮肉推定に取り組むことが挙げられる．また皮肉推定に使用する文脈情報やメタデータを拡大することにより，推定精度の向上を目指す．

謝辞

本研究を進めるにあたり，御指導，御鞭撻を賜りました森直樹教授に深く感謝の意を表します．岡田真助教には直接御指導頂き，研究方針を初めとした多大な御助言を頂きました．心より御礼申し上げます．

諸先輩方からは研究に関して建設的な意見を頂きました．また同期の皆さんとは共に切磋琢磨し励ましあいながら研究に取り組みました．心より感謝致します．

2022 年 2 月 25 日

参考文献

[1] Raymond W. Gibbs and Jennifer O'Brien. Psychological aspects of irony understanding. Journal of Pragmatics, Vol. 16, No. 6, pp. 523–530, 1991.

[2] Henk Haverkate. A speech act analysis of irony. Journal of Pragmatics, Vol. 14, No. 1, pp. 77–109, 1990.

[3] オーゼロヴァ・アナスタシーア. 言語的皮肉の現象についての理論とその原理 : 日本語における皮肉の分析を中心に. 日本語・日本文化研究, Vol. 26, pp. 147–157, 2016.

[4] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In EMNLP, pp. 704–714, 2013.

[5] Aniruddha Ghosh and Tony Veale. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 482–491, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[6] Rishabh Misra and Prahal Arora. Sarcasm detection using hybrid neural network. 2019.

[7] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A Large Self-Annotated Corpus for Sarcasm. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, May, 2018. European Language Resources Association (ELRA).

[8] Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. CASCADE: Contextual Sarcasm Detection in Online Discussion Forums. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1837–1848,

Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[9] Devin Pelser and Hugh Murrell. Deep and dense sarcasm detection. 2019.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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), pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017. In Advances in Neural Information Processing Systems, Vol. 30.

[12] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhut- dinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE international conference on computer vision (ICCV 2015), pp 19–27, Los Alamitos, CA, USA, December 2015. IEEE Computer Society.

[13] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's Neural Machine

Translation System: Bridging the Gap between Human and Machine Translation. 2016.

[14] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, pp. 2623–2631, New York, NY, USA, 2019. Association for Computing Machinery.

[15] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. Journal of Machine Learning Research, Vol. 9, No. 86, pp. 2579–2605, 2008.