コメントの親子関係を利用した ネットいじめコメントの検出

李 子怡[1]　川本 淳平　フォン　ヤオカイ[2]　櫻井 幸一[2]

概要：ネットいじめは，被害者に悪影響を与えることから，青少年の間で深刻な問題になっている． SNS における投稿からのネットいじめを検出する手法はテキストベースとソーシャル情報ベースのものがある．前者では統一の索引語の集合を用いるため，各被害者のいじめ投稿の傾向に対応できないことが多い．後者ではユーザーペア間でのイントラクションが少ない SNS で有効ではない．本研究では，コメントとその返事の親子関係をスライド形式で捉え，ユーザー間のインタラクションを用いてネットいじめコメントを検知する．ネットいじめでは第三者は被害者を助ける行為を取る傾向であることから，各 SNS 利用者のニーズに応じたネットいじめ検知手法を提案する．

キーワード：ソーシャルネットワークサービス, ネットいじめ, テキストマイニング

Cyberbullying Detection Using Parent-Child Relationship between Comments

Ziyi Li[1]　Junpei Kawamoto　Yaokai Feng[2]　Kouichi Sakurai[2]

Abstract: Cyberbullying poses significant threat to mental and physi- cal health on its victims, leading to a worrisome social issue. Previous research of automated cyberbullying detection on SNS is mainly textual-based, in which cyberbullying con- tent is identified through a set of textual features. Those methods are straightforward but are more likely to suffer from the subjectiveness of the researcher. Moreover, each content is evaluated with same standard, not catering for individual's variance. Therefore, in this article we propose a automated cyberbullying detection method that utilises the parent-child relationship between comments to capture the reaction from a third party to detect cyberbullying com- ments. We were able to improve the effectiveness of cyber- bullying detection using only publicly available data.

Keywords: SNS, cyberbullying, data mining

1. はじめに

近年，ソーシャルネットワークサービス (SNS) のユーザーは急速に増えている．ユーザーは親密またはプライベートな情報を気楽に共有している．このようなコミュニケーションは，悪意を持つ人に SNS を乱用する隙を与えた．投稿されたメッセージには，罵倒や失礼な内容を含む場合があり，更にはネットいじめにエスカレートするケースもある．ユーザーが安心して SNS を利用するためには，ネットいじめに該当するメッセージを自動検知する方法が必要である．

ネットいじめとは、インターネット上で故意に他人を侮辱、脅迫、困惑させる，苦しめる攻撃である．通常，成人は SNS に存在する危険を意識し，より安全なコミュニケーションをとることができる．それに対して，未成年者は脅威への認識力が低いため，身体的と精神的に大きな悪影響を受け，生死に関わる事件に至る可能性が高い．Ditch

[1] 九州大学大学院システム情報科学府
Kyushu University
[2] 九州大学大学院システム情報科学研究院
Kyushu University

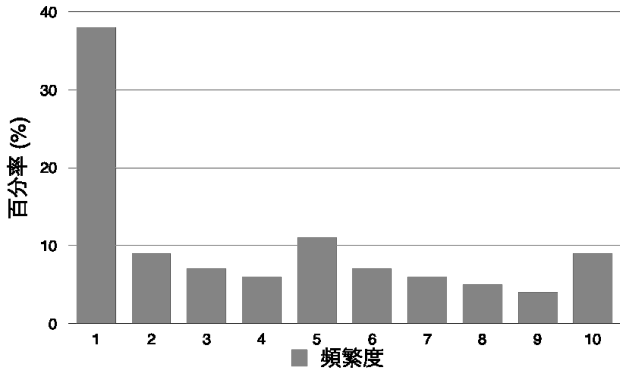


図 1: ネットいじめの頻度. (Ditch the Label 調べ)

the Label 組織が 3023 名のイギリス学生を対象として行った 2015 年度いじめ状況サーベイ (ABS; Annual Bullying Survey)[*1] では，いじめにあった結果，30 %が自殺する考えを持ち，29 %が自傷，27 %が授業に欠席，14 %が摂食障害を患い，12 %が家出したなどの直接な被害が報告されている．インターネット以前の伝統ないじめの被害者は，ネットいじめに比べてより厳しいいじめ行為を受け普段の生活に影響を感じることが多い．それに対して，ネットいじめの被害者は，物理的ないじめ行為を受けることはないが，社会生活が困難になり不安や意気消沈を感じる傾向にあることが報告されている [13].

ネットいじめは伝統的ないじめよりも固執である．伝統的ないじめは学校などでの接触を必要とするが，ネットいじめはそれらの環境を必要とせず被害者に一刻の猶予も与えない．特に SNS では、被害者はオンラインコミュニティ全体の前で傷つけられる [2]．また，「Once on the Internet, always on the Internet」と言われるように，一旦インターネットにアップロードされたものは，いかなる形式であっても永久に存在する．そのため，伝統ないじめと違う形式で，オンライン行為も反復性を持っていると言え，被害者に大きな悪影響を与える [10]．ABS によれば，調査に参加した学生のうち 43% もの学生がいじめを経験したと述べている．また，図 1 に報告にあったネットいじめの被害頻度を示す．図中 1 は経験したことがない，5 は度々，10 は高頻度でネットいじめの被害を受けていることを表している．62% ものいじめ被害者がネットいじめにさらされており，そのうちの 9% は頻繁にネットいじめされていることがわかる．このように，ネットいじめは無視できない影響と範囲を持っており，深刻な社会問題となっている．

Dadvar らによると，SNS におけるネットいじめ対策は主に人手に依存している．各サイトに担当者がおり，サイトに投稿された情報を監視しネットいじめに関係する投稿を手動で削除している [14]．しかし，大量の情報を処理する必要があり，手が及ばないことがある．このような状況を改善するために，計算機を用いた自動検知が期待されている．

既存のネットいじめに関係する投稿の自動検知手法は，主に辞書ベースの手法 [2][3] と，ソーシャル情報を用いた手法 [4][12] に分けられる．辞書ベースの手法は，簡単に実現できるが，いじめに関係する単語の選定や重み付け方法に辞書作成者である研究者の主観的な考えに影響されてしまう可能性がある．加えて，悪意ある単語を使わないネットいじめメッセージは検知することができない．特に英語は一つの単語でもたくさんの意味持っており，インターネット上でさらに様々な意味が派生されているため，ブラックリスト辞書を用いた検出は難しい．ソーシャル情報，すなわちユーザ間の関係を表したソーシャルグラフを用いた手法やユーザー背景情報を用いた手法は，実際の投稿以外にもユーザ間のインタラクション履歴 [4] やユーザーの背景情報 [12] など各種情報を必要とする．この様な情報はユーザーのプライバシーに関わるので，取得が比較的に難しい．また，ユーザーペア間のイントラクションが少ない SNS での検知も比較的に困難である．

本研究では，SNS のうち動画共有サイトを対象に，いじめ投稿を検出する方法を提案する．本研究で対象とする動画共有サイトでは，動画投稿者一人に対してコメントが集まる形式を考え，コメント投稿者は動画投稿者について背景知識を持っていると仮定する．Bastiaensens らの研究によると，ネットいじめにおいては，第三者はいじめ行為を目撃した際に被害者を助ける行動を取ることが多い [7]．実際，我々が動画共有サイトの一つである Youtube から取得したデータにおいても，第三者はいじめコメントに対して非難を向けるなど被害者救済行動が見て取れた．特に SNS では，こうした第三者はいじめ被害者についての背景知識を持っていることが多く，どのような言葉が被害者にとっていじめに成り得るのかを考え，いじめ行為者へ非難を向ける，被害者を慰めるといった救済行動を行う．本研究では，こうしたネットいじめにおける人々の行動特徴を元に，単純な辞書式手法では発見できないが被害者にとっては傷つくような投稿を発見する．具体的には，コメントを単独ではなく，スレッドとして扱い親子関係を利用する．そして，いじめコメント投稿者と善意の第三者達とのインタラクションを用いてネットいじめコメントを発見する．また， SNS は投稿者を焦点とした短文や画像の共有する目的の他にも，物事を伝えるためにも使われている．そのため，コメントの対象はビデオの内容宛てと投稿者宛ての二種類が考えられる．そこで我々の提案手法では，コメントの対象を調べ，投稿者に対するいじめコメントを発見する．以降，本論文では英語のデータセットを用いて議論を行う．しかし，提案手法は言語に限らず使える方法である．

本研究で提案された手法を用いて，少量のコメント情報と研究者の主観知識でネットいじめコメントを 80.6% の精

[*1] http://www.ditchthelabel.org/the-annual-bullying-survey-2015-is-here/

表 1: 平均 DASS スコア

いじめの形式	憂鬱	不安	ストレス	総計
ネット	11.16	8.23	11.36	30.84
伝統	7.72	6.00	9.29	23.02
両方	14.62	11.73	15.35	41.70
経験なし	5.92	4.75	6.9	17.57

度で検出することが出来た．ネットいじめコメントとして抽出されたものには，通常のいじめ話題ではなく，あるビデオ投稿者だけにとっていじめである話題であるコメントを含んでおり，既存のテキストベース手法では検知できないものを抽出できることが判明した．

本論文の構成は以下のとおりである．第 2 節では，SNS におけるネットいじめ分析に関する先行研究を紹介する．第　3 節では，動画共有サイトにおけるコメントの扱い方を述べる．第 4 節では，本論文における提案手法を説明する．第 5 節では手法の実装を述べ，6 節では実験の結果と考察について述べる最後に本稿をまとめる．

2. 既存研究

2.1 社会調査

ネットいじめ研究の初期段階では，ネットいじめの現象を理解するために、社会科学の専門家は、いじめの加害者と被害者の両方の心理的要因，人格と社会的関係に焦点を当ててきた。そのため，多くの大規模な調査を行い，いじめ現象のスコープを発見した．Bastiaensens らは，第三者のいじめに対する行動を調査した [7]．その結果，第三者はネットいじめを目撃した時に，被害者を助ける行動をとる可能性が高いことがわかった．ネットいじめの程度も第三者の行動と関連しており，いじめがより過酷であれば，援助行動を取る可能性も高くなる．一方，その第三者がいじめ加害者と友達であれば，被害者を助けるよりもいじめに加入する可能性が高いことがわかっている．

Campbell らが 9 から 19 歳の学生 3112 名を対象として行ったサーベイでは，ネットいじめと伝統ないじめの被害者の心理被害を調査している [13]．このサーベイでは，抑うつ不安ストレススケール（DASS，Depression Anxiety Stress Scales）を用いて，参加者の憂鬱，不安，ストレスのレベルを自己評価の形式で調査した．DASS は，大人の抑うつ，不安，ストレスの症状を測る尺度で，1 項目 0 3 点の 4 段階で計算され，各 3 スケールの最低スコアが 0 で，尺度の最高スコアが 42 で，値の高さが問題と判断される。その結果を表 1 に示してある．いじめの経験者はいじめ経験なしの参加者と比べて，明らかに DASS スコアが高い．そして，ネットいじめと伝統いじめ両方を経験している参加者は最も心理的な負担を抱えており，伝統いじめの被害者はいじめ経験者の中でより低い心理問題を持つことがわかる．

このように，ネットいじめに関する社会分析は数多く行われており，ネットいじめの悪影響と迅速な対処の必要性を示している．その一方で，ネットいじめ行為を自動的に検知する研究はい未だ少ない．

2.2 テキストコンテンツに対するネットいじめ検知

近年，コンピュータサイエンスの研究者を始めとし，ソーシャルネットワークサービスにおいてのネットいじめの自動検知手法が提案されるようになった．これらは主に，SNS でのメッセージとコメントなどのテキストコンテンツを扱う手法である．インターネット上での会話にいじめ関連のキーワードが含まれているか調べることで，ネットいじめに関するメッセージを識別する Bag-of-words が基準線とされている [11]．

Yin らは感情分析と文脈の特徴を結合したモデルを提案し，Bag-of-words より良い精度が得られることを示した [11]．メッセージを tf-idf を用いたベクトル空間モデルで表し，感情情報を加えたものを特徴として用いることで，39.4% の精度，61.9% の再現率でネットハラスメントを検知した．Dinakar らは動画共有サイトの一つである Youtube に投稿されたコメントに含まれるネットいじめコメントを検知する方法を提案した [2]．彼らの手法では，まずコメントがいじめ関係する敏感トピック群に属するものか分類する．そして，さらにサポートベクターマシンを使ってどの敏感トピックに属するかを分類している．敏感トピックとは，性別，人種，知能や物理属性など人の簡単に変えられない特性である．彼らの手法は，ネットいじめであるコメントを検知できる精度は 66.7% である．Reynolds らは，コメントに対するラベル付けの中で，ネットいじめに関するメッセージは悪意ある単語を含む可能性が高いこと発見した [3]．そして，その情報を利用し，精度 81.7%，真陽性率 61.6% の自動分類器を提案している．単なる Bag-of-words ではなく，著者らは `www.noswearing.com` から得た 296 個の悪意ある単語をその悪意の度合いによりそれぞれ重み付けをした．各メッセージが含んでいる悪意ある単語の数及び密度を特徴とし，C4.5 決定木で分類している．

しかし，ネットいじめは社会的な現象であり，テキストで捉える情報だけでは不完全であることから，これらのテキストベースによる検出方法の精度は限られている．

2.3 ソーシャル情報を用いたネットいじめ検知

ここ何年か，テキストコンテンツのみを分析するのではなく，ネットいじめメッセージが交換された社会的な背景，ユーザーの背景や同時に投稿された画像を扱う手法も提案された．Huang らはソーシャルネットワークのグラフとしての特徴を分析することでネットいじめを検知する方法を提案した [4]．ノードとエッジ数，次数中心度，リンク数や k コアスコアなどを用いて，ネットいじめの加害

者と被害者両方のソーシャル背景を特徴付けた．比較として，大文字，感嘆符や悪意ある単語の密度などのテキスト特徴も使用した．ユーザー間のソーシャルネットワーク構造の特徴とテキスト特徴両方を用いて分類した結果，ソーシャル情報を用いた手法はテキスト情報のみよる検知手法の検知率を改善できることが判明された．Huang らの手法における最も良い真陽性は 76.3% である．また，ネットいじめと検知されたソーシャルグラフを観察したのち，ユーザー間でのインタラクションすなわちリンクが比較的に多い場合は，ネットいじめであることも比較的に可能性が高いと述べられている．加害者と被害者の間では，予想以上にインタラクションが多いことも分かっている．Dadvar らはユーザーがネットいじめを受けた後の行為をネットいじめの自動検知手法の特徴として導入した [12]．ユーザーがネットいじめを受けたものとは異なるソーシャルネットワークサービスで行動を調査することで，ネットいじめの検知の精度を向上できることが示されている．

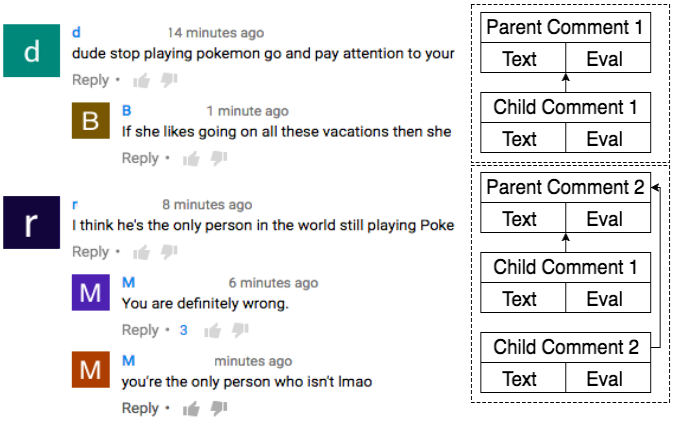
以上の研究は，すべてコメントやメッセージは人を対象として，それぞれ単独な物であると仮定している．しかし，コメント間の関連は，ネットいじめメッセージが交換された際の情報を捉えられる．また，SNS は投稿者を焦点とした短文や画像の共有する目的の他にも，物事を伝える情報共有基盤でも使われている．物事に対しての意見と考えは人それぞれであり，誰もがその考えを伝える権利を持っている．そのため，ある物または事についてきつい事を言うのは建設的な意見とも考えられるので，全てのコメントやメッセージの対象を人だと仮定してネットいじめを検知するのは不完全である．

3. 動画投稿サイトにおけるコメント

本稿では，スレッド形式のコメント投稿を受け付ける動画投稿サイトを対象としている．本節では，そのデータ構造を定義する．

スレッドとはある特定の話題に関する投稿の集まりのことを言う．新たな話題を提供するためにコメントを投稿することを「スレッドを立てる」と言い，そのコメントを親コメントと呼ぶ．この親コメントに対する返信コメントや，返信コメントに対するさらなる返信コメントが連なりスレッドが形成される．そのため，親コメントと返信コメント間には親子関係があり，木構造の一種として扱うことができる．本節では，このコメント構造の形式的定義を与える．

返事文から得られる感情のほか，多くの SNS で取り扱われている高評価ボタンが押された回数も使用した．高評価ボタンは，図 2a に示されているように，コメント欄の下方に位置しており，コメントが有用，または同感を感じた時に使われる．低評価ボタンも存在するが，その負の評価値自体がいじめに鳴りうるため，多くの SNS では 0 以



(a) Threaded comment section example.

(b) Matching comment trees.

図 2: Comment structure examples.

上の値のみを許している．本文では，各コメントには正の評価値が付いているものと仮定する．

定義 1：コメント．コメント c はコメント自身のテキスト情報 t とユーザーからの評価値 $e \in \mathbb{Z}^{*}$[*2], i.e. $c = (t, e)$ を含む構造体である．

定義 2：コメントツリー．コメントツリーはスレッドを表示する有向木である．

$$T = (V, A), \tag{1}$$

where V is a set of c within a thread, and

$$A = \{(a, b) | a, b \in V, \text{ a is the direct reply of b}\} \tag{2}$$

定義 3：親コメント．親コメントとはスレッドを立てるコメントであり，コメントツリーのルートに相応する．

定義 4：一人の発行人から捉えたデータセット．コメントツリーからできる森と表せる。

$$F = (\mathbb{V}, \mathbb{E})$$

と示す．

以上の定義を用いて，図 2a に示されているスレッド型コメント欄を図 2b と書くことができる．

ツリー構造は、ノードにあるテキストデータを示すことができるほか，重要な投稿間の関連もあらわにすることができる．

4. 親子関係を用いたコメント解析

本稿では，動画共有サイトにおいてコメントの対象分岐を解決するために，投稿コメントの対象を人と物に分けて扱い，またコメントの親子関係を利用したネットいじめ検知手法を提案する．投稿コメントの対象を区別することで，物事を対象とした建設的な意見をネットいじめであると誤検知することを防ぎ，コメントの親子関係を考慮する

*2 $\mathbb{Z}^{*}$ denotes a set of integers greater or equal to 0.

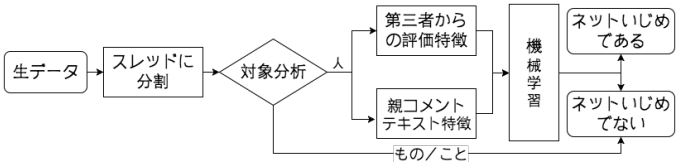


図 3: 提案手法の手順

ことでネットいじめメッセージが投稿された際のユーザー間のインタラクションを捉える．

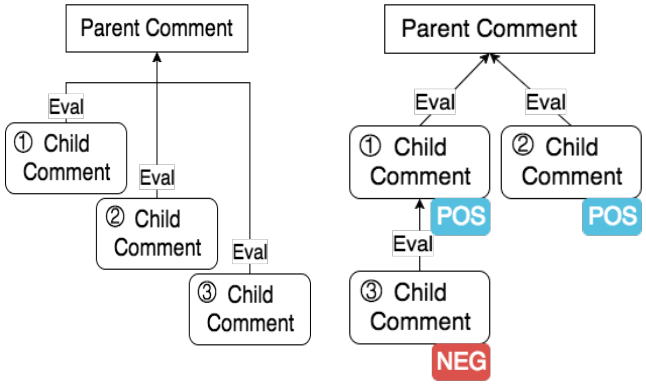
提案手法の手順を図 3 に示す．動画投稿サイトから取得したコメントをスレッド単位に分割する．そして，親コメントの対象を調べ，対象がユーザであるコメントをネットいじめコメントとして抽出する．その後，各スレッドごとに子コメントに対する感情分析を行う．最後に，得られた結果を元に親コメントに対するスコアを求める．

子スレッド投稿されたコメントは，親コメントに対するポジティブまたはネガティブな反応を含んでいることがあり，ネットいじめにおいては，善意の第三者はネットいじめに関係するコメントに対しては否定的な態度をとる [7] ことから，子コメントが多くネガティブな反応を含んでいる場合，親コメントはネットいじめに関する内容を含んでいる可能性が高いと判断できる．特に，動画の視聴者の一部は動画投稿者についてよく知っていると仮定できる．すなわち，その投稿者の動画を購読しているような視聴者は，投稿者の好みやどのような言葉に傷つくのかといった背景知識を共有していると仮定する．

対象分析では，コメントの対象を調べ，投稿者に対する悪意あるコメントのみを抽出する．物事を伝える情報基盤として使われているソーシャルネットワークサービスでは，コメントの対象はビデオの内容と投稿者とで分かれている．感情分析で得た低評価のトップレベルコメントを物事に対した建設的な意見であるか，またはユーザーに対したネットいじめである．すなわち，動画で語られている物事に対する否定的なコメントや一見いじめに見えるコメントは，それらの物事に対する意見であり，個人攻撃とは限らない．本稿では，動画投稿者へのいじめコメントの発見が目的である．そのため，対象分析を行い，投稿者宛てのコメントのみを抽出している．

感情分析では，コメントをスレッド形式で捉え，コメント投稿者間の反応を解析することで視聴者からの反応を調べる．テキストベースとユーザーリアクションベースの手法を用いて感情分析を行う．テキストベースの感情分析では，子コメントのテキスト情報 t に対して感情分析を行い，返事の全体がネガティヴ評価であるか，ポジティブ評価であるかの感情スコア $S(t)$ を得る．その評価に基づき，ツリーの根である親コメントを評価する．

リアクションベースの感情分析方法では SNS で頻繁に用いられている高評価ボタンを使ってユーザーの感情を抽



(a) 二層コメントツリー　　(b) 多層コメントツリー

図 4: コメントツリーの例

出する．他ユーザーによる高評価ボタンは、返信テキストを書くように、コンテンツへの反応を表現する定量的な代替手段であり，視聴者がコメントへの意見を取得する簡単な方法である．ネットいじめ検知を目的として、我々は合意レベルの指標として評価ボタンを用いる。高評価ボタンが多いほど多くの視聴者に認証されたと言え，各コメントの重みとして使用する．

親コメントの評価については，二種のコメントツリーがあるとする．

一つ目のコメントツリーは図 4a のように，すべて二層木構造であるとする．即ち，ここではすべての子コメントは親コメントに対したものと仮定する．また，コメントツリー T は以下の条件を満たす．

$$\text{Indegree}(\text{子コメント}) = 0$$
$$\text{Indegree}(\text{親コメント}) = \text{子コメントの合計数}$$

なので，親コメントの得点は単純にすべての子コメントの加重合計とする．他ユーザーからの評価値を重みとして用いる．親コメント得点 $= \sum_{i=1}^{\text{子コメントの合計数}} S(t) \times e.$

二種類目のコメントツリーでは，コメントツリーを多層木として扱う．即ち，子コメントの対象は親コメントに限らず，子コメント同士のやりとりも存在すると仮定する．つまり：

$$0 \leq \text{Indegree}(\text{子コメント}) \leq n-1$$
$$1 \leq \text{Indegree}(\text{親コメント}) \leq n$$
$$n = \text{子コメントの合計数}$$

ここでは，各子コメントの感情スコアをそのもっとも近い先行への評価とする．多層木で親コメントの得点を計算するため，コメントツリーの葉から層ごとに得点を求めていき，その得点を用いてさらにその上の層の得点に反映し，根の親コメントまでコメントツリーを走査する．

先行への反映方法は二種類用いた．まずは，単純に後継者の得点をその先行者の得点に足す方法である．つまり，

あるコメントの後継者の感情スコアがネガティブであれば，その得点は減ることになる．もう一つの方法は，後継者の得点を重みとして用いる手法である．後継者の感情スコアが高いほど，このコメントがより重要とみなされ，親コメントの評価でより大きい役割を果たす．

5. 実装と評価実験

5.1 コーパス

本研究で用いたデータセットはソーシャルネットワークサービスの www.youtube.com から抽出したもである．YouTube とはグーグル社が運営する，世界で最も利用者が多い動画共有サービスである．YouTube という SNS をデータ源として選んだ理由は以下が挙げられる．

(1) YouTube 利用者の人口分布は一般インターネット利用者の分布と一致している [1]．そのため，得られた結果はこのサイトに限らず，インターネット上においてのユニバーサルなものである可能性が高い．

(2) ネットいじめであるコメントが存在する情報源である．消費者が内容を生成していくメディア消費者生成メディアであることから，ビデオ投稿者個人に対したネットいじめ行為や嫌がらせコメントを投げ出すプラットフォームとして乱用されるケースもある．YouTube は動画所有者に動画からコメントを削除する権利を与えますが，動画投稿者が実際にコメントを見た時点で被害を受けていると言えるため，ネットいじめ行為が緩和できるとは言えない．特に，論争のトピックに関するビデオは，多くの場合，不快と失礼なコメントがある．

(3) コメントの対象はユーザーに限らず様々である．投稿された動画のトピックは幅広く，投稿者の生活動画はもちろんのこと，オンライン講義など物事を伝える情報基盤としても使われている．また，基本的には実名サイトではないので，人間関係は比較的に薄く，ユーザー間の直接なメッセージのやり取りは少ない．

(4) コメント量が豊富である．YouTube は 2013 年末から，スレッド形式のコメント欄を導入しており，10 億ものユーザーがいるため，十分なデータが得られる．

著者らの知る限りでは，ネットいじめ検知に使えるオープンな YouTube コメントのデータセットはない．そこで，独自に YouTube からコメントデータを取得した．本研究は，投稿者一人一人異なる不快な言説も含めていじめ関連コメントを発見することが目的である．そのため，セクシュアリティ，人種と文化と知性などの一般的に議論を招くものをトピックとした動画のほか，美容，ファッションに関する動画からもコメントを取得した．なお，動画投稿者が特定できないパブリックアカウントからの動画は排除した．また，Youtube 自体は多言語で構成されており，投稿動画の言語もそれぞれだが，本研究では英語でのコメントのみを取得した．

次に，得られた動画から，トップレベルコメント，コメントに対した返事と押された高評価ボタンの回数を取得した．本稿で使用するデータセットは，合計で 15683 件の親コメントを採集した．

本研究で使用した情報は，コメントのテキスト文とその高評価数である．YouTube を含む動画共有サイトからは一般的に，その他のデータも収集することができる．本研究では，最少限度のデータを使用し，それらの関係を考慮することでネットいじめコメントを分析できるかを知るべく，上記の情報のみを用いることにした．

スレッド形式のコメントの例として，YouTube の二つのトップレベルコメントとそれに対した返事を図 2a に示す．

図 2a に示す例では，動画視聴者 A と E がそれぞれスレッドを立て，そのコメントに対した返事コメントが投稿されている．もし，その返事が直接トップレベルコメントの投稿者に対するものでなければ,「+」記号の後にそのコメントの対象ユーザーを示している．また，コメント投稿者は別のスレッドを立てて，他のユーザーのコメントについて評価する事もできる．しかし，比較的に少ない行為であるため，今回は独立したコメントとして扱う．我々が調べたところ，実際には無関係のコメント投稿，例えば，スパムや絵文字などの本研究に取って無意味な投稿が存在するが，極性分析には影響が少ないことが分かっている．

5.2 対象分析

対象分析として，各トップレベルコメントに以下の操作を行った：

(1) 重要でない文字列の除去:「lollllll」,「HAHAHAHAHA」と「Wow」なの文字列はデータセットから除去した．

(2) 略記を書き戻す：オンラインでよく使われている略記を元に書き直す．例えば,「u」を「you」に直し，以後の処理を容易にする．

(3) コメントを句ごとに分割．nltk の PunktSenteceTokenizer を用いた．

(4) 対象抽出．Pattern.en は，Python プログラミング言語 [16] のための Web マイニングモジュールである．これは、データマイニング，自然言語処理，機械学習，ネットワーク分析とキャンバスの可視化のためのツールがあります．本文では，分析のみにパターンから文の主語をテキスト分析モジュールを使用した．

(5) 直接ビデオ投稿者に対したのネットいじめコメントを抽出対象とする．有名人，ある製品や地域などをエンティティとしたトップレベルコメントは取り除き，エンティティがビデオ投稿者であるもののみをネットいじめコメントとして登録する．

表 2: 対象分析の結果

クラス	もの，こと対象	ユーザー対象	合計
ネットいじめ	95	558	653
通常	11938	3092	15030
合計	12033	3650	15683

5.3 感情分析

VADER を用いて感情分析を行った．VADER（Valence Aware Dictionary and sEntiment Reasoner）我々の研究に適しているソーシャルメディアで表現感情に特に敏感である語彙とルールベースの感情解析ツールである。

5.4 親コメントの評価

二層木として扱う場合は子コメントの加重得点で親コメントの得点を計算した．多層木の場合では，YouTube のコメント欄は二層のみであるため，時間線とユーザータグを用いて多層構造のスレッドを立直した．ここでは，時間線上もっとも近い，ユーザータグと一致する作者のコメントを先行とする．

二種類の得点反映方については，それぞれシグモイド関数とログ関数を用いた．

$$T(c) = S(c) + f(a)$$
$$f(a) = \sum_{i=0}^{n} \frac{S(a) \times e}{1 + |S(a) \times e|}$$
$$a, c \in \mathbb{V},\ e \in a,\ (a, c) \in E$$

$$T(c) = S(c) * f(a)$$
$$f(a)' = 2 \cdot \log\left(\frac{\sum_{i=0}^{n} \mathrm{S}\,(a) \cdot e}{\sum_{i=0}^{n} e} + 5\right) - 0.4$$
$$a, c \in \mathbb{V},\ e \in a,\ (a, c) \in \mathbb{E}$$

6. 結果

対象分析実験で得た真陽性率を表 2 に示す．対象分析は，多くの無関係のデータを除外することができ，感心するネットいじめコメントは多く残されている．スレッドごとの感情分析のみで抽出されたコメントリストでは，ビデオで討論された話題に関するものや人が対象であるコメントが多く含まれている．例として，古代残酷な統治者についてのビデオでは，「Caligula was not insane.. he was just an asshole to his consuls and the senate.」というコメントが残された．ビデオ内容には反論し，悪意ある単語も含まれているが，ネットいじめコメントとは言えない．

Weka 3.8.0　を用いて分類を行った．またここでは非常に不均衡データを扱うため，SMOTE で事前処理を行った．SMOTE では正例を人工的に作成（オーバーサンプリング）し，負例をアンダーサンプリングする処理である．

違う親コメントの評価法と先行研究で提案された特徴（親コメントの感情スコア [11]，悪意ある単語の密度 [2]，大文字の密度 [2]，疑問符と感嘆符の数 [4]）を用いた分類結果を表 3 に示している．表から，本文で提案された方法は以前の特徴に比べ，精度と F 値を大幅に向上できたことがわかる．さらに，すべての特徴を結合した場合，真陽性率を含むすべての評価基準でより良い結果を得たことが証明できた．この結果はさらに情報ゲイン特徴選択アルゴリズムでランクされたトップ特徴，表 4，によって証明された．

表 3: 正例に対した分類結果

方法	精度	真陽性率	F 値
1. 二層木	0.838	0.608	0.705
2. 多層木 (1)	0.821	0.631	0.714
2'. 多層木 (2)	0.749	0.646	0.693
3. 既存手法	0.754	0.655	0.701
$1+3$	0.828	0.735	0.779
$2+3$	0.835	0.729	0.778
$2'+3$	0.823	0.743	0.781
1 木の深さ >2	0.788	0.630	0.700
2 木の深さ >2	0.771	0.680	0.723
2′木の深さ >2	0.656	0.732	0.692
3 木の深さ >2	0.714	0.588	0.645

表 4: ランク上位の特徴

ランク	特徴
1	子コメントによる評価
2	悪意ある単語の密度
3	親コメントの感情スコア

本手法では，従来の手法では検知が難しい，明らかなテキスト特徴を持たないネットいじめコメントを検知できた．例えば，あるビデオ発行者が自分の性的傾向を公開するビデオでは，「This is why Alice adopted」というコメントが残されている．（ビデオ投稿者のユーザー名を控えるため，ここでは Alice とする）．この一件悪意ある単語を持たないコメントは，Alice が捨てられても当然という意味を持っており，ビデオ発行者にとってとても傷つく発言である．このコメントに対して，多くの第三者が「shut up」などの声をあげており，子コメントでは非常にネガティブなフィードバックを得ているので，本手法で検知することができた．

しかし，第三者が皮肉な表現でコメントを否定するとき本手法での検知は難しい．例えば，「you are the worst person in the world」に対して，「you are」と返事する場合，本手法で検知することができない．また，極端な意見を持つビデオ発行者は時々大勢の人にネットいじめコメン

トを投稿される場合での性能も衰える場合がある．例として，あるビデオ発行者は菜食主義者であり，肉を食べる人は皆キラーであると主張し，発行するビデオで他の肉を食べるビデオ発行者を侮辱している．これを原因とし，多くの視聴者から悪意あるコメントを残され，またこのようなコメントは多くの人に賛同されている．このような状況では，本手法を用いたネットいじめコメントの検知も難しい．

また，コメントツリーの深さを考慮すると，深さが 2 以上の親コメントの検知率は本文で提案された三つの手法ともに既存手法より優れていることがわかる．原因としては，構造が深いネットいじめコメントは大抵論争を招くようなトピックに関したものであり，簡単な呪うコメントではないからではないかと考えている．提案手法同士で比べて見ると，多層木としてモデリングされた手法の方がより優れた真陽性率を得ているが，精度では衰えている．多層木を再構築するときの誤差が原因として考えられる．一方，多層木として得られて特徴と従来の特徴を組み合わせた特徴がもっとも優れた分類性能を得ているので，コメントツリーを多重構造として考える方が有効だと言える．

7. まとめ

本研究では，近年 SNS 上で問題視されているネットいじめの自動検知手法として，コメントの対象とコメント間の親子関係に着目した．トップレベルコメントとその返事の間の親子関係を考慮したのち，辞書ベースでは検知できない各被害者に対した特定話題のネットいじめコメントを抽出することが出来た．また，コメントの対象をユーザーのみのものに絞ることで，ある話題への建設的な意見を排除でき，80.6% の正解率を得た．

本稿で提案された手法は，コメントのテキスト以外の情報と研究者の主観への要求が少ないのが利点である．一方，不足な部分もいくつかある．コメントの返事を基準として，コメント自身を評価しているので，返事を持たないネットいじめコメントは無視されてしまう．また，返事の中でのネットいじめコメントは検知できない．これらの不足点を解決するのが今後の課題である．本研究で使われたのはコメントのテキスト文とその高評価数であるが，明らかに，YouTube サイトからさらなる豊富なデータを収集することができる．データフィールドを拡大し，プロファイル情報などを用いてツリー構造を改善し，より細かなコメントの分析に対応させる．

参考文献

[1] Cha, Meeyoung, et al. "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system." Proceedings of the 7th ACM SIGCOMM conference on Internet measurement. ACM, 2007.

[2] Dinakar, Karthik, Roi Reichart, and Henry Lieberman. "Modeling the detection of Textual Cyberbullying." The Social Mobile Web. 2011.

[3] Reynolds, Kelly, April Kontostathis, and Lynne Edwards. "Using machine learning to detect cyberbullying." Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on. Vol. 2. IEEE, 2011.

[4] Huang, Qianjia, Vivek Kumar Singh, and Pradeep Kumar Atrey. "Cyber Bullying Detection Using Social and Textual Analysis." Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. ACM, 2014.

[5] Wilson, Theresa, et al. "OpinionFinder: A system for subjectivity analysis." Proceedings of hlt/emnlp on interactive demonstrations. Association for Computational Linguistics, 2005.

[6] Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." Journal of Computational Science 2.1 (2011): 1-8.

[7] Bastiaensens, Sara, et al. "Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully." Computers in Human Behavior 31 (2014): 259-271.

[8] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005.

[9] "Alchemy api,". [Online]. Available: www.alchemyapi.com

[10] Langos, Colette. "Cyberbullying: The challenge to define." Cyberpsychology, Behavior, and Social Networking 15.6 (2012): 285-289.

[11] Yin, Dawei, et al. "Detection of harassment on web 2.0." Proceedings of the Content Analysis in the WEB 2 (2009): 1-7.

[12] Dadvar, Maral, et al. "Improved cyberbullying detection using gender information." (2012).

[13] Campbell, Marilyn, et al. "Victims' perceptions of traditional and cyberbullying, and the psychosocial correlates of their victimisation." Emotional and Behavioural Difficulties 17.3-4 (2012): 389-401.

[14] Dadvar, Maral, et al. "Improving cyberbullying detection with user context." Advances in Information Retrieval. Springer Berlin Heidelberg, 2013. 693-696.

[15] Kansara, Krishna B., and Narendra M. Shekokar. "A Framework for Cyberbullying Detection in Social Network." International Journal of Current Engineering and Technology 5.1 (2015).