Vol.2018 No.9, B1-1
2019/2/21

多特徴を用いたTwitter上のネットいじめの自動検出

大友泰賀[†1]　張建偉[†1]

パソコンやスマートフォンの普及に伴い、ネット上のいじめが深刻な問題となっている。本研究ではTwitter上のテキストを対象とし、多くの特徴量を用いていくつかの機械学習手法と組み合わせ、ネットいじめの自動検出に最適なモデルの構築を図った。構築したモデルを用いていじめ文、非いじめ文の分類の精度評価を行ったところ、最も精度の良かったモデルでは90%を超える精度評価を得ることができた。

1. はじめに

近年、パソコンやスマートフォンの普及に伴い、ネット上のいじめが深刻な問題となっている。平成30年10月の文部科学省の調査では、平成29年度のネットいじめの認知件数は1万2632件で、年々増え続けている。日本の現状では、ネットいじめの発見は主に被害者からの報告やアンケート調査などによる手動のものである。ネットいじめに関する研究は主に教育分野で検討されており、条例や教材などの開発に留まっている。しかし、膨大なウェブデータ中に存在するネットいじめの発見には、これらの手法の効果が限られていると思われる。一方、情報分野では、ネットいじめの自動検出に係る研究はあるが、ほとんどは英語データの分析であり、技術も未熟である。日本語のデータに対して、ネットいじめを精度良く自動的に検出できる技術が求められている。

本研究ではネットいじめの自動検出を目的とする。目的の実現のために機械学習を用いる。機械学習には特徴量の抽出と機械学習手法の選択が必要である。Twitterのテキスト(以後、ツイートと呼ぶ。)を対象とし、ネットいじめの検出に貢献度の高い特徴量の抽出を検討する。特徴量には主にテキストマイニングの手法を用いて、Nグラム、Word2vec、Doc2vec、ツイートの感情値、Twitterの特有の特徴を使用した。また、複数の機械学習手法を用い、特徴量と組み合わせて、最適なモデルの構築を図る。収集したツイートをもとにモデルを構築し、いじめ文か非いじめ文かの自動検出の精度評価を行った結果、最も良かったモデルでは90%を超える精度を得られた。

本論文の構成を述べる。第2章では、関連研究を紹介する。第3章では提案手法の流れについて説明する。第4章ではデータ収集の方法と収集したツイートに対する前処理の方法について述べる。第5章ではいじめ文か非いじめかの分類に用いた特徴量について述べる。第6章ではいじめ文か非いじめ文かの分類に用いた機械学習手法について述べる。第7章では収集したツイートをいじめ文か非いじめ文かに正しく自動分類できるかの実験とその結果についての考察を述べる。第8章では本論文のまとめとこれからの課題について述べる。

†1 岩手大学 Iwate University

2. 関連研究

中村ら[1]は、Twitter上でネットいじめに関連する単語を組み合わせて検索し、ネットいじめに関連する投稿を発見できないかを試みている。さらにネットいじめ加害・被害ユーザ検出アルゴリズムを検討し、それらの流れについて述べている。三島ら[2]は、中高生向けソーシャルメディア上でソーシャルグラフを生成し、仲間集団の中で発生したネットいじめを検出している。

英語の攻撃的なテキストの自動検出に関する研究は盛んに行われている。Nobataら[3]はYahoo!金融とYahoo!ニュースのコメントからNグラム、言語、構文、分布意味論の特徴を使って有害なコメントの自動検出を試みている。Hosseinmardiら[4]はInstagramと呼ばれる写真共有サービスから写真とその写真につけられたコメントをセットで取得し、写真の情報、投稿したユーザー、コメント間の時間、Nグラムの特徴を用いてネットいじめやネット攻撃の自動検出を試みている。Rafiqら[5]はVine(現在はサービス終了)と呼ばれる動画共有サービス上から動画とその動画につけられたコメントをセットで取得し、動画の情報、投稿したユーザー、コメントの感情、Nグラムの特徴を用いてネットいじめやネット攻撃の自動検出を試みている。Chatzakouら[6]は、Twitter上からハッシュタグを利用してツイートを取得し、それらのツイートをしたユーザーらをいじめユーザー、攻撃的ユーザー、スパムユーザー、普通のユーザーの4クラスに分類し、ユーザー、テキスト、ネットワークの特徴を用いて自動で正しく分類できるかを試みている。Burnapら[7]は、Twitter上からとある殺人事件についてのツイートをハッシュタグを用いて取得し、Nグラム、単語の依存関係の特徴を用いて人種や宗教の点で攻撃的なツイートの自動検出を試みている。

本研究ではNグラムという基本的な特徴量だけではなく、Word2vec[8]、Doc2vec[9]、感情辞書による感情値[10][11]や、Twitter特有の特徴などを特徴量を用いてTwitter上の日本語のツイートに対して、ネットいじめの自動検出を試みるという点がこれらの研究とは異なっている。

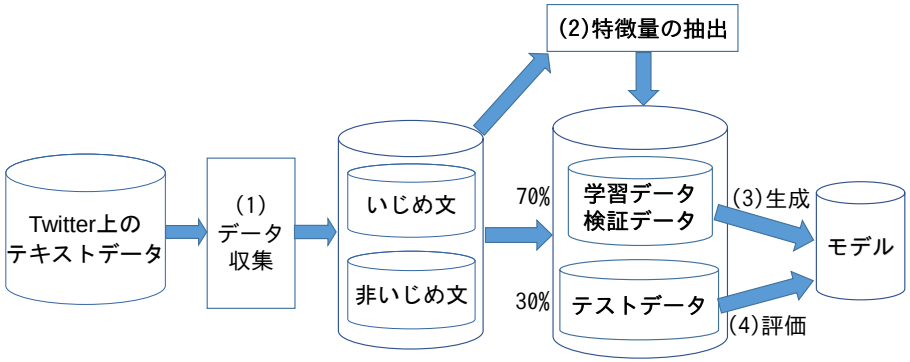

図 1 提案手法の流れ

3. 提案手法の概要

提案手法の流れを図 1 に示す。

(1) データ収集 (4.1 節で説明)

同研究室のメンバーのいじめ単語を用いて収集する方法 (主にいじめ文) と著者のランダムに収集する方法 (主に非いじめ文) の 2 つの方法によって Twitter 上からツイートを収集し、ツイートの中身に応じていじめ文、または非いじめ文のラベルを付ける。

(2) 特徴量の抽出 (4.2 節、5 章で説明)

形態素解析に不要と思われる単語を排除してから MeCab[12] を用いて形態素解析をし、収集したツイートから各特徴量を抽出する。特徴量には N グラム、Word2vec、Doc2vec、ツイートの感情値、Twitter 特有の特徴を用いる。

(3) モデルの生成 (6 章で説明)

収集したツイートを学習・検証データとテストデータに分け、学習・検証データと各特徴量と各機械学習手法を用いてモデルを生成する。機械学習モデルには線形サポートベクトルマシン、ロジスティック回帰、決定木、ランダムフォレスト、勾配ブースティング回帰木、パーセプトロンを用いる。また、検証データでの学習には交差検証とグリッドサーチを用いてより良いモデルの生成を目指す。

(4) 評価 (7 章で説明)

テストデータを用いて生成したモデルがどれだけ正しくいじめ文か非いじめ文かを分類できるかの評価を行う。評価基準には正解率、適合率、再現率、F 値を用いる。

4. データ収集と前処理

4.1 データ収集

始めに Twitter 上のテキストであるツイート文の収集を行う。ツイート文はまず、同研究室のメンバーが収集したものを用いようとした。収集方法は石坂ら [13] と新田ら [14] と畠山ら [15] の研究で用いられていた 36 個のいじめ単語を使用して 2,349,052 ツイートを収集。次にツイートの中に含まれているいじめ単語の数に応じてスコア付与を行い、3,450 ツイートに絞る。最後に Yahoo!クラウドソーシングを利用して 1 ツイートにつき 3 人の利用者にツイートがいじめ文か非いじめ文かを判断してもらい、3 人一致でいじめ文と判断した 1,395 ツイートにいじめ文のラベルを付け、3 人一致で非いじめ文と判断した 282 ツイートに非いじめ文のラベルを付けた。

しかし、この時点でいじめ文のラベルと非いじめ文のラベルの数に大きな差があり、このまま機械学習をさせてもいいモデルが出来上がらない。そこで非いじめ文と思われるツイートを再び収集することにした。収集には Twitter API を用いた。Twitter API とは Twitter 社が公開しているものであり、様々な条件でツイートやユーザーを取得可能であり、そのツイートやユーザーに関する様々な情報を確認することができる。これを用いて公開されている全ツイートからランダムに収集するものとした。この方法によって 1,113 ツイートを収集し、それらのツイートに非いじめ文のラベルを付けた。その結果、非いじめ文のラベルの数が 1,395 ツイートとなり、いじめ文のラベルの数と非いじめ文のラベルの数を均等にした。

いじめ文 1,395 ツイート、非いじめ文 1,395 ツイートの計 2,790 ツイートを実験に使用する。

4.2 形態素解析

特徴量の抽出のために収集した 2,790 ツイートの形態素解析を行う。なお、事前に形態素解析に不要と思われる特定の文字を排除する。排除する文字は Twitter 特有の単語である “RT”、“お気に入り”、“まとめ”や URL、半角記号、数字、英字、全角記号、である。さらに改行文字、全角空白、複数の連続した半角空白は全て 1 つの半角空白に変換した。

形態素解析には MeCab を使用した。MeCab とは京都大学情報学研究科コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンである。一行一文を前提として解析を行い、ツイートを単語単位に分解し、表層形 (ツイート上の単語の形) から品詞、細かい品詞分類、活用形、原形、読み、発音を解析してくれる。なお、5 章で出てくる「単語」という言葉は基本的に表層形のことを指す。

本研究では多数の Web 上の言語資源から得た新語を追加することでカスタマイズした MeCab 用のシステム辞書である mecab-ipadic-NEologd を用いた。この辞書の利点は MeCab の標準のシステム辞書では正しく分割できない固有表現などの語の表層とフリガナの組を約 310 万組採録してある。さらに開発サーバ上で辞書の更新が Web 上の言語資源の活用によって毎週 2 回更新されてるため、比較的新しい単語やネット上でよく使われる単語に対しても対応することができる。

5. 特徴量の抽出

機械学習を用いるいじめテキストの検出には、分類に有効な特徴量の選択が重要である。本研究では収集した 2,790 ツイートから N グラム、Word2vec、Doc2vec、ツイートの感情値、Twitter の特有の特徴を抽出し、これらを特徴量として使用した。

5.1 N グラム

任意の文字列や文書を連続した N 個の文字列または単語で分割し、それがどの程度出現するかを調査する言語モデルである。文字列や単語の発生確率が直前の文字列や単語に依存すると仮定して扱われる。いじめに使われるような単語を抽出し、分類に役立ってくれるのではと推測し、取り入れた。

本研究では文字単位での分割 (以下文字 N グラム) と単語単位での分割 (以下単語 N グラム) をともに使用し、文字 N グラムでは N=2∼5、単語 N グラムでは N=1∼5 とした。文字 N グラム、または単語 N グラムを生成したら、各文字、単語の連なりを要素とした行列ベクトルを生成し、ツイート内の各要素の出現頻度を各ベクトルの値とする。

しかし、このまま各要素一つ一つを全て特徴量として使ってしまうと膨大な数の特徴量となってしまい、明らかに全く重要ではない要素も多数存在している。さらに文字 N グラムに関しては、それらのほとんどが「文章的には意味のない文字の組み合わせ」であり、ノイズの発生にもなってしまう。この問題を解決するために主成分分析という方法を用いた。

主成分分析とは、データセットの特徴量を相互に統計的に関連しないように回転する手法である。回転したあとの特徴量から、データを説明するのに重要な一部の特徴量だけを抜き出す。アルゴリズムとしてはまず最も分散が大きい方向を見つけ、それに「第 1 成分」というラベルを付ける。データはこの方向に対して最も情報を持つ。つまりこの方向は特徴量が最も相互に関係する方向である。次に「第 1 成分」と直交する方向の中から、最も情報を持っている方向を探す。それに「第 2 成分」というラベルを付ける。このようにして見つけていく方向がデータの分散が存在する主要な方向であり、「主成分」と呼ぶ。主成分はもとの特徴量と同じ数だけ存在するが、主成分はデータを説明するのに重要な順にソートされており、パラメータで主成分の残す数を決めることで、重要度の大きい主成分のみを残しながら次元削減をすることができる。本研究では各 N グラムごとに 100 次元まで次元削減を行った。

5.2 Word2vec

Word2vec とは、大量のテキストデータを解析・学習し、各単語の意味をベクトル表現化する手法である。各単語のベクトル同士の内積を計算することで意味の近い単語を知ることができたり、単語のベクトル同士の加算減算などをすることによって単語間の関係性を理解することができる。

本研究では各単語のベクトルを利用することでいじめ文の分類に役立てようとした。学習には 4.1 節で述べた 36 個のいじめ単語を用いて収集した 2,349,052 ツイートを用いた。各単語の単語ベクトルは 100 次元に設定した。1 ツイート単位で分類をするため、1 ツイートに含まれる各単語の単語ベクトルの平均をそのツイートの Word2vec の特徴量とした。

5.3 Doc2vec

Doc2vec とは、大量のテキストデータを解析・学習し、ベクトル表現化する手法である。Word2vec と異なる点は、ベクトル表現化の際に文中の単語の語順を考慮する点と、ベクトル表現化するのが単語単位だけではなく、文書単位でもあるという点である。これらの点により文書間の関係性も理解することが出来る。学習には Word2vec と同じく 36 個のいじめ単語を用いて収集した 2,349,052 ツイートを用いた。文書 (ツイート) ベクトルは 100 次元に設定した。1 ツイート単位で分類をするため、ベクトル表現化も 1 ツイート単位で行い、そのベクトルをそのツイートの Doc2vec の特徴量とした。

5.4 ツイートの感情値

ツイートの感情値を感情辞書を用いて算出し、特徴量として利用する。感情辞書とは、単語とその単語が表す感情の種類とその程度を数値で表すものである。いじめ文はネガティブな単語が多く含まれており、文章としてもネガティブなものになっているのではないかと推測をし、特徴量に取り入れた。本研究では熊本らの「3 軸感情辞書」[10] と高村らの「単語感情極性対応表」[11] の 2 つの感情辞書を使用した。なお、「3 軸感情辞書」[⋆1]、「単語感情極性対応表」[⋆2]はともに Web 上で公開されている。

⋆1 3 軸感情辞書 http://www.zl.cis.iwate-u.ac.jp/∼zjw/wiki/index.php?%E6%84%9F%E6%83%85%E8%BE%9E%E6%9B%B8
⋆2 単語感情極性対応表 http://www.lr.pi.titech.ac.jp/∼takamura/pndic_ja.html

5.4.1 3 軸感情辞書

3 軸感情辞書とは、熊本ら [10] が「ある感情を有する単語はその感情を表現する感情語群と共起しやすく、逆の感情を表現する感情語群とは共起しにくい」という仮定のもと、新聞記事データを用いて、ある単語と対比的な感情を有する 2 つの感情語群との共起の仕方を調べ、数値化したものを、その単語の感情値として辞書形式でまとめたものである。

感情軸には、感情特性を多変量解析手法を用いて分析することにより、感情語 42 語から新聞記事の感情を表現するのに適した感情軸を抽出することによって、楽しい⇔悲しい、うれしい⇔怒り、のどか⇔緊迫の 3 軸で構成されている。単語の感情値はそれぞれ 0〜1 の値で、単語の感情値が 1 に近いほど、単語の感情が「楽しい」「うれしい」「のどか」に近いものとなり、単語の感情値が 0 に近いほど、単語の感情が「悲しい」「怒り」「緊迫」に近いものとなる。

5.4.2 単語感情極性対応表

単語感情極性対応表とは、高村ら [11] が各単語が一般的に良い印象を持つか (ポジティブ) 悪い印象を持つか (ネガティブ) を-1〜1 の値で表したものである。「岩波国語辞書 (岩波書店)」をリソースとして、語彙ネットワークを利用して自動的に計算されている。単語の感情値が 1 に近いほど、単語の感情が「ポジティブ」に近いものとなり、単語の感情値が-1 に近いほど、単語の感情が「ネガティブ」に近いものとなる。

5.4.3 ツイートの感情値の算出方法

まず、MeCab で形態素解析をしてから各単語の原形を取得する。次に各単語の表層形、または原形が各感情辞書の中に含まれているかどうかをチェックする。含まれていれば対応する感情値をリストに追加し、含まれていなければ各感情辞書の値の範囲の中央値 (3 軸感情辞書ならば 0.5、単語感情極性対応表ならば 0) を感情値としてリストに追加する。ツイート内の全ての単語を確認し終えたら、リストの平均値を求め、その値を文章の感情値とする。

5.5 Twitter 特有の特徴

Twitter には他の人のツイートを再びツイートする「リツイート」、他の人の気に入ったツイートを登録する「お気に入り」、自分のツイートをカテゴライズして検索を容易にする「ハッシュタグ」という機能が存在しており、ツイートにはこれらの情報も含まれている。本研究ではこれらの数とツイート中の URL の数を特徴量として利用した。

6. 機械学習モデルの選択

機械学習を用いるいじめテキストの検出には、機械学習手法の選定も重要である。本研究では単純なモデルである線形モデルの (1) 線形サポートベクトルマシンと (2) ロジスティック回帰、木構造のモデルである (3) 決定木と (4) ランダムフォレストと (5) 勾配ブースティング回帰木、複雑なモデルであるニューラルネットワークの一種である (6) パーセプトロンを使用した。

7. 評 価 実 験

7.1 実験の設定

収集したツイートを用いて、各特徴量、各機械学習手法によって学習させ、どれだけ正しく分類できるか、そしてどの特徴量、どの機械学習手法が分類に役立つかの実験を行う。

分類に用いるツイートは 4.1 節で述べたいじめ文のラベルが付いた 1,395 のツイートと非いじめ文のラベルが付いた 1,395 のツイートである。

特徴量に関してはまず、1 種類ずつ使用した。単語 N グラム (N=1〜5)、文字 N グラム (N=2〜5)、Word2vec、Doc2vec、感情 (楽しい⇔悲しい、うれしい⇔怒り、のどか⇔緊迫、ポジティブ⇔ネガティブ)、Twitter 特有の特徴量 (リツイート数、お気に入り数、URL 数、ハッシュタグ数) の 6 種類である。(表 1)

次に最も適切に分類できていた特徴量をベースとして、ツイートの感情値の特徴量と組み合わせて使用した。最後は全ての特徴量を用いて使用した。ただし、あまりにも多い次元数とならないように文字 N グラムと単語 N グラムに関してはそれぞれ単体で試した時に決定木を図示しておき、分類に貢献していた文字 2 グラム、文字 4 グラム、単語 1 グラムのみ用いている。

機械学習に関しては全ての機械学習手法共通で、実験に用いるツイートの 70%を学習・検証データ、30%をテストデータとした。そして層化 5 分割交差検証を利用し、学習データを各分割内でのラベルの比率が全体の比率と同じになるように 5 個に分割し、5 個のうち 4 個を学習データ、1 個を検証データとし、モデルの訓練と評価を行う。一度評価と訓練が終わったら 4 個のうち 1 個の学習データと検証データを入れ替え、再びモデルの訓練と評価を行う。これを 5 回行い、5 回の平均分類精度を出し、その値を分類精度として使うことで、データの分割によらない頑健な評価を可能にしている。

各機械学習手法のパラメータの調整に関しては、グリッドサーチを用いてあらかじめ指定していたパラメータの全ての組み合わせに対して前述した交差検証をを行い、もっとも良い分類精度を示したパラメータのモデルを生成する。

最後に生成した各モデルでテストデータをどれだけ正しく分類できるかを試した。テス

表 1 特徴量の種類分け

特徴名	含まれている特徴量
単語 N グラム	単語 1 グラムを主成分分析したもの(100 次元)
	単語 2 グラムを主成分分析したもの(100 次元)
	単語 3 グラムを主成分分析したもの(100 次元)
	単語 4 グラムを主成分分析したもの(100 次元)
	単語 5 グラムを主成分分析したもの(100 次元)
文字 N グラム	文字 2 グラムを主成分分析したもの(100 次元)
	文字 3 グラムを主成分分析したもの(100 次元)
	文字 4 グラムを主成分分析したもの(100 次元)
	文字 5 グラムを主成分分析したもの(100 次元)
Word2vec	各単語ベクトルの平均 1〜各単語ベクトルの平均 100(100 次元)
Doc2vec	文書ベクトル 1〜文書ベクトル 100(100 次元)
ツイートの感情値	楽しい⇔悲しい
	うれしい⇔怒り
	のどか⇔緊迫
	ポジティブ⇔ネガティブ
Twitter 特有の特徴	リツイート数
	お気に入り数
	URL 数
	ハッシュタグ数

トデータの評価に関しては正解率 (Accuracy)、適合率 (Precision)、再現率 (Recall)、F 値 (F-measure) の 4 つの値を利用した。

正解率 (Accuracy) は、モデルがいじめ文や非いじめ文と判断したデータのうち、実際にそうであるものの割合である。

適合率 (Precision) は、モデルがいじめ文と判断したデータのうち、実際にいじめ文であるものの割合である。

再現率 (Recall) は、実際にいじめ文であるデータのうち、いじめ文であると予測されたものの割合である。

F 値 (F-measure) は、適合率と再現率の調和平均をとった値である。

それぞれの値を求める式は、以下のようになる。

表 2 分類精度

	単語Nグラム	文字Nグラム	Word2vec	Doc2vec	ツイートの感情値	Twitter特有の特徴	文字Nグラム+ツイートの感情値	全ての特徴
線形サポートベクトルマシン	-	-	0.827 0.915 0.731 0.812	0.725 0.886 0.530 0.663	0.727 0.728 0.745 0.736	0.697 0.652 0.873 0.747	-	-
ロジスティック回帰	0.892 0.935 0.848 0.889	**0.929** **0.936** **0.925** **0.930**	**0.913** **0.921** **0.908** **0.915**	0.872 0.902 0.841 0.870	0.737 0.741 0.745 0.743	0.702 0.657 0.871 0.749	**0.933** **0.932** **0.936** **0.934**	**0.934** **0.934** **0.936** **0.935**
決定木	0.864 0.879 0.852 0.865	0.908 0.903 0.918 0.910	0.818 0.872 0.754 0.809	0.727 0.731 0.738 0.734	0.745 0.708 0.852 0.774	0.715 0.664 0.897 0.763	0.908 0.903 0.918 0.910	0.902 0.882 0.932 0.906
ランダムフォレスト	0.841 0.845 0.843 0.844	0.890 0.878 0.911 0.894	0.880 0.912 0.848 0.878	0.836 0.837 0.843 0.840	**0.769** **0.752** **0.817** **0.783**	**0.719** **0.668** **0.894** **0.765**	0.892 0.884 0.908 0.896	0.919 0.918 0.925 0.922
勾配ブースティング回帰木	**0.912** **0.917** **0.911** **0.914**	0.925 0.919 0.936 0.928	0.904 0.914 0.897 0.905	**0.875** **0.880** **0.876** **0.878**	0.755 0.737 0.808 0.771	0.715 0.663 0.899 0.763	0.918 0.912 0.929 0.921	0.933 0.924 0.946 0.935
パーセプトロン	0.879 0.899 0.859 0.879	0.919 0.914 0.929 0.922	0.719 0.961 0.469 0.631	-	-	0.545 0.566 0.476 0.517	0.913 0.927 0.901 0.914	0.909 0.909 0.913 0.911

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision} \tag{4}$$

TP:モデルがいじめ文と判断したいじめ文
TN:モデルが非いじめ文と判断したいじめ文
FP:モデルがいじめ文と判断した非いじめ文
FN:モデルが非いじめ文と判断したいじめ文

7.2 実験結果

結果を表 2 に示す。表中のセルの行が使用した特徴量、列が使用した機械学習手法、数字は上から正解率、適合率、再現率、F 値を表している。正しい分類に大きく貢献している特

徴量、機械学習手法、各特徴量ごとに最もいい精度をしたセルを太字にしている。また、-がついている箇所は明らかにそれぞれの値がおかしなもの (0.000 など) となっており、なぜそうなったかの原因も分からなかった。今後の課題とする。

最も良い精度を示したモデルは特徴量は全て、機械学習手法はロジスティック回帰を用いたもので、全ての評価基準で 90%を超えていた。

特徴量別に見てみると全ての特徴が全体的に良い精度を示している。個々の特徴量に限ってみてみると文字 N グラムが最も良く、それに次いで Word2vec と単語 N グラムが良い精度となっている。Doc2vec は良い精度を示すモデルもあったが、決定木や線形サポートベクトルマシンを機械学習手法に用いたモデルでは低い数値が出た。一方、感情や Twitter はそれほど良い数値が出なかった。したがって、文字 N グラム、単語 N グラム、Word2vec の特徴量が正しい分類に大きく貢献していることが分かった。

機械学習手法別に見ると、勾配ブースティング回帰木とロジスティック回帰が安定して非常に良い精度を示した。それらに次いでランダムフォレスト、決定木となった。パーセプトロンは良い精度を示す特徴量もあったが、特徴量によっては非常に低い数値が出てしまうということがあった。線形サポートベクトルマシンに関しては特に良い精度を示す特徴量もなかった。また、各特徴量ごとに最もいい精度のモデルに使われている機械学習手法はロジスティック回帰、ランダムフォレスト、勾配ブースティング回帰木のいずれかであり、これらの機械学習手法が正しい分類に大きく貢献していることが分かった。

7.3 考　　察

特徴量について考察する。N グラムが良い精度を示したのは本研究でいじめ文のラベルが付けられたツイートは特定のいじめ単語を用いて収集されたものであり、非いじめ文のラベルが付けられたツイートにはそれらの単語はほとんど入っておらず、その明確な違いによるものだと考えられる。しかし、友人同士で冗談で本来いじめに使われるような単語を言い合ったり、いじめ単語を使わずに皮肉のような表現で相手を侮辱するようなツイートを判別するとなった場合、この特徴では正しく分類できないと考えられる。

Word2vec、Doc2vec の場合も良い精度を示したのは N グラムが良い評価を示したのと同じ理由だと考えられる。しかし、今回本研究で用いたいじめ単語だけではネットいじめに使われるような単語を網羅しきれておらず、そのような単語を用いてネットいじめがされていた場合、現在の Word2vec モデル、Doc2vec モデルでは対応しきれないと考えられる。そのため、新たないじめ単語を抽出し、その単語が含まれているツイートも含めて新たな Word2vec モデル、Doc2vec モデルを構築することが必要である。

感情があまり良い精度を示さなかったのは、本研究で用いた感情辞書のリソースは新聞記事や国語辞書であり、Twitter 特有の単語や最近生まれた単語の感情が分からないという問題といじめで使われる単語以外にも悲しい、怒り、緊迫、ネガティブを表す単語は多数存在する問題があった。そのため、Twitter に対応した感情辞書を新たに作成する、いじめ単語に特化した辞書を作成するといった対策が考えられる。

Twitter 特有の特徴が良い精度を示さなかったのは、リツイート、お気に入り、URL、ハッシュタグが多いということには様々な要因が絡んでおり、単純なこれらの数でいじめをどうかを判断するのは好ましくない。ハッシュタグに関しては、いじめ単語のハッシュタグの数や、あるいじめ単語のハッシュタグが含まれているかどうかという形式に変えれば分類性能が向上する可能性がある。

8. まとめと今後の課題

本研究では Twitter 上のテキストを対象とし、特徴量には N グラム、Word2vec、Doc2vec、ツイートの感情値、Twitter 特有の特徴、機械学習手法には線形サポートベクトルマシン、ロジスティック回帰、決定木、ランダムフォレスト、勾配ブースティング回帰木、パーセプトロンを使用し、モデルを生成することで、ネットいじめの自動検出を試みた。本研究では用いた特徴量や機械学習手法によっては非常に高い分類精度が出たが、分類に用いたテキストが限定的なものであり、まだ課題が残っている。

今後はより広い範囲のテキストに対しても正しく分類出来るよう考察でも述べた通り、新たないじめ単語の抽出、その単語を使用して新たないじめツイートの収集、ネット上の単語や最近の単語に対応した感情辞書や、いじめ単語に特化した辞書の作成に取り組んでいきたい。

参　考　文　献

1) 中村 健二，寺口 敏生．CGM 解析に基づくネットいじめ被害の検出手法の検討．大阪経大論集，第 66 巻第 5 号，2016 年.
2) 三島 浩路，本庄 勝．技術的観点からのネットいじめ対策．通信ソサイエティマガジン，No.34 秋号，2015 年.
3) Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, Yi Chang. Abusive Language Detection in Online User Content. WWW 2016, pp.145-153, 2016.
4) Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra. Detection of Cyberbullying Incidents on the Instagram

Social Network. AAAI 2015, 2015.
5) Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, Sabrina Arredondo Mattson. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. ASONAM 2015, pp.617-622, 2015.
6) Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali. Mean Birds: Detecting Aggression and Bullying on Twitter. WebSci 2017, pp.13-22, 2017.
7) Pete Burnap, Matthew L. Williams. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. Policy & Internet, vol.7, no.2, pp.223-242, 2015.
8) Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013, pp.3111-3119, 2013.
9) Quoc Le, Tomas Mikolov. Distributed Representations of Sentences and Documents. ICML 2014, pp.1188-1196, 2014.
10) 熊本忠彦，河合由起子，張建偉．ユーザ印象評価データの分析に基づく印象マイニング手法の設計と評価．情報処理学会論文誌データベース，vol.6，no.2，pp.1-15，2013 年.
11) 高村大也，乾孝司，奥村学．スピンモデルによる単語の感情極性抽出．情報処理学会論文誌ジャーナル，vol.47，no.2，pp.627-637，2006 年.
12) Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. EMNLP 2004, pp.230-237, 2004.
13) 石坂達也，山本和英．Web 上の誹謗中傷を表す文の自動検出．言語処理学会第 17 回年次大会発表論文集，2010 年.
14) 新田大征，桝井文人，プタシンスキ ミハウ，山本和英．有害表現抽出に対する種単語の影響に関する一考察．第 30 回人工知能学会全国大会，2016 年.
15) Suzuha Hatakeyama, Fumito Masui, Michal Ptaszynski, Kazuhide Yamamoto. Statistical Analysis of Automatic Seed Word Acquisition to Improve Harmful Expression Extraction in Cyberbullying Detection. IJETI, vol.6, no.2, pp.165-172, 2016.