静岡大学　大学院総合科学技術研究科　情報学専攻

2025 年度修士論文

対照学習による文脈依存語の皮肉判別可能な埋め込み表現の獲得

THET SAUNG AYE

指導教員　綱川　隆司

2025 年 8 月 18 日提出

概要

皮肉表現の検出は、文脈への高度な依存性や暗示的な意味合いにより、自然言語処理における困難な課題の一つである。本研究では、文脈に応じて文字通りにも皮肉的にも解釈され得る語に着目し、対照学習を用いて皮肉検出に有効な埋め込み表現の獲得を試みる。まず、文脈に依存して意味が変化する語を抽出する手法を提案する。次に、そうした語が出現する皮肉的な文と文字通りの意味の文の意味的対比に基づいて対照学習を行い、文の埋め込み表現を学習する。皮肉検出による埋め込み表現の有用性評価の結果、対照学習により得られた埋め込み表現は、従来の分類ベースの手法と同等の性能（F1 スコア差 0.0145 以下）を達成することを示す。さらに、対照学習と分類の双方による手法について皮肉検出の性能に差異が出る場合についての定性分析を行った。

目次

図表目次

第１章　　はじめに

皮肉を含む発話を文字通りの意味で扱うと、その発話のもつ本来の感情が正しく分析できない可能性がある。たとえば、「What a lovely rain to have a picnic.」という皮肉的な発話は、文字通りではポジティブの意味を持つが、皮肉であることを検出できなければ、本来のネガティブの感情を見逃してしまう。皮肉表現を適切に検出することは、感情を正確に分析するうえで重要である。

皮肉かどうかを判断するためには、単語が文脈によってどのように使われているかが重要である。前出の文における「lovely」は、たとえば「It is a lovely afternoon to sit outside and read.」における「lovely」とは異なり、皮肉として解釈される。皮肉の手がかりとなる単語やフレーズに着目した手法も提案されている[1][2]。しかし、これらの手法では、語の意味変化が暗黙的にしか扱われていなかった。

本研究では、文脈の違いをモデルに直接与えることで、その差異を明示的に学習させることを目指す。対照学習を用いて、文脈に応じて意味が変化する語に対し、皮肉的な用法と文字通りの用法を区別可能な埋め込み表現を獲得することが目的である。まず、文脈に応じて皮肉的に用いられる単語を抽出する手法を提案する。次に、抽出された文脈依存語に対し、皮肉表現と文字通り表現の意味的対比に基づいた対照学習を行う。

文脈依存語の抽出として、教師あり学習とコサイン類似度を用いた手法を提案する。本研究では、あるモデルに対して皮肉検出を目的した学習をする際には、文脈によって意味が変化する単語がそのモデルの重みにより影響を与えるという仮説を立てる。具体的には、事前学習済み言語モデルに対して皮肉検出タスクで再学習を行い、その際に最も大きな影響を与えた単語群を抽出する。これらの単語を、皮肉的にも文字通りにも解釈可能な文脈依存語と定義する。

文脈依存語の対照学習およびその評価を行うためには、各単語に対して対照学習用と皮肉検出用のデータセットの両方が必要である。皮肉検出用データセットとしては、既存の皮肉データセットSARC [3]を用い、対象単語が出現する文を抽出する。対照学習用データセットは、対象単語が含まれる皮肉文と文字通りの意味の文を対応づけて構築する。まず、対照学習用データセットを用いて事前学習済み言語モデルを再学習し、その後、皮肉検出用データセットを用いて評価を行う。

対照学習したモデルを皮肉検出タスクで評価するにあたり、学習済みモデルのパラ

メータを固定し、分類器を追加して学習を行い、その性能を測る。ベースラインとなる皮肉検出の分類タスク手法と比較した結果、同等の性能が得られることを示す。また、定性分析により、対照学習と分類タスク手法がそれぞれ得意とする文の傾向を分析する。

第２章　　関連研究

皮肉検出は、文章が皮肉を含むか否かを識別する二値分類問題として扱われてきた[4-5]。皮肉の暗黙的な性質により、発話自体だけでなく、発話に関連する情報を活用する手法が提案されている。たとえば、He ら[6] は、常識知識ベースと発話内容との不整合性に着目する手法を提案した。Babanejad ら[5]は、文脈に含まれる感情的側面をエンコードする方法を示した。Hazarika ら[4]は SNS における発言者の過去の投稿や言語スタイルから抽出したパーソナリティ特徴を皮肉検出に活用した。また、Ghosh ら[7]は、皮肉が他者によって再解釈される現象に注目し、皮肉の解釈に用いられる言語的戦略をモデル化した。

皮肉を検出するために、皮肉の手がかりとなる特定な語や表現パターンに注目した手法も提案されている。Riloff ら[2]は、皮肉的な文に見られるポジティブな感情語の出現に着目し、それらを統計的に抽出する手法を提示した。Ghosh ら[1]は皮肉的発言とその説明文の対比に基づき、皮肉を誘発する語句を同定するアプローチを採用している。これらの手法では特定の語句への限定や、皮肉の再解釈に他者の関与を前提とする場合がある。これに対し、本研究では、発言者によってラベル付けされたデータを用いることで、第三者の解釈を介せずに皮肉検出において重要となる単語を抽出する手法を提案する。ただし、抽出された単語の質は、皮肉検出タスクにおけるモデルの学習性能に依存する。

対照学習は主に意味的類似文判定、ロバストな意味表現の獲得などに有効性を示している[8-9]。皮肉検出に関する対照学習の応用という点では、自然言語の意味変化を直接学習するという特性が、自然言語の離散的な性質に対応する手法として注目されている。Wang ら[10]は、わずかな単語の違いによって意味が大きく変化する文章に対応できるために、意味的に類似または対照的な文を用いて対照学習を行った。Qiu ら[11]は、皮肉検出において深い意味理解を評価するため、単語とラベルの疑似相関を排除する目的で対照学習を導入している。

本研究では、同一の単語を含む皮肉的な文と文字通りの意味の文の意味的対比を利用した対照学習の枠組みを提案する。従来の分類タスク手法と比べて、皮肉の文脈依存性をより明確に抽出できると考えられる。

第３章　　関連技術

本章では、事前学習済み言語モデル、皮肉検出の分類手法、そして対照学習について説明する。さらに、モデルから文の埋め込み表現を抽出する手法および、その表現を目的タスクにおいて評価する手法としてのプロービング手法について述べる。

3.1.　事前学習済み言語モデル

事前学習済み言語モデルは、大規模なコーパスで事前に学習されており、大きく分けて一方向学習と双方向学習の二つのカテゴリに分類される。一方向学習では、ある単語の前または後の文脈に基づいて次の単語を予測する形式で学習される。この手法は主にテキスト生成を目的としたモデルに用いられ、代表的な例として GPT 系のモデル [12]が挙げられる。双方向学習では、対象となる単語の前後の文脈を同時に考慮して学習が行われる。この手法は、マスク言語モデルや次文予測などのタスクを通じて事前学習され、代表的なモデルとして BERT[13]が知られている。BERT を下流タスクに再学習（ファインチューニング）することで、文分類や質問応答などの多様なタスクにおいて高い性能を得られることが報告されている[13-14]。

言語モデルにおいては、文脈情報は埋め込みベクトルとして表現される[15]。各単語や文に対応する埋め込みベクトルは、そのトークンの意味および周囲の文脈情報を含み、文脈に応じた意味の変化を捉えることが可能である。これらのベクトルは通常、ランダムに初期化された後、学習を通じて最適化される。事前学習済みモデルのファインチューニングにおいては、事前に学習されたチェックポイントから重みを読み込み、目的タスクに適したパラメータへと更新される。

3.2.　皮肉検出の分類手法

皮肉検出タスクは、与えられた発話が皮肉的か文字通りかを判別する二値分類問題として定式化されてきた[4]。このタスクに対する分類手法としては、事前学習済み言語モデルを皮肉検出タスクにファインチューニングしたモデルが、ベースラインや比較対象として広く用いられている[5-6]。皮肉検出タスクのファインチューニングは、皮肉を含む文とそのラベルを用いて、事前学習済み言語モデルの重みを最適化する教師あり学習の手法である。本研究においても、ベースライン手法としてファインチューニングによる分類モデルを用いる。

事前学習済み言語モデルは、入力層、複数のトランスフォーマー層（隠れ層）、および最終層から構成されている[13]。分類タスクのファインチューニングするにおいては、最終の隠れ層の出力に対して、分類クラス数に対応するノードを持つ線形出力層を追加し、全体を通して学習を行う。皮肉検出におけるファインチューニングでは、「皮肉」と「文字通り」の 2 クラスを予測する線形出力層を用い、二値分類タスクとして学習を行う。

クラスに対するモデルの予測と正解ラベルの差を評価するために、損失関数が用いられる。分類タスクにおいては、交差エントロピー損失が一般的に用いられており、モデルの出力（予測確率）が正解ラベルに近いほど損失は小さく、誤ったラベルに近いほど損失は大きくなる[4-5]。本研究では、皮肉検出のベースライン手法として BERT と RoBERTa[16]に対し、交差エントロピー損失関数に基づいてモデルのパラメータを更新する。

3.3. 対照学習

対照学習は、類似するサンプルの表現を互いに近づけ、非類似なサンプルの表現を遠ざけることで、表現空間上で有意義な特徴表現を獲得することを目的とした手法である[17]。本研究における対象語の皮肉性に着目して対照学習を適用することは、例えば、同一語彙が用いられていても、皮肉的文と文字通りの意味の文（以下、文字通り文と呼ぶ）を異なる意味として区別し、表現空間上で引き離すように学習させることに相当する。

対照学習における正例および負例の設計のために、既存の教師あり皮肉検出データセット SARC [3]を活用した。SARC データセットは、発話文と、それが皮肉を含むか否かを示すラベルから構成されている。本研究ではこれらのラベル情報を対照学習用のペアデータ構築に用い、文同士の意味的な類似あるいは非類似の関係を定義する。ただし、対照学習そのものは分類を直接行うのではなく、文表現間の関係性に基づいて表現の学習を行う枠組みである。

対照学習における表現間の類似尺度は、一般にコサイン類似度が用いられ、ベクトル間の角度的な距離に基づいて、文同士の意味的な近さを測定する[8-9]。本研究では、Gao ら[8]によって提案された、コサイン類似度に基づく対照学習の目的関数を採用し、BERT と RoBERTa の表現を学習する。

3.4. 文埋め込み表現の抽出手法

文埋め込み表現とは、文の意味や文脈情報を保持した上で、固定長の数値ベクトルとして表現する方法である[9]。本研究の評価では文頭の特殊トークン（例：BERT における[CLS]）に対応する隠れ状態を利用する。意味的類似度の計算などを目的とするタスクでは、特に最終の二つの隠れ層のすべてのトークン出力を平均化して文埋め込みとする手法の方が効果的であることが報告されている[9][18]。本研究では、対照学習済みモデルに対して両方の文埋め込み抽出手法を適用し、皮肉検出タスクにおける性能を比較した。その結果、F1 スコアにおける差は最大でも 0.0024 と小さく、埋め込み抽出手法による性能差は限定的であることが示唆された。文埋め込み表現の抽出手法ごとの性能比較の詳細は付録 A に示す。

3.5. プロービング手法

プロービング手法とは、言語モデルの中にエンコードされた情報を分類器を通じて分析し、モデル内部にどのような言語的あるいは意味的情報を保持しているかを明らかにする手法である[19]。適用例として、BERT の隠れ層にエンコードされた情報を調べるために、学習済み BERT のパラメータを固定し、単純な分類器を学習する手法が提案されている[20]。本研究では、対照学習済みモデルの皮肉検出性能を分析する目的で、対照学習後のパラメータを固定し、モデル出力（文埋め込み表現）を入力とした 2 層の全結合ニューラルネットワークによる分類器を学習する。

第４章　　文脈依存語の抽出

表 1　学習前後のコサイン類似度が低い単語ランキング

(1) wonderful	(6) creators	(11) pathetic	(16) johnson
(2) 300	(7) 40	(12) amir	(17) 390
(3) confused	(8) lovely	(13) 28	(18) dropped
(4) arrogant	(9) interesting	(14) important	(19) shocked
(5) ##5	(10) horror	(15) michael	(20) children

本研究では、皮肉検出の学習において、文脈に応じて意味が変わる単語が、モデルのパラメータに影響を与えると考える。こうした単語を抽出するために、皮肉検出タスクにファインチューニングしたBERTモデルを用いる。まず、両モデルに対して、文中の各単語に対応する最初の隠れ層の埋め込み表現を取得する。次に、それぞれのモデルから得られた単語埋め込み表現間のコサイン類似度を各単語ごとに計算する。最後に、コサイン類似度が低い単語、すなわち2つのモデル間で埋め込み表現が大きく異なる単語を上位から抽出する。なお、同一の単語が複数の文脈で出現する場合、それぞれを別の出現として扱い、ランキングするときに代表値として類似度の最小値を用いる。

皮肉データとして、SARCデータセット[3]を学習用、検証用、テスト用と分割して用いた。学習データには、皮肉文115,754件と、文字通り文115,619件を用い、BERTモデルを再学習した。テストデータの皮肉文32,222件および文字通り文32,222件に含まれる単語に対して、コサイン類似度を算出した。モデル学習は、バッチサイズ32、学習率 $2\text{x}10^{-5}$ 、エポック数4で実施した。単語埋め込み表現の抽出には、単語に対応するトークンを用い、入力テキストの各トークンに対して、モデルの最初の隠れ層から768次元の埋め込みベクトルを取得した。実装には PyTorch (1.12.0+cu116) および Transformers (4.46.3) を使用した。NVIDIA Quadro RTX 6000 GPU (CUDA Version 11.6) を搭載した計算機上で実行した。

皮肉検出の学習前後においてコサイン類似度が低い上位20単語を表1に示す。文脈に応じて皮肉的または文字通りに解釈できる単語が見られた。例えば、「wonderful」や「lovely」はネガティブな状況に対して用いると皮肉的に聞こえる。例えとして、「It is wonderful to wait in this heat.」が挙げられる。また、「important「は状況にそぐわない使われ方をすることが多い。例えば、「Game is way more important than homework.」である。抽出した単語の中でネガティブな意味を持つ単語もふくま

れており、「arrogant」や「pathetic」がそれに該当する。望ましい状況に対してネガティブな単語を用いることで皮肉が表現される。例えば、「It’s just pathetic that you only got second place.」である。これら抽出した上位 20 単語のうち、対照学習に使用する単語として、数字や人名を除いた上位 9 語を選定した（表 2)。

第５章　　文脈依存語の対照学習

本研究では、第４章の手法より抽出した文脈に応じて皮肉的にも文字通りにも解釈可能な９単語（表 2）に着目し、対照学習を行う。それら単語を含む文を基に、 皮肉検出用および対照学習用のデータセットを構築し、これらのデータセットを用いて、事前学習済み言語モデルである BERT および RoBERTa のパラメータを最適化する。図１より、対照学習と皮肉検出の概要図を示す。対照学習では、たとえば、lovely を含む文に対して対照学習を行う場合、皮肉文および文字通り文それぞれの文埋め込み表現におけるコサイン類似度を計算し、SimCSE[8]の損失関数（式１）を用いて対照学習を行う。最後に、対照学習済みモデルのパラメータを固定し、それらの文埋め込みを２層ニューラルネットワークによる分類器を学習させることで皮肉検出タスクの性能を評価し、ファインチューニング手法との比較を行う。

5.1.　対照学習の損失関数

本研究では Gao ら[8]が提案した対照学習の目的関数を採用し、BERT と RoBERTa に対して対照学習を行う。ある単語が文字通りに使われている２つの異なる文 x_i，x_i^+ と、同じ語が皮肉的に使われている文 x_i^- に対し、それぞれの文埋め込み表現であるモデルの[CLS]トークンに対応する隠れ状態を h_i，h_i^+ ，h_i^-とする。N 文からなるバッチに対して、目的関数は以下のように定義される。

$$-\log \frac{exp(cos(\mathrm{h}_i,\mathrm{h}_i^+)/\gamma)}{\Sigma_{j=1}^{N}\left(exp\left(cos\left(\mathrm{h}_i,\mathrm{h}_j^+\right)/\gamma\right)+\alpha^{\mathbb{I}_i^j} exp\left(cos\left(\mathrm{h}_i,\mathrm{h}_j^-\right)/\gamma\right)\right)} \qquad (1)$$

γ は温度スケーリングのハイパーパラメータであり、cos(h₁，h₂）は文埋め込み表現間のコサイン類似度を表す。 $\mathbb{I}_i^j \in \{0,1\}$ は i=j のとき 1、それ以外のとき 0 となる指示関数である。

α は、皮肉文と文字通り文の対比が損失関数に与える影響度を調整するための重みである。直感的には、α を高く設定することで、皮肉との違いがより強調されると考えられる。複数の値を設定して比較実験を行った結果、α=20 のときに最も良好な性能が得られたため、本研究では以下の実験でこの値を用いる。α の値および性能に関する詳細は、付録 C に記載する。

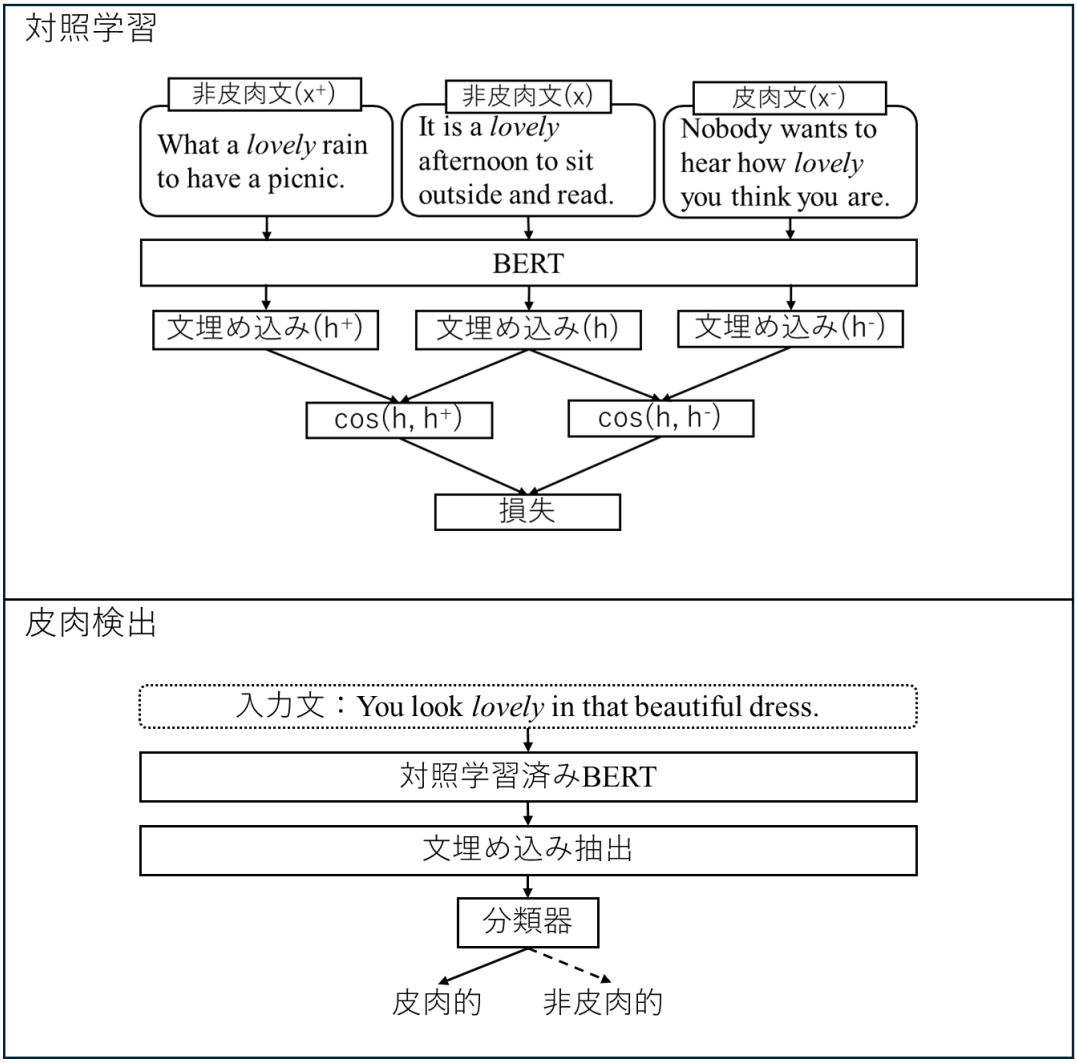

図 1　提案手法における対照学習および皮肉検出の概要図

5.2.　データセットの構築

対照学習で用いる正例ペアと負例ペアの作成には、すでに発話者によって皮肉に関する正解ラベルが付与されている皮肉検出用データセットである SARC[3]を使用する。まず、ある文脈依存語を含む文を SARC データセットから抽出する。次に、ランダムに文字通り文２つと皮肉文１文からなる３つ組（x^i, x^{i+}, x^{i-}）を抽出する。その作業を全ての文脈依存語に対して行い、それぞれの対照学習用データセットを構築する。表２には、文脈依存語と、学習に採用したその出現文数を示す。

皮肉検出用データセットは対照学習の評価（第 3.5 節)、およびベースライン手法のファインチューニング手法の学習に必要である。このデータセットは、SARC データセットから文脈依存語を含む文を抽出し、その皮肉ラベルを用いて構築する。

表 2　文脈依存語とその出現文数

important (19371),	interesting (5589),	wonderful (5316),
confused (3411),	lovely (2283),	horror (2223),
pathetic (1932),	creators (1461),	arrogant (879)

5.3.　ベースラインモデル

ベースラインとして、皮肉検出の性能評価のために、従来の研究において比較対象として広く用いられているファインチューニング手法を採用する［5-6]。これらの手法では、事前学習済みモデルに対して、発話とそれに対応する皮肉ラベルを用いて再学習を行う。本研究においても同様に、文脈依存語を含む文とその皮肉ラベルから構成される皮肉検出用データセットを用い、BERT と RoBERTa に対してファインチューニングを行う。

第6章　　実験

6.1.　実験設定

実験には、SARCデータセット[3]から抽出した、文脈依存語9語（表2）に対するデータ合計 42,465 件を用いた。対照学習では、文字通り文と皮肉文を組み合わせて、 3つ組を作成した。学習には、最も多い単語である「important」に対する4519組、最も少ない「arrogant」に対する205組のデータを使用した。ベースライン手法のファインチューニングでは、発話文とそれに対応する皮肉ラベルで学習を行った。学習には、皮肉と文字通り文数が等しくなるように調整し、「important」で9038文「arrogant」で410文の範囲のデータを使用した。各単語データセットを、学習用80%、検証用 10%、テスト用 10%に分割して使用した。

モデルは、BERTおよびRoBERTaに対して、HuggingFaceライブラリが提供する事前学習済みチェックポイントを初期値として用い、再学習を行った。対照学習において、学習率は{$1x10^{-4}$, $1.5x10^{-5}$, $2x10^{-5}$, $3x10^{-5}$, $4x10^{-5}$, $5x10^{-5}$}、バッチサイズは{16, 32}の範囲から最適化した。バッチサイズ32、エポック数20を用い、BERT-baseとRoBERTa-baseには学習率5×10^{-5}、BERT-largeには学習率3×10^{-5}、RoBERTa-largeには$2x10^{-5}$で学習した。ベースライン手法のファインチューニングには、学習率2×10^{-5}、バッチサイズ 32、エポック数4を用いた。対照学習済みモデルの皮肉検出性能を評価するためのプロービングに用いる、2層の全結合ニューラルネットワークによる分類器は、学習率$5x10^{-5}$、バッチサイズ16、エポック数4で学習した。実装には PyTorch（1.12.0+cu116）および Transformers（4.46.3）を使用した。NVIDIA Quadro RTX 6000 GPU（CUDA Version 11.6）を搭載した計算機上で実行した。

6.2.　結果と考察

文脈依存語のデータに対する、対照学習とファインチューニング手法による皮肉検出の性能を比較した。あわせて、事前学習済みBERT-baseモデルにおいて、最終2層の隠れ層出力の平均化により得られた文埋め込み表現を用いた場合の性能とも比較を行った。簡単のために、全ての文脈依存語のマクロ平均性能で、適合率（Avg P.）、再現率（Avg R.）、F1スコア（Avg F1）を報告する。また、データサイズ2000件以上の単語（表2）に対する性能（Filter F1）およびポジティブの意味を持つ単語{important, interesting, wonderful, lovely}とネガティブの意味を持つ単語

{confused, horror, pathetic, arrogant}別の性能（Pos. F1, Neg. F1）も述べる。

表３は、各手法による皮肉検出性能の評価を示している。提案手法である対照学習を用いたモデルには「*」を付けている。実験結果から、対照学習モデルはファインチューニングモデルと比較して、同等の性能を示した。F1 スコアの差は、BERT-base では 0.0145、RoBERTa-base では 0.0053 といずれもわずかであり、BERT-large および RoBERTa-large でも同様の傾向が見られた。

同一の分類器を用いた比較において、BERT-base における最終 2 層の隠れ層出力の平均化により得られた文埋め込みよりも、対照学習で再学習した埋め込みの方が F1 スコアで 0.0320 以上向上した。この結果は、対照学習が文脈依存語に対して、皮肉的な用法と文字通りの用法を区別可能な埋め込み表現を学習できていることを示唆している。なお、RoBERTa-large モデルにおける、各単語データセットの学習曲線を付録 C に示す。

対照学習およびファインチューニングの両手法において、ポジティブな意味を持つ単語の方が、ネガティブな単語よりも皮肉の検出性能が高い傾向が見られた。これは、ネガティブな語を用いた皮肉表現は文字通りの否定的な意味と区別がつきにくく、皮肉として成立させるのが人間にとっても難しい傾向があるためだと考えられる。また、RoBERTa の方が BERT よりも感情の極性に対する反応が強いことが示された。ファインチューニングおよび対照学習のいずれにおいても、RoBERTa ではポジティブおよびネガティブな意味をもつ単語間での性能差が、BERT よりも顕著であった。

学習データ数が検出性能に与える影響について、データ数が 2000 件以上ある単語に限定して性能を比較した結果、RoBERTa は BERT よりもデータサイズの影響を受けやすい傾向が示唆された。RoBERTa-base モデルでは、データサイズによる F1 スコアの変動が大きく、対照学習とファインチューニングの差はそれぞれ 0.0149 および 0.0361 であった。一方、BERT-base モデルではそれぞれ 0.0053 および 0.0078 と、変動幅は小さかった。

表 3　モデルの皮肉検出性能の評価

Model	Avg P.	Avg R.	Avg F1	Pos. F1	Neg. F1	Filter F1
$BERT_{base}$ Avg.	0.7506	0.6944	0.7134	0.7191	0.6854	0.7180
Finetune-$BERT_{base}$	**0.7646**	0.7564	**0.7599**	**0.7748**	**0.7237**	**0.7677**
＊ContrastiveLearning-$BERT_{base}$	0.7003	0.7976	0.7454	0.7641	0.7023	0.7507
＊ContrastiveLearning-$BERT_{large}$	0.6673	**0.8282**	0.7372	0.7610	0.6963	0.7551
Finetune-$RoBERTa_{base}$	**0.7259**	0.7547	**0.7363**	**0.7811**	0.6734	**0.7724**
＊ContrastiveLearning-$RoBERTa_{base}$	0.7229	0.7709	0.7310	0.7516	**0.6819**	0.7459
＊ContrastiveLearning-$RoBERTa_{large}$	0.6534	**0.8133**	0.7198	0.7602	0.6547	0.7487

6.3. 定性分析

対照学習とファインチューニング手法それぞれの得意な例文を分析するために、予測結果を比較し、一方の手法が正解でもう一方の手法が誤りだった文を抽出した。抽出した文から、文脈に応じて意味が変わる単語の多様性、明確さと解釈のしやすさに基づいて、サンプルを選択した。表4にその一部を示す。全てのサンプル文は付録Dに示す。

ファインチューニングは、皮肉に特有の語彙的なパターンをより学習している傾向が見られた。例えば、ファインチューニング手法では、強調語である「no other」、「what a」、「extremely」を含む皮肉文を正しく検出できていた（第1〜3行)。一方でファインチューニングは、皮肉によく用いられる語句を含む文字通りの表現を誤って皮肉と判断する傾向がある。例えば、強調語 「no other」 、悲劇的や滑稽な表現の「suicide」 や 「big perky titted」を含む文字通り文を皮肉と誤判定していた（第10〜12行)。これに対して、対照学習は文全体の意味をより正確に捉えることができており、これらのような例を正しく識別できていたと考えられる。

対照学習は、比較的長い文脈で皮肉が成立する文にも対応できている。例えば、誇張表現 「shoved up my ass」 の皮肉的な意味を正しく捉えることができた（第7行)。また、「it is residents responsibility to not get old or sick or confused」 のように、強調語などを含まない皮肉文も正しく検出できていた（第8〜9行)。このような例では、ファインチューニング手法では、捉えにくい文全体として皮肉が成り立っている文を、対照学習が捉えられている可能性がある。

一方で、対照学習は、意味の対比を含む文字通り文や、皮肉的に意味が近い文には失敗する傾向がみられる。例えば、第4〜5行の文では、「shithole」と「lovely」、あるいは「wonderful」と「madness」といった語の対比が見られ、このような対比が皮肉であると判断する要因になったと考えられる。また、第6行の文「What are the most common misconception? That reddit is important.」は、皮肉を含まない疑問と回答であるが、皮肉文と意味的に近いため、対照学習では皮肉と誤って分類された可能性がある。一方、皮肉文とそのラベルとの関係性をもとに学習するファインチューニング手法では、このような皮肉表現ではない対比を無視して正しく分類できると考えられる。

表 4　一方のモデルが正しく予測し、もう一方のモデルが誤判定した例文。「S.」は皮肉文、「NS.」は文字通り文を意味する。

Model	**Example texts**	**Pred.**	**Truth**
Finetuning $BERT_{base}$	This kids show will fix everything! That's all to talk about in this piece of news relevant to the middle east, there's no other *horrors* relevant to the situation or directly related to this situation that should be talked about	S.	S.
	What a *lovely* reddit thread. Much better then pizza gate accusations	S.	S.
	I know, right? This is extremely *interesting*!	S.	S.
	Yet her country is slowly being turned into a shithole, all thanks to their acceptance of refugees. Their women and children are being raped, men beat in the street while she sits and had tea in her guarded safe spot. Erm, I was in Munich before Christmas and I have never felt safer. It was *lovely*.	NS.	NS.
	The most *wonderful* time of the year, March Madness is here!	NS.	NS.
	What are the most common misconceptions? That reddit is *important*.	NS.	NS.
Contrastive Learning $BERT_{base}$	Then how the fuck do you want to be treated? Wow, this is so powerful and so *important* that I want this shoved up my ass, so that it'll be with me forever.	S.	S.
	It is residents responsibility to not get old or sick or *confused*.	S.	S.
	Good Christian Family Man Bill O'Reilly Loses Custody Of His Kids, Was Accused Of Domestic Violence This is very shocking and I am very surprised that such a good Christian man would do something so *pathetic*.	S.	S.
	It changes proportions, i remember using the steam one and getting *confused* at big perky titted grannies	NS.	NS.
	Damn, still look *lovely*! Don't feed the trolls they just want the attention of a beautiful woman, and they know no other way besides insults.	NS.	NS.
	Which entire fandom are you able to offend with a single sentence? The *creators* of Steven Universe don't want you to bully people to the point of suicide attempts.	NS.	NS.

第７章　　おわりに

本研究では、文脈に応じて皮肉的にも文字通りにも解釈できる単語の抽出手法を提案し、その単語が出現する文に基づいて対照学習を行った。実験結果より、対照学習によって皮肉検出に有効な埋め込み表現が得られたことを示した。また、定性分析より、ファインチューニング手法が典型的な語彙パターンを学習していることが示唆されたのに対し、対照学習が文全体の意味に基づいて皮肉検出を行っている傾向が見られた。さらに、皮肉検出における語彙パターンの認識と意味の理解両方の必要性を論じた。将来の展望として、対照学習と分類目的を組み合わせる手法で性能の向上が期待できる。

謝辞

まず、研究全体にわたり丁寧にご指導くださった指導教員の綱川隆司准教授に、深く感謝いたします。研究テーマの設定から議論の方向性まで、的確な助言を頂き、大きな支えとなりました。また、貴重なご意見をくださった西田昌史教授に、厚く御礼申し上げます。先生の幅広い知見は本研究に多くの示唆を与えてくださいました。次に、研究活動において支え合った綱川研究室のメンバーの皆様に感謝いたします。皆様の存在が研究を続ける励みとなりました。最後に、研究環境と設備を提供してくださった静岡大学に感謝申し上げます。

参考文献

[1] Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words. Empirical Methods in Natural Language Processing, pages 1003–1012. Association for Computational Linguistics.

[2] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. Empirical Methods in Natural Language Processing, pages 704–714. Association for Computational Linguistics.

[3] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A Large Self-Annotated Corpus for Sarcasm. Language Resources and Evaluation 2018. European Language Resources Association.

[4] Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual Sarcasm Detection in Online Discussion Forums. Computational Linguistics, pages 1837–1848. Association for Computational Linguistics.

[5] Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and Contextual Embedding for Sarcasm Detection. Computational Linguistics, pages 225–243. International Committee on Computational Linguistics.

[6] He, Yuanlin, Mingju Chen, Yingying He, Zhining Qu, Fanglin He, Feihong Yu, Jun Liao, and Zhenchuan Wang. 2023. Sarcasm Detection Base on Adaptive Incongruity Extraction Network and Incongruity Cross-Attention. Applied Sciences 13, no. 4: 2102.

[7] Debanjan Ghosh, Elena Musi, and Smaranda Muresan. 2020. Interpreting Verbal Irony: Linguistic Strategies and the Connection to theType of Semantic Incongruity. Computation in Linguistics 2020, pages 82–93. Association for Computational Linguistics.

[8] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. Empirical Methods in Natural Language Processing, pages 6894–6910. Association for Computational Linguistics.

[9] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992. Association for Computational Linguistics.

[10] Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding. Association for Computational Linguistics and Natural Language Processing, pages 2332–2342. Association for Computational Linguistics.

[11] Ziqi Qiu, Jianxing Yu, Yufeng Zhang, Hanjiang Lai, Yanghui Rao, Qinliang Su, and Jian Yin. 2025. Detecting Emotional Incongruity of Sarcasm by Commonsense Reasoning. Computational Linguistics, pages 9062–9073. Association for Computational Linguistics.

[12] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Association for Computational Linguistics: Human Language Technologies, pages 4171–4186. Association for Computational Linguistics.

[14] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification? Chinese Computational Linguistics, pages 194–206. Springer-Verlag.

[15] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. Association for Computational Linguistics: Human Language Technologies, pages 2227–2237. Association for Computational Linguistics.

[16] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv abs/1907.11692 (2019): n. pag.

[17] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. Computer Vision and Pattern Recognition (CVPR), pages 1735–1742. IEEE.

[18] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. Empirical Methods in Natural Language Processing (EMNLP), pages 9119–9130.

[19] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. Association for Computational Linguistics, pages 2126–2136. Association for Computational Linguistics.

[20] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What Happens To BERT Embeddings During Fine-tuning?. Analyzing and Interpreting Neural Networks for NLP, pages 33–44. Association for Computational Linguistics.

付録

A　文埋め込み表現の抽出手法

表 A.1 に、文埋め込み表現として、[CLS]トークンより抽出する手法および最終２つの隠れ層の平均を取った手法における、皮肉検出性能を示す。事前学習済み BERT では、平均化手法が[CLS]手法より 0.0431 高い F1 スコアを示した。対照学習した BERT では[CLS]の埋め込み表現の方が F1 スコアで 0.0024 高い結果を示した。

B　対比損失重み係数αの皮肉検出性能評価

文脈依存語 important および pathetic に対して、エポック数 20 までにおける α を {0, 0.5, 5, 15, 20} に設定し、皮肉検出性能を比較した。モデルの学習には BERT-base を用い、バッチサイズ 32、学習率 5×10^{-5}で実施した。表 B.1 は、各単語データセットにおける異なる α 値およびエポック数 {3, 10, 20} に対する F1 スコアを示す。important データセットにおいては、エポック数を 20 に設定した場合に、同様に α の増加による性能向上が見られた。pathetic データセットでは、α を 0 から 20 に上げることで全体的な性能向上が確認された。

C　RoBERTa-large における学習曲線

RoBERTa-large モデルに対して、学習率 2×10^{-5}、対比損失重み係数（α）20、バッチサイズ 32 で設定し、各文脈依存語データセットに対する学習の様子を示す。図 C.1 には、各単語データセットにおける対照学習のエポック数に対する、学習データおよび検証データにおける皮肉検出精度の推移を示す。

D　対照学習およびファインチューニング手法の得意文

表 D.1 および表 D.2 に、対照学習およびファインチューニングで再学習した BERT-base における、一方の手法が正しく判定し、もう一方の手法が誤判定していたものを示す。

表　A. 1　文埋め込み表現抽出手法の皮肉検出性能

Pooler	Precision	Recall	F1
Pretrained $BERT_{base}$			
[CLS]	0.7692	0.6267	0.6703
Top 2 avg.	0.7506	0.6944	0.7134
Contrastive Learning-$BERT_{base}$			
[CLS]	0.7003	0.7976	0.7454
Top 2 avg.	0.7081	0.7848	0.7430

表　B. 1　対比損失重み係数αによる皮肉検出性能

	0	0.5	5	15	20
3	0.7817	0.7619	0.7227	0.5234	0.7383
10	0.7842	0.7849	0.7790	0.7045	0.7790
20	0.7851	0.7834	0.7962	0.7995	0.7967

important

	0	0.5	5	15	20
3	0.0952	0.0588	0.3830	0.5410	0.5438
10	0.5206	0.5266	0.5641	0.6319	0.6793
20	0.5522	0.5204	0.5491	0.6545	0.6718

pathetic

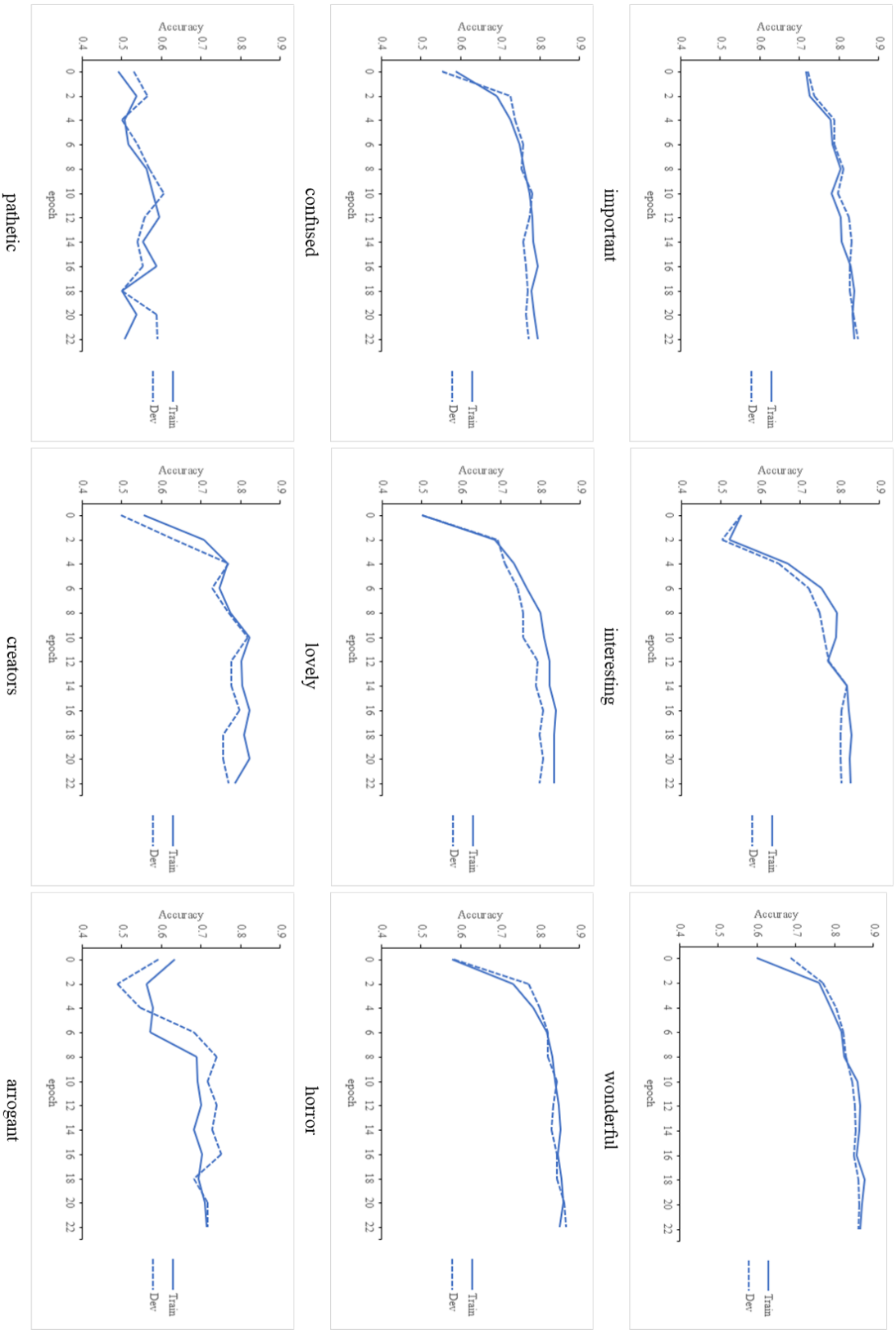

図　C．1　単語の対照学習エポック数に対する皮肉検出精

表　D. 1　対照学習が正しく予測し、ファインチューニングが誤って予測した皮肉的な文

Sarcastic example texts

- Then how the fuck do you want to be treated? Wow, this is so powerful and so important that I want this shoved up my ass, so that it'll be with me forever.
- It's because the majority are uneducated. Sounds racist, sure, but it's true and sometimes people don't like to hear it. Blacks should like Jews a lot more for what we've done for them when Jews had their own problems and didn't have to help. Not to mention, they have a rough history of experiencing racism, so they should empathize with the fact that the world seems to hate us the most. But, instead they feel for Palestinians because they're portrayed as browner (which they aren't), not to mention the Arabs orchestrated the slave trade. Fuck facts though, emotion is more important.
- Yeah, it was there right in the opening credits. That was the most exciting part of the series for the kids—the wonderful opening credits.
- Thank you. Haha. I wish I could take credit for it, but I'm not that clever. Someone used it on me once. I'm sure that you're now married to that gentleman with three lovely young children.
- Fillers for November and December! Fuck yes. I was hoping we could get some more filler episodes because the story is so much more interesting and the animation is so much more better.
- It is residents' responsibility to not get old or sick or confused.
- Drug raid in rural Georgia ends in a homeowner dead, no drugs found, and no police punished. Why isn't the color of this man's skin mentioned at all in this article?? I'm so confused.
- What the hell am I looking at—the true horror experience, starring the Wormbilly.
- Good Christian family man Bill O'Reilly loses custody of his kids, was accused of domestic violence. This is very shocking and I am very surprised that such a good Christian man would do something so pathetic.
- Phil Lord, director of 21 Jump Street, Cloudy with a Chance of Meatballs, and son of a Cuban refugee, writes an open letter in response to Jay-Z. Two brilliant creators of culture exchange poignant thoughts, awe the public with their insight.

Non-sarcastic example texts

- Well said. That's the point I was trying to make. People have that mentality, and perhaps it's true in some instances, but no person should have to earn not sleeping under a tree or being able to eat regularly. There are enough empty houses in the US to end its homeless problem, but nobody wants to help because there's no money in it for them. And food production is such that we could feed the entire world and there would be no hunger—but again, it's all about the money. So sad that money is more important than humanity and compassion.
- Hmmm, interesting.
- Congratulations! What a wonderful project.
- It changes proportions. I remember using the Steam one and getting confused at big perky titted grannies.
- Damn, still look lovely! Don't feed the trolls. They just want the attention of a beautiful woman, and they know no other way besides insults.
- It's straight out of a horror film. Poor guy—talk about a wrong place at the wrong time scenario.
- Good. That's a pathetic response to someone's career being ended by injury. Or any career ending, really.
- Which entire fandom are you able to offend with a single sentence? The creators of Steven Universe don't want you to bully people to the point of suicide attempts.
- Reddit, how do you impress a girl? You impress girls by being able to have conversations with them that most guys can't. Also, not letting any confidence come out as arrogance because... well, screw arrogant people.
- How is this arrogant or obnoxious? Your face looks arrogant and obnoxious. Your comment further proves that you are, in fact, both arrogant and obnoxious. Case closed.

表　D．2　ファインチューニングが正しく予測し、対照学習が誤って予測した皮肉的な文

Sarcastic example texts

- How to get subscribers on your YouTube channel? The most important thing is to spam your channel in the comments of more popular channels.
- I know, right? This is extremely interesting!
- I'm bi. And when drunk at a party, it's only safe to flirt with the women. Or grind on them. Can't grind on dudes when in a relationship—what a wonderful double standard.
- Cam Newton's face as he watches his Super Bowl dreams crushed—isn't it wonderful watching someone's hard work result in failure? Hahaha. He's such an idiot. That's what he gets for putting his heart into his craft.
- What a lovely Reddit thread. Much better than Pizzagate accusations.
- Until recently, my apartment complex thought I only had one cat. One is a short-haired brown mackerel tabby, and one is a medium-haired brown classic tabby. How can people get them confused?
- At least she recognizes that what you eat has a lot to do with what you weigh. I'm confused—I thought if the people in India were starving, they would enter starvation mode and gain weight.
- This kids' show will fix everything! That's all to talk about in this piece of news relevant to the Middle East. There's no other horrors relevant to the situation or directly related to this situation that should be talked about.
- Has he really? What a tool. And we are obviously the pathetic ones. We are just too undesirable, so we deserve it.
- Remember this sort of shit the next time a theist tells you that atheists are the arrogant ones. Why do you call religious people arrogant? It's not like Christians claim to be the specially chosen, most loved, personal favorite, divinely royal, holy child of the creator of everything in existence or anything like that. ^

Non-sarcastic example texts

- What are the most common misconceptions? That Reddit is important.
- It would be interesting for anyone with 0-cost minions and Coin, and fuckin' terrible for everyone else.
- Hillary Clinton's cabinet picks have been leaked, and she was going to appoint John Podesta as Secretary of State!! This is relevant to Pizzagate because Podesta would have had clearance to engage in child trafficking around the globe, under the guise of being the chief diplomat of the USA. Anna Wintour is an interesting one. We all know that the fashion world is known for—
- The most wonderful time of the year: March Madness is here!
- Hope you already have a girlfriend. Married to a wonderful woman with a beautiful daughter, but thanks for your judgment!
- So what the heck do the Classic, Compact, and Complete refer to?! That confused me as well. They are the names that go along with the C1, C2, and C3 names. For example, every C2 is also Compact.
- Still waiting on someone to send someone an ounce of weed. That would have been lovely, lol.
- Yet her country is slowly being turned into a shithole, all thanks to their acceptance of refugees. Their women and children are being raped, men beaten in the street—while she sits and has tea in her guarded safe spot. Erm, I was in Munich before Christmas and I have never felt safer. It was lovely.
- You're pathetic for pretending that he would've joined us. This has naught to do with Arsenal.
- I watched it again about six months ago with my cousin, and we died of laughter when Yugi's grandpa had to go to the hospital after losing a card game. Kid anime logic... so cringe. The creators must really love kids to come up with weird situations like this.