様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １５ 日現在

機関番号：１０１０６
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４６００００１

研究課題名（和文）世評・感情・倫理を考慮して柔軟に有害表現を検出する技術の開発とその応用

研究課題名（英文）Development of a method to detect cyber-bullying entries based on reputation, emotion and ethics

研究代表者
桝井　文人（Masui, Fumito）

北見工業大学・工学部・准教授

研究者番号：８０３２４５４９

交付決定額（研究期間全体）：（直接経費）　　4,200,000円

研究成果の概要（和文）：本研究では，世評，感情，倫理という要素に着目し、学校非公式サイトに書き込まれた有害表現を効率よく検出する技術を提案した．これらの要素に基づいて有害極性値を計算・判定するモジュールを設計・開発した．インターネットから収集した有害書き込みを用いて評価データを作成し、開発した技術の性能を検証した．
検証の結果，世評と倫理情報は有害表現検出に有効であること，感情情報は有害表現とは相関がないこと，総合的にはネットパトロール活動の過負荷を軽減できること，をそれぞれ明らかにした．

研究成果の概要（英文）：This research focuses on factors such as reputation, emotion, and ethics to realize a new method to detect cyber-bullying entries in unofficial school internet board.
On the basis of each factor, we designed modules to calculate harmful polarity. We conducted experiments for evaluation of our implemented modules with a test dataset based on harmful entries gathered from the internet. The results said that reputation and ethics are effective to calculate harmful polarity value although emotion has no correlation with harmful polarity.
On the whole, our method performs enough to reduce the load of net-patrol.

研究分野：知識工学

キーワード：有害表現　世評　感情情報　倫理判断　ねっと上のいじめ　学校非公式サイト　ネットパトロール

１．研究開始当初の背景

「ネット上のいじめ」が新しい「いじめ」の形態として問題視され，青少年に対する影響が懸念されている．「ネット上のいじめ」とは，携帯電話やパソコンを通じてインターネット上のいわゆる学校非公式サイトの掲示板などに特定個人に対する悪口や誹謗・中傷を書込むなどして，有害情報によるいじめを行うものである[1].

このような「ネット上のいじめ」では，短期間で深刻化するケースも多い上に，当事者は容易に被害者にも加害者にもなり得る．そのため，深刻化を見過ごすと事件にまで発展する危険性があり，早期発見早期対応に向けた取組みが急務である.「ネット上のいじめ」問題は，海外でも“Cyber-bullying”として注目されつつあり，研究開始以降，世界規模で研究が進展すると予想された（現在では世界各国で多くの研究が進められている）．

これらの有害情報は，青少年保護という観点に基づいてネットパトロール活動にて対応されている．ネットパトロールとは掲示板を巡回，監視する行動であり，侮辱や誹謗中傷など有害な書込みを発見した場合に該当掲示板の管理人あるいはプロバイダに削除を依頼するプロセスを踏む.

しかしながら，これらの活動は多くの場合，教育委員会や学校教職員，PTA，外部委託の教育アドバイザーなどがボランティアベースで行っている．ネットパトロール作業は，書込み内容を記録した後，改めて詳細な内容チェックを行うなど，大部分を手作業で行っている現状があり，その負荷が担当者の健康や生活に与える影響も看過できない．一部では，ネットパトロールサービスをビジネスとする企業へ委託するケースも見られるが，作業を人手に頼っている現状は同様であり，作業者への負担の問題は変わらない．よって，ネットパトロール活動を支援する仕組みの実現は急務である.

２．研究の目的

「ネット上のいじめ」問題は，ネット上のデータを扱うという性格上，当然ながら情報工学的アプローチによる有害表現検出技術が鍵となる．既にいくつかの基礎研究もみられるが，文字列や単語といった表層的な要素を扱っている状態であり，世評や感情，倫理などの要素を考慮した上で有害表現を扱った例は見られない．したがって，本研究で試みようとする「ネット上のいじめ」への取り組みは先駆的である.

本研究では，ネットパトロール活動による監視担当者にかかる負荷を軽減することを目的として,学校非公式サイトに書き込まれた有害表現を効率よく検出する技術の開発に取り組んだ.

３．研究の方法

本研究では，図１に示すように，４つのモジュール（統合判定，世評に基づく有害極性判定，感情に基づく有害極性判定，倫理に基づく有害極性判定）を開発し，それら統合して一つのシステムとすること，調査および評価のために有害表現検出テストコレクションを構築することを目指す.

具体的には，以下に述べるような方法によって研究を進めた.

まず，インターネット上の学校非公式サイトや掲示板，ブログを対象として，それらに書き込まれる書き込みデータを収集する．さらに，収集したデータ群に含まれる有害表現の割合や傾向，特徴などを調査・分析し，有害表現と世評，感情，倫理（常識）判断との関連について知見を得た.

次に，上記で得られた知見に基づき，有害表現検出メソッドを設計する．本研究では，世評に基づく判断，感情に基づく判断，倫理に基づく判断を独立として捉え，それぞれ異なる検出メソッドの設計を試みた．このとき，同時にメソッドを構成するため，あるいはメソッドの訓練データとして利用するために必要な基本データも，調査対象データから得ておく．また，可能であればメソッドを評価する際に用いる評価データも得ておく.

上述した設計に基づき，各メソッドを構築する．さらに，構築したメソッドを評価データに適用し，メソッドの有効性を検証する．評価結果を吟味しその有効性と課題について考察する.

４．研究成果

(1) インターネット上の有害表現の調査

インターネット上に存在する学校非公式サイトや掲示板上に記述される書き込みを収集し,有害表現について調査を実施した．しかし．当初予定した通りの十分な規模のデータを収集できなかった．その理由は以下の２点で説明できる.

第一に，有害書き込みの数が一般的な書き込みに対して圧倒的に少ないという点である．実際に，学校非公式サイトと呼ばれる書き込みの中に有害書込みが含まれる割合を調べたところ，平均で約 12%という結果が得られたこともこのことを示唆している.

第二に，書き込み先が学校非公式サイトやインターネット掲示板といったオープな空間から SNS やプロフサイトといったクローズドな空間へ移行したことである.

上記で収集したデータを対象として，有害性と関連の強い語句や有害性を判定するための要素などの知見を見出した.その結果，時間の推移とともに変化する有害表現と，時間の推移に関わらず頻出する有害表現が存在することがわかった.この知見に基づけば，時間の推移に関わらず頻出する有害表現を

基本知識とした有害性判定を考えることで，時間の推移による表現の変化にも強い処理の実現可能性が考えられる．

(2) 世評に基づく有害極性判定

インターネット上の学校非公式サイトやブログに書き込まれたテキスト情報の統計的な性質を利用することで，インターネット上の世評を表す語句や表現を検出する有害極性判定メソットを設計・構築した．

構築しようとするメソッドの有害極性判定エンジンは，評判情報抽出において用いられている PMI-IR（SO-IR）法を応用する．

エンジンは，対象とする有害表現を，誹謗中傷語・暴力誘発語・卑猥語という３つのカテゴリ（文部科学省の定義による[1]）に区別し，それぞれに関連する有害極性と有害極性値を計算する．

(1)で構築した評価データを用いて，本メソッドの有効性を検証した．その結果，F 値 0.67 の性能が得られ，従来手法[]の性能（F 値 0.46）を上回った．さらに，初期知識の規模が性能に影響を与えるかどうかを知るために，複数の初期知識を用いた実験を行った．その結果，初期知識の規模が大きくなれば検出性能に寄与する傾向はあるものの，初期知識の選択の方が性能に大きく影響することがわかった．

評価データは現実に即した状況を再現するために，有害書込み混合率 12%のデータを作成して用いた．松葉ら [2] が用いたデータ 2,998 件（混合率 50%）から有害書込み 60 件，非有害書込み 440 件を無作為に取り出し，混合率 12%の評価データ 500 件を作成した．この試行を 5 回繰り返し，同様の評価データを 5 組（計 2,500 件）作成した．なお，評価データ中の書込みには方言やスラング，誤送信による不完全な文章が含まれ，これらは形態素解析誤りを引き起こし，適切な評価が行えない．この問題を回避するために，評価データ 2,998 件中 1,430 件に対し，人手による正規化を行った．

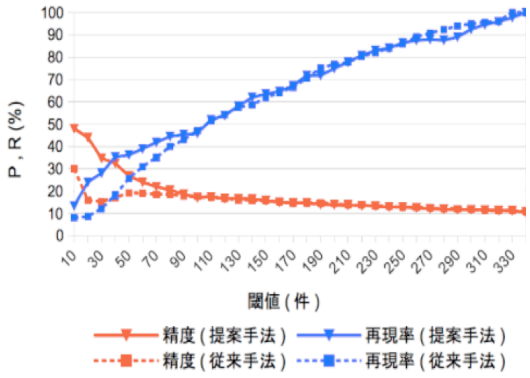

図 1 世評に基づく有害表現検出の性能

実験の結果を図 1 のグラフに示す．x 軸は判定しきい値，y 軸は検出性能を示す．5 組の評価データに対する検出性能でみると，精度は 0.09～0.48，再現率は 0.13～1.00 であった．この結果は先行研究[2]（精度 0.09～0.16，再現率 0.08～1.00）と比較しても優位な結果であるといえる．

(3) 感情に基づく有害表現検出判定

まず，松葉らによる混合率 50%の評価データ (2,998 件)を対象として，有害表現と感情極性の関連を調査した．その結果，負の感情極性を持つ有害書き込みは全体のわずか 2%であることがわかった．

次に，感情文は全て有害表現である（emotive=harmful）という仮定に立ち，感情極性によって有害表現判定をするメソッドを構築してその性能を評価した．評価結果を図 2 のグラフに示す．評価結果からは，精度，再現率とも 0.5（F 値も 0.5），感情極性と性能の相関は 0 が得られ，有害表現と非有害表現の中にはほぼ同数の感情極性語が含まれていることが明らかとなった．このことは，感情情報と有害極性との間にはほとんど相関が無いことを意味しており，感情情報処理を用いた有害表現検出は不可能であると結論できる．

以上の結果を受け，感情に基づく有害極性判定メソッドの構築は見送ることにした．

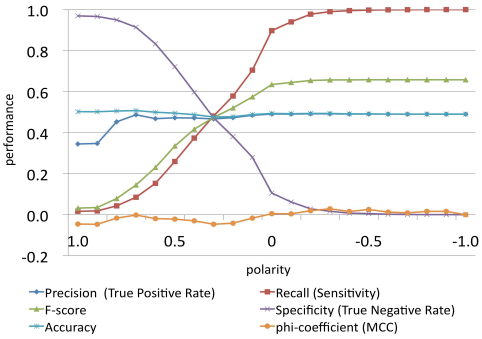

図 2 感情に基づく有害表現検出の性能

(4) 倫理・常識に基づく有害極性判定

GENTA プロジェクトの常識解析エンジン[3]を応用して有害表現と倫理判断による有害表現検出メソッドを構築した．基本知識として，先行研究で提案された 2 種類の感情語辞書のエントリ，Kohlberg[4]による社会的通念（叱れる⇔褒められる，罰される⇔受賞する）関連表現，JApprisal 評価表現 [5]を用いて複数のモジュールを構築した．

これらのモジュールの有効性を検証するために，松葉ら[2]による混合率 50%の評価データ (2,998 件)を用いて評価実験を行った．

実験の結果，最も良い結果（全ての組み合わせ)でも F 値 0.24 という結果であった．評

価データ中に現れる表現の感情極性を確認したところ，感情極性が負でかつ有害な書き込みは全体のわずか 2%しか存在しないことがわかった．

そこで，固有名詞や特定の頻出有害語や伏せ字に用いられる記号，特定の有害語に重みを持たせるヒューリスティクスを適用して処理の精緻化を試みた．これらの評価結果を表 1 に示す．結果として，ヒューリスティクスは大きな効果をもたらした．全てのヒューリスティクスを組み合わせた場合に最も高い性能が得られ，ヒューリスティクスを適用しない場合に比べ F 値では約 8 ポイント（精度では約 5 ポイント，再現率では約 10 ポイント）性能が向上している．

このことから，倫理（常識）判断に基づく有害表現検出は有効であるが，高い性能を確保するためには詳細なヒューリスティクスや詳細な判定知識の設定が必要であることがわかる．

表 1 倫理・常識に基づく有害表現検出の性能

Heuristics	*Precision*	*Recall*	*F-Score*
No heuristics	57.56	45.03	50.53
Weighting「○」	59.91	45.86	51.95
Weighting 「死ね」	57.82	45.36	50.84
Weighting Proper Nouns	61.49	53.90	57.44
All heuristics	**63.44**	**55.40**	**59.15**

これらのヒューリスティクスを全て適用した場合，倫理判断に基づく有害極性判定の性能は F 値で約 0.59(精度 0.63,再現率 0.55)という最高値を示した．この結果から，倫理判断処理は有害表現抽出に対して一定の効果があることがわかった．ただし，結果のエラー分析からは，より高い判定性能を得るためには，文脈処理や固有名詞の曖昧性解消など当初予想されたものより難易度の高い処理が必要であり，精緻化のためには研究計画年度を越えた継続的な取り組みが必要であることもわかった．

(5)　以上より，本研究における目標のうち，世評に基づく有害極性判定モジュール開発はほぼ予定どおり成功した．倫理・常識に基づく有害極性判定モジュール開発については，目標とする性能には至らなかったが有効な性能を得ることには成功した．

感情に基づく有害極性判定モジュール開発については，当初有効と考えていた感情情報と有害極性には関連性が無いという新たな知見を得ることとなった．

総合判定については，上記二つの単純な結合では性能向上が得られないことがわかったため，引き続き相補的な結合による判定手法を検討する必要がある．

本研究で設計・開発した各有害極性判定モジュールを評価するために，インターネットから収集した評価データには，書き込み内容，有害・非有害の有無，キーワードなどが付与されており，テストコレクションデータとして利用することができる仕様となっている．現在，処理結果を自動評価するプログラム（スコアラー）を設計・開発中であり，完成後はテストコレクションの一部とする予定である．本データの内容は，個人情報保護の観点から当面は非公開とするが，データセットの仕様やスコアラーについては順次公開する予定である．

＜引用文献＞

[1] 文部科学省:「ネット上のいじめ」に関する対応マニュアル事例集(学校・教員向け)，文部科学省 (2008)

[2] 松葉達明，桝井文人，河合敦夫，井須尚紀: 学校非公式サイトにおける有害情報検出を目的とした極性判定モデルに関する研究，言語処理学会第 17 回年次大会発表論文集，P2-26，2011.

[3] Rafal Rzepka and Kenji Araki: Polarization of consequence expressions for an automatic ethical judgment based on moral stages theory, IPSJ SIG Notes 2012-NL-207 (14), pp. 1-4, 2012.

[4] Lawrence Kohlberg: The Philosophy of Moral Development, Harper and Row, 1981.

[5] 国立国語研究所コーパス開発センター: 日本語アプレイザル評価表現辞書(JAppraisal 辞書) ～態度評価編～ Version1.2 仕様説明書,及び，評価表現分類表，2012.

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）

① Michal Ptaszynski, Pawel Lempa, Fumito Masui: A Modular System for Support of Experiments in Text Classification, Technical Transactions: Mechanics, Wydawnictwo Politechniki Krakowskiej, 2015 (to appear).
http://www.ejournals.eu/Czasopismo-Techniczne/

② Rafal Rzepka and Kenji Araki: Society as a Teacher - Automatic Recognition of Instincts Underneath Human Actins by Using Blog Corpus, Computer Science, Social Informatics, Vol. 8238, 2013, pp.370-376, 2013.
DOI:10.1007/978-3-319-03260-3

〔学会発表〕（計 12 件）

① Michal Ptaszynski, Fumito Masui, Yasutomo Kimura, Rafal Rzepka, Kenji Araki: Brute Force Works Best Against Bullying, the International Workshop

on Intelligent Personalization in the 25th International Joint Conference Artificial Intelligence (IJCAI-15), 2015.08, Buenos Aires, Algentina (accepted).

② Michal Ptaszynski, Fumito Masui, Rafal Rzepka, Kenji Araki: Detecting emotive sentences with pattern-based language modelling, the 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2014), 2014.09, Gdynia, Poland.

③ 福島裕斗, 桝井文人, Ptaszynski Michal Ptaszynski, Fumito Masui: SPASS: A Scientific Paper Writing Support System, the Third International Conference on Informatics Engineering and Information Science (ICIEIS2014), 2014.09, Lods, Poland.

④ Michal, 中島陽子, 渡辺桂祐, 河石良太郎,新田大征, 佐藤亮弥: 表層的言語情報から読み取れる直接性に着目したツイートログの分類, 第 28 回人工知能学会全国大会,2014.05, 松山.

⑤ Michal Ptaszynski, Fumito Masui, Rafal Rzepka and Kenji Araki: Emotive or Non-emotive: That is The Question, ACL / the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2014), 2014.06, Baltimore, USA.

⑥ Yuuto Fukushima, Fumito Masui, Michal Ptaszynski, Yoko Nakajima, Keisuke Watanabe, Ryotaro Kawaishi, Taisei Nitta and Ryoya Sato: Macroanalysis of Microbologs: An Empirical Study of Communication Strategies on Twitter, AAAI2014 Spring Symposium on Big Data Becomes Personal: Knowledge into Meaning, 2014.03, Palo Alto, USA.

⑦ Rafal Rzepka and Kenji Araki: Possible Usage of Sentiment Analysis for Calculating Vectors of Felific Calculus, IEEE 13th International Conference on Data Mining Workshop "SENTIRE", 2013.12, Dallas, USA.

⑧ 新田大征, 桝井文人, Ptaszynski Michal, 木村泰知, Rzepka Rafal, 荒木健治: カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出, 第 27 回人工知能学会全国大会,2013.06, 富山.

⑨ Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, Kenji Araki: Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization, the International Joint Conference on Natural Language Processing (IJCNLP2013) , pp.579-586, 2013.10, Nagoya.

⑩ Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, Kenji Araki: Cyberbullying Detection Based on Category Relevance Maximization, Language Processing and Intelligent Information Systems 2013 (LP&IIS'13) , 2013.06, Warsaw, Poland.

〔図書〕(計０件)

〔産業財産権〕
○出願状況 (計１件)
名称：インターネット上の有害書き込み検出方法と装置
発明者：桝井文人, ミハウ・プタシンスキ, 新田大征
権利者：北見工業大学
種類：特許
番号：特願 2013-245813
出願年月日：2013 年 11 月 13 日
国内外の別： 国内

６．研究組織
(1)研究代表者
桝井　文人 (MASUI, Fumito)
北見工業大学・工学部・准教授
研究者番号：80324549

(2)研究分担者
Rzepla Rafal (RZEPKA, Rafal)
北海道大学大学院・情報科学研究科・助教
研究者番号： 80396316

(3)研究分担者
木村　泰知 (KIMURA, Yasutomo)
小樽商科大学・商学部・准教授
研究者番号： 50400073

(4)研究分担者
プタシンスキ・ミハウ (PTASZYNSKI, Michal)
北見工業大学・工学部・助教
研究者番号： 60711504

(5)研究協力者
荒木　健治 (ARAKI, Kenji)
北海道大学大学院・情報科学研究科・教授
研究者番号：50202742