E-028

返信コメント予測を用いた SNS 投稿メッセージの炎上可能性判定手法
Potential Flaming Prediction of SNS Post Message using Reply Comment Generation

七條 旭澄人‡　佐川 雄二‡　田中 敏光‡
Asuto Shichijo　Yuji Sagawa　Toshimitsu Tanaka

1. はじめに

近年,ソーシャルネットワークサービス（SNS）の普及により,ユーザ間のコミュニケーションは益々活発になっている．しかし，SNS は便利である一方で適切に利用しなければ炎上を招く恐れがある．

炎上とは,SNS 上である投稿やトピックが注目を集め，激しい論争や批判が巻き起こる現象である．炎上によって，個人や企業が多大な被害を受ける場合があり，社会問題の一つとなっている．SNS の炎上に関する研究は，入力テキストに対して SVM モデルを用いた炎上可能性の判定[1]や返信コメントの感情分析による炎上の検出[2]などが行われてきたが,前者は投稿のみでの判定なので限界があり，後者は実際に投稿した後でないと判定できないという問題点があった．

そこで本報告では，入力テキストから返信コメントを予測生成し，ネガティブな返信コメントが多いと予測される場合,炎上可能性が高いと判定する手法を提案し, 過去の投稿を用いた実験結果について報告する．

2. 提案手法

2.1 投稿とユーザの属性を入力

ユーザは,炎上可能性を判定する投稿とユーザの属性を入力する.「属性」とは性別や年齢,職業など人の持つ特徴であり,属性の違いによって返信コメントの内容や炎上の原因が異なる場合があるため入力する.属性はこの後 GPT で処理されるため,返信コメントに影響のありそうな情報を単語や文章など自由な形式で入力することとする．

2.2 入力した情報を基に返信コメントを予測生成

返信コメントの予測生成には,OpenAI 社により公開されている ChatGPT API[3]を利用する．本研究では,GPT-3.5-turbo モデルを使用した.ChatGPT にユーザの属性をプロンプトとして与えてから,投稿を入力することで返信コメントを生成する.返信コメントは 50 文生成するように設定した．

2.3 投稿と返信コメントに PN 判定を行う

本研究では,投稿に対してネガティブな意見を持つ返信コメントが炎上の原因であるとして,そのネガティブなコメントを抽出するために,投稿と ChatGPT によって生成した返信コメントに対して,極性値判定（PN 判定）を行う.PN 判定には,Amazon Web Services 社により提供されている Amazon Comprehend の感情分析 API[4]を使用する.感情分析 API は,文章をポジティブ,ネガティブ,肯定的でも否定的でもない中立なニュートラル,肯定と否定が含まれるミックスの 4 つに分類する．

2.4 類似度判定

元の投稿がネガティブな内容を含む場合は,返信コメントもネガティブなものが多くなるが,それらは必ずしも炎上につながるわけではない.そこで投稿がネガティブと判定された場合に,返信コメントが投稿に対して批判的か共感的かを判断するために類似度判定を行い,類似度が閾値未満と判定された場合は,投稿と異なる意見を持った批判的な返信コメントであるとして,炎上の原因に含める．

類似度判定に必要な文章ベクトルの計算は,Word2Vec と tf-idf を組み合わせた方法を用いた. 本研究では,東北大学の乾・鈴木研究室が公開している訓練済み Word2Vec[5]を使用した．

2.5 炎上可能性と批判的なコメントを表示

最後に,炎上可能性と批判的なコメントを表示する.炎上可能性は,抽出された批判的なコメントの数に応じてユーザに表示するメッセージが異なり,「炎上する可能性は低いでしょう！」,「投稿する際は気を付けましょう。」,「投稿する際は十分に気をつけましょう。」と段階的に表示する．

ユーザはシステムを使用することによって,SNS で実際に文章を投稿する前に炎上の可能性を判定し,炎上につながる可能性のある具体的な予測返信コメントを確認できるので,投稿文の問題点を具体的に把握し修正することが出来る．

3. 実験

3.1 実際の投稿を用いた評価

3.1.1 炎上していない投稿に対する炎上可能性の判定

炎上していない過去の投稿を入力してシステムを実行し,炎上可能性を調べた.炎上していない投稿は,X に投稿されて 1 ヶ月以上が経ち,ネガティブな返信コメントが付いていないものから無作為に集めた 50 文を使用した.属性は,アカウントに記載されている情報を可能な限り入力した．

表 1 には,炎上していない投稿を用いて炎上可能性判定を行った結果を示す.また,炎上していない投稿から抽出したネガティブな返信コメント数を図 1 に示す.ネガティブな返信コメント数全ての炎上していない投稿のうち,64%が「炎上する可能性は低いでしょう」と判定された.一方で,36%の投稿で 1 文以上のネガティブな返信コメント生成された.13 件の判定結果で,抽出した返信コメントの中に炎上の原因として相応しくないコメントが含まれていた.そのうち,10 件で抽出されたコメント全てが誤判定という結果となり,投稿文がネガティブと判定されたのは 6 件,それ以外の判定は 4 件あった.投稿文がネガティブの場合,類似度判定で返信コメントを分類しきれなかったため誤抽出が発生した.しかし,類似度のみで返信コメントを完全に分類するのは難しいものの,

‡名城大学大学院 理工学研究科 情報工学専攻, Graduate School of Science and Technology, Meijo University

表 1　炎上していない投稿を用いた
炎上可能性判定結果

判定結果	投稿の数
炎上する可能性は低いでしょう	32
投稿する際は気をつけましょう	18
投稿する際は十分に気をつけましょう	0

表 2　炎上した投稿を用いた炎上可能性判定結果

判定結果	投稿の数
炎上する可能性は低いでしょう	3
投稿する際は気をつけましょう	45
投稿する際は十分に気をつけましょう	2

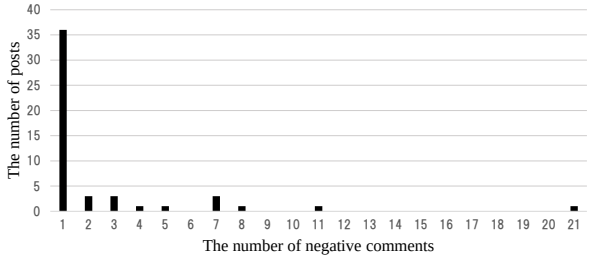

図 1　炎上していない投稿から抽出した
ネガティブな返信コメント数

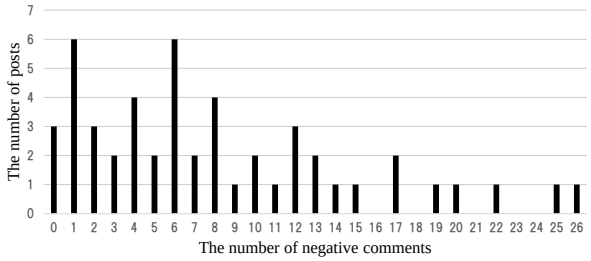

図 2　炎上した投稿から抽出した
ネガティブな返信コメント数

類似度によって共感のコメントの誤抽出を防いだケースは多く存在したため,期待した効果が得られたと言える.改善案としては,類似度の閾値設定を変更することや,分類器などの異なる返信コメントの分類方法を用いることがあげられる.投稿文がポジティブ,ニュートラル,ミックスの場合,批判的なコメント以外がネガティブと判定されるケースが多少存在した.具体例を挙げると,「風邪で寝込んでいたが、元気になったので頑張ろう」という内容のポジティブな投稿文に対して,「体調が悪かったですか？無理せず休むことが大切ですよ。」や「大変ですね。ゆっくり休んで充電してください。」といった投稿者を心配するコメントが生成され,それらがネガティブと判定された.これは,投稿者がポジティブであっても,ネガティブなコメントが付くという本手法では対応できないケースであった.対処法としては,投稿文がネガティブ以外であっても類似度判定をすることが挙げられるが,コメント抽出に与える影響が大きく,通常のケースで批判的なコメントが抽出されにくくなってしまう可能性があるため,投稿文がネガティブな場合と同じようにするのではなく,閾値をより高く設定するなどの工夫が考えられる.

3.1.2 実際に炎上した投稿に対する炎上可能性の判定

システムの炎上可能性判定機能の性能を確かめるために,インターネット上で実際に炎上した投稿を入力してシステムを実行し,判定された炎上可能性を調べた.炎上した投稿は,X（旧 Twitter）等の短文投稿サイトや,ニュースサイト記事から集めた 50 文を使用した.属性は,アカウントや記事に記載されている情報を可能な限り入力した.

表 2 に,実際に炎上した投稿を用いて炎上可能性判定を行った結果を示す.また,炎上した投稿から抽出したネガティブな返信コメント数を図 2 に示す.全ての炎上した投稿のうち,94%の投稿で 1 文以上のネガティブな返信コメントが生成されており, 3.1.1 の結果と比較すると,ネガティブな返信コメントが生成された投稿数が多く,妥当な結果が得られたといえる.一方,3 件の投稿ではネガティブな返信コメントが生成されなかった.これらは,実際に炎上したとき,常識から外れた部分や一部に対する配慮不足を指摘されていた.そのため,コメント生成の際には文章から読み取れる情報から更に考えを広げる必要があるが,ChatGPT は表面的な意味のみを受け取ってしまった為,批判的な意見を生成することが出来なかったと考えられる.

図 2 から,生成されたコメント数の多くは 0～15 文程度であり,過半数に達する 26 文以上が抽出された投稿は 2 件のみであった.今回の実験では,炎上可能性をネガティブな返信コメントが 0 文,1～25 文,26～50 文の 3 段階に分けて判定したが,実際に炎上した投稿から抽出されたネガティブな返信コメント数が図 2 の分布であったことから,炎上可能性の判定の基準設定が適切ではなかったと考えられる.「投稿する際は十分に気をつけましょう」の判定となるコメント数を過半数ではなく 15 文程度まで下げることや,判定をより細かい段階数で分けることが必要であると感じた.

4. まとめ

本研究では,入力テキストから返信コメントを予測生成し分析することで,入力テキストの炎上可能性を判定するシステムを提案した.また,実際の投稿を用いて実験を行うことによって,システムの有効性を示した.

今後の課題として,ネガティブな返信コメント抽出機能の改良や過去の投稿を考慮したコメント生成などが挙げられる.ネガティブな返信コメント抽出機能の改良は,3.1.1 で述べたポジティブな投稿にネガティブな返信コメントが付くケースへの対応やコメントの抽出精度の改善を検討している.過去の投稿を考慮したコメント生成は,ユーザの過去の投稿が炎上の一因となる場合があり,それらを考慮した適切なコメント生成を行う必要がある.方法として ChatGPT に与えるプロンプトを改良することが挙げられる.

参考文献

[1] 大西真輝，澤井裕一郎，駒井雅之，“ツイート炎上抑制のための包括的システムの構築”，人工知能学会全国大会論文集 29，pp. 1-4，(2015).
[2] 高橋直樹, 檜垣泰彦，“Twitter における感情分析を用いた炎上の検出と分析”，電子情報通信学会技術研究報告 = IEICE technical report : 信学技報 116 (488), 135-140, (2017).
[3] ChatGPT, https://openai.com/chatgpt/
[4] Amazon Web Services,https://docs.aws.amazon.com/comprehend/latest/dg/how-sentiment.html
[5] https://github.com/singletongue/WikiEntVec/releases