ネットいじめに関わる投稿に応じた警告文の自動作成方法の提案およびウェブアンケート調査を用いた方法の妥当性の検証

立命館大学

西原陽子　山西良典

安心ネットづくり促進協議会　研究発表会

2020年2月26日

研究背景：中高生のスマートフォン所持によるトラブルの発生

- 近年，中高生のスマートフォン所持率が急上昇
  - 2011年には14.6 %だったが，2016年には81.4 %にまで上昇

利便性が増す一方で，トラブルも多発

ネットいじめの防止の需要

投影のみ

- ネットいじめの特徴
  - 匿名性，相手の顔が見えない
  - 攻撃的な投稿が生まれやすい
- ネットいじめの特徴に関する研究は多数
  - 技術的観点から抑止するような研究は少ない

システムを介していじめの芽を摘み取る

研究代表者等の既存研究の紹介
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

- ネットに投稿されるメッセージに含まれる有害表現を自動判定する手法を提案

- 明らかな有害表現：キーワードマッチングで判定可能
  - 既存手法にも存在する（ニコニコ動画のNGワードリストなど）

- 造語や隠語、文脈に依存する有害表現：時系列深層学習器を用いた手法で判定可能

有害表現の言い換え表現の獲得手法(1)

ドラッグストア

大阪から野菜売ります 

1g 4000~7000円　仕入れ時期などにより値段変わります

野菜以外にも売人紹介します

＃警察警戒中

（Twitterのあるユーザのプロフィールより引用）

ドラッグストアでも野菜売ってるんだー

え？高すぎない？不作で高騰かな？

売人ってなに？
え、なんで警察！？

有害表現の言い換え表現の獲得手法(2)

ドラッグストア

大阪から大麻売ります 

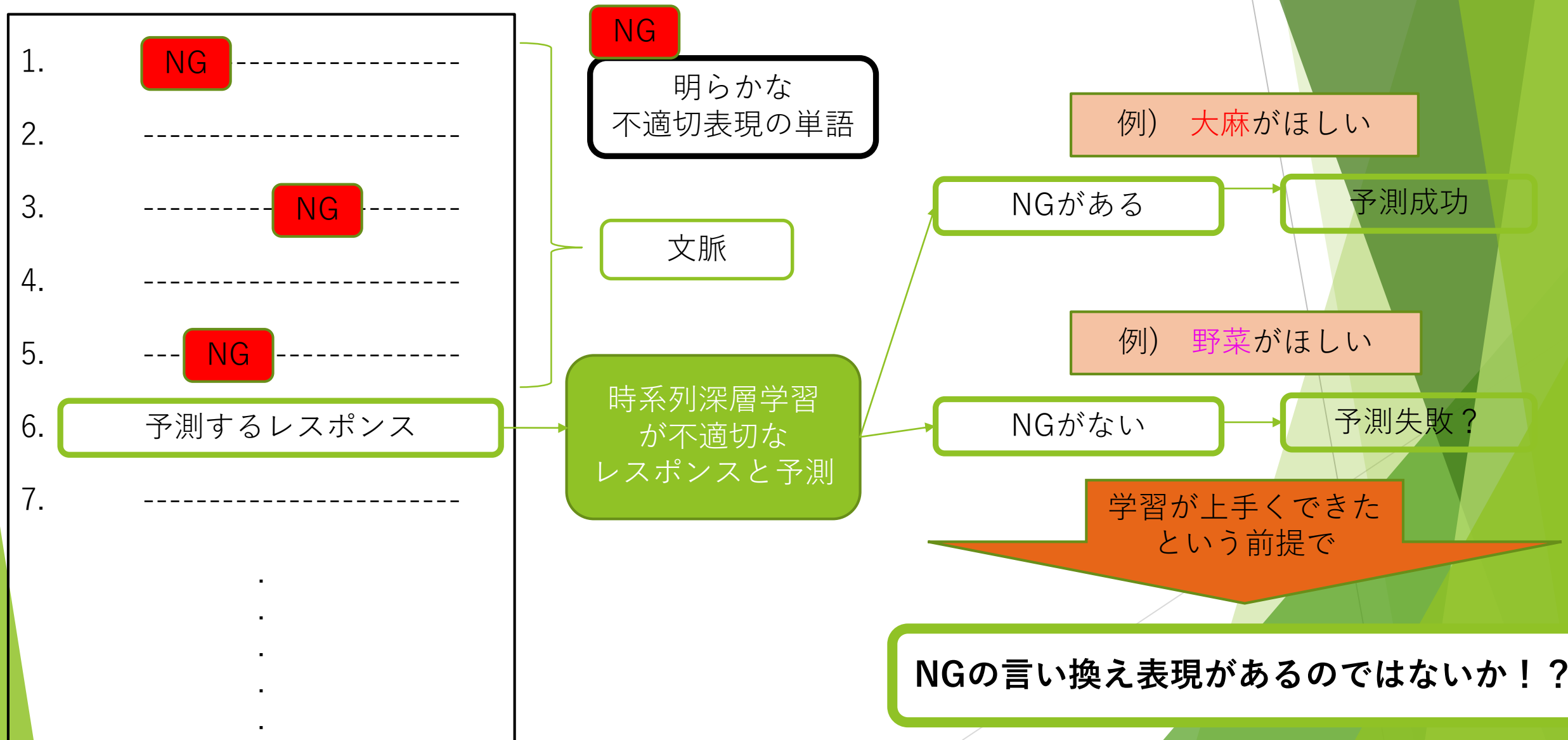
1g 4000~7000円　仕入れ時期などにより値段変わります

大麻以外にも売人紹介します

＃警察警戒中

大麻に置き換えても
違和感がない！

有害表現の言い換え表現の獲得手法(3)



有害表現の言い換え表現の獲得手法(4)

- 正しく分類されたレスポンスの数は1,556+14,413=15,969
  - 全体の約8割に相当した
- 1,207件のうち，言い換え表現が含まれたレスポンスは 677件あった
  - 677/1207 => 約56%

	レスポンス内に明らかな不適切表現がある	レスポンス内に明らかな不適切表現がない
LSTMの出力では不適切である	1,556	1,207
LSTMの出力では不適切ではない	2,413	14,613
合計	3,969	15,820

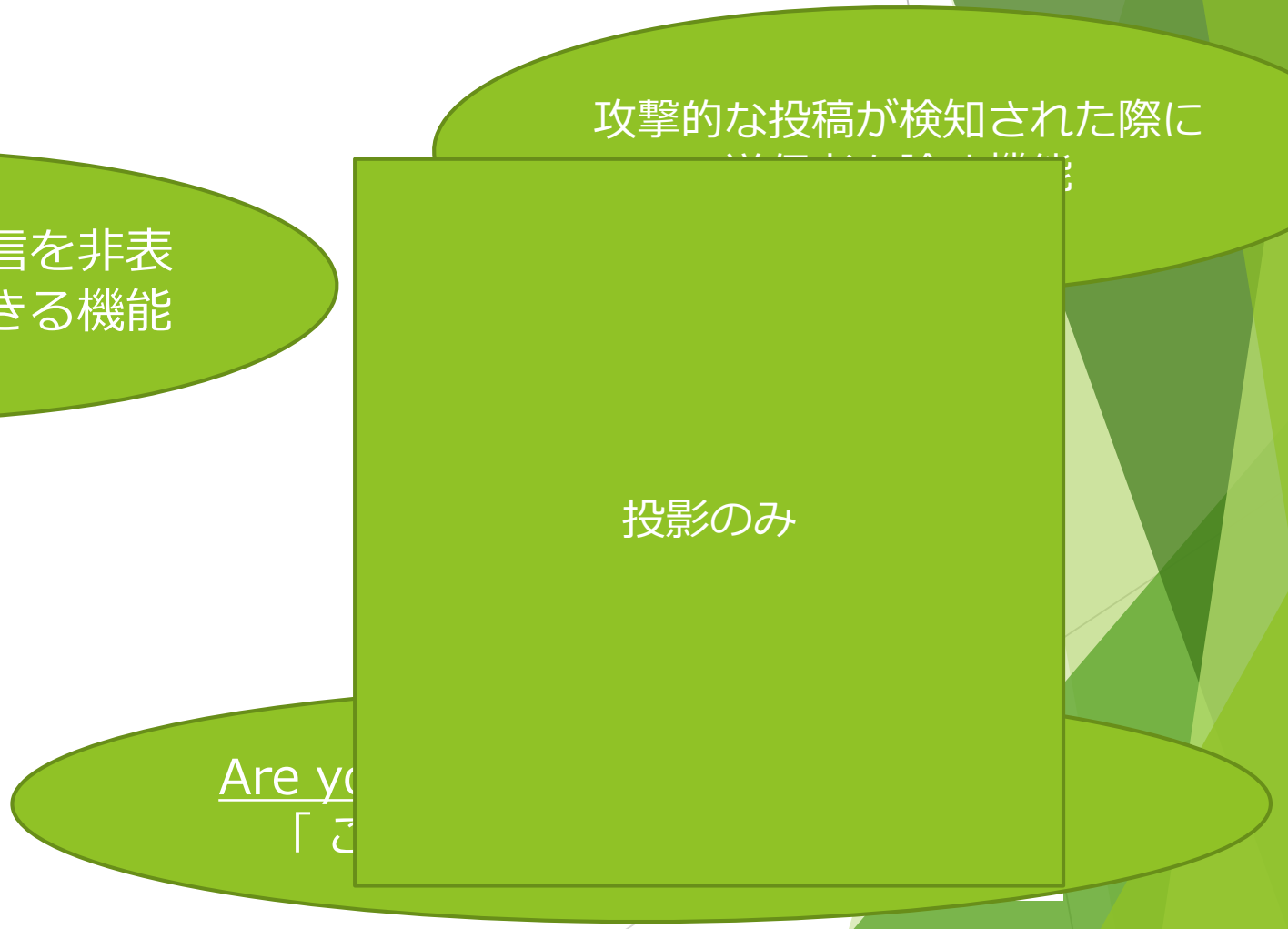
有害表現を判定するだけでは ネットいじめを防ぐことはできない

- 仮に、有害表現が含まれる投稿をフィルタアウトしても、投稿者はフィルタの目をかいくぐり、新たな表現で投稿する可能性が残されている
- 投稿された後にフィルタアウトするのではなく、投稿される前に投稿者に投稿の取り下げを検討してもらうことが肝要

攻撃的なメッセージに対する各SNSの対応

- Twitter

- Instagram



既存のSNSの対応の問題点

- TwitterやInstagramの手法の問題点

  Twitter：

  - ✓ 隠すだけでは問題解決になっていない
  - ✓ 隠された側を刺激してしまう恐れも

  Instagram：

  - ✓ フィードバック文が単調
  - ✓ 情報システムにより提示されるメッセージは無視されやすい

投稿内容に応じた効果的なメッセージを検討する

ここでひとつ

関連研究

- Greenwald「説得の認知反応プロセス」（1968年）
- セルフトーク効果：「心の中でのひとりごと」

✕：人はメッセージを受け取ることで態度変容

〇：メッセージを受け取った後に行う**セルフトーク**により態度変容

本稿では，セルフトーク効果に着目した
メッセージ文の有用性を検討する

実験目的

1. 悪口の種類は取り下げに寄与するか
2. 悪口の具体的な指摘は取り下げに寄与するか
3. メッセージの種類は取り下げに寄与するか

上記3つについて検証するために
情報理工学部内にて予備実験を行った後に
中高生対象に**Webアンケート形式**で評価実験を行った

実験手順

① 被験者が、SNSのグループ内でメンバーに悪口を投稿しようとしていると仮定

② 取り下げを促すメッセージを示し、取り下げたいと思うかを評価

③ 1で実験条件を変えて、20問（4セクション＊5メッセージ）繰り返し、終了

- 被験者（株式会社クロス・マーケティングを通じて実施）
  - 年齢：15歳～18歳（平均16.81歳）
  - 調査対象：中学生・高校生
  - 調査時期：2019年12月
  - 回答件数：400件（内，男性200件，女性200件）

投影のみ

実験条件

セクション	条件１：露骨な悪口	条件２：悪口の指摘
1	○：「馬鹿」「死ね」を含む	あり
2	○：「馬鹿」「死ね」を含む	なし
3	×：あからさまな言葉は含まない	あり
4	×：あからさまな言葉は含まない	なし

赤文字の部分が具体的な指摘

- 条件1：SNS内での投稿に含まれる悪口が露骨／そう

投影のみ

- 条件2：メッセージで悪口を具体的に指摘する／具体的に指摘しない

悪口の指摘	フィードバック文
あり	**この投稿には「馬鹿」，「死ね」という悪口が含まれています．**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？
なし	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？

メッセージ文について

- 各セクションにつき5種類のメッセージ文を提示
- 1から4は**セルフトーク型**、5は**禁止型**

分類	メッセージ文
1. 自己投影（主語が投稿者）	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？
2. 自己投影（主語が受信者）	あなたの投稿は，トークルームの人や，相手を不快にしていませんか？
3. 未来提示（被害の提示）	もし投稿を取り下げれば，周囲や相手を傷つけません。
4. 未来提示（被害未提示）	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？
5. 禁止	悪口を投稿することはやめてください。

被験者は，各メッセージに対して取り下げたいと
「**強く思う：+2**」から「**全く思わない：-2**」までの5段階で評価

実験目的

1. **悪口の種類は取り下げに寄与するか**
2. 悪口の具体的な指摘は取り下げに寄与するか
3. メッセージの種類は取り下げに寄与するか

上記3つについて検証するために
情報理工学部内にて予備実験を行った後に
中高生対象にWebアンケート形式で評価実験を行った

実験結果　1．悪口の種類は取り下げに寄与するか

●悪口の種類（ 露骨な悪口かどうか）は取り下げに**寄与する**

フィードバック	露骨な悪口	露骨ではない悪口
1．自己投影（主語が投稿者）	投影のみ	
2．自己投影（主語が受信者）		
3．未来提示（被害の提示）		
4．未来提示（被害未提示）		
5．禁止		

露骨な悪口の場合：
悪口を言っている自覚のもと
フィードバックを受けた場合，
再考する機会が生まれている

露骨ではない悪口の場合：
悪口を言っているという自覚が薄いため，
フィードバックを受けても
「**無視**」されてしまう，
「**機械の誤認識**」と考えられてしまう

実験目的

1. 悪口の種類は取り下げに寄与するか
2. **悪口の具体的な指摘は取り下げに寄与するか**
3. メッセージの種類は取り下げに寄与するか

上記3つについて検証するために
情報理工学部内にて予備実験を行った後に
中高生対象にWebアンケート形式で評価実験を行った

実験結果　2．悪口の具体的な指摘は取り下げに寄与するか

●**限定された場面において**，悪口の具体的な指摘は取り下げに**寄与する**

露骨な悪口の場合

FB	指摘あり	指摘なし
1．自己投影（主語が投稿者）		
2．自己投影（主語が受信者）		
3．未来提示（被害の提示）		
4．未来提示（被害未提示）		
5．禁止		

投影のみ

露骨な悪口を用いた場合：
具体的な指摘がある方が効果大

露骨な悪口ではない場合

FB	指摘あり	指摘なし
1．自己投影（主語が投稿者）	0.068	**0.228**
2．自己投影（主語が受信者）	0.185	**0.295**
3．未来提示（被害の提示）	0.098	**0.198**
4．未来提示（被害未提示）	0.120	**0.158**
5．禁止	**0.310**	0.275

露骨な悪口ではない場合：
具体的な指摘がない方が効果大

- 投稿者に悪口を送るという自覚がない場合，具体的に指摘を行うことは，「**機械の誤認識**」という捉え方を**助長してしまう**ことが考えられる

実験目的

1. 悪口の種類は取り下げに寄与するか
2. 悪口の具体的な指摘は取り下げに寄与するか
3. **メッセージの種類は取り下げに寄与するか**

上記3つについて検証するために
情報理工学部内にて予備実験を行った後に
中高生対象にWebアンケート形式で評価実験を行った

実験結果　3．メッセージの種類は取り下げに寄与するか

●メッセージの種類は取り下げに**寄与する**

	露骨な悪口		露骨ではない悪口	
メッセージ	指摘あり	指摘なし	指摘あり	指摘なし
1．自己投影（主語が投稿者）	投影のみ			
2．自己投影（主語が受信者）				
3．未来提示（被害の提示）				
4．未来提示（被害未提示）				
5．禁止				

露骨な悪口・
具体的な指摘あり

1の自己投影が最高値

露骨な悪口・
具体的な指摘なし

2の自己投影が最高値

露骨ではない悪口・
具体的な指摘あり

5の禁止が最高値

露骨なではない悪口・
具体的な指摘なし

2の自己投影が最高値

4セクション中
3つの条件下で
セルフトークの効果が高いことがわかった

セルフトーク型と禁止型の比較

- 露骨ではない悪口の場合における**禁止型の評価値が高かった**

送信者に**悪口を送ろうとしている**という自覚がないと
セルフトーク型では取り下げにまでは至らず
禁止型による文面の強さが評価値に影響したと考えられる

●ここから考えられることは,

セルフトーク型のフィードバック文の効果は,
送信者の意識に依存することが考えられる

今後の展望

- 今回の実験は，Webアンケート形式による**一度きりの調査**であった

  今後の課題：**1ユーザに対して**本システムを
  繰り返し提示した場合における効果の検証

- 送信者が悪口だと思っていない投稿（露骨ではない悪口）では，
  取り下げの効果が低かった

  これらの投稿をどのように防ぐか
  どのようにして送信者の気持ちを変えるか

本研究の結論

- 本研究では，悪口を含む投稿に対する投稿抑止を促すフィードバック文の解明を行った
- 実験は，中高生を対象にWebアンケート形式で調査を行った
- 実験の結果
  - 投稿する**悪口の種類（ 露骨な悪口／そうでない悪口）によってはフィードバック提示効果に違いがある**
  - 悪口の具体的な指摘は**限定された場面で効果がある**
  - 提示する**メッセージの種類は取り下げに寄与する**

ことがわかった。