皮肉表現を含めた悪口の判定について

熊本高等専門学校熊本キャンパス
3年　杉田 浩明
3年　宮原 敦也
3年　藤條 照平

1. 実験の概要

三菱総合研究所が2022年に行った統計[1](総務省「プラットフォームサービスに関する研究会」発表)によれば50.1%の人が誹謗中傷をネット上で見たことがあると回答した. この統計からSNSの発達により, SNS上での誹謗中傷が常態化していることがわかる. これら誹謗中傷を制限するシステムは存在しているが皮肉的な悪口には対応できていないのが現状である.

そこで今回我々はSNSの投稿を収集してラベル付けを行い, 自然言語処理モデルBERTで機械学習させ, その識別制度を検証した.

そして約84%の割合で正解できる精度を持つAIを作成することができた.

検証データ作成の流れを図1に表す.

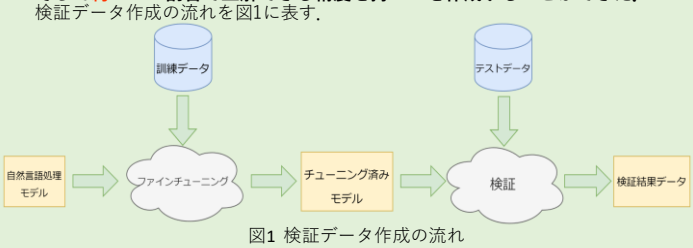

図1 検証データ作成の流れ

2. 実験内容の設計

モデルにはBERTの改良モデルであるdeberta-v2-base-Japanese[2]を用いる. このモデルをファインチューニングすることによって, 遠回しな誹謗中傷を判定する. その実験内容を以下に示す.

(1)訓練データの調査

どのデータが訓練データとして適しているのかを調べた.

統計[1]によれば誹謗中傷を目撃したサービス(複数回答)としてX(Twitter)が52.6%と最も多く, 次いで掲示板サービス(5ちゃんねるなど)が39.7%, Yahoo!コメントが32.0%, YouTubeが28.2%であった.

X(Twitter)のAPIでは現在, 仕様変更が生じており今回はツイートを訓練データとして使わない. またYahoo!コメントについては利用規約よりスクレイピングが禁止されていたためデータを取ることができない. よってYouTubeのコメントと5ちゃんねるの書き込みを取得した. これらを用いて学習させどちらのデータが良い結果になるのかを検証する. またこの時点では悪口を判定するのに適したデータを選んでいるので皮肉的悪口は学習させない.

(2)訓練データの作成

(1)の検証でよい結果を得られた方のサービスからデータを取り, 皮肉的な悪口を含めた訓練データを作成する. ラベルは非悪口を0, 悪口を1, 皮肉的な悪口を2とする. ラベル付けの例を表1に示す.

表1　ラベル付けの例

ラベル	文
0(非悪口)	学生時代ボクシング部に所属していた
1(悪口)	ダサい絡みしながら死んでいく哀れな30代のカス
2(皮肉的な悪口)	この顔で人前に出ようと思えるのすごいわ

(3)エポック数の決定

(2)で作成したデータを使って[2]のモデルをファインチューニングする. その際どのぐらいのエポック数で学習させると性能が上がるかを検証する. 検証の方法は1つ1つエポック数を変化させ, 性能の変化を調べ, 過学習, 学習不足にならないように値を決める.

(4)2人で構成される皮肉的悪口の判定

SNS上で実際に交わされている遠回しな悪口は前後の文章と組み合わせて2人の書き込みで構成されている場合もある. そのような場合, 何も目印がない状態ではどの文章がどの文章とつながって皮肉的な悪口になるか判定することは難しい. よって「@システム」によって考える.

「@システム」とはYouTubeのコメントなどで実装されている返信先のユーザー名を指定するシステムの事と定義する. このシステムを用いて返信先の指定があれば高精度に皮肉的な悪口の判定ができるのかについて検証する.

3. 実験の過程・考察

(1)訓練データの調査

YouTubeのコメント, 5ちゃんねるの書き込みでそれぞれ学習させた. その際のエポック数はいずれも10にした. 訓練データの構成を表2に示す. テストデータを用いて性能を調べた結果を表3に示す. なお, テストデータは非悪口が42件, 悪口が49件で構成される.

表2 訓練データの構成

訓練データ	非悪口	悪口
5ちゃんねる	518件	886件
YouTubeコメント	636件	771件

表3　訓練データによる性能の違い

訓練データ	accuracy	F1	recall	precision
5 ちゃんねる	0.7143	0.7127	0.7143	0.7140
YouTubeコメント	0.6923	0.6893	0.6923	0.6925

表2より5ちゃんねるのデータで学習させた方がYouTubeコメントで学習させるより性能が少し良いことがわかる. そのため実用的に考えると5ちゃんねるの方が訓練データとして適していると考えられる.

(2)訓練データの作成

5ちゃんねるの書き込みから皮肉的な悪口を含めた訓練データを作成した. テストデータについても21件の皮肉的な悪口を追加した.
表4に訓練データの構成を示す.

表4 皮肉的な悪口を含めた訓練データの構成

非悪口(0)	悪口(1)	皮肉的な悪口(2)
1435件	3473件	468件

(3)エポック数の決定

エポック数に対する性能の変化を図2に示す. F1値はマクロ平均で算出した.

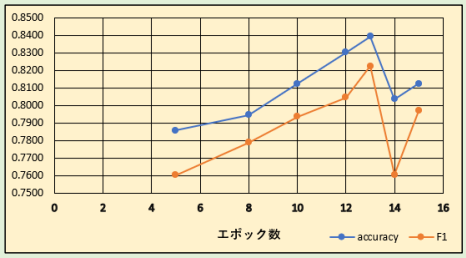

図2 エポック数による精度の推移

図1よりグラフの推移をみるとエポック数が13の時に過学習でも学習不足でもない, ちょうどよいエポック数になっておりaccuracyが約84%あることが分かる.

(4) 2人で構成される皮肉的悪口の判定

以下の文章の流れで実験を行った.
左のアルファベットをユーザー名とし[@(ユーザー名)]となるようになっている. 「」の中が書き込みの内容である.

A:「この俳優マジでかっこいい！」
B:[@A]「なんでこの人こんなに汚いんだろう」
C:[@A]「それ分かるわーー」
D:[@B]「朝から泥水に浸かってきたからだよ」

このDの書き込みはBの書き込みに関連する皮肉である.
(1)のモデルにDのこの1文だけで予測させると非悪口(0)と判定してしまう. また, もし「@(ユーザー名)」などの返信先を指定する方法がない場合, 前後の2文をまとめCとDを結合して判定することになり, その場合もこのモデルは非悪口(0)と判定してしまう. しかし, 「@(ユーザー名)」を利用して返信先を確認できればBの書き込みとDの書き込みの関係性を特定でき, この2文を結合して判定させることで皮肉(2)と判定できる. 仮にBの書き込みが複数ある場合でもDが[@B]とつけていればすべてのBの書き込みと結合して判定させ, BとDの書き込みの関係性を確かめることができる.

4. 実験のまとめ・参考文献

この実験ではAIによる皮肉的な悪口の判定を行い, 約84%の精度を出すことに成功した. (4)の実験では返信先が正しい人になっているかの確認など実施できていない事柄もあり, 精度向上の余地はまだある. 今後も研究を引き続き行い, さらなる精度の向上に努め, 多くの人が使いやすいSNSを作成したい.

参考文献

[1] 三菱総合研究所 (2022)「**インターネット上の誹謗中傷情報の流通実態に関するアンケート調査結果**」https://www.soumu.go.jp/main_content/000813680.pdf
[2] 京都大学言語メディア研究室 「**deberta-v2-base-japanese**」https://huggingface.co/ku-nlp/deberta-v2-base-japanese