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　３ 年　 ６ 月　 ９ 日現在

機関番号：３４３１５
研究種目：若手研究(B)
研究期間：2017～2020
課題番号：１７Ｋ１３２５４

研究課題名（和文）子どものネットいじめを防止するための造語・隠語と文脈に対応した有害表現の自動判定

研究課題名（英文）Automatic detection of toxic expressions considering coined/hidden words and contexts for preventing cyber bullying among children

研究代表者
西原　陽子（Nishihara, Yoko）

立命館大学・情報理工学部・准教授

研究者番号：７０５１２１０１

交付決定額（研究期間全体）：（直接経費）　　2,000,000円

研究成果の概要（和文）：本研究では子どものネットいじめを防止するために、テキストメッセージなどで使われる不適切な表現を自動的に判定する手法を研究した。不適切な表現は直接的な表現と間接的な表現の2種類があるとした。直接的な表現とは、「アホ」「バカ」といった誰が見ても不適切と分かる表現を指す。間接的な表現とは、隠語や造語など文脈によっては不適切となる表現を指す。直接的な表現を判定する手法として辞書ベースの手法を提案した。間接的な表現を判定する手法として、時系列深層学習により文脈を表現し、間接的な不適切表現を判定する手法を提案した。子どもに不適切表現を含むメッセージの投稿の取り下げを促す手法も提案した。

研究成果の学術的意義や社会的意義
不適切な表現の判定は情報フィルタリングの分野で研究が進められてきた。既存研究にも直接的な不適切表現をフィルタリングする手法は提案されているが、隠語や造語を用いることでフィルタリングを回避することは可能であった。本研究では文脈によって不適切な表現となりうる隠語や造語の判定手法を実現した点に意義がある。不適切な表現が含まれるかどうかを判定することはできるが、結局のところ投稿をするしないはユーザの判断に任されており、不適切な表現の判定だけではネットいじめの防止は難しい。本研究では不適切な表現が含まれる時に取り下げを促す手法を提案し、有用性を確認した点に意義がある。

研究成果の概要（英文）：In order to prevent cyber bullying among children, this study investigated a method to automatically detect toxic expressions used in text messages and bulletin boards on the Web. I assumed that there are two types of toxic expressions: direct and indirect expressions. Direct expressions are those that anyone can see as toxic, such as "stupid" or "idiot. Indirect expressions are those that can be toxic depending on the context, such as hidden words and coined words. I proposed a dictionary-based method for detecting direct expressions. As a method to detect indirect expressions, I proposed a method to represent the context and detect indirect harmful expressions by using time-series deep learning. Using the obtained automatic detection method, I also proposed a method to encourage children to take down messages containing toxic expressions.

研究分野：子ども学

キーワード：ネットいじめの防止

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

子ども達がデジタルデバイスを用いてインターネットへ頻繁にアクセスするようになった。それに伴い、現実世界でのいじめがネットの世界にも持ち越され、「ネットいじめ」が起こるようになった。ネットいじめとは、インターネット上におけるいじめおよび嫌がらせを指し、ネットいじめに対する有効な対策の検討が早急に求められている。

ネットいじめに関する国内外の研究では、ネットいじめにおける行動的側面の実態解明（Smith et al., 2008）や、被害者と加害者の心理的側面の実態解明（原, 他, 2015）が進められている。一方で、ネットいじめを防止する方法については、余り研究が進められていない。ネットいじめでは、主に、いじめの被害者に対し不適切な表現が投稿される。不適切な表現とは、誹謗中傷表現、暴力表現、個人情報を含む表現などがある（下表参照）。不適切な表現を含む投稿を阻止することにより、ネットいじめを防止することができる。しかし、ネットいじめに関わる不適切な表現を含む投稿を阻止する方法として、効果的なものは少ない。

表．不適切な表現の分類。

不適切な表現の種類	不適切な表現の形式	
A. 誹謗中傷表現	a. 直接的な表現	
B. 暴力表現	b. 間接的な表現	b-1. 伏字
C. 個人情報を含む表現		b-2. 当て字
D. その他		b-3. 直喩
		b-4. 隠喩
		b-5. 隠語
		b-6. 造語

ネットいじめを防止するには、不適切な表現を含む投稿を阻止することが重要である。不適切な表現を自動判定する研究は、情報フィルタリング技術の分野で研究が進められている。既存の情報フィルタリングの技術では、不適切な表現となる可能性が高い表現を集め、リストを作成し、不適切な表現を判定する。しかし、ネットいじめでは、リストに乗っていないような表現が用いられる場合がある。また、ある表現が不適切な表現となるかどうかは、表現が含まれる文脈により変化する。ネットいじめを防止するためには、不適切な表現を文脈にも対応して判定する手法が求められている。

２．研究の目的

本研究の目的は、子どものネットいじめの防止のために、文脈に対応した不適切な表現の自動判定の方法を明らかにすることである。この目的を達成するためのマイルストーンとして、以下の(1) から(6)の小目的を設定し、達成する。

(1) 子どものネットいじめに関わる不適切な表現の収集、および言語特徴の解明。不適切な表現の種類と形式は上表に示すものを想定する。収集した不適切な表現を種類と形式に応じて分類し、それぞれの言語特徴を明らかにする。
(2) 不適切な表現の言語特徴を用い、不適切な表現を自動判定するための言語モデルの作成
(3) 表現が含まれる文脈を表す言語特徴の解明
(4) 文脈を考慮した上で、不適切な表現を自動判定するための言語モデルへの改良
(5) 改良した言語モデルを評価する実験の実施
(6) 投稿の取り下げを促すメッセージのアンケート調査

(1)から(5)で自動判定の方法を確立し、(6)では自動判定された投稿をどのように取り下げてもらうかの方法を確立する。

３．研究の方法

(1) 子どものネットいじめに関わる不適切な表現の収集、および言語特徴の解明。

ネットいじめで用いられる不適切な表現を収集する。すでに収集済みの不適切な表現を元にし、インターネット上の検索エンジンを通じて収集する。

(2) 不適切な表現の言語特徴を用い、不適切な表現を自動判定するための言語モデルの作成

(1) で得られる言語特徴を用いて、不適切な表現を自動判定するための言語モデルを作成する。不適切な表現を直接的な表現と間接的な表現の 2 つに分類し、直接的な表現に対して辞書ベースでの判定手法を提案する。

(3) 表現が含まれる文脈を表す言語特徴の解明

表現は、それが含まれる文脈により不適切／適切が決定されることがある。例えば、「恐ろしい」という単語はそれ単体では不適切な表現になるとは限らない。しかし、なにかしらの議論を行っているときに誰かが出した意見に対し、別の誰かが「恐ろしい考え方だな」と言ったとする。このときに「恐ろしい」は人によっては不適切な表現になることがありうる。

テキストメッセージの投稿により作られる文脈は、単語の並びにより作られると考えられる。近年単語の使われ方から単語の特徴量を分散表現として表す手法が提案されてきた。多くの場合はある単語の周辺に存在する単語の情報を用い、その単語の特徴を表現する。その際に単語の並びも考慮して表現することが多い。本研究では、文脈は単語の並びにより作られると仮定し、単語の並びを時系列深層学習の LSTM で学習することにより文脈を表現する。

(4) 文脈を考慮した上で、不適切な表現を自動判定するための言語モデルへの改良

LSTM を用いて文脈を表現し、リストにかからない不適切な表現を判定する言語モデルを作成する。掲示板にテキストメッセージが投稿されるとする。この投稿に含まれる単語の流れを文脈とする。テキストメッセージの中には明らかな不適切な表現が含まれるものとそうでないものがあり、明らかな不適切な表現には辞書を用いてラベルが付与されているとする。LSTM には N 個の連続するテキストメッセージを与え、N+1 個目のメッセージが不適切な表現を含むかどうかを予測可能となるように学習をさせる。仮に、LSTM が N+1 個目のメッセージに不適切な表現が含まれると判定した場合、そこに直接的な不適切な表現が含まれていれば、単に予測成功とする。反対に直接的な不適切な表現が含まれていない場合は、通常の機械学習のタスクであれば予測失敗とみなすが、本研究では予測は正しく行われていると考え、N+1 個目のメッセージには間接的な不適切な表現、つまり隠語や造語の不適切な表現が含まれていると判定する。

(5) 改良した言語モデルを評価する実験の実施

(4) で得られた言語モデルを評価する実験を行う。実験結果と成果は 4. 研究成果で示す。

(6) 判定手法を用いた不適切な表現を含む投稿の取り下げを促すメッセージシステムの開発

直接的な不適切な表現、間接的な不適切な表現を判定する手法は得られるが、現在のテキストメッセージアプリや電子掲示板では、投稿をするしないを決定するのはユーザに委ねられており、判定をするだけではネットいじめを防ぐことは難しいと考えられる。そこで、不適切な表現が含まれる場合に投稿の取り下げを促すメッセージを提示するシステムを開発する。直接的な不適切な表現が含まれる場合と、間接的な不適切な表現が含まれる場合に分け、複数のメッセージを提示し、最も取り下げの効果が高いメッセージを中学生、高校生を対象とするアンケート調査により明らかにする。

４．研究成果

(5)に関する実験を行った。実験では不適切な表現を含む投稿がなされる電子掲示板のデータを用いた。用いたデータはスレッドが約 9,000 件で、その中に含まれる投稿は、明らかな不適切な表現が含まれるものが約 108 万件、そうでないものが約 565 万件であった。9,000 件のスレッドのデータを LSTM で学習させ、投稿に不適切な表現が含まれるかどうかを予測するモデルを作成した。予測の際に除外した 20 件のスレッドを評価データとして用いた。評価データには、直接的な不適切な表現が含まれる投稿が約 4,000 件、そうでないものが約 1,500 件あった。実験の結果は下図の通りになった。

実験の結果、正しく予測できた投稿の割合は全体の 8 割になった（1,556+14,613=16,169 件）。このことから概ね LSTM での学習はできていると考えられる。このため投稿内に直接的な不適切な表現は含まれないが、不適切な表現があると予測された 1,207 件の中には間接的な不適切表現を含むものが含まれると考えられる。1,207 件の投稿を手作業で精査したところ、677 件に間接的な不適切表現が含まれた。割合は約 56%であった。全投稿は 19,789 件であり、そこから間接的な不適切表現を含む可能性がある投稿を 1,207 件までに絞り込むことができた。この結果から、間接的な不適切表現の判定が一定程度できることが確認された。

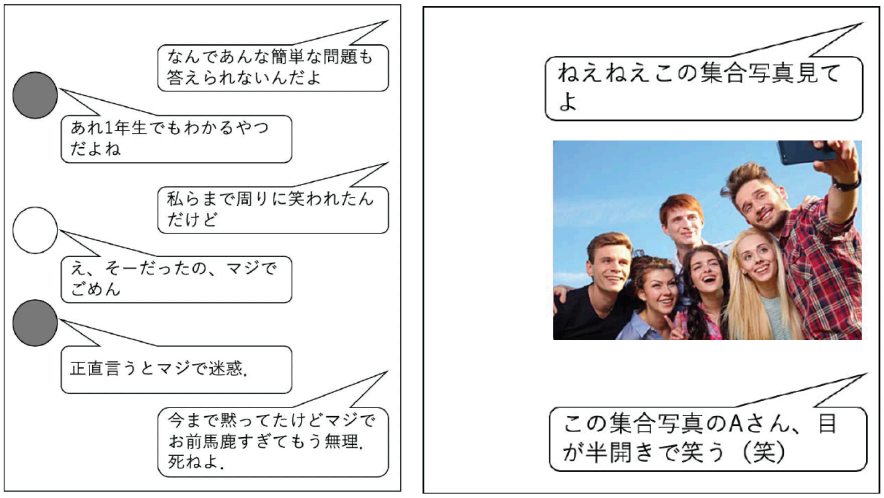
表．実験結果

	投稿に該当表現がある	投稿に該当表現がない
LSTM の出力では NG	1,556	1,207
LSTM の出力では OK	2,413	14,613
合計	3,969	15,820

(6)に関するアンケート調査を行った。アンケート調査では、直接的な不適切表現として露骨な悪口、間接的な不適切表現として露骨ではない悪口を取り上げた。そして、悪口が提案手法により判定された場合にそれを指摘するかしないかで 4 つの場合分けを行った。場合分けを表に示す。そして、露骨な悪口を含む投稿の例と露骨ではない悪口を含む投稿の例を作成し、アンケートに用いた。投稿の例を図に示す。

セクション	悪口の種類	悪口部分の指摘
1	露骨な悪口	有り
2	露骨な悪口	無し
3	露骨ではない悪口	有り
4	露骨ではない悪口	無し

表．悪口の種類、悪口部分の指摘での 4 種類の場合分け



図．アンケート調査に用いた投稿の例

	メッセージ文
ST1	「もし投稿を取り下げれば，周囲や相手を傷つけません」
ST2	「もしその投稿をあなたが受けたら，あなたは不快に思いませんか？」
ST3	「あなたの投稿は，トークルームの人や，相手を不快にしてしまいませんか？」
ST4	「あなたの投稿の後，その先のトークはどう進んでいくと思いますか？」
禁止	「悪口を投稿することはやめてください」

表．取り下げを促すメッセージ文

取り下げを促すメッセージとして 4 種類のメッセージを作成した。比較を行うために禁止のメッセージをベースラインとして加え、合計 5 種類のメッセージに対してアンケート調査を行った。

アンケート調査の結果、2 つのことが明らかになった。1 つは、露骨な悪口に対しては悪口部分を指摘することで、取り下げの効果が上がるであった。もう 1 つは悪口の相手の立場に立ち考えることを促すメッセージを提示することで、取り下げの効果が上がるであった。

本研究により、ネットいじめで用いられる不適切な表現の自動判定が実現できた。自動判定は直接的な不適切表現、間接的な不適切表現のいずれにも対応できる。さらに、自動判定された後で、投稿の取り下げを促す効果的なメッセージの解明も行えた。これらの成果を合わせることにより、ネットいじめの防止を実現できる。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 藤堂 悠杜，山西 良典，西原 陽子
２．発表標題 SNS上の悪口を含む投稿に対する取り下げを促すフィードバック文の自動生成方法の検討
３．学会等名 第23回インタラクティブ情報アクセスと可視化マイニング研究会
４．発表年 2019年

１．発表者名 Ryuichi Omi, Yoko Nishihara, and Ryosuke Yamanishi
２．発表標題 Extraction of Paraphrases using Time Series Deep Learning Method
３．学会等名 International MultiConference of Engineers and Computer Scientists 2019（国際学会）
４．発表年 2019年

１．発表者名 近江 龍一，西原 陽子，山西 良典
２．発表標題 時系列深層学習を用いた言い換え表現の獲得
３．学会等名 ARG 第13回Webインテリジェンスとインタラクション研究会
４．発表年 2018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
研究協力者	山西　良典 (Yamanishi Ryosuke) (50700522)	関西大学・総合情報学部・准教授 (34416)	

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