オンラインニュースコメントを対象とした アスペクトベースのコメントフィルタリングシステム

笠原璃音[1] 大和淳司[2] 菊井玄一郎[2]
[1]工学院大学大学院工学研究科情報学専攻 [2]工学院大学情報学部
em23013@g.kogakuin.jp, yamato@cc.kogakuin.ac.jp, kikui@fw.ipsj.or.jp

概要

SNS やオンラインニュースのコメント欄では，誹謗中傷など知りたくない情報に触れることが多く，精神的な負担を感じる人が増えている．NG ワード設定によるフィルタリングの手法もあるが，攻撃的なコメントを完全に防ぐことは難しい．また，既存のコメント表示アルゴリズムは古いコメントを優位に扱う仕様であるため，新しく投稿された有益なコメントが埋もれてしまい，得たい情報を得ることができないという問題もある．これらの課題を解決するために，誹謗中傷表現を含まない建設的なコメントを抽出し，建設的度合いを閲覧者が調節できるシステムを提案した．本研究では，コメントのアスペクト情報を基に建設的度合いを推定する推定器を作成し，GPT4 と比較した実験で提案手法の有効性を確認した．

1 はじめに

SNS のコメント投稿では匿名性の高さから，誹謗中傷や攻撃的なコメントが増加している。このようなコメントは，炎上のきっかけとなるだけでなく，コメントを閲覧する者(以下閲覧者と称する)に不快感を与えることも多く SNS 疲れの要因となっている[1]．誹謗中傷やネガティブなコメントが氾濫する状況において，SNS 利用者は見たくない投稿を避けるための自衛策を講じる必要があるのではないかと考えている。現状，SNS 疲れを軽減する方法として，デジタルデトックスという物理的にデジタルツールから距離を取る方法が提案されている。しかし，SNS が人々の生活の一部となっている現代では完全に断つことは難しく，断念した例も多い[2]．そこで，ストレスなく SNS を利用できる環境が求められるのではないかと考えた。SNS 疲れを感じる閲覧者には，誹謗中傷は見たくない，建設的な内容であれば批評的なコメントを見たい[7]，状況に応じて知りたい情報を見たい[1,6]というニーズがある。このニーズから，本研究では誹謗中傷表現を含まない建設的なコメントを表示させるコメントフィルタリングシステムを提案した．コメントの建設的度合いは閲覧者が調節可能とし，設定した程度に応じて表示内容をカスタマイズできる．コメントの建設的度合いを提示することで，閲覧者の個別ニーズに応じたフィルタリングが実現できると考える．このシステムは，Yahoo!ニュースのコメント欄を対象としている。Yahoo!ニュースは誹謗中傷や炎上が多く、文字数制限が 400 文字と長文から短文まで多様なコメントを収集できるため採用した[3]．また，現行のコメント表示アルゴリズムはいいね数やコメント数を基準としており，新しく投稿された建設的なコメントが埋もれる問題がある。Yahoo!もアルゴリズムの見直しを行っているが，依然として古いコメントが優位な仕様となっている[9]．本システムの導入より，時系列やインプレッション数に依存することなく多くのコメントに触れることができると考える。また，閲覧者が見たいコメントを優先的に表示することで，誹謗中傷表現を目にする機会を減らし，有意義な議論の活性化と SNS 疲れ軽減が望めると考えられる．

2 事前実験

システムを作成にむけ，閲覧者がどのようなコメントを閲覧したい・したくないと感じるかを調査した．Yahoo!ニュースのコメントを被験者 5 名に提示しアンケートを取った．全員が閲覧したいコメント，全員が閲覧したくないコメント，評価が分かれたコメントの例を表 1 に示す．閲覧したいコメントには具体的なメリットが述べられたもの，閲覧したくないコメントには，攻撃的かつ内容がないものだった．評価が分かれたコメントには，「情弱」とった攻撃的な表現を含みつつも後半は客観的な意見が書かれ

表 1. 閲覧したいコメントの例

閲覧したいコメント	マンションは旬な新しいのを借りて住むのが一番．飽きたらまた新しいのに住み替え．これがマンションの醍醐味です．
評価が分かれたコメント	タワマンとかいう情弱が住む特大団地．住民と周囲の街の新陳代謝が落ちなければ，団地よりはまだもつ可能性が高いが経年劣化の限界はある．
閲覧したくないコメント	こういう無駄に歳だけとった人間が日本をここまでグダグダにしてきたんだろうね．皆さっさと死んでもらわないと世の中良くならないと思うな．

表 2. 閲覧したいコメントの定義

主条件	記事に関連し，誹謗中傷を含まない
サブ条件	自分の意見をもとに議論を促している
	客観的な根拠の提示
	新しい解決策の提示
	珍しい体験談の提示

ていた．このようなネガティブな表現の中に建設的な意見が混在しているようなコメントは，人によっては「建設的なことを述べているから見たい」もしくは「表現がきついから見たくない」と感じるようだ．よって，一律にネガティブな表現をフィルタリングすればよいという訳ではなく，閲覧したいコメントの程度を調節する機能のニーズがあるのではないかと考えられる．

3 先行研究

Napoles ら[4]は，建設的な対話を促進する要因を特定することを目的とし，Yahoo!ニュースコメントの分析を通じて建設的なコメントと非建設的なコメントの特徴を明らかにした．建設的なコメント欄には質の高い対話や議論が促されている一方で，非建設的なコメント欄は対立や誤解を生む対話が行われていると指摘している．この研究で明らかになった特徴を参考に，後述する閲覧したいコメントの定義を設定した．

4 提案手法

図 1 に本研究で作成した推定器の概略図を示す．コメントから抽出した 5 つの特徴量を機械学習モデルに入力し，4 段階の建設的度合い[無・弱・中・強]を評価している．本研究で設定した閲覧したいコメントの定義に基づき，特徴量を選定した．先行研究[4,5,8]を基に設定した閲覧したいコメントの定義を表 2 に示す．主条件を必ず満たし，サブ条件のうちいずれか一つを満たしていれば閲覧したいコメントであるとした．

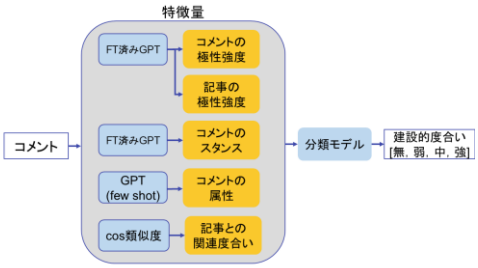

図 1. 推定器の概略図

定義をもとに選定した特徴量は以下のとおりである．主条件の「記事との関連性」を推定するためにコメントと記事本文との関連度合いを採用し．「誹謗中傷を含まない」を推定するために，コメントの極性強度を採用した．サブ条件の内どれに該当しているかを推定するために，コメントの属性判定の結果を採用した．属性は 4 つのサブ条件に加えて非建設を加えた 5 項目のことを指す．加えて，記事内容がネガティブな場合，コメントにネガティブな表現があっても建設的と判断されやすい傾向があるため，記事の感情強度とコメントのスタンスも特徴量として選定した．このようなコメントの一例を付録 1 に示す．次節では特徴量の抽出方法について述べる．

4.1 記事との関連度合い

記事本文とコメントとのコサイン類似度を算出した結果を特徴量として使用している．本文とコメントの分散表現の取得には自然言語ライブラリである GiNZA を使用した．

4.2 極性強度

WRIME データセット[10]を編集し fine tuning させた GPT-4o-mini に，ニュートラル(0)を挟んだ強いネガティブ(-3)から強いポジティブ(3)までの 7 段階の極性強度を判定させた結果を用いている．WRIME データセットとは，文章に含まれる 8 つ感情と極性強度を書き手 1 名と読み手 3 名が判定した値が記録されているものである．このデータセットから文章

と読み手の極性強度を抽出し，fine tuning させた．fine tuning 済み GPT モデルにコメントを 1 件ずつ提示し，コメントに含まれる極性強度を zero-shot で判定させた．記事の極性強度も同様に判定させた．

4.3 属性判定

GPT-4o-min に属性を判定させた結果を用いている．属性判定ではモデルの fine tuning を行わず，建設的なコメントの定義と，建設的なコメントの例文を提示し判定させた．建設的なコメントの例文は GPT4o に出力させたものを使用し，例文をもとに段階的に考えるように指示を出した[13]．プロンプトの例を付録 2 に示す．

4.4 スタンス検出

Xstance データセット[11]を編集し fine tuning させた GPT-4o-mini に，コメントの賛成・反対・中立を判定させた結果を用いている[12]．Xstance データセットとは，政治的な質問とそれに対するコメントのペアから構成され，各コメントに賛成・反対ラベル付けされたデータセットである．質問はドイツ語，フランス語，イタリア語，英語で記載され，コメントは英語を除く 3 言語で記載されている．このデータセットから質問，コメント，ラベルを抽出し，質問とコメントを GPT4o で日本語訳し再構成させた．fine tuning 済み GPT モデルに記事本文とコメントを提示し，コメントの賛否を zero-shot で判定させた．

5 実験

本実験は特徴量からコメントの建設的度合いを推定することができるのかを検証した．特徴量を説明変数とし，機械学習モデルで建設的度合い無，弱，中，強の 4 段階で評価している．

5.1 コメントセット作成

実験にあたりコメントセットを作成した．Yahoo! ニュースの 7 つの記事ジャンル（地域, 国内, 国際, 経済，科学，スポーツ，IT）から各ジャンル 2 つの記事を選定し，コメント欄から 102 件のコメントを収集した．コメントの並びは時系列順に設定し，上位 34 件，中間 34 件，下位 34 件をランダムに収集した．合計で，14 記事から 1,428 件のコメントを収集した．

20 代女性のアノテータ 3 名が，1 件のコメントに対し「建設的か否か」の二値ラベルを付与した．建設的なコメントの定義を提示し，それをもとに判定させた．アノテーション結果をもとに 4 段階の建設的度合い[無，弱，中，強]を設定した．3 名のアノテータのうち, 全員が建設的ではないと判定したら無, 1/3 が建設的だと判定したら弱, 2/3 が判定したら中, 全員が判定したら強というように割り振った．

5.2 実験設定

回帰モデルはランダムフォレストを用い，テストデータに対し 5 回のクロスバリデーションを実施した．以降の結果では，平均絶対誤差（MAE）の 5 回の平均を示す．テストデータの異なる 2 種類の分割方法によって実験条件を変えている．

実験条件 1 では，一つのコメントセットを 8:2 の割合で学習データとテストデータに分割した．

実験条件 2 では，記事単位で学習データとテストデータを分割している．テストデータにはある一つのコメントセットを用い，学習データにはそれ以外のコメントセットを用いている．実験条件 2 を行った意図としては，多様なジャンルのコメントを学習させることで，未知のコメントに対しても分類できるかどうかといった，実装を想定した検証を行うためである．

6 結果

結果では，推定器と GPT4 による回帰精度を示している．GPT4 には，記事本文とコメントを提示し，zero-shot で 4 段階の建設的度合いを推測させた．実験条件 1 の結果を表 3 に, 実験条件 2 を表 4 に示す．GPT4 の方が精度を上回った値を赤字下線で，GPT4 より精度が下回った値を青字下線で示している．全体の MAE の平均は 0.72 であり，1 段階未満の誤差に収まる結果となった．実際の予測結果では，1 段階のミスが多く見られた一方で，2 段階は全体の 40%，3 段階は全体の 5%程度だった．誤差の方向性として，実際の値より高く予測するミスが全体的に多かった．

7 考察

本項では両実験において GPT4 の精度を上回ったコメントセットと，下回ったコメントセットについて論じる．両実験において GPT4 の精度を上回ったコメントセット 6 件を便宜上「優位コメントセット」と呼ぶ．精度を下回ったコメントセット 4 件を「下位コメントセット」と呼ぶ．本章では，アノテータ

表 3. 実験条件 1 の結果（model:推定器の回帰結果　GPT:GPT4 による回帰結果）

	covid	disaster	nepal	harris	honda	yen	drug	life	ICT	otani	mudflow	weather	harass	AI
model	0.69	0.90	0.64	1.06	0.59	0.75	0.75	0.84	0.60	0.70	0.87	1.10	0.71	0.69
GPT	0.84	0.71	0.89	0.84	0.89	0.76	0.77	0.77	0.49	0.85	0.77	0.79	0.80	0.77

表 4. 実験条件 2 の結果

	covid	disaster	nepal	harris	honda	yen	drug	life	ICT	otani	mudflow	weather	harass	AI
model	0.57	1.10	0.70	1.19	0.43	0.76	0.43	0.71	0.67	0.57	0.65	0.86	0.95	0.43
GPT	0.84	0.71	0.89	0.84	0.89	0.76	0.77	0.77	0.49	0.85	0.77	0.79	0.80	0.77

が付けた建設的度合いの分布がモデルの学習に影響を与えている可能性について考察を行った．

コメントセットが建設的度合い無・強に分布している割合と MAE との間に相関があると考えた．全コメントセットの分布と MAE には 0.45 と中程度の正の相関があった．優位コメントセットの間には 0.55 と中程度の正の相関があり，全コメントセットより強い相関を持っていることから，建設的度合い無・強のような極端なコメントを多く含むコメントセットは，モデルによる判定が容易であることが示唆された．続いて，建設的度合い弱・中に分布している割合と MAE との相関を調査した．全コメントセットと MAE には-0.09 と相関が見られなかったことから，建設的度合い弱・中の割合が高いことに回帰精度が寄与していないことが分かる．一方，下位コメントセットと MAE との間には 0.87 と強い正の相関があったことから，建設的度合い弱・中のコメントを多く含むコメントセットは，モデルの精度が落ちることが示唆された．このことから，建設的度合い弱・中といったアノテータ間でも意見が分かれるようなコメントはモデルでも判定が難しいのではないかと考える．これを検証するために，アノテータ一致率をもとに議論する．

表 5 内の” annotator”列には 3 人のアノテータ間の一致率を，” model”列には 3 人のアノテータとモデルの一致率を示している．本研究では，下位コメントセットにおける一致率を抜粋して掲載した．また，14 件のコメントセットにおける一致率の平均値を表中の「14_ave」に示した．いずれの一致率も Fleiss' Kappa を用いて算出している．14_ave を下回った値を青字下線で示した．14 件の平均値を下回るコメントセットが多いことから，アノテータ間でも一致率が低いコメントセットはモデルによる判定も困難であることが示唆された．

表 5. アノテータ間とモデルの一致率

	annotator	model
disaster	0.25	0.31
harris	0.38	0.47
madflow	0.24	0.42
weather	0.13	0.14
14_ave	0.27	0.38

建設的度合い弱・中程度のコメントには，誹謗中傷表現ではないが棘のある表現や，ネガティブな内容の記事に対してネガティブな表現を使って賛同しているものが多い．建設的度合い中程度のコメントの例を表 6 に示す．このようなコメントを捉えるのは GPT の方が得意な可能性があると考える．一方で，GPT4 による精度を上回ったコメントセットも多くあり，6 割を超える精度を記録していることから，GPT4 で建設的度合いを推定させるよりも特徴量抽出器として使う方が良いことが示唆された．

8 まとめ

本研究では，オンラインニュースコメント欄において，ユーザが閲覧したいコメントの建設的度合いを選択できるコメントフィルタリングシステムを提案し，特徴量ベースで建設的度合いを推定する推定器を作成した．実験の結果，GPT4 で建設的度合いを直接推定させるよりも，GPT4 を特徴量抽出器として利用し，特徴量を用いた推定器の方が有効であることが示唆された．しかし，今回作成した推定器では推定が難しいコメントがいくつか観測され，推定器と GPT4 がそれぞれ得意とするコメントの傾向があることが分かった．今後はシステムの有用性の検証を行いたい．被験者にシステムを実際に使用して貰い，システムの使用感や人間の感覚に近いかどうかといった満足度の調査を行いたいと考えている．

表 6. 建設的度合い中程度のコメント

貧困層が増えていくばかりの日本では，これからも乱用者が増えていくだろう．貧困とドラッグはセットである

参考文献

1. 株式会社アップデイト."SNS 疲れに関する実態調査".otalab.2023.09.10. https://otalab.net/press_snstsukare/ ,(閲覧日 2024.4)
2. 上東伸洋, 坂部創一, 山崎秀夫" SNS 交流と共感力との関係性"第 30 回環境情報科学学術研究論文発表会 2016 p. 273-278
3. 国際大学グローバル・コミュニケーションセンター2023."Innovation Nippon 2022　わが国における誹謗中傷の実態調査"
4. Napoles, Tetreault, Pappu, Rosato, Provenzale,"Finding Good Conversations Online:The Yahoo News Annotated Comments Corpus" Association for Computational Linguistics2017
5. Kolhatkar, Taboada"Constructive Language in News Comments" Association for Computational Linguistics2017
6. Nicholas Diakopoulos and Mor Naaman. 2011. Towards quality discourse in online news comments. In Proceedings of the ACM 2011 conference on Computer supported cooperative work. 133–142
7. 森丈弓，名取洋典，小崎茉貴."SNS 疲れを測る(1)受動的ストレスイベント尺度の作成".第 78 回日本心理学会大会 2014 発表論文集 P61
8. 田渕,小林,村尾."Yahoo!ニュースにおける建設的コメント順位付けモデルの導入"言語処理学会 第 25 回年次大会 2019
9. LINE ヤフー"コメント欄の「おすすめ順」で多様な意見が上位に表示されやすくなる機能の導入について"newsHACK2023.04.18 https://news.yahoo.co.jp/newshack/information/comment_20230418.html(閲覧日 2023.10)
10. 梶原智之"WRIME:主観と客観の感情強度を付与した日本語データセット"自然言語処理 28 巻.2021-9
11. Vamvas,Sennrich "X-stance: a multilingual multi-target dataset for stance detection". 5th SwissText & 16th KONVENS Joint Conference 2020
12. Iain J. Cruickshank, Lynnette Hui Xian Ng" Prompting and Fine-Tuning Open-Sourced Large Language Models for Stance Classification" ACM Transactions on Intelligent Systems and Technology, Special Issue on Evaluations of Large Language Models2023
13. Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa "Large Language Models are Zero-Shot Reasoners", NeurIPS 2022

A 付録

付録 1. 極性が反転するコメントの一例（元スポーツ選手が受けたパワハラに関する記事に投稿されたコメント．すべてのアノテータが建設的だと判定した）

胸糞悪い．本当に学校の部活は解体的出直しが要るな．いや，何十年経って現状も似たようなもんなんだからもう原則いらないんじゃないの．

付録 2. 属性判定のユーザプロンプトの例（システムプロンプトは「あなたは日本語学者です。」と指定）

以下の文章を次の属性の 1 つまたは複数に分類してください．
文章: 「{text}」
1. 自分の意見をもとに議論を起こそうとしている
2. 客観的で根拠が提示されている
3. 新たな考え方や解決策を提供している
4. 記事に関する珍しい経験談である
5. 当てはまらない

##出力形式：1〜5 のいずれかの番号のみをカンマで区切って入力（0 は含まない）

以下に各属性の例文と，その属性に当てはまる根拠を提示します．
例文と根拠を参考に段階的に考えて文章を分類してください．
出力するのは 1~5 の属性番号のみです．
##例文
1. 自分の意見をもとに議論を起こそうとしている
「私は、リモートワークが普及しても対面でのコミュニケーションの重要性は変わらないと考えています。」
→この部分で、筆者は自分の意見を述べている。「皆さんはどう思いますか？」と問いかけていることで、他の読者や聞き手に対して議論を喚起し、反応を促そうとしている。以上よりこの文章は個人の考えに基づく意見の表明が議論の起点となっていると言える。