一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NLC2022-8(2022-09)

汎用言語モデルと複数の特徴を組み合わせた攻撃的な文の判定

李　振銘[†]　嶋田　和孝[††]

[†] 九州工業大学大学院 先端情報工学専攻 知能情報工学分野
〒820-8502 福岡県飯塚市川津 680-4
[††] 九州工業大学 大学院情報工学研究院 知能情報工学研究系
〒820-8502 福岡県飯塚市川津 680-4
E-mail: [†]li.zhenming714@mail.kyutech.jp, [††]shimada@ai.kyutech.ac.jp

あらまし　SNS 上では他人に対する侮辱や攻撃的な言葉がしばしば見られる。このような言葉は他人を精神的に傷つける。したがって、自然言語処理技術を用いて、高精度かつ自動的に攻撃的な言葉を判断してくれるシステムを作る必要性が高まっている。本研究では、(1) 単語レベルでの特徴と (2) 文レベルでの特徴を考える。単語レベルの特徴として、辞書ベースのベクトル化と通常の BOW の二つを考える。文レベルの特徴としては、大規模事前学習言語モデルである BERT と SNS のデータから学習された Emoji 予測モデルの DeepMoji の二つを利用したベクトル化を考える。得られた 4 つのベクトルを連結し、Curious Cat というデータセットを用いて学習する。分類器には Support Vector Machines を用いる。実験結果より、4 つのベクトルをすべて利用した場合の精度が最も高いことが確認され、提案手法の有効性が示された。また、BERT よりも DeepMoji から得られるベクトルの方が分類に貢献することも確認された。

キーワード　攻撃性判定、複数特徴、汎用言語モデル、SNS 投稿

Offensive language detection using the combination of generative language model and machine learning features

LI ZHENMING[†] and Kazutaka SHIMADA[††]

[†] Department of Advanced Informatics, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
[††] Department of Artificial Intelligence, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
E-mail: [†]li.zhenming714@mail.kyutech.jp, [††]shimada@ai.kyutech.ac.jp

Abstract　Nowadays, people often express their abusive and offensive thoughts to others on the Internet much easier. The abusive and toxic comments hurt others seriously. Therefore abusive and toxic comments should be detected correctly by using natural language processing. In this paper, we focus on two types of features in offensive language: word-level and sentence-level. We use lexicon-based and standard bag-of-words features as the word-level. We also introduce BERT-based and DeepMoji-based features as the sentence-level. We apply the four features to a machine learning approach: support vector machines. We evaluate the method with combinations of four features with a dataset, Curious Cat. The best F1 score was generated by the method with all features. This result shows the effectiveness of our proposed method. In addition, the experimental result indicates that DeepMoji generated from Twitter data is better than BERT generated from written language, for an offensive language detection task about SNS data.

Key words　Offensive language detection, Features for machine learning, Pre-trained language model, SNS comments

1. はじめに

SNS 上では他人に対する侮辱や攻撃的な言葉がしばしば見られる。このような言葉は他人を精神的に傷つける。Munro [11] の研究では SNS での侮辱的な発言が児童に対して影響することが報告されている。その研究ではネット上で攻撃的の言論に接したことのある児童は焦燥や絶望などの精神的な影響を受けやすいと報告さていれる。Pew Research Center の 2017 年の調査

と報告書 [6] によると、アメリカでは 40% の成人が実際に SNS で攻撃されており、その中で 18% が悪質なセクハラに遭遇している。ネットユーザの心身の健康を守るために、SNS 上にある有害で攻撃的な投稿を正確に検出する必要が高まっている。

ある投稿に攻撃性があるかどうかをの判断は、SNS に関する知識を一定量持った人間が行うのが最も精度が高いだろう。しかしながら、SNS に投稿される文章の数は爆発的に増加しており、すべての投稿を人間が判断することは実質的に不可能である。したがって、自然言語処理技術を用いて、高精度かつ自動的に攻撃的な言葉を判断してくれるシステムを作る必要性が高まっている。

文の攻撃性の自動判断にはいくつかの観点を考慮する必要がある。本研究では、大きく、(1) 単語レベルでの特徴と (2) 文レベルでの特徴を考える。

単語レベルの特徴として最もシンプルな考え方は、単語の表層を用いるものである。これは、特定の単語が文の攻撃性を顕著化するという仮定である。例えば、

例 1)　彼は馬鹿だ！

という例では “馬鹿” という単語が攻撃性を持っていると考えられる。単語の攻撃性が分かっていれば、その単語の有無によって文が攻撃的かどうかを判断できる。このような処理をするためには単語の攻撃性の情報を付与した辞書を使うことが効率的である。一方で、一般にこのような専門的な特徴が付与された辞書に含まれている語は限定的である。つまり、辞書だけでは、カバレッジが十分に高いモデルを作ることは困難であることが多い。自然言語処理では、古くからこのような表層特徴を Bag-of-words（BOW）として利用してきた。本論文でもその考え方を利用し、辞書情報だけではなく、通常の BOW も利用し、攻撃性の判断モデルを構築する。

単語レベルで判断できることも多いが、文は単語の繋がりであり、文中の文脈を考慮しなければならない場面も多い。攻撃性を持つ単語が文に出現しても文全体に攻撃性が示されていない場合も存在する。例えば、

例 2)　彼は馬鹿正直なぐらい誠実な人だ。

という例では、“馬鹿” という単語は存在するが、攻撃性は極めて低いと考えられる。このように、文全体を通した何らかのモデル化が必要である。本論文では、ニューラルネットワークを用いた文脈モデル（BERT [4]）を攻撃性判断に利用する。加えて、文レベルの特徴として、文の持つ感情情報も考慮する。具体的には、DeepMoji [7] と呼ばれる双方向 LSTM を組み合わせた文から絵文字を予測するモデルを利用する。例えば、例 1 の文が入力されるとネガティブな感情を意味する絵文字が、例 2 の文が入力されるとポジティブな感情を意味する絵文字が出力されることが想定される。また、DeepMoji を利用する理由として、BERT が一般に通常のテキストから学習されるのに対し、DeepMoji は SNS のデータから学習されており、本研究で対象とする SNS データとの親和性も高いという点も挙げられる。

本研究での手法の概要図を図 1に示す。本論文では、前述のように単語レベルおよび文レベルで合計 4 つの特徴についてベクトル化する。それらを連結し、一つのベクトルとして、機械学習を用いて、入力文の攻撃性の有無を判断する。機械学習のモデルには Support Vector Machines（SVM）を利用する。単語レベルと文レベルの特徴を組み合わせることに加えて、SNS の特性を考慮した手法（DeepMoji）を組み合わせることで、SNS を対象とした攻撃性判定のモデルの有効性を実験的に検証する。

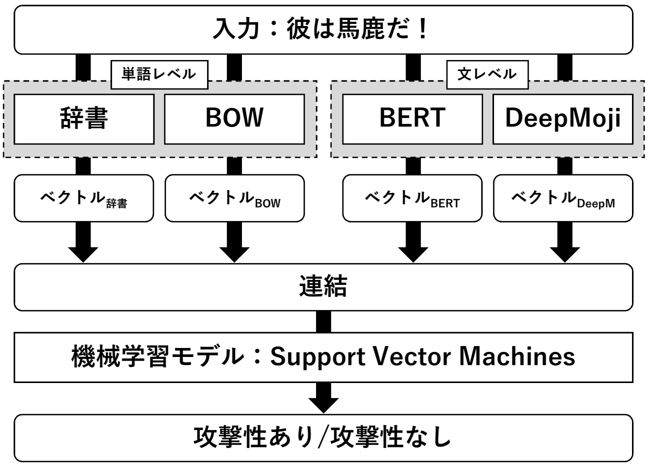

図 1　4 つの特徴量を組み合わせた手法．単語レベルと文レベルの手法から 4 つのベクトルを作成し，それを連結したものを機械学習モデルで学習する．

2. 関連研究

単語レベルの関連研究としては、語彙ベース (Lexicon-based) の手法が挙げられる。Razavi ら [14] は 2700 単語からなる Insulting and abusing language dictionary を構築した。その辞書を分類器に使用し、文の攻撃性を分類している。Njagi ら [12] の研究においても Razavi らと同様の手法が用いられている。

近年はニューラルネットワークを用いた手法が主流になっている。ニューラルネットワークでは文をトークン（単語、文字列など）に分割し、系列から文の攻撃性を学習することが多い。すなわち、単語レベルの手法と比較して文脈的な要素を加味できる場合が多い。Djuric ら [5] の研究は、攻撃性検出のタスクを対象にニューラルネットワークを用いた初期の研究の一つである。彼らは paragraph2vec [10] を用い、Yahoo ファイナンスの Web サイトから収集した英語のコメントデータセットから特徴を抽出し、その特徴ベクトルをロジスティクス回帰分類器で学習し、攻撃性検出をした。Badjatiya ら [1] は、約 1.7 万ツイートに対して人種差別（Racism）や性差別（Sexism）というラベルを付与したデータセット [17] を対象として、複数のニューラルネットワークの組み合わせを試している。最も精度が高かった手法は LSTM と Random Embedding、勾配ブースティング決定木（GBDT）の組み合わせであると報告している。

近年主流なモデルは、大規模汎用言語モデルに基づくものである。汎用言語モデルは大規模なコーパス（Wikipedia など）を学習データとし、言語の汎用的な特徴を学習するモデルである。この汎用言語モデルを特定のタスクでファインチューニングすることでそれ以前の LSTM などの手法と比較して高い精度が得られることが数多く報告されている。その代表は Google 社が

公開している BERT モデルであり [4]、これを基に攻撃性検出のタスクに特化した BERT を構築、利用する研究も多く存在する [2], [3], [9]。

本論文では、上記のような先行研究を踏まえ、旧来の語彙ベースの手法と大規模汎用言語モデルなどを複数組み合わせ、攻撃性検出のタスクでの有効性を検証する。

3. 提案手法

本節では、提案手法に組み込まれる 4 つの特徴を順に説明する。得られた特徴を使用して学習する分類器には図 1に示したように、Support Vector Machines（SVM）を利用する。

3.1 単語レベルの特徴

本節では、単語レベルの特徴量である辞書と Bag-of-Words（BOW）に基づく二つの特徴を説明する。

3.1.1 辞書

単語には攻撃性のあるものとないものがある。そこで、単語の攻撃性の度合いが付与された辞書を用いて入力文をベクトル化する。今回は Michael ら [19] が作成した Abusive Lexicon を用いる。

先行研究での辞書の作り方を簡単に説明する。彼らはまず人手で否定的な表現の基盤語彙リストを作成し、それに攻撃性の有無のアノテーションをしている。結果的に、その基盤語彙リスト中の約 33% に攻撃性ありのラベルが付与された。そして、いくつかの特徴量を算出し、各語彙を分類する分類器に SVM を使って、モデルを構築している。その学習済みの SVM をさらに用い、ラベルなしの negative word list の分類を行い、その list に含まれる単語の攻撃性を判断している。その結果を利用し、スコアリングをすることで −5〜4 弱までの実数値が付与された攻撃性の度合いを含む辞書を構築し、公開している。辞書の一部を以下に示す。

```
horrible_noun  3.6796008
disgusting_adj  3.4936825
moron_noun  3.4696771
bastard_noun  3.3992381
・・・
scorching_noun  0.00056426093
treacherous_adj  0.00020639443
unfucked_adj   -0.00061546087
knout_noun   -0.0017596102
・・・
doubt_noun   -5.1261741
doubt_verb   -5.1454248
```

本論文ではこの公開されている辞書[1]を利用する。この辞書には 8478 単語が含まれている。なお、辞書中で正の値を持つ攻撃性ありは 2989 個であり、残りは攻撃性なし（負の値を持った単語）である。辞書中の単語は上記の例からも分かるように、品詞情報が付与されているが、提案手法では、表記のみを用い、品詞情報は利用しない。つまり、品詞は異なるが表記が同じ単語は一つにまとめる。例えば、一番下にある、`doubt_noun` と `doubt_verb` は doubt として扱われる。その結果、7045 個の単語が選ばれ、この 7045 単語に対して、入力文中に辞書中の各単語が存在するかしないかの二値ベクトル相当でベクトル化する。

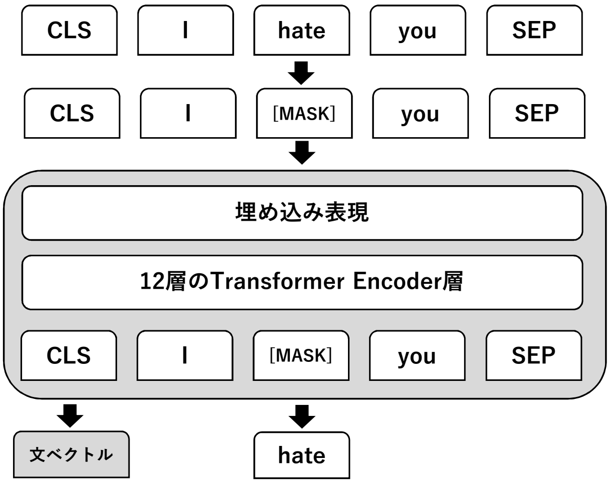

図 2　BERT による Masked Language Model のイメージ．[CLS] トークンを文ベクトルだと見なして SVM の入力に用いる．

3.1.2 BOW

3.1.1節で示した辞書は攻撃性判断に有用な単語群を含んでいる。一方で、攻撃性に関係するすべての単語を含んでいるわけではない。つまり、辞書の単語だけではモデルのカバレッジが十分ではない可能性があるという問題が残る。

この問題を解消するために、提案手法では BOW も併用する。BOW は自然言語処理において古くから使用されている文のベクトル化の手法である。コーパス内に出現する単語を基に重複のないワードリストを作成し、そのリストから単語の有無や出現回数で文のベクトルを生成する。本論文では、ワードリストの作成の際には scikit-learn の CountVectorizer を使い、stop word を除去している。提案手法が用いる BOW のベクトルサイズは 4972 次元である。また、ベクトルの値は単語の有無（バイナリ）とする。BOW では各事例中に出現した単語そのものを特徴として利用しているため、攻撃的な文に頻出する単語が SVM によって効率的に学習されることが期待される。

3.2 文レベル特徴

本節では、文レベルの特徴量である BERT と DeepMoji に基づく二つの特徴を説明する。DeepMoji に関しては SNS としての特徴を加味できる可能性がある。

3.2.1 BERT

BERT は Google 社に発表された大規模汎用事前学習言語モデルである。BERT の事前学習は Mask Language Model (MLM) で実現される。MLM ではまず入力文をトークン単位に分割して、全データセットからランダムに 15% のトークンを選んで [mask] というトークンに置き換える。続いて学習の時に、[mask] のところの本来の単語を正しく予測するように誤差を逆伝搬し、学習を進める。BERT の学習には BooksCorpus [20] と Wikipedia

---

(注1)：`https://github.com/uds-lsv/lexicon-of-abusive-words`

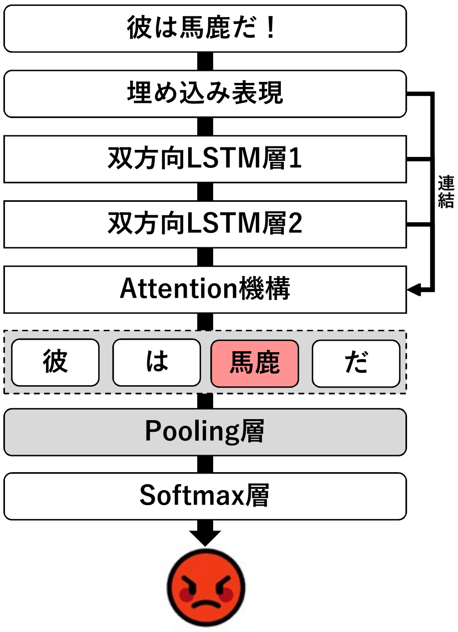

図 3　DeepMoji のイメージ．実際には出力される絵文字を使うわけではなく，Pooling 層で得られたベクトルを SVM の入力として利用する．

の二つのコーパスが使用されている。

BERT の中身は複数の Transformers である (図 2は 12 層の Transformers 層からなる BERT-base である)。入力文がトークンに分割され、埋め込み表現に変換された後、文の先頭に [CLS] と呼ばれる特殊トークンを挿入する。また、複数の文を入力する場合は、[SEP] と呼ばれる特殊トークンを挿入し、文を繋ぐ。複数の Transformer 層は Attention 機構 [16] の計算に基づき、双方向からの文脈を考慮した単語のベクトル表現を生成する。一般的に、出力層の [CLS] トークンが文全体の意味を表現したものになると考えられており、ある文を入力したときの [CLS] トークンを分類器（本論文では SVM）の入力として扱うこととする。

3.2.2　DeepMoji

攻撃的な意味を持つ文はネガティブな感情を伴うことも多く、攻撃性と感情は大きな関係があると考えられる。そこで、入力文の感情推定モデルを利用し、攻撃性判定の特徴として使う。文の感情分析については枚挙に暇がないが [13]、本論文では対象が SNS であるという前提から、SNS によく使われる Emoji について着目する。

Emoji は Oxford Languages の定義によると、アイデアやエモーションを伝えるのに用いられる小さいデジタル画像またはアイコンである。実際に Emoji は SNS 上で多くの人に使用されており、投稿当時の気持ちをシェアし、喜怒哀楽の感情を幅広く表現できる。その中で攻撃性と強く関連する Emoji も存在しており、それらの Emoji に含まれる SNS の文脈情報を利用し、文の攻撃性を検出することができる。

DeepMoji [7] は Twitter から収集された 12 億文のテキストからなる超大規模なコーパスを学習データにする事前学習モデルである。DeepMoji は二つの双方向 LSTM 層（BiLSTM 層）と Attention 機構、Pooling 層、Softmax 層から構成される。その構成を図3に示す。BiLSTM 層は Encoder に相当し、入力文を単語ごとに意味ベクトルを付与する。Attention 機構では BiLSTM 層から抽出された単語意味ベクトルに重みを付与する。例えば、図3では、入力中の “馬鹿” という単語に大きな重みが付与されていることを意味している。最後に Pooling 層で単語意味ベクトルから文全体の意味ベクトルを生成し、Softmax により文の意味ベクトルの次元数を Emoji カテゴリの数に揃え、予測する。

本論文での目的は、実際に Emoji を予測するのではなく、攻撃性判定のための入力データを得ることである。そこで、Emoji を予測するために必要な情報が集積されている Pooling 層に着目する。具体的には、2304 次元から成る Pooling 層を SVM への入力として扱うことにする。

DeepMoji の導入は感情予測を攻撃性判定に組み込むことが第一の目的であるが、SNS 特有の情報を考慮できる可能性がある点も導入の目的である。BERT は非常に強力な大規模事前学習言語モデルであるが、その事前学習には3. 2. 1節でも述べたように Wikipedia のような書き言葉のコーパスが利用されている。一方で、今回の論文で扱うデータセットは SNS を想定している。話し言葉に近い SNS と Wikipedia のような書き言葉のデータとの親和性はそれほど高くない。しかしながら、DeepMoji は主要な SNS の一つである Twitter のデータから学習されており、データの種類としても、今回のタスクと親和性が高い。したがって、DeepMoji の導入により、感情のみならず、学習媒体の有用性も検証することが可能になる。

4.　実　　験

本節では既存のデータセットを用いて提案手法の有効性を検証し、考察する。

4.1　データセット

本研究では Niloofar らが構築した英語の攻撃文のデータセット [15] を使用する。このデータセットでは、まず、Curious Cat というサイトから約 50 万文の質問-応答ペアを収集している。次に、著者らが設計した悪意を検出するアルゴリズムを用いてその 50 万のペアから 2482 ペアを抽出する。最終的にこの 2482 ペアに対して人手で攻撃意思のありなしをアノテーションして構築された。

実験の方法もこの論文の方法に倣う。具体的には、先行研究で分割された学習、開発、検証の三つを本研究ではそのまま使用し、質問-応答ペアの文を区分なく実験に使用する。ただし、本論文での実験に際し、いくつかの文を文字コードの問題から事前に削除している。このデータセット中の文には Unicode 文字（ASCII 文字以外の文字や記号）が含まれている場合がある。このような Unicode 文字（たとえば漢字やハングルなど）は顔文字や表情記号として使われることが多い。しかしながら、Unicode 文字をそのまま利用すると英語で学習されたモデルや辞書と上手く適合しないなどの問題が生じる。そこで、今

表 1　実験結果．下記の結果は「攻撃性あり」の評価値である．

文レベル	単語レベル	Precision	Recall	F1
BERT	なし	0.726	0.528	0.611
	+ 辞書	0.706	0.500	0.585
	+BOW	0.603	0.676	0.638
	+ 辞書 +BOW	0.609	0.685	0.645
DeepMoji	なし	0.728	0.644	0.683
	+ 辞書	0.644	0.620	0.632
	+BOW	0.650	0.741	0.693
	+ 辞書 +BOW	0.613	**0.778**	0.686
BERT+DeepMoji	なし	0.604	0.769	0.676
	+ 辞書	0.735	0.644	0.686
	+BOW	0.671	0.755	0.710
	+ 辞書 +BOW	**0.751**	0.699	**0.724**

表 2　最高精度となった組み合わせのハイパーパラメータ．

C	degree	gamma	kernel
14.4837	5	2.7589	rbf

回はFelboらが設計したアルゴリズム[2]を用い、Unicode文字が含まれる文はデータセットから除外した。その数は687文である。したがって、全データ（2482 × 2）からこの687文を除いた4277文を使って実験を行う。なお、4277文のうち、攻撃性ありというラベルが付いたデータは834文であり、残りの3443文が攻撃性なしとラベル付けされた不均衡なデータ分布になっている[3]。

4.2 実験設定

図1で示したように、本論文での提案手法では分類器としてSVMを用いている。実験では、BERTの出力のみをSVMの入力に使う場合、DeepMojiの出力のみをSVMの入力に使う場合、BERTとDeepMojiの二つの出力を連結したものをSVMの入力に使う場合の三つを基盤とし、それらのモデルに辞書とBOWを追加した場合の合計12パターンの精度評価を行う。評価にはPrecision、Recall、F1値を用いる。なお、「攻撃性あり」とラベル付けされたもの側のみを評価の対象とした。つまり、次節以降で出てくる評価値は、2つのクラス（攻撃性ありとなし）の重み付き平均などではない。

また、SVMのハイパーパラメータはBayesian Optimizationを用いて探索した。Bayesian OptimizationはGaussian ProcessとBayesian Inferenceの理論をベースに、連続のパラメータ空間を短時間で探索することのできるアルゴリズムである。このハイパーパラメータの探索には開発データセットを用い、12パターンの組み合わせそれぞれで設定をした。

4.3 結果と考察

表1に実験の結果を示す。表中の太字は、各評価尺度で最も高い値を意味する。4.1節でも説明したように、今回の実験ではデータセットが不均衡であるため、F1値を主な指標にし、考察を行う。表中で、“+ 辞書” となっている部分は、基盤となる文レベルの特徴量に加えて、辞書中の単語の有無をベクトルとして追加したことを意味している。表から分かるように、4つの要素をすべて利用した手法（BERT+DeepMoji+ 辞書 +BOW）が一番高いF1値となった（表1の最下部）。この結果から、本論文での提案手法の有効性が確かめられた。このときのハイパーパラメータを表2に示す。

まず､辞書とBOWの有効性について考察する。F1値でみると、辞書に含まれる単語の追加(表中の“+ 辞書”)はBERT+DeepMojiの場合は有効であったが、BERT単体を使う場合とDeepMoji単体を使う場合では数値が低下する結果となった。たとえば、DeepMojiの場合ではF1値が0.683から0.632と大幅に低下している。F1値の上がったBERT+DeepMoji+ 辞書の場合も含め、どの組み合わせでもRecallが低下している。これは、辞書だけではカバレッジが十分ではないという問題と関連がある可能性がある。BOWを利用した場合（表中の “+BOW”）はすべての場合で「単語レベル」の特徴が “なし” と比較してF1値が向上している。特にRecallが改善されており、これは前述の辞書のみの場合の問題をBOWによって改善できていることを示唆している。また、辞書単体、BOW単体と比較して、辞書 +BOWの追加の方がF1の改善に繋がっており、二つの単語レベルの特徴を組み込むことの有効性が示されている。

次に文レベルの特徴（BERTとDeepMoji）について考察する。BERT（単語レベルなし）とDeepMoji（単語レベルなし）を比較すると、F1値はDeepMojiの方が大幅に高い（0.611 vs. 0.683）。これは一般的な汎用言語モデルであるBERTよりも感情推定を目的として作られたDeepMojiの方が、攻撃性の判断には有効であるという可能性を示唆している。加えて、DeepMojiはBERTとは異なり、SNSのデータで学習されているという点も精度向上に繋がった可能性がある。

BERT+DeepMoji（単語レベルなし）に関しては、F1値では、DeepMoji（単語レベルなし）よりも低い値になったが、それ以外（すなわち単語レベルを加える場合）では、BERT単体およびDeepMoji単体よりもF1値が向上している。この結果は、文レベルの特徴も複数を組み合わせることが有効であることを意味している。また、単に複数の特徴量を組み合わせることが重要であるわけではなく、前述のようにBERTは書き言葉で、DeepMojiは話し言葉に近いコーパスで学習されているという点が重要であると考えられる。つまり、精度のさらなる向上のためには、多様な特性を持った特徴量を組み合わせることが重要であるといえる。

5. おわりに

本論文では、SNSを対象とした攻撃性の有無を判定するタスクに取り組んだ。攻撃性判定のために「単語レベルの特徴」と「文レベルの特徴」という観点を導入し、辞書ベースのベクトル化、BOWによるベクトル化、BERTを用いたベクトル化、DeepMojiを用いたベクトル化の4つの特徴ベクトル作成を経て、機械学習によりモデルを構築した。実験結果より、BOWは辞書ベースのみのベクトル化で生じるカバレッジの問題を改善している可能性が実験的に示された。また、BERTとDeepMoji

---

（注2）：`https://github.com/bfelbo/DEEPMOJI`

（注3）：ただしこれは今回のUnicode削除処理で生じた不均衡さではなく、オリジナルのデータセットがそもそも持っていた不均衡さである。

を比較した場合、DeepMoji の方が有効であり、これは感情が攻撃性に強く関係している可能性と学習したコーパスの差異が精度に影響する可能性を示唆している。実験全体で見ると、4 つのベクトルをすべて利用した手法が最も高い F1 値を得ており、本論文での提案手法の有効性が確認された。

提案手法から、複数の多様性を持ったベクトルを組み合わせて学習することの有効性が示された。精度向上のために、新たな特徴の追加が必要になる。一つのアイデアとしては、今回利用した文レベルと単語レベルの特徴に加えて、文字レベルの特徴を加えることである。固有表現抽出や関係抽出のタスクでは文字レベルの Embedding である Contextual String Embeddings（CSE）が一定の有効性を示している [8], [18]。CSE の導入は今後の課題の一つである。

また、今回は分類器として SVM を用いたが、SVM が最適である保証はない。他の学習アルゴリズムとの比較やファインチューニングされた BERT との統合など、分類器そのものの工夫や改善も精度向上のための重要な手段である。

今回は SNS の投稿文を対象とし、そのためのデータセットの一つである Curious Cat を用いた。提案手法が他のデータセットでも同様の精度を得ることができるかを検証する必要がある。また、今回は英語のデータを対象としたが、日本語など他の言語での実験も興味深い課題である。さらに、今回のデータセットは比較的明示的な攻撃表現が含まれるデータであったが、日常生活であからさまに攻撃的な投稿や発話をするわけではない。明示的な表現が使われないが、ある種の攻撃性を含んでいるような言動は少なからず存在する（たとえば皮肉や無礼な表現など）。今後はそのようなさらに深い言語処理・言語理解の研究を進めたい。

謝辞　本研究の一部は科研費 20K12110 の助成を受けたものです。

文　献

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep Learning for Hate Speech Detection in Tweets. June 2017.

[2] Tommaso Caselli, Valerio Basile, Jelena Mitrovi?, and Michael Granitzer. HateBERT: Retraining BERT for Abusive Language Detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, August 2021. Association for Computational Linguistics.

[3] Camilla Casula, Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. FBK-DH at SemEval-2020 Task 12: Using Multi-channel BERT for Multilingual Offensive Language Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1539–1545, Barcelona (online), February 2020. International Committee for Computational Linguistics.

[4] 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)*, pages 4171–4186, 2019.

[5] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 29–30, New York, NY, USA, 2015. Association for Computing Machinery. event-place: Florence, Italy.

[6] Maeve Duagan. Online harassment 2017, 2017.

[7] Bjarke Felbo, Alan Mislove, Anders Sogaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, 2017. arXiv:1708.00524 [cs, stat].

[8] Satoshi Hiai, Kazutaka Shimada, Taiki Watanabe, Akiva Miura, and Tomoya Iwakura. Relation extraction using multiple pre-training models in biomedical domain. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 530–537, Held Online, September 2021. INCOMA Ltd.

[9] Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online, January 2020. Association for Computational Linguistics.

[10] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[11] Emily Munro. The protection of children online: A brief scoping review to identify vulnerable groups, 06 2011.

[12] Dennis Njagi, Z. Zuping, Damien Hanyurwimfura, and Jun Long. A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230, April 2015.

[13] Bo Pang and Lillian Lee. *Opinion mining and sentiment analysis*, volume 2. Foundations and Trends in Information Retrieval.

[14] Amir Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. Offensive Language Detection Using Multi-level Classification. pages 16–27, May 2010.

[15] Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. Attending the Emotions to Detect Online Abusive Language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88, Online, 2020. Association for Computational Linguistics.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, December 2017. arXiv:1706.03762 [cs].

[17] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics.

[18] Taiki Watanabe, Akihiro Tamura, Takashi Ninomiya, Takuya Makino, and Tomoya Iwakura. Multi-task learning for chemical named entity recognition with chemical compound paraphrasing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6244–6249, Hong Kong, China, November 2019. Association for Computational Linguistics.

[19] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a Lexicon of Abusive Words ? a Feature-Based Approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[20] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books, June 2015. arXiv:1506.06724 [cs].