E-004

ソーシャルメディアにおける発話の攻撃性推定に向けたデータセット構築
Construction of a Dataset for Estimating Offensiveness in Social Media Comments

藤原 知樹* 伊藤 彰則* 能勢 隆*
Tomoki Fujihara Akinori Ito Takashi Nose

1 はじめに

近年ソーシャルメディアにおける攻撃的な表現への対策が重要視されており, ヘイトスピーチやハラスメントのような攻撃的な発言の自動検出やそれに向けたデータセット構築の研究 [1–7] が盛んに進められている. 既に Facebook, Twitter, YouTube などの SNS (Social Networking Service) では, 攻撃的なコメントを自動検出する機能が利用されている. 本研究では, 文脈によって攻撃的ともそうでないとも受け取られる可能性があるグレーゾーンの発言を自動的に検出し, ソーシャルメディアにおける会話をサポートするシステムの実現を目指す. 提案する会話補助システムのイメージを図 1 に示す. 本システムでは, グレーゾーンの発言を検出すると, システムが会話に参加して読み手と書き手の双方をサポートする. そして, 本稿では攻撃的な発言や無害な発言とグレーゾーンの発言の区別を目指し, SNS 上の発話について攻撃性を詳細に評価したデータセットを構築する. また, グレーゾーンの発言を効率よく収集するために, 4 種類の発話収集手法を比較する.

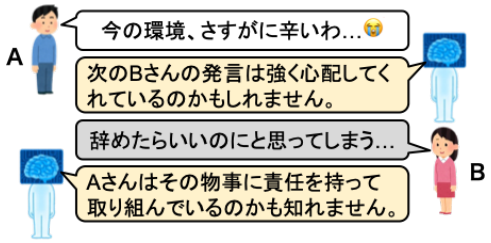

図 1 提案する会話補助システム

2 攻撃性推定に関する先行研究

ここでは, 先行研究で用いられている攻撃性の評価方法と発話の収集方法について説明する.

2.1 ラベル方式

多くの先行研究では攻撃的か否かというような単一の二値ラベル [4], または攻撃の種類に合わせた多クラスラベル [1, 5, 6] が用いられている. 前者はアノテータの作業負担が小さく, 複数名のアノテータの多数決によって正解ラベルを決定できるが, 攻撃の細かな区別が困難である [5]. また, 後者は攻撃の種類を細かく区別できる一方, アノテータにとってそれぞれの攻撃を明確に区別することは容易ではない.

特徴的なラベル方式を提案している研究の一つに Zampieri らの研究 [2] がある. この研究では, 攻撃的か否かの二値ラベルに加えて, 攻撃の標的が特定か不特定か, 個人か集団かなどを区別するラベルが提案されている. 攻撃の標的を区別することで, 例えば特定の個人が標的の攻撃はいじめに, 特定の集団が標的の攻撃はヘイトスピーチに該当すると推測される. このラベル方式を採用しているコーパスは複数の言語 [8–10] で公開されており, コンペティション [10, 11] でも使用されている.

また, 最新の研究では文脈情報を考慮することの重要性が議論されている. 文脈情報には主に前後の会話が使用される [7]. このような具体的な文脈情報を提示する手法には, 会話に含まれている一部の文脈情報しか参照できない [7], アノテータの作業負担が大きいなどの課題がある. 一方で, 先行研究 [3] では具体的な会話は提示せず, アノテータが想像する文脈を誘導するという手法が提案されている. 具体的には,「攻撃的でないと評価され得る文脈を考慮して攻撃的か否かを評価する」というような文脈の誘導によって, 実験では 100 発話のうち 1 割程度の発話で攻撃的か否かの評価結果が変化している.

2.2 発話の収集手法

実環境との乖離が小さいデータセットを構築するには, 発話はランダムに収集される必要がある. しかしながら, ランダム収集では攻撃的な発言の収集される割合が数%ほどと極端に少なく [1], 十分な量の攻撃的な発言を収集するために大量の発話へのラベル付与が必要となる. そのため, ほとんどの先行研究では NG ワードなどのキーワードを用いた発話の収集が行われている. その他, キーワードと極性分析を併用する手法 [1] や, 攻撃検出モデルのアンサンブルによって大量の発話へのラベル付与を行う先行研究 [12] などが存在する.

3 構築したデータセット

本稿では, 攻撃的な発言や無害な発言とグレーゾーンの発言との区別に向けて, 攻撃性評価の新たなラベル方式を提案する. また, グレーゾーンの発言を効率よく収集するために, 4 種類の発話収集手法を比較する.

3.1 提案するラベル方式

本研究では攻撃的な発言を次のように定義する. 攻撃的な発言とは, 発言の対象が特定か不特定かに関わらず, 対象を傷つける, 侮辱する, 脅迫する, または対象に不快感を与えるような発言である. そして, 発話の攻撃性評価のための 2 種類各 4 段階のラベル方式を提案する.

ラベル A: 攻撃の可能性 与えられた文が攻撃的な発言か? 攻撃的な発言ではないか? **4)** 攻撃的な発言である; **3)** 文脈によってどちらともとれるが, やや攻撃的な発言である; **2)** 文脈によってどちらともとれるが, やや攻撃的な発言ではない; **1)** 攻撃的な発言ではない.

ラベル B: 攻撃の強さ 与えられた文が「他者に向けられた攻撃的な発言である (独り言でない)」と仮定した場合に, 攻撃的に感じる度合いが強いか? 弱いか? **4)** とても強い; **3)** やや強い; **2)** やや弱い; **1)** とても弱い.

ラベル A では攻撃的な発言である可能性の高さを多段階評価する. また, ラベル B では攻撃的な発言であるという文脈を想定させて, そのうえで攻撃的に感じる強さを多段階評価する. 本稿では, 攻撃の可能性は高いが攻撃の強さは弱い, または可能性は低いが強さは強いと評価される発話をグレーゾーンの発言とする.

* 東北大学大学院工学研究科 Graduate School of Engineering, Tohoku University

3.2 4 種類の発話収集手法

比較する 4 種類の発話収集手法について説明する．

3.2.1 キーワードによる発話の収集

キーワードによる収集 (Keyword; KEY) では, 設定したキーワードのいずれかを含む発話を収集する．キーワードの選定にあたり, 日本語の攻撃検出用擬似データセットを構築してナイーブベイズ分類器を学習させ, 単語ごとに攻撃的である確率を算出した．攻撃検出用擬似データセットは, 攻撃的な発言をネット上の悪口やハラスメントの例文から収集し, 攻撃的でない発言を対話コーパス Twitter-Triple [13] からランダム収集して構築している．そして,「おまえ, 野郎, チビ, 消えろ, デブ, ブス, 死ね, ハゲ, カス, いけよ」の上位 10 個のキーワードについて各 20 文, 合計 200 文を収集する．

3.2.2 攻撃判定による発話の収集

攻撃検出モデルの攻撃判定による収集 (Offensive; OFF) では, 2 種類の検出モデルが攻撃的と判定した発話を収集する．使用する検出モデルは, 英語のコーパス [2] を日本語に機械翻訳*したデータセットで Fine-tuning (FT) した東北大 BERT-v2 (BERT-FT) と, 翻訳データセットおよび複数の外国語コーパス [10] で FT した多言語 RoBERTa (XLMR-FT) である．両者が攻撃的と判定した発話からランダムに 200 文を収集する．

3.2.3 不確実性サンプリングによる発話の収集

不確実性サンプリングとは, 能動学習で用いられる学習中のモデルの決定境界付近にあると予想されるサンプルを収集する手法である．仮説として, 不確実性サンプリングによってグレーゾーンの発言を効率よく収集できる可能性があると考えた．本実験で用いる不確実性サンプリング手法は Least Confidence (LC) [14] と, Contrastive Active Learning (CAL) [15] の 2 種類である．

LC は分類モデルの予測する確率分布においてクラス間での予測確率の差が小さい発話を収集する手法である．各発話 x_i の予測された確率分布を $p(x_i)$ として, 不確実性スコアの定義式を式 1 に示す．

$$Score_{LC} = 1 - max(p(x_i)) \tag{1}$$

CAL は既にラベル付与されている発話と比較して, 埋め込み表現の類似度は高いが, 分類モデルの予測する確率分布の差異が大きい発話を収集する手法である．不確実性スコアの定義式を式 2 に示す．ラベルが付与されていない発話 $x_s \in D_{source}$ と, 既にラベルが付与されている発話 $x_l \in D_{lab}$ の埋め込み表現の類似度を算出する．そして, x_s と類似度が最も高い K 個のラベル付き発話 $N_{x_s} = \{x_l^{(k)}, k = 1, ..., K\}$ を選択する．そして, それらの予測確率分布 $p(x_s)$ および $p(x_l^{(k)})$ の間で KL ダイバージェンスを算出し, K 個の平均値をとったものが x_s の不確実性スコアとなる．

$$Score_{CAL} = \frac{1}{K} \sum_{k=1}^{K} Div_{KL}(p(x_l^{(k)}) || p(x_s)), \tag{2}$$

$$Div_{KL}(p||q) = H(p,q) - H(p) = \sum_i p_i \log \frac{p_i}{q_i}$$

BERT-FT, XLMR-FT それぞれに対する不確実性スコアを算出し, その平均値を各発話の不確実性スコアとする．そして, $Score_{LC}$ および $Score_{CAL}$ のそれぞれ上位 200 文を収集する．

∗ DeepL: https://www.deepl.com/translator

表 1 アノテータ間の一致度

ラベル	評価の段階数	$\alpha \uparrow$
A. 攻撃の可能性	4	0.43
	3	0.37
	2	0.43
B. 攻撃の強さ	4	0.33
	3	0.30
	2	0.30
文献 [6] (Hate, Offensive, or Not)		0.46
文献 [9] (Offensive or Not)		0.55

4 実験

4.1 実験条件

ツイート, リプライ, それに対するリプライという 3 つ組の会話文を収集している対話コーパス Twitter-Triple から発話を収集した．本実験では 2020 年 8 月 14 日から 2020 年 8 月 19 日の期間のツイートデータを使用した．コーパス本来の前処理には NG ワードを削除する処理が含まれているため, 該当する処理を取り除いて前処理を行った．また, アノテータの作業負担を軽減するため, ツイート単位ではなく文単位で発話を収集した．合計約 237 万文の発話から 3.2 節の各手法により計 800 文を抽出してラベル付与を行った．アノテータは著者が所属する研究室の実験当時学部 4 年生の 3 名で, 3 名とも 20 代前半の男性, ネイティブな日本語話者である．第一著者がラベル付与した発話 10 文を例として提示し, 評価方法に不明点がないか事前に確認した．

4.2 アノテータ間の一致度

一致度の評価には Krippendorff's α を用いる．ラベル A, B のアノテータ間の一致度と, 先行研究における一致度を表 1 に示す．また, 提案法のラベル A, B の評価結果を, それぞれ 4 段階評価から 3 または 2 段階まで統合した場合の一致度を併せて示す．ラベル A の一致度は先行研究 [6] に近いが, 一致度としてはかなり低いと言える．一致度が低くなる要因として, 4 段階評価という評価の複雑さが考えられる．しかしながら, 評価結果を統合して段階数を減らした場合も一致度に変化はない．異なる要因としては, アノテータ間の主観の差異が大きいことが考えられる [3, 7]．ラベル A に比べてラベル B の一致度が低いことは, アノテータ間の主観の差異がより反映されていることを示唆している．

4.3 提案法のラベル分布

3 名のアノテータの評価結果を統合して正解ラベルを決定した．初めにラベル A, B それぞれについて多数決による統合を試み, 多数決で統合できない場合は 1~4 の評価値の平均値を四捨五入して正解ラベルを決定した．正解ラベルの分布を図 2 に示す．ラベル A, B 間の相関係数は 0.79 と強い正の相関がある．分析を単純化するため, 図中の上下左右によって発話を 4 つのグループに分割する．本稿でグレーゾーンの発言と定めるのは, 図中の右下または左上のグループに含まれる発話である．

ワードクラウドを用いて各グループで出現頻度の高い単語を比較する (図 3)．ワードクラウドは, MeCab を用いて発話を形態素解析し, 単語の出現頻度が高い順に強調表示した図である．また比較対象として, 構築したデータセットの 800 文に加え, Twitter-Triple から発話 79, 200 文を追加でランダム収集し, 8 万文からなるソー

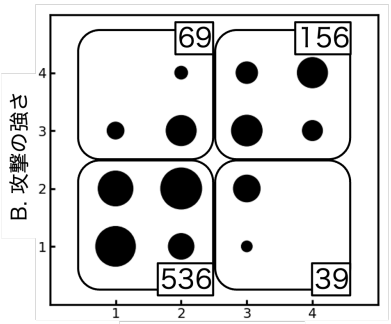

図 2　正解ラベルの分布

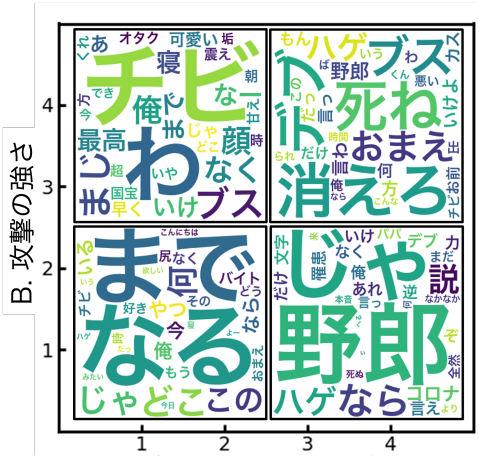

図 3　各グループのワードクラウド

スデータセットを作成した．そのワードクラウドを図 4 に示す．ソースデータセットにおける出現頻度が上位の 50 単語をストップワードとしている．

左下のグループは他の 3 グループとは異なり，発話の収集に用いたキーワードの出現頻度が比較的低い．しかしながら，大部分がランダム収集であるソースデータセットともやや異なっており，発話の偏りが示唆される．そして，他の 3 グループは共通してキーワードの出現頻度が高い傾向にある．特に右上のグループはその傾向が非常に強く，攻撃性が高い発話にはキーワードに用いたような露骨な攻撃表現やそれと共起しやすい単語が含まれやすいことが示唆された．

次に，右下のグループは上側の 2 グループに比べると露骨な攻撃表現の出現頻度は低い．「ハゲ」，「デブ」という単語の出現頻度は高いが，発話を確認すると明示的に自分自身の話題として発言している自虐的な発言であった．その他の特徴として，「コロナ」，「罹患」という単語の出現頻度がやや高い．この結果は，自虐的な内容や触れずらい話題に関する発話がこのグループに分類されやすいことを示唆している．

また，左上のグループは「チビ」，「顔」，「ブス」，「オタク」といった個人の容姿や個性に関わる単語の出現頻度が高い．容姿や個性に関する発言は読み手によっては強く不快に感じる可能性があり，ハラスメントになり得る発言である．加えて，「最高」，「可愛い」，「国宝」といった肯定的な単語の出現頻度が高く，これらは文脈によっては皮肉的に受け取られる可能性がある．この結果は，ハラスメントや皮肉になり得る発話がこのグループに分類されやすいことを示唆している．

図 4　ソースデータセットのワードクラウド

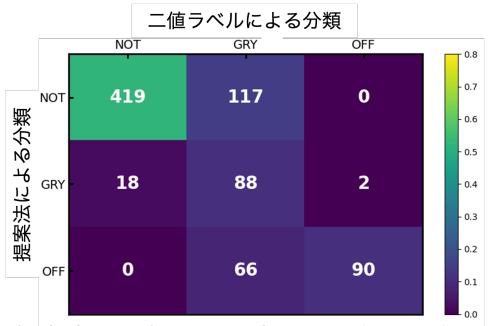

図 5　提案法と二値ラベル (意見の一致/不一致) での分類結果の比較

4.4　二値ラベルによる攻撃性分類との比較

提案法のラベル方式と先行研究で最も使用されている攻撃的か否かの二値ラベルの比較を行う．各手法で発話を攻撃的でない (not offensive; NOT), グレーゾーンである (gray area; GRY), 攻撃的である (offensive; OFF) の 3 クラスに分類し，分類結果を比較する．提案法では，ラベル A, B が共に 1 または 2 である発話を NOT, 共に 3 または 4 である発話を OFF, それ以外を GRY と分類した．また，ラベル A が従来の評価指標に近いと考え，アノテータ毎にラベル A の 4 段階評価の結果を 2 段階まで統合して擬似的な二値ラベルを作成した．その上で，3 名のアノテータ間で意見が完全に一致した発話を NOT または OFF と分類し，意見が割れた発話を GRY と分類した．

両者の分類結果の混合行列を図 5 に示す．二値ラベルで GRY と分類された発話のうち，提案法で NOT または OFF と分類された発話は約 67%である．逆に，提案法で GRY と分類された発話のうち，二値ラベルで NOT または OFF と分類された発話は約 19%と比較的少ない．ここで，グレーゾーンか否かの分類結果が異なる発話の例を表 2 に示す．提案法で GRY, 二値ラベルで NOT と分類された発話は皮肉になり得る発言である．また，提案法で GRY, 二値ラベルで OFF と分類された発話は愚痴のような他者を下げる発言に受け取れるが，他者を酷く貶めるような発言ではなく，攻撃の強さはやや低いと評価されている．以上のことから，提案法の方がより正確な攻撃性の評価および分類がなされている可能性がある．

また，提案法では NOT, 二値ラベルでは GRY と分類された発話は自虐的な発言であると同時に，文脈によっては他者を茶化すような発言になり得る．そして，提案法では OFF, 二値ラベルでは GRY と分類された発話は，発言の対象や読み手が不快に感じる可能性がある発言である．しかし，ヘイトスピーチのような過激な攻撃表現ではなく，比較的グレーゾーンの発言に近いと考えられる．ここで，これらの発話のラベル A, B の値は共に 2 または

表 2　グレーゾーンの発言の比較

提案法 (A, B)	二値ラベル	発話
GRY (2, 3)	NOT	幸せな感覚のまま死ねたらいいのに。
		でも脳はまだ寝てるんすわ。
GRY (3, 2)	OFF	訴えて勝つからなあの野郎、あれ？
		バイトしんどい割に給料安い。
NOT (2, 2)	GRY	ウツが健康だったらこの際ハゲてもいいです。
		私は今オコリザル暴れたい気分。
OFF (3, 3)		下手に声かけて逆に傷つけても嫌やし。
		元彼巻き込み被害くらってほんま笑った。

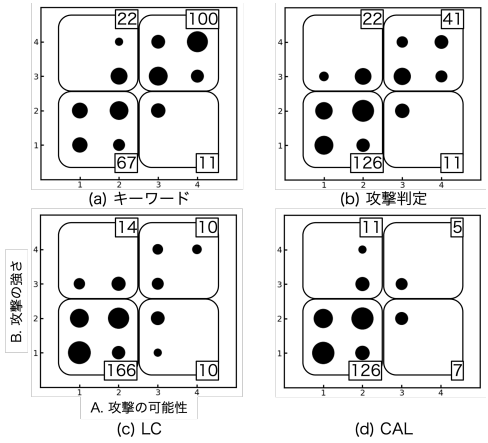

図 6　各サンプリング手法の正解ラベルの分布

3 である．そのため, 1 または 4 と評価された発話とは区別する, または意見の不一致を併用して分類するなど, より細かな攻撃性の分類方法を今後検討する必要がある．

4.5　4 種類の発話収集手法の比較

各手法で収集された発話の正解ラベルの分布を図 6 に示す．仮説として, 不確実性サンプリングによってグレーゾーンの発言が効率よく収集されることを期待した．しかしながら, 各手法で収集された発話のうちグレーゾーンの発言の割合は, キーワードおよび攻撃判定では共に 16.5%, LC では 12%, CAL では 9%であり, キーワードまたは攻撃判定による収集の方が収集効率が高いという結果になった．要因として考えられるのは, 学習データのクラス間の発話数の偏りである．多くの攻撃検出用コーパスでは実環境に比べて攻撃的な発話の割合が多い．そのため, コーパスには実環境で出現頻度の高い攻撃的でない発言が十分に含まれていない可能性があり, 学習不足により攻撃的でない発言の不確実性スコアがより高くなった可能性がある．

5　おわりに

本稿では, ソーシャルメディアにおける発話の攻撃性推定に向けたデータセットの構築を行なった．提案法のラベル方式によって, 攻撃的な発言や無害な発言から, ハラスメントや皮肉になり得る発言, 自虐的な内容や触れづらい話題に関する発言などが区別できることが示唆された．構築したデータセットの課題は, ラベル A, B に基づくより細かな攻撃性の分類方法の検討と, 発話やアノテータの偏りの解消である．また今後は, 攻撃性推定モデルの構築, およびグレーゾーンの発言に対する適切な補助文生成手法の検討を行う必要がある．

謝辞

本研究の一部は, JSPS 科研費 JP21H00895 および JP19H05589 の支援を受けた．

参考文献

[1] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proc. ICWSM 2018*, Vol. 12, No. 1, 2018.

[2] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Predicting the type and target of offensive posts in social media. In *Proc. NAACL 2019*.

[3] K. Arhin, I. Baldini, D. Wei, K. N. Ramamurthy, and M. Singh. Ground-truth, whose truth?–examining the challenges with annotating toxic text datasets. *arXiv preprint arXiv:2112.03529*, 2021.

[4] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In *Proc. ACL-IJCNLP 2021*.

[5] F. Bianchi, S. HIlls, P. Rossini, D. Hovy, R. Tromble, and N. Tintarev. "it's not just hate": A multi-dimensional perspective on detecting harmful speech online. In *Proc. EMNLP 2022*.

[6] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proc. AAAI 2021*, Vol. 35, pp. 14867–14875, 2021.

[7] P. Fortuna, M. Dominguez, L. Wanner, and Z. Talat. Directions for NLP practices applied to online hate speech detection. In *Proc. EMNLP 2022*.

[8] J. Deng, J. Zhou, H. Sun, C. Zheng, F. Mi, H. Meng, and M. Huang. COLD: A benchmark for Chinese offensive language detection. In *Proc. EMNLP 2022*.

[9] Y. Jeong, J. Oh, J. Lee, J. Ahn, J. Moon, S. Park, and A. Oh. KOLD: Korean offensive language dataset. In *Proc. EMNLP 2022*.

[10] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proc. SemEval 2020*.

[11] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proc. SemEval 2019*, pp. 75–86, 2019.

[12] S. Rosenthal, P. Atanasova, G. Karadzhov, M. Zampieri, and P. Nakov. SOLID: A large-scale semi-supervised dataset for offensive language identification. In *Find. ACL-IJCNLP 2021*.

[13] 山崎善啓, 千葉祐弥, 能勢隆, 伊藤彰則. 言語・F0 特徴量系列を考慮したニューラル音声応答生成の検討. 日本音響学会 2021 年春季研究発表会講演論文集, pp. 809–812.

[14] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In B. W. Croft and C. J. van Rijsbergen, editors, *SIGIR 1994*, 1994.

[15] K. Margatina, G. Vernikos, L. Barrault, and N. Aletras. Active learning by acquiring contrastive examples. In *Proc. EMNLP 2021*.