オンライン動画の音声認識結果に基づく炎上予測手法の検討

小林　瑞貴† 　桂井麻里衣†

† 同志社大学理工学部　〒 610–0394 京都府京田辺市多々羅都谷 1-3
E-mail: †cgud0029@mail4.doshisha.ac.jp, ††katsurai@mm.doshisha.ac.jp

あらまし　オンライン動画の投稿者にとって，批判や誹謗中傷の集中は避けたい事象の一つであり，炎上リスクの高いシーンの検出が望まれる．本稿では，視聴者がコメントする前に検知することを目的とし，動画の音声認識結果を入力とした炎上予測の実現可能性について検討する．まず，炎上・非炎上データセット構築のため，タイムスタンプつきのコメントを感情分析する．これにより，ネガティブなコメントが多く集まっているシーンが特定できる．それらを炎上シーンとし，ネガティブコメントが少ないシーンを非炎上シーンとみなす．次に，データセットの各シーンに字幕を付与し，これを教師データとして，2 クラス分類器を学習する．実際の動画を用いた実験では，音声認識のみによる炎上予測の効果と性能限界についてそれぞれ考察する．

キーワード　オンライン動画，炎上予測，感情分析

1 はじめに

YouTube[1]，ニコニコ動画 [2]，BiliBili [3]などの動画共有プラットフォームの利用者が急拡大を遂げている．中でも YouTube は，インターネット上で群を抜いて最大のビデオ共有ウェブサイトであり，2020 年時点で 20 億人をこえる YouTube ユーザがいるとされている [1]．多くの動画共有プラットフォームでは，ユーザがコメント欄に動画に対する感想や意見を書き込むことができる．演者の発言や行動によっては批判や誹謗中傷が多く書き込まれることになる．このようにインターネット上のコメント欄などにおいて批判や誹謗中傷などを含む投稿が集中することを炎上という．Moor ら [2] によると，YouTube では炎上が頻繁に起こっており，一部のユーザにとっては個人的な動画のアップロードを控える理由になりうるという．炎上する一般的な理由は，動画にがっかりしたり，動画や別のコメント投稿者に気分を害したりすることであった．これらの理由を考慮すると，炎上は，イメージダウンや視聴者離れに繋がる危険性があり，避けたい事象の一つである．

そこで本稿では，視聴者がコメントする前，つまり動画公開前に各シーンの炎上可能性を予測することを目的とし，動画の音声認識結果を入力とした予測の実現可能性について検討する．提案手法は，炎上・非炎上データセットの構築と音声認識結果を入力とした炎上予測からなる．具体的には，まずデータセット構築のために，各動画のタイムスタンプつきのコメントを収集する．集めたコメントを極性分類し，タイムスタンプを用いて，一定時間区間ごとにネガティブコメントの数を集計する．ネガティブコメントが多く集まっているシーンを炎上シーン，少ないシーンを非炎上シーンとみなす．次に，データセットの各シーンに字幕を付与し，これを教師データとして，炎上・非炎上の 2 値分類器を学習する．

本稿の構成は以下のとおりである．2 章では，コメントを活用した動画分類に関する従来研究と，炎上検出に関する従来研究を説明する．3 章では，オンライン動画における炎上・非炎上データセットの構築手法を提案する．4 章では，3 章で構築したデータセットを用いて，オンライン動画の音声認識結果に基づく炎上予測器の構築手法を提案する．5 章では，炎上・非炎上データセットと炎上予測器の評価を行う．最後に，6 章で本文のまとめ，今後の方向性について検討する．

2 関連研究

ユーザの心情把握のため，コメントデータを利活用する研究は盛んに行われている．Bhuiyan ら [3] は YouTube コメントの感情分析を行うことで，動画の品質や人気度を予測する方法を提案した．Novendri ら [4] は YouTube の映画予告編に関して，好意的・批判的なコメントの量によって視聴者の需要の有無を推し量るため，ナイーブベイズを用いてコメントを感情分析する方法を示した．

炎上に関する研究も盛んに行われている．炎上は動画共有プラットフォーム以外にも Facebook[4], Twitter[5], Instagram[6]などのオンラインソーシャルメディアでも起こりうる．Ozawa ら [5] は，Twitter の投稿について，ネガティブコメントの数を監視することにより，炎上事象を正しく検出できることを示した．Yoshida ら [6] は Twitter, Facebook, ブログといった SNS メディアのコメントを感情分析することにより，炎上が検出できることを示した．Lingam ら [7] は YouTube における炎上動画のコメントを手動で分類することによって炎上の種類の多様性を示し，そのうえで炎上を特定することの重要性を述べた．

上記の関連研究は炎上の検出や炎上の種類の特定を達成して

1：https://www.youtube.com/
2：http://www.nicovideo.jp/
3：https://www.bilibili.com/
4：https://www.facebook.com/
5：https://twitter.com/
6：https://www.instagram.com/

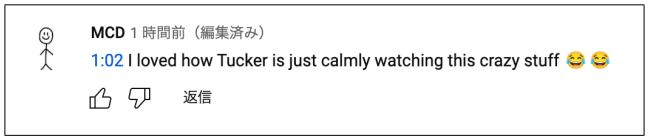

図 1　タイムスタンプつき YouTube コメントの例.

いるが，いずれの研究も受け手の反応を分析するものであり，投稿前に炎上を予測することが課題と考えられる．これを実現するには，投稿に対するコメントではなく投稿内容自体を分析し，炎上確率を算出することが望ましい．本稿では，オンライン動画におけるネガティブコメントの数から炎上シーンを特定し，音声認識結果に基づく炎上予測手法の検討に取り組む.

3　炎上・非炎上データセット構築

YouTube のコメントを分析して炎上・非炎上データセットを構築する．動画は複数のシーンから構成されており，動画内には炎上シーンと非炎上シーンが混在しうる．そのため提案手法では，タイムスタンプつきコメントの極性を分類し，ある時間区間でネガティブと予測されたコメントの件数を集計し，件数が多いシーンを炎上シーン，少ないシーンを非炎上シーンとする.

3.1　タイムスタンプつきコメントの収集

タイムスタンプつきコメントとは図 1 に示すように，動画内の何分何秒に関するコメントであるかをコメントの投稿者が記載したものである．コメントの投稿後，動画内時刻の文字列は YouTube 上でハイパーリンクとなり，クリックすると動画内の該当の場面に遷移する機能を持つ．このタイムスタンプつきコメントを用いることにより，動画内で炎上可能性の高いシーンが抽出できる．コメントの収集には YouTube Data API[7]を用い，そこからタイムスタンプつきコメントのみを抽出する．コメント収集先とした動画チャンネルは，トーク系，企画系，ゲーム実況の 3 種類である．なおキッズ系の動画はコメントがオフになっていることが多く，音楽系の動画は発言の音声認識が困難であるため，これら 2 種類のチャンネルは対象外とした．本稿では，361 個の動画チャンネルから収集した計 124,441 本の動画を対象に実験を行う.

3.2　タイムスタンプつきコメントの極性分類

各動画の一定時間区間ごとのネガティブコメントの数を算出するために，3.1 で収集したタイムスタンプつきコメントの極性分類モデルを 2 つの方法で構築する．1 つ目は，Wikipedia と BookCorpus [8] で事前学習済みの BERT [9] を IMDb データセット [8]を用いてファインチューニングし，極性分類モデルを構築する．IMDb データセットとは，英語で書かれた 5 万件の映画のレビュー文章に対してポジティブまたはネガティブのラベルが付与されたものである．構築したモデルを用いて各タイムスタンプつきコメントのポジティブ，ネガティブの予測確率を算出し，予測確率が高いコメントのみを採用する．2 つ目は，valence aware dictionary and sentiment reasoner (VADER) [10] を用いて極性分類を行い，BERT モデルのときと同様にネガティブコメントを抽出する．VADER とは，「lol」といった頭字語，「sux」といった俗語，utf-8 でエンコードされた絵文字などのソーシャルメディア特有の表現を処理でき，ソーシャルメディアで表現された感情に特化した感情分析ツールである.

3.3　炎上シーン抽出および字幕テキスト収集

タイムスタンプつきコメントの極性分類によって各時間区間のネガティブコメントの件数が算出できる．本研究では，5 秒間に 10 件以上のネガティブコメントが付与されたシーンを炎上シーンとみなす．3.2 章で構築した 2 つの極性分類モデルを用いてシーンを分類し，双方から炎上と判定されたものを抽出する．そして，YouTube Transcript API[9]を用いて字幕テキストを収集する．このとき，発言内容に無関係なタグである [Music] と [Applause] を削除したうえで，字幕が 15 字以下となるシーンは除外する．以上により，炎上検出のための正例候補が獲得される.

モデルの負例となる非炎上シーンは，タイムスタンプつきコメントが 300 件以上ある動画のうち，5 秒間に付与されたネガティブコメントが 3 件未満のシーンとする．タイムスタンプつきコメントの条件は，およそコメントの総数が 5000 件以上ある動画に匹敵する．この処理によって，炎上可能性があるが，コメント数が少ないために正例には含まれないシーンを除外する．3.2 章で構築した 2 つの極性分類モデルを用いてシーンを分類し，双方から非炎上シーンと判定されたものを抽出する．字幕データには正例のときと同様の前処理を適用し，発言内容が極端に少ないシーンを除外する．以上により得られたデータセットの内訳を表 1 に示す.

表 1　構築したデータセットの内訳.

	データ数
炎上シーン	3,592
非炎上シーン	3,087,662

4　字幕テキストに基づく炎上シーン検出

3 章で構築したデータセットを用いて動画シーンの炎上を予測する．表 2 の通り，炎上・非炎上シーンのデータ数は 1 : 1 となるように，アンダーサンプリングする．これらのデータをランダムに訓練・検証・テストデータに 6 : 2 : 2 の割合で分け，訓練・検証データを用いて学習する．ファインチューニングするモデルとして，Wikipedia と BookCorpus で事前学習済みの BERT（以降，WikipediaBERT）と，ツイートで事前学習した BERTweet [11] の二種類を比較する.

7：APIhttps://developers.google.com/youtube/v3/getting-started?hl=ja

8：http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz

9：https://github.com/jdepoix/youtube-transcript-api

表 2　炎上予測に用いるデータの内訳.

	データ数
炎上シーン	2,932
非炎上シーン	2,932

表 3　BERT を用いた極性分類モデルの性能.

ラベル	precision	recall
ネガティブ	0.92	0.88
ポジティブ	0.89	0.92

5 実　　験

5.1 データセット構築方法に関する評価

はじめに，正例を擬似的に獲得するための炎上シーン抽出に用いた極性分類モデルの性能を評価する．BERT モデルに関して，IMDb データセットのテストデータにおける分類性能を表 3 に示す．ただし，ポジティブ，ネガティブの予測確率が 8 割以上のもののみを採用した．ネガティブ，ポジティブ共に高い precision が得られた．これはファインチューニングに用いたデータ量が大量であった点，閾値を予測確率が 8 割以上と高い値で設定した点が理由と考えられる．また，高い閾値にも関わらず，recall も比較的良好な結果となった．

上記のモデルと VADER を用いてそれぞれタイムスタンプつきコメントを極性分類し，炎上シーンを特定した．実際に炎上シーンとして検出されたシーンの例を表 5 に示す．炎上種別とその詳細は該当シーンのコメントや映像を基に著者が手動で付与した．表 5 の通り，偏った発言，人種差別，政治的発言，侮辱などが含まれていた．特定可能性を避けるため一部を伏せ字とする．また，発言が原因で炎上しているものに加え，字幕テキストのみでは状況が不明瞭であり，画像情報や文脈の理解が必要なシーンの存在を確認した．

5.2 炎上予測の評価

本研究で独自に構築したデータセットに基づく炎上予測器の性能評価を行う．それぞれ 80 エポック学習し，検証データにおける損失が最も低いモデルを採用する．WikipediaBERT，BERTweet を用いた炎上予測器におけるテストデータの分類結果を表 4 に示す．提案手法は音声認識結果のみを基に予測する点，さらには独自で構築した炎上・非炎上データセットを目視による精査なしで使用するといった制約を考慮すると，良好な結果が得られたといえる．これは，正例・負例の選定に，厳格な条件を採用したことが理由と考えられる．さらなる精度向上を目指すには，音声認識のみでは性能に限界があると考えられ，予測に用いる特徴量に発言以外の画面情報を考慮した特徴量を導入することが有効と考えられる．また，WikipediaBERT と BERTweet を比較すると，BERTweet を用いた方が良い性能が得られた．これは，ツイートを事前学習に用いた方が口語的な表現を考慮でき，YouTube データに適しているためと考えられる．

表 4　炎上分類結果.

ラベル	BERT		BERTweet	
	precision	recall	precision	recall
炎上	0.71	0.79	0.76	0.82
非炎上	0.77	0.67	0.80	0.73

5.3 炎上シーン予測結果

未知のラベルを持つ音声認識結果に対して，4 章のモデルを用いて推論を行う．BERTweet モデルで推論した際の炎上確率が高い音声認識結果の例を表 6 に示す．特定可能性を避けるため一部を伏せ字とする．事実誤認，非人道的，人種差別などの観点から炎上可能性が高いシーンの存在を確認した．一方，炎上確率が高い音声認識結果の中には，その理由が特定困難なシーンも存在した．そのため，炎上・非炎上データセットに対し，発言以外が原因の炎上シーンを予め除外する処理が必要と考えられる．

6 まとめと今後の課題

本稿では，動画公開前のリスク予防手段として，動画の音声認識結果を入力とした炎上予測の実現可能性について検討した．提案手法では，まず，タイムスタンプつきコメントの極性分類によってネガティブコメントが多く集まっている炎上シーンを検出した．その後，非炎上シーンを追加し，各シーンに字幕を付与した．これらを教師データとして炎上予測器を作成した．提案手法は，全てのプロセスが自動化されていることから，ジャンル別に特化した新たなモデルの構築や，モデルに用いるデータセットの更新が容易である．また，音声認識結果のみから炎上を予測するため，処理が高速という運用上のメリットがある．実験から，コメントに基づく炎上シーン検出の効果を確認でき，かつ炎上予測に関しては制約の中で一定の成果が得られたが，さらなる精度向上のためには，以下の 3 点の検討が望まれる．1 点目に，構築した炎上・非炎上データセットには発言内容以外が原因の炎上も含まれている．この問題に取り組むには，発言に加えて画面の視覚的情報を考慮したマルチモーダル手法を検討したい．2 点目に，同じ発言であってもジャンルごとに炎上可能性が異なることである．解決策として，ジャンル別の学習用データセット構築が考えられる．3 点目に，炎上予測器の学習時にアンダーサンプリングを用いたが，オーバーサンプリングの採用や重みづけの工夫など，様々な不均衡データ対策が有効と考えられる．

文　　献

[1] Osman, M. Eye-Opening YouTube Stats and Facts (2nd Most-Visited Site), 2019, Last accessed: 12/26/2022.

[2] Moor, P. J., Heuvelman, A. and Verleur, R. Flaming on YouTube. *Computers in Human Behavior*, Vol. 26, No. 6, pp. 1536–1546, 2010.

[3] Bhuiyan, H., Ara, J., Bardhan, R. and Islam,M. R. Retrieving YouTube video by sentiment analysis on user comment. *Proc. 2017 IEEE Int. Conf. Signal Image Process. Appl.*

表 5　炎上・非炎上データセットの例．XX は著者らによる伏せ字を表す．

字幕	炎上種別	炎上詳細	ネガティブコメントの数 (BERT, VADER)
many of these teams are big they have lots of sponsors and at the same time they're being racist XX and trying to convince other children to	発言	人種差別	(110, 59)
governments have killed XX million people in the 20th century you know XX whatever people say oh that'll never happen here we all live in	発言	人種差別	(11, 11)
i'm not a XX i'm not a XX can i ask you what you think about XX i mean there's no life there's no life yet there's no life like there's a as a beating little thing where it's a cell growing	発言	事実無根	(13, 14)
man it looks painful yeah I mean to me when I see someone that's tattooed like that it's just self-expression it doesn't bother me too much on the face anyway I'm so badly	発言＋発言と画面の乖離	タトゥー批判	(42, 23)
XX are you crazy are you any mind he could have been a XX could have	発言＋怒鳴る	侮辱＋怒鳴ったこと	(59, 181)
personally I'm XX I'm a XX and I'm XX and those are related knowing we have someone in the White House who calls XX aside	発言	政治的発言	(15, 17)
i'm here wait maybe i'll do this	画面	心霊現象	(493, 348)
I basically want to do that but to the	画面	誹謗中傷	(23, 22)
you know they've actually done wind tunnel testing and they had you know blew you know 50 60 mile an hour winds across to see how they behave but	画面	鳥が死んだ	(840, 1017)
you do live around here yourself yeah I do all right yeah right now it's just going for a little jog today a quick jog in Nice well what are	音声	字幕に拾えていない音声	(445, 500)

表 6　BERTweet を用いた炎上予測モデルの推論における炎上予測確率が高い字幕の例．XX は著者らによる伏せ字を表す．

字幕	炎上確率	炎上可能性
XX are starving and poor that is except of course for some people who live in a place called XX who for some reason like to watch XX each other in some kind of treatment show style gladiatorial	0.9904	事実誤認
other guy looks at the XX and he goes you look like someone put your XX in a blender that's	0.9898	非人道的
great idea let's put the XX on the planet in the XX outfits possible they look like the XX of	0.9897	人種差別
looks because he's not a complete XX hides his face because it's mushy and gross and he looks like if XX and XX had a XX so if	0.9895	侮辱

ICSIPA 2017, pp. 474–478, 2017.

[4] Novendri, R., Callista, A. S., Pratama, D. N., Puspita and C. E. Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes. *Bulletin of Computer Science and Electrical Engineering*, Vol. 1, No. 1, pp. 26–32, 2020.

[5] Ozawa, S., Yoshida, S., Kitazono, J., Sugawara, T. and Haga, T. A sentiment polarity prediction model using transfer learning and its application to SNS flaming event detection. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7. IEEE, 2016.

[6] Yoshida, S., Kitazono, J., Ozawa, S., Sugawara, T., Haga, T. and Nakamura, S. Sentiment analysis for various SNS media using Naïve Bayes classifier and its application to flaming detection. In *2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD)*, pp. 1–6. IEEE, 2014.

[7] Lingam, R. A. and Aripin, N. Comments on fire! Classifying flaming comments on YouTube videos in Malaysia. *Malaysian Journal of Communication*, Vol. 33, No. 4, pp. 104–118, 2017.

[8] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A. and Fidler, S. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 19–27, 2015.

[9] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[10] Hutto, C. and Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8, pp. 216–225, 2014.

[11] Dat, Q. N., Thanh, V., and Anh, T. N. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 9–14, 2020.