E-046

BERT を用いた攻撃的表現の検出

Detecting Offensive Expression Using BERT

三浦彩茄 †　　井上浩孝 †
Miura Ayaka　Hirotaka Inoue

1.　はじめに

近年，LINE や X(旧 Twitter)，Instagram などの SNS の普及が進み，日本における SNS 普及率は 2021 年時点で 82%まで拡大した [1]．SNS はコミュニケーションや情報共有などを簡単に行うことができ，個人生活や企業の経済活動にも深く浸透している．

SNS の匿名性は自由に発言ができるというメリットになるが，誹謗中傷が容易となり，拡散性も高いため個人が投稿した情報が急速に広まるようになり，生活を脅かされる，経済的損失を被るなどとした実害を受けてしまうことが問題となっている．著名人が SNS 上で不特定多数の人から誹謗中傷を受け自殺に追い込まれた事件が起きるなど SNS の誹謗中傷は社会問題化している．そのため，迅速に攻撃的な表現を含む投稿を検出し，被害者の保護を行うことが重要である．

本研究では，入力された文章が攻撃的なものであるか判定することを目的とする．システムは Google が公開している自然言語モデル BERT（Bidirectional Encoder Representations from Transformer）に基づき構築する．

2．手法

本研究では Google Colaboratory 上で開発を行った．

2.1　データ収集

本研究では X（旧 Twitter）の投稿を収集した．リプライ，引用リポストを含むポストを収集対象とし，攻撃的な文章を含む投稿と攻撃的な文章を含まない投稿をそれぞれランダムに収集を行った．最終的に 400 個の投稿が集まった．収集した投稿の分布を表 1 に示す．

表 1　投稿の分類

投稿の内容	投稿数
攻撃的な表現を含む	115
攻撃的な表現を含まない	285
合計	400

2.2　アノテーション

収集したデータに対し，アノテーションを行った．1 リプライにつき攻撃的な表現を含むかどうか「0：含まない」，「1:含む」の二つに分類を行い，データセットを構築した．

2.3　BERT（Bidirectional Encoder Representations from Transformer）

BERT は 2018 年に Google から発表された深層学習モデルである．Transformer がベースとなっており，ファインチューニングを利用することにより様々な自然言語処理に対応可能となっている．図 1 は BERT の学習を示す図である．

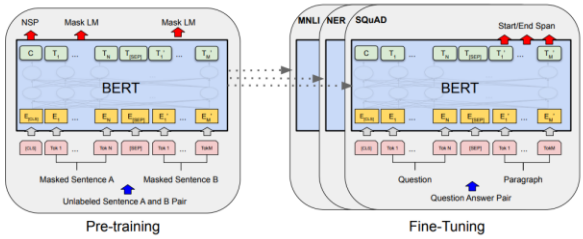

図 1　BERT の学習 [2]

BERT は，事前学習とファインチューニングの二つのフェーズで構成されており，事前学習のフェーズでは，文章を入力しモデルの訓練を行い，事前学習により得られたパラメータを初期値としてファインチューニングを行うしくみとなっている．

本研究では日本語の攻撃的な文章の検出を行うため，日本語 BERT 訓練済みモデルを使用した．このモデルは東北大学自然言語処理研究グループによる，日本語版 Wikipedia をもとにデータベース化し訓練した汎用言語モデル BERT の訓練済モデルで，現在は MeCab(ipadic) と WordPiece で単語分割したモデル，文字単位で単語分割したモデルの 2 種類を公開中である．

2.4　分類器の作成

本研究では事前学習済みモデルとして日本語 BERT 訓練済みモデルを使用し，データとしては 2.2 で作成したデータセットを用いた．この日本語 BERT 訓練済みモデルをファインチューニングすることにより分類器を作成した．

2.5　ファインチューニング

BERT の学習には非常に時間がかかるため，ファインチューニングを使用する．

2.6　評価

攻撃的な表現の判定に，損失関数，正答率を使用して評価を行う．損失関数は BERT で既に定義されているクロスエントロピー誤差を使用する．モデルの予測値を t_k，正解ラベルの one-hot-encoding 表現を y_k とした時のクロスエントロピー誤差は次のように定義される．

$$CrossEntropyLoss = -\sum_{k=1}^{K} t_k \ln y_k$$

正答率は予測値と正解ラベルを比較し，各バッチの正答率を計算し，訓練中の正答率を計算することにより評価を行う．

† 呉工業高等専門学校専攻科　Advanced Course, National Institute of Technology, Kure College

3．攻撃的表現の判定実験

作成したデータセットを訓練用・テスト用に 8：2 の割合で分割を行った．

分類器のファインチューニングにはバッチサイズが 16，オプティマイザーに勾配のスケーリングと重みの減衰処理が互いに干渉しない AdamW を用いる．AdamW の学習率 lr = 0.00001 と設定した．

訓練用のデータセットには 320 個のデータが存在し，そのデータセットから 16 個ずつのデータを取り出し，20 個のミニバッチを作成した．

今回 Epoch 数は 3 に設定し，学習を行った．その学習の様子を図 2，図 3 に示す．図 2 は学習中モデルの損失関数を示している．不安定であるが，学習回数が多くなるにつれ，損失が小さくなり，バッチ数が 50 回以上で損失が 0.3 を下回る結果となった．図 3 は学習中の精度を示す．学習回数が 40 回以上で精度が 0.9 を上回る結果となった．

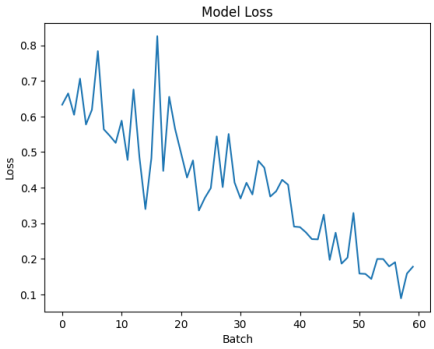

図 2　作成したモデルの損失関数

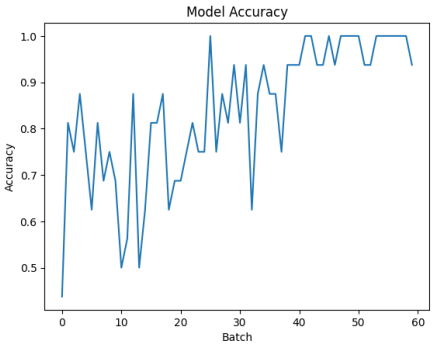

図 3　作成したモデルの精度

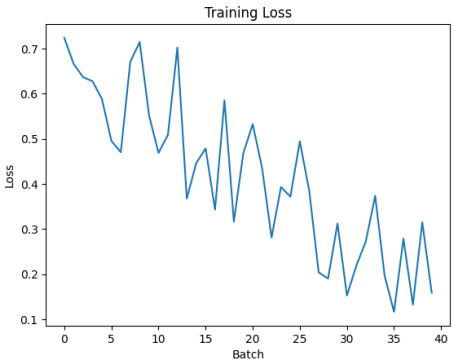

図 4　テストデータの損失関数

テストのデータセットには 80 個のデータが存在し，そのデータセットから 8 個ずつのデータを取り出し，10 個のミニバッチを作成した．そのときの結果を図 4，図 5 に示す．テスト結果も同様に学習回数が多いほど損失関数は小さくなり，精度も高くなった．

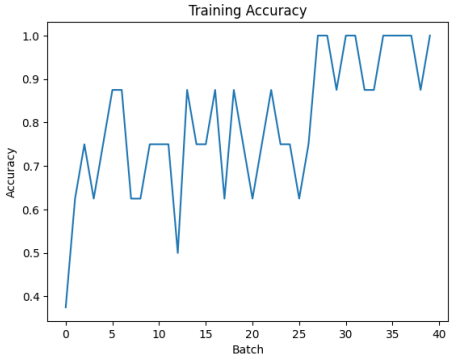

図 5　テストデータの精度

4．考察

図 2，図 3 では学習回数が多くなるにつれ，損失関数が小さくなり，精度が高くなっていることから，モデルの訓練ができていると考えられる．しかし，どちらのグラフものこぎり型のグラフとなっているため非常に不安定なものとなっている．作成したデータセットは X 上で攻撃的な表現を検索しヒットした投稿を収集したため，限定的な内容となっている．そのため，作成したモデルに攻撃的な文章を入力しても，モデルが判断できず，攻撃な文章でないと出力される文章が存在する．データセットを構築する際にはより多くの投稿を収集することに加え検索により得る投稿だけでなく，同じアカウントから投稿される別の投稿を収集することにより広い範囲の投稿の収集を行うことにより，精度の向上が期待できる．

5．まとめ

本研究では，BERT を用いて攻撃的表現を含む文章の判定を行った．SNS の一つである X から攻撃的な表現を含む文章と含まない文章の両方を収集し，ラベル付けを行い 400 個の投稿から成るデータセットを作成した．日本語 BERT 訓練済みモデルを利用することによりファインチューニングを利用し，分類器を作成した．このモデルの学習を行い評価を行ったところ，訓練回数が増えるにつれ損失関数の値は小さくなり，精度が高くなるという結果が得られた．しかし構築した分類器に SNS 上の攻撃的文章を入力したところ、攻撃的でない文章と判定した。今後はより多くの投稿を収集し，精度のよいモデルの作成を行いたい．

参考文献

1) ICT 総研　2022 年度 SNS 利用動向に関する調査 https://ictr.co.jp/report/20220517-2.html/

2) 我妻幸長：BERT 実践入門　PyTorch＋GoogleColaboratory で学ぶあたらしい自然言語処理技術，株式会社翔泳社，2023 年