様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ２０ 日現在

機関番号：１１２０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2023
課題番号：１９Ｋ１２２３０

研究課題名（和文）ソーシャルメディア分析に基づくネットいじめの高精度・網羅的・早期的自動検出技術

研究課題名（英文）High-accuracy Comprehensive and Early Detection of Cyberbullying Based on Social Media Analysis

研究代表者
張　建偉（Zhang, Jianwei）

岩手大学・理工学部・准教授

研究者番号：２０６３５９２４

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究課題では，ネットいじめの高精度・網羅的・早期的自動検出技術の開発を目的とした．具体的には，以下の３つのテーマに絞って研究開発を行った．A)いじめ表現辞書の構築によるネットいじめの高精度識別技術の開発，B)コンテキスト分析に基づく潜在的ネットいじめの網羅的検出技術の開発，C)時系列的ソーシャル分析に基づくネットいじめの早期発見技術の開発に取り組んでいた．３つのテーマとも，国内研究会だけではなく，国際会議や論文誌に論文が採択された．

研究成果の学術的意義や社会的意義
ネットいじめ自動検出の精度が向上できると，誤検出が減少し，ネットパトロール労力の軽減が予測できる．潜在的なネットいじめを網羅的に検出できると，より多くのユーザをいじめから保護できる．ネットいじめの早期発見が実現できると，いじめの拡散やエスカレートを迅速に阻止できる．高精度・網羅的・早期的なネットいじめの自動発見は，SNSサービスを提供する企業にも，ネットパトロールを実施する学校にも，被害を避けたいユーザにも有用な研究で，安全・安心なインターネット環境の整備に重要な技術になると考える．

研究成果の概要（英文）：The purpose of this research project was to develop a highly accurate, comprehensive, and early automatic detection technology for cyberbullying. Specifically, we focused our research and development on the following three themes. We were working on the A) development of high-accuracy identification technology for cyberbullying by constructing a bullying expression dictionary, B) development of comprehensive detection technology for latent cyberbullying based on context analysis, C) development of early detection technology for cyberbullying based on time-series social analysis. Our work for all the three themes was summarized and evaluated with the papers accepted not only by domestic conferences but also by international conferences and journals.

研究分野：ウェブ情報学およびサービス情報学

キーワード：ネットいじめ　いじめ表現辞書　早期検出　皮肉検出　機械学習　深層学習　事前学習言語モデル　説明可能性

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

インターネットの発達やスマートフォンなどの普及により，Twitter などのソーシャルメディアの利用者は年々増加し続け，若年層も利用する機会が多くなっていた．それと共に，特定の人に対する誹謗中傷が行われるなどのネットいじめが深刻な問題になっていた．実際，ネットいじめが世界中の若者のソーシャルメディアユーザの半数以上に影響を及ぼすと報道されていた．日本でも，平成 30 年 3 月の総務省の調査によると，平成 28 年度のネットいじめの認知件数は 1 万 779 件で，年々増え続けている状況であった．その対策として，SNS サービス提供企業や学校などが，書込みに対して人手によるネットパトロールを実施していたが，データ量が膨大であるため人手による監視には限界があった．

人手による監視の負担を軽減するため，機械学習を用いたネットいじめの自動検出技術の開発は着手されていたが，その性能はまだ十分とはいえなかった．具体的には 3 つの問題点があった．第 1 に，誹謗中傷語を含む明示的いじめ文に対しても，十分な識別精度が達成できておらず，誤検出が存在する点にあった．第 2 に，明示的いじめ語を含まない皮肉などの潜在的いじめに対応できず，検出漏れがある点にあった．第 3 に，従来技術は既出のネットいじめの識別を行うものであり，検出のタイミングが遅れてしまい，早期発見に至っていない点にあった．

２．研究の目的

本研究課題では，以下の 3 つの学術的「問い」(RQ: Research Questions)を設定し，ネットいじめの高精度・網羅的・早期的自動検出技術の開発を本研究の中核とした．

RQ-1: テキスト特徴量を補強する方法でネットいじめの自動識別の精度向上が可能か？
RQ-2: 文脈や背景知識に基づく方法で潜在的ネットいじめを網羅的に自動検出可能か？
RQ-3: ユーザ SNS 活動の時系列変化を分析する方法でネットいじめを早期に発見可能か？

上記を踏まえ，本研究課題では以下の 3 つのテーマに絞って研究開発を行った．

A) いじめ表現辞書の構築によるネットいじめの高精度識別技術の開発
B) コンテキスト分析に基づく潜在的ネットいじめの網羅的検出技術の開発
C) 時系列的ソーシャル分析に基づくネットいじめの早期発見技術の開発

３．研究の方法

テーマ A)について，Twitter 上のテキストを対象とし，いじめ表現辞書を構築した．いじめ表現辞書とは，単語とその単語がどれだけいじめと関連するかの程度を数値で表すものである．辞書に登録する単語は，特定の単語を使って収集したツイートに含まれている単語とし，その単語につける値は，SO-PMI というものを用いて算出した．他にも N グラム，Word2vec，Doc2vec といった特徴量を使用した．また，複数の機械学習手法を用い，特徴量と組み合わせて，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．

テーマ B)について，潜在的ネットいじめの網羅的検出を目的とし，皮肉表現の自動抽出に着目した．BERT を用いた文脈理解，及び文中の絵文字を考慮した皮肉文の検出方法を提案した．日本語 Twitter データセットを用い，テキスト及び絵文字それぞれの特徴ベクトルを抽出した．また，事前学習言語モデル（BERT 及び RoBERTa，DeBERTa）を用いた文脈理解，テキスト類コンテキスト及び非テキスト類コンテキストを考慮した皮肉文の検出手法を提案した．既存研究に多く利用されたハッシュタグで収集したデータセット以外に，皮肉投稿者により収集したデータセットにも手法を適用し，検出パーフォーマンスを比較した．

テーマ C)について，コンテンツとそれに関連するツイートデータを用いた，機械学習による 2STEP の特定情報の早期検出手法を提案した．両方の STEP において，予測確率が閾値以上の情報に対してラベルを確定させて追跡対象から取り除くことで，検出の精度と早期性の両立を図った．また，GCN という深層学習モデルとユーザ特徴・テキスト特徴を用い，ソーシャルメディア特有のネットワーク構造とその変化を考慮した深層学習モデルの学習を行うことで早期検出手法の提案を行った．

また，当初の提案書に記載しなかったモデル説明可能性の課題に向けて，解決手法を検証した．モデルの説明とは，いじめと予測した投稿のテキストにおいて，その根拠となる部分を抜き出したものである．既存モデルの事前学習モデルの変更，根拠学習の拡張，さらに，投稿の対象となるグループを予測するターゲット分類といじめ検出のマルチタスク学習を用いた説明可能なネットいじめ検出モデルを提案した．

４．研究成果

テーマ A)について，ラベル付けされ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，どれだけ正しく検出できるか，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った（表 1～表 3）．構築したモデルを用いていじめ文、非いじめ文の分類の評価を行ったところ，多くの機械学習手法でいじめ表現辞書が正しい検出に貢献することが分かった．また，最も良かったモデルでは 90%を超える評価を得ることができた．

また，異なる時期からツイートを収集し，新たないじめ表現辞書を特徴量として用いると評価はどう変わるか，異なる時期から収集したデータセットを用いると評価はどう変わるかの追加実験を行った．いじめ表現辞書の構築時期やデータセットの収集時期がずれていても，分類評価に大きな影響を及ぼさないことが分かった．辞書の構築時期の変化より，辞書に登録される単語の数の方が，分類評価に影響を及ぼすのではないかと考えられる．

表 1　分類評価（1 種類ずつと全ての特徴量）

	いじめ表現	文字 N グラム	単語 N グラム	Word2vec	Doc2vec	全ての特徴量
線形サポートベクトルマシン	0.822	0.870	0.812	**0.902**	0.881	0.819
ロジスティック回帰	0.838	0.870	0.813	0.882	**0.883**	**0.921**
決定木	**0.805**	0.759	0.669	0.774	0.706	**0.827**
ランダムフォレスト	0.801	0.802	0.695	**0.839**	0.767	**0.840**
勾配ブースティング回帰木	0.828	0.830	0.760	**0.868**	0.810	**0.892**
パーセプトロン	0.845	0.869	0.816	**0.910**	0.882	**0.914**

表 2　分類評価（評価の良い特徴量と「評価の良い特徴量＋いじめ表現」）

	Word2vec	Word2vec+いじめ表現	Doc2vec	Doc2vec+いじめ表現
線形サポートベクトルマシン	**0.902**	0.894	**0.881**	0.872
ロジスティック回帰	0.882	**0.899**	0.883	**0.899**
決定木	0.774	**0.812**	0.706	**0.804**
ランダムフォレスト	0.839	**0.853**	0.767	**0.813**
勾配ブースティング回帰木	0.868	**0.890**	0.810	**0.844**
パーセプトロン	**0.910**	0.907	0.882	**0.892**

表 3　分類評価（全ての特徴量と「全ての特徴量 - いじめ表現」）

	全ての特徴量	全ての特徴量-いじめ表現
線形サポートベクトルマシン	0.819	**0.833**
ロジスティック回帰	**0.921**	0.920
決定木	**0.827**	0.783
ランダムフォレスト	**0.840**	0.824
勾配ブースティング回帰木	**0.892**	0.881
パーセプトロン	**0.914**	**0.914**

テーマ B)について，従来と比較してより文脈の理解に長けた BERT と呼ばれる言語モデルを日本語テキストに対して適用し，皮肉文検出精度の向上を図った．また，文中の重要な要素である絵文字を考慮し，その特徴をテキスト特徴と同時に用いることで検出精度の向上を図った．評価実験には Twitter 上の日本語投稿文を用い，クラウドソーシングによってより信頼性の高い皮肉文データセットを構築した．その結果，どちらの手法においても従来よりも検出精度が向上する結果となった．

また，従来と比較してテキスト類及び非テキスト類コンテキストを利用することで，皮肉内容の検出精度の向上を図った．複数の事前学習言語モデルの比較をし，異なるモデルにおける検出結果で DeBERTa が皮肉内容の検出により優れることがわかった（表 4）．収集方法が異なるデータセットで同様な手法を適用し検出精度を確認した．皮肉内容を投稿したユーザから収集したデータセットは劣った結果を示した．

表 4　事前学習言語モデルの検出性能の比較結果

	適合率	再現率	F 値
BERT	0.82	0.85	0.83
RoBERTa	0.85	0.84	0.84
DeBERTa	0.86	0.85	0.85

テーマ C)について，提案手法を用いることにより，特定情報の拡散初期において，ベースラインと比較して精度と F 値の向上が確認された．SNS ユーザのデータをソーシャルコンテクストとして用いた実験では，提案手法を用いることにより，拡散から 5 分後と 10 分後のタイミングにおいて，ベースラインよりも精度と F1 値が向上することが確認された（図 1，図 2）．リプライツイートを用いた実験では，提案手法による精度と F1 値の向上は見られなかったものの，リプライツイートを用いることが検出率の向上に役立つことが確認された（図 3，図 4）．

GCN を用いた実験の結果，提案手法では最大シーケンス長が 30 の時点で精度 0.6 を超え，F 値も 0.65 を超えることができた（図 5）．実験データのシーケンス長の中央値が 211.5，平均シーケンス長が 356.2 であることを考慮すると，少ないデータ数でもある程度の学習と予測が可能であることが示せたと考えられる．

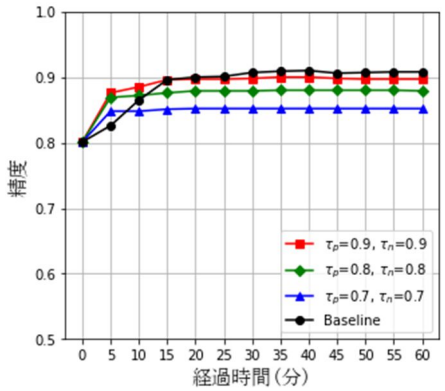

図 1　ユーザデータを用いた実験
（5 分毎の精度）

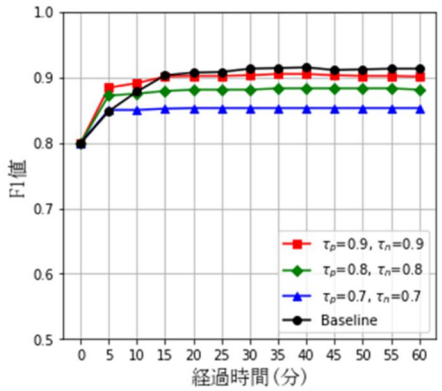

図 2　ユーザデータを用いた実験
（5 分毎の F1 値）

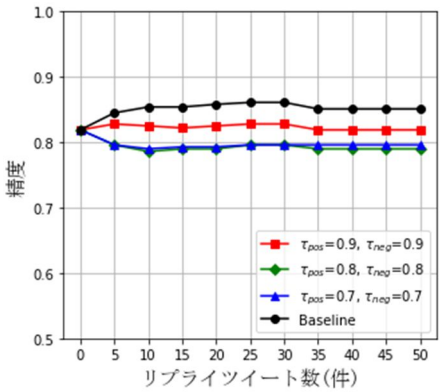

図 3　リプライツイートを用いた実験
（5 件毎の精度）

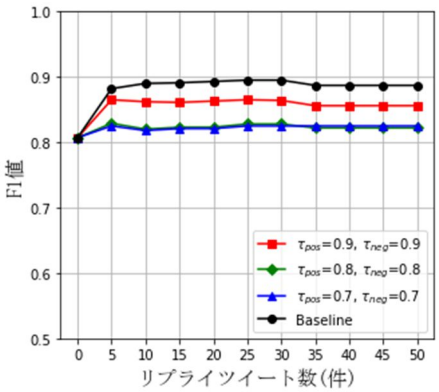

図 4　リプライツイートを用いた実験
（5 件毎の F1 値）

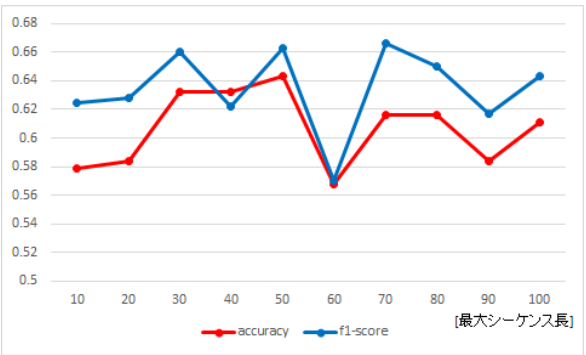

図 5　最大シーケンス長毎の性能比較結果

説明可能性の評価について，各提案手法に基づき３つの実験を行った．モデルの説明の内容を評価する説明可能性の観点で，すべての提案手法において，既存手法より向上する結果となった（表 5〜表 7）．

表 5　事前学習モデルの置き換えによる説明可能性の評価結果

		Plausibility			Faithfulness	
モデル	説明手法	IOU-F1 ↑	Token-F1 ↑	AUPRC ↑	Comp ↑	Suff ↓
Base	Attn	37.0	62.9	55.2	53.7	15.7
1st HateBERT	Attn	41.8	**63.5**	**57.3**	47.8	12.2
2nd HateBERT	Attn	31.6	57.3	51.5	40.6	16.3
All HateBERT	Attn	**42.1**	**63.5**	55.6	51.5	**9.5**
Base	LIME	54.7	69.3	77.3	53.3	-2.7
1st HateBERT	LIME	54.9	70.0	78.5	50.0	-3.8
2nd HateBERT	LIME	**56.7**	**70.9**	78.4	46.2	-1.3
All HateBERT	LIME	55.6	70.3	**78.6**	**53.7**	**-4.9**

表 6　根拠学習の拡張による説明可能性の評価結果

		Plausibility			Faithfulness	
モデル	説明手法	IOU-F1 ↑	Token-F1 ↑	AUPRC ↑	Comp ↑	Suff ↓
Base	Attn	37.0	62.9	55.2	53.7	15.7
根拠 3 値分類モデル	Attn	**39.3**	**64.3**	**57.0**	**57.8**	**13.8**
Base	LIME	54.7	69.3	77.3	53.3	-2.7
根拠 3 値分類モデル	LIME	**56.7**	**71.2**	**79.5**	**56.6**	**-4.0**

表 7　マルチタスク学習による説明可能性の評価結果

		Plausibility			Faithfulness	
モデル	説明手法	IOU-F1 ↑	Token-F1 ↑	AUPRC ↑	Comp ↑	Suff ↓
Base	Attn	37.0	62.9	55.2	53.7	15.7
Multi-task ($\alpha = 0.7$)	Attn	**41.2**	**65.3**	57.6	**61.1**	14.6
Multi-task ($\alpha = 0.5$)	Attn	40.8	64.5	61.7	58.8	**12.1**
Multi-task ($\alpha = 0.3$)	Attn	38.6	64.0	**63.2**	56.9	12.4
Base	LIME	54.7	69.3	77.3	53.3	-2.7
Multi-task ($\alpha = 0.7$)	LIME	57.2	72.4	79.9	**57.5**	-0.5
Multi-task ($\alpha = 0.5$)	LIME	57.5	72.1	79.8	55.0	-0.8
Multi-task ($\alpha = 0.3$)	LIME	**58.3**	**72.8**	**80.1**	54.1	-1.2

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 Zhang Jianwei、Li Lin、Nakajima Shinsuke	４．巻 未定
２．論文標題 Constructing Japanese Bullying Expression Dictionary for Automated Cyberbullying Detection on Twitter	５．発行年 2022年
３．雑誌名 Vietnam Journal of Computer Science	６．最初と最後の頁 1～24
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S2196888822500373	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計19件（うち招待講演　2件／うち国際学会　5件）

１．発表者名 賀毅，張建偉
２．発表標題 事前学習言語モデルを用いた皮肉検出におけるコンテキストの有用性の検証
３．学会等名 2022年度情報処理学会東北支部研究会（岩手大学
４．発表年 2022年

１．発表者名 谷聡馬，佐々木裕多，張建偉
２．発表標題 ニュースコンテンツとソーシャルコンテクストを用いたフェイクニュースの早期自動検出
３．学会等名 第15回データ工学と情報マネジメントに関するフォーラム（DEIM 2023）
４．発表年 2023年

１．発表者名 佐々木裕多，張建偉，白石優旗
２．発表標題 Commonsense-aware AttentionとDiscrepancy Resolution Lossを用いたユーモア検出手法の提案
３．学会等名 第15回データ工学と情報マネジメントに関するフォーラム（DEIM 2023）
４．発表年 2023年

１．発表者名
Yoshio Okimoto, Kosuke Suwa, Jianwei Zhang and Lin Li

２．発表標題
Sarcasm Detection for Japanese Text Using BERT and Emoji

３．学会等名
The 32nd International Conference on Database and Expert Systems Applications (DEXA 2021)（国際学会）

４．発表年
2021年

１．発表者名
須藤広平，張建偉

２．発表標題
ユーザとテキストの時系列を考慮した深層学習によるTwitter上の偽情報の早期発見

３．学会等名
第18回テキストアナリティクス・シンポジウム

４．発表年
2021年

１．発表者名
須藤広平，張建偉

２．発表標題
Graph Convolutional Networkを用いたソーシャルメディア上の偽情報の早期発見

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム（DEIM 2022）

４．発表年
2021年

１．発表者名
佐々木裕多，張建偉，白石優旗

２．発表標題
Commonsense生成モデルを用いたユーモア検出の性能評価

３．学会等名
2021年度情報処理学会東北支部研究会（山形大学）

４．発表年
2021年

１．発表者名
松本蓮史，張建偉

２．発表標題
ネットいじめ検出における事前学習言語モデルの検証

３．学会等名
2021年度情報処理学会東北支部研究会（山形大学）

４．発表年
2021年

１．発表者名
鈴木春，張建偉，沖本吉生

２．発表標題
日本語テキスト分類における事前学習言語モデルの性能評価

３．学会等名
2021年度情報処理学会東北支部研究会（山形大学）

４．発表年
2021年

１．発表者名
Guowei Wang, Lin Li, and Jianwei Zhang

２．発表標題
meanNet: A Multi-layer Label Mean based Semi-supervised Neural Network Approach for Credit Prediction

３．学会等名
The 4th APWeb-WAIM International Joint Conference on Web and Big Data (APWeb-WAIM 2020)（国際学会）

４．発表年
2020年

１．発表者名
Jianwei Zhang, Jinto Yamanaka, and Lin Li

２．発表標題
Early Automatic Detection of False Information in Twitter Event Considering Occurrence Scale and Time Series

３．学会等名
The 22nd International Conference on Information Integration and Web-based Applications & Services (iiWAS 2020)（国際学会）

４．発表年
2020年

１．発表者名
諏訪光輔，張建偉

２．発表標題
BERT及び絵文字を利用した日本語文における皮肉の検出

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム（DEIM 2021）

４．発表年
2021年

１．発表者名
Jianwei Zhang, Taiga Otomo, Lin Li and Shinsuke Nakajima

２．発表標題
Automatic Cyberbullying Detection on Twitter Using Bullying Expression Dictionary

３．学会等名
13th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2021)（国際学会）

４．発表年
2021年

１．発表者名
Jianwei Zhang, Taiga Otomo, Lin Li, and Shinsuke Nakajima

２．発表標題
Cyberbullying Detection on Twitter using Multiple Textual Features

３．学会等名
The 10th IEEE International Conference on Awareness Science and Technology (iCAST 2019)（国際学会）

４．発表年
2019年

１．発表者名
張建偉

２．発表標題
WEBマイニングの事例研究

３．学会等名
京都産業大学（招待講演）

４．発表年
2019年

１．発表者名
山中仁斗，張建偉

２．発表標題
機械学習を用いたSNSにおける偽情報の早期自動検出

３．学会等名
令和元年度情報処理学会東北支部研究会（岩手大学）

４．発表年
2019年

１．発表者名
大友泰賀，張建偉，中島伸介，李琳

２．発表標題
いじめ表現辞書を用いたTwitter上のネットいじめの自動検出

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（DEIM 2020）

４．発表年
2020年

１．発表者名
山中仁斗，張建偉

２．発表標題
発生規模と時系列を考慮したTwitterイベントにおける偽情報の早期自動検出

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（DEIM 2020）

４．発表年
2020年

１．発表者名
Jianwei Zhang

２．発表標題
Towards the Automatic Detection of Cyberbullying

３．学会等名
Huazhong University of Science and Technology（招待講演）

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	中島　伸介 (Nakajima Shinsuke) (90399535)	京都産業大学・情報理工学部・教授 (34304)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
中国	Wuhan University of Technology	Huazhong Univ. of Science and Technology		