動画サイトにおける視聴者コメントの特徴抽出

堺, 雄之介
九州大学大学院システム情報科学府

伊東, 栄典
九州大学情報基盤研究開発センター

出版情報：JSAI Technical Report, SIG-KBS. 124, pp.17-22, 2021-11-15. The Japanese Society for Artificial Intelligence
バージョン：

動画サイトにおける視聴者コメントの特徴抽出

Feature extraction from user comments on YouTube

堺　雄之介[1*]　伊東 栄典[2†]　Yunosuke Sakai[1]　Eisuke Ito[2]

[1] 九州大学システム情報科学府
[1] Graduate School of ISEE, Kyushu University
[2] 九州大学情報基盤研究開発センター
[2] Research Institute for IT, Kyushu University

Abstract: In recent years, slander and bullying have become serious problems on SNS and video sharing service such as YouTube. Our research purpose is detection of comment flaming and extraction features using machine learning. We focus on YouTube, because YouTube is the largest video sharing site, massive viewers are watching movies everyday, and a lot of comments are posted. In this paper, we report comment collection method, flame judgment of comments, vectorization, and flaming detection. Additionally, we also report feature extraction method using machine learning.

1　はじめに

近年，SNS での誹謗中傷やいじめ，それを原因とする自殺が問題になっている．動画サービスでも，視聴者が投稿するコメントが荒れ，誹謗合戦になることも発生している．2020 年 5 月 23 日，SNS 上での誹謗中傷を受けて女子プロレスラーの木村花さんが自殺し社会問題となった [Wikipedia 20]．2020 年 12 月 31 日，Youtuber として活動していた『うごくちゃん』が誹謗中傷を受けて自殺しニュースとなった．動画サイトの利用者は多いため，動画サイトにおける炎上の発見や誹謗中傷コメントの分析には意味がある．本研究では誹謗中傷合戦やネットいじめ状態にある，いわゆる炎上動画の判別器の作成した．さらに炎上動画のコメントの特徴抽出を目指す．

我々はニコニコ動画を対象に炎上動画の自動検出について研究してきた [竹内 21]．その際，次の手順で炎上動画検出器の作成を試みた．まず，人力でコメントが荒れている動画を正例として 847 件収集した．次に，ニコニコデータセットの視聴回数とコメント総数を用いて，正例の動画と同程度の視聴回数とコメント総数を持つ動画絞り込み負例として 847 件選定した．その後，コメントの感情分析 API 等を用いて動画を数値ベクトルに変換した．数値ベクトルに対して，SVM (Support Vector Machine)，決定木，MLP (Multi Layer Perceptoron) を用い学習モデルを作成し炎上動画分類器とした．作成した炎上動画分類器に対して，正解率，適合率，再現率，F 値を用いて性能を評価した．

本研究では日本語の YouTube 動画を対象にする．以前の研究では，ニコニコ動画に特有の映像上に流れる弾幕コメントを用いた [竹内 21]．YouTube には映像の上を流れる弾幕コメントは存在しない．そこで，動画コメントが炎上している動画の検出を目指した．本論文では初めに，YouTube 動画のメタデータ取得方法と，教師あり機械学習における正例の訓練データとなる，検出対象である炎上動画の選出手法を述べる．その後，訓練データのベクトル化手法と炎上動画分類器の作成手法と実験手法を述べる．

本論文の構成を述べる．第 2 節では関連研究について述べる．第 3 節で，YouTube からのメタデータ収集について述べる．第 4 節では学習用の炎上動画選定を説明する．第 5 節では動画の数値ベクトル変換を述べる．第 6 節では教師あり機械学習 MLP，LightGBM を用いた炎上分類器作成を説明する．第 7 節では炎上動画によるコメントの特徴抽出として，異なるタイプのアイコンのユーザーによるコメントに，異なる特徴があるか否かを検出する実験の検討内容を説明する．最後に第 8 節でまとめと今後の課題を述べる．

2　関連研究

日本には「他人の不幸は蜜の味」という言い回しが有る．同義語にドイツ語のシャーデンフロイデ (Schaden-

*連絡先：九州大学システム情報科学府
〒 819-0395 福岡県福岡市西区元岡 744
E-mail: y.sakai.a96@s.kyushu-u.ac.jp
†連絡先：九州大学情報基盤研究開発センター
〒 819-0395 福岡県福岡市西区元岡 744
E-mail: ito.eisuke.523@m.kyushu-u.ac.jp

freude) もある．Wikipedia では「自分が手を下すことなく他者が不幸、悲しみ、苦しみ、失敗に見舞われたと見聞きした時に生じる、喜び、嬉しさといった快い感情」と説明している．脳科学者の中野信子は著書「シャーデンフロイデ [中野 18]」の中で，この感情は人類が長い年月の間で獲得したヒトに備わる反応と述べている．科学技術や情報通信が発達した現代においてもヒトの脳は古いままであるため，ネット上での誹謗中傷やいじめ発生して炎上になるのであろう．

炎上検出や，それに類する状態の検出に関する研究は行われている．2ch に代表される掲示板サイトにおける炎上検出が研究された．投稿頻度や時間を使う手法や，自然言語処理を用いた迷惑メール検出技術を援用した炎上検出が行われた．Twitter などの SNS 利用が普及すると，SNS でも炎上が頻出し，テキスト処理や自然言語処理を用いた炎上検出が行われている．近年では動画サイトを対象とした炎上検出も研究されている．

Salawu らは，ネットいじめ (Cyberbulling) の自動検出手法について報告している [Salawu 17]．ネットいじめの自動検出手法では，テキスト群にたいする自然言語処理と機械学習の組合せが多いと述べている．Salawu らの論文では、ネットいじめ検出手法のアプローチに関する論文では，自動検出手法は教師あり機械学習手法・辞書ベース手法・ルールベース手法・混合イニシアチブ手法の 4 つがあると述べている．教師あり機械学習に基づく手法では，SVM やナイーブベイズなどの分類器を使用する．辞書（字句）ベースの手法では，ネットいじめ用語の辞書を作成し，辞書に登録された単語の有無を利用する．ルールベースの手法では，ネットいじめと判定するためのルールを事前に定義する．混合イニシアチブ手法では，人間が定義した推論を前述のアプローチの 1 つ以上と組合せている．また，ネットいじめ検出の研究では，ラベル付けされたデータセットの欠如が問題だと述べている．

李らは，YouTube を対象にコメントの親子関係を用いたネットいじめコメントの検出を研究している [李子 16]．李らは辞書ベースやルールベースの手法を用いていない．元コメントとその返信の親子関係に着目し，コメント投稿者の間のインタラクションを用いて，ネットいじめの検出を試みている．

Mori らは，個人への誹謗中傷やいじめではなく，ネット上での企業に対する炎上について，炎上後の企業行動および企業株価の変化をまとめている [Mori 19]．2009 年から 2018 年の間に発生した日本の上場企業を対象とした 154 件の炎上を対象にしている．154 件の炎上イベントのうち，70 件では企業は何もせず，残りの 74 件では反応をしている．反応した 74 件のうち，49 件は公式謝罪を，8 件は異議の提示，7 件はコメントを削除している．企業が謝罪またはコメント削除すると，短期的には株価は下落するものの，数日後には株価が戻ると述べている．一方，会社が炎上たいし反対的な行動をすると，株価は炎上発生の数日後から継続的に下落する傾向があると述べている．Mori らの研究は，本研究が対象とする炎上検出ではない．しかしながら炎上が発生した際の対応指針になる．

Rajapaksha らはニュースサイトにおける炎上検出について調査している [Rajapaksha 19]．ニュース記事に対する SNS や Web サイトでの投稿コメントを対象に，否定的コメントを分析することで，炎上の監視と特定が可能だと述べている．Word2Vec または FastText による単語のベクトル化とコメント全体をベクトル化し，深層学習ニューラルネットワーク (NN) モデルで，コメント文の感情を 5 つのクラス「非常にポジティブ，ポジティブ，ニュートラル，ネガティブ，非常にネガティブ」に分類する分類器を学習させている．炎上検出では,「ネガティブ」と「非常にネガティブ」に分類されたコメントが対象となる．実際に Facebook の 3 つの人気ニュースメディア（BBCNews、CNN、FoxNews）に投稿された記事を対象に，機械学習と炎上検出を試している．その結果，提案手法が炎上が検出できたこと，炎上検出に利用できる主な特徴 (feature)，および炎上になる記事のトピックについて述べている．Rajapaksha らの手法は，本研究で考えているコメント文に着目した炎上検出と近く，参考になる部分が多い．

冨永らは，Twitter 上でのユーザーの特性とアイコン画像の関係を調査している [冨永 14]．アイコン画像の分類は経験的に行われており，13 種類に分類している．また分類する人によって結果に差異が生まれないか，ほぼ全てのユーザーのアイコンを 13 種類のいずれかに分類できるかを検証している．調査結果として，13 種類のアイコンごとにユーザーのフォロー数・フォロワー数，及びツイート数を示している．13 種類全ての間に目立つ差は見られなかったとしているが，いくつかの差は発見されたとしている．そのうちの 1 つが最初から設定されている標準アイコンのユーザーと他アイコンユーザーとの差である．標準アイコンのユーザーはフォロー数，フォロワー数，ツイート数の全てにおいて最小値である．また，アニメやゲームの画像を使用するユーザーはツイート数が他より多いという差も発見している．YouTube においてもユーザーはアイコンを設定できるため，動画に対するコメントに差があるのではないかと予想される．上記の文献に示された，標準アイコンのユーザーとアニメやゲームの画像を使用したアイコンを設定しているユーザーのコメントを優先的に分析したいと考えている．

3　YouTube からのデータ収集

我々は以前ニコニコ動画を対象に炎上動画の検出を試みた．本論文では YouTube での炎上動画の検出を試みる．YouTube を対象とする理由は 3 つ有る．1 つ目は利用者数である．YouTube は世界で利用者が最も多い動画共有サービスであるため，対象動画，対象利用者が多い．そのため炎上動画の数も多いであろう．2 つ目の理由は若い世代の利用者数である．若い世代のほぼ全員が YouTube を利用するのに対し，ニコニコ動画の利用は少ない．若い世代も対象とするには YouTube の方が良い．3 つ目の理由は世界対応である．本論文では日本語の動画を対象とするものの，日本語の動画で上手く炎上を検出できれば，英語などの言語でも炎上動画を検出可能であろう．

3.1　動画メタデータおよびコメント収集

YouTube の動画メタデータおよびコメントの収集には，YouTube が提供する Data API [1]を用いる．

視聴者の少ない動画は炎上の可能性も低いし，また社会的な影響も小さいと判断し，再生回数の多い人気動画を対象にすることとした．まず初めに日本向け YouTube 動画のカテゴリごとに，再生回数の多い人気動画のメタデータを取得した．そこから各動画の投稿チャンネル ID を収集した．収集した約 1,800 件のチャンネル ID を用いて，各チャンネルの投稿動画 ID リストを取得した．取得した動画 ID の数は約 46 万件である．収集した 46 万件の動画の中には日本語でないコメントが多数を占める動画も多い．API から動画の情報を収集する際にコメントの言語を絞り込むことはできないが，日本語のコメントが多い動画を抽出するため，動画メタデータの default audio language という項目が日本語に設定されている動画に絞り込んだ．この結果約 10 万件の動画メタデータを収集した．最後に動画のカテゴリを用いた絞り込みも行った．収集した 10 万件の動画の内，最も投稿数が多かったゲームカテゴリに投稿された約 3 万件の動画を本研究の対象とした．

各動画に付随するコメントも，Data API を用いて取得できる．図 1 に YouTube の各動画におけるデータの構造を示す．動画は，動画 ID，動画メタデータ，映像データ，コメント群から成る．動画メタデータには，動画タイトル，投稿者・チャンネル ID，動画投稿日時，動画長，高評価/低評価の数が含まれる．本研究では，対象とする約 3 万件の動画について，それぞれ最大 100 件のコメントを収集した．

動画に付随する視聴者からのコメント群は，文献 [李子 16] で李らが記載しているように，木構造になっている．

[1] https://developers.google.com/youtube/v3/getting-started?hl=ja

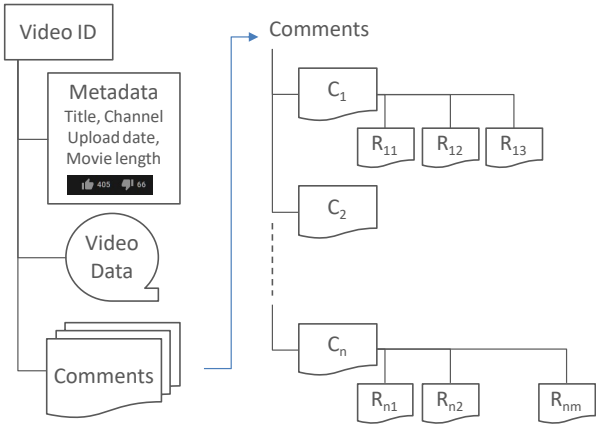

図 1: YouTube におけるデータの構造

図 1 の右側に示すように，トップレベルコメントと，トップレベルコメントへの返信コメントの 2 層構造で構成される．図 1 の右側では，トップレベルコメントを $C_1, C_2, ..., C_n$ とし，トップレベルコメント C_i への返信コメントを $R_i1, R_i2, ...$ としている．また，各コメントには高評価のスコアが有る．今回はトップレベルコメントのみを使用した．

4　学習用の炎上動画の選定

炎上動画か否かの分類は 2 値分類であり，これを教師あり機械学習で解く場合，はじめに正例と負例データを収集する必要がある．ただし YouTube に投稿されている動画数は膨大であり，少数の炎上動画を探すの難しい．そのため本研究では正例の選定の前に候補を絞り込んだ．

炎上している動画では，ある程度の数の視聴者が多数のコメントを投稿しており，さらに動画に対する低評価数の割合が，通常の動画より高いと予想した．そのため，炎上動画候補の絞り込みには動画メタデータの内，再生回数，コメント数，評価数の 3 つの指標を用いた．本研究における絞り込み条件は，再生回数 1 万回以上，コメント数 100 件以上，低評価数が高評価数の半分以上とした．絞り込んだ後，目視で 544 件の正例を選定した．また，再生回数 1 万回以上，コメント数 100 件以上，高評価数が低評価数の半分以上の動画からランダムに 544 件を負例とした．

今回絞り込み条件として使用した数値は適切な値を設定できているとは断言できない．今後多数の動画を調べることで，より適切な絞り込み値を設定できるようになると考えている．

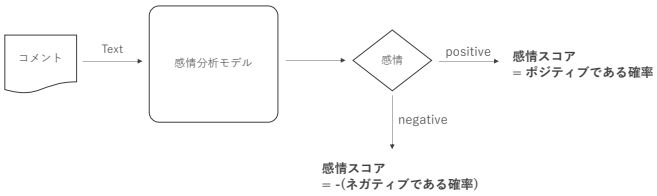

図 2: 1 つのコメントから感情スコアを算出する流れ

5　動画の数値ベクトル化手法

教師あり機械学習で分類タスクを解く際には，学習データのベクトル化手法も重要である．本研究では視聴者によるコメントが荒れている動画を炎上動画とみなしているため，動画本体のデータではなく，動画に付随するメタデータとコメントデータを扱う．これらをベクトル化する上で課題となるのは，コメント本文の数値ベクトル化手法である．

文書の数値ベクトル化手法には，単語ごとの出現頻度のカウント，辞書ベースの手法，Doc2Vec や BERT を用いた手法等があるが，本研究ではセンチメント分析を利用した．炎上していると判断する動画には，ネガティブなコメントが多いように見えたためである．

コメント本文のセンチメント分析には Hugging Face で公開されている，autonlp-japanese-sentiment-59363 モデルを使用した [abhishek 21]．これは BERT を使用したセンチメント分析を行うモデルであり，日本語テキストを対象としている．このモデルは文章を入力として与えると，感情を表すラベルと推定確率を出力する．ラベルは”positive”と”negative”の 2 通りであり，確率は 0.0 から 1.0 の範囲の実数である．本研究ではコメントテキストを入力し，出力された確率を感情スコアとして利用した．この際，出力されたラベルが”negative”の場合には確率を負の値に変換している．その流れを図 2 に示す.

動画をベクトル化するにあたり，動画の再生回数，コメント数，各コメントの感情スコアと高評価数の 4 つの属性を利用した．コメントは 1 つの動画につき先頭 50 件を用いた.

6　分類器の作成と性能評価

分類問題を解く教師あり機械学習手法にはいくつかある．本研究では Python 言語用の Tensorflow を利用して作成した MLP (Multi Layer Perceptron) モデルと，Microsoft が公開している LightGBM[Ke 17] を利用して作成したモデルの 2 種類を用いて，炎上動画分類器を作成した．作成した分類器は 4 つの指標 (Accuracy, Precision, Recall, F-measure) で性能を比較した.

本研究ではコメントのセンチメント分析に注目して動画をベクトル化したため，各モデルにつき 2 通りの手法で動画をベクトル化して実験を行った．1 つ目は今回着目した，各コメントの感情スコアのみを利用した場合である．2 つ目は各コメントの感情スコアに加えて，動画の再生回数とコメント数，各コメントの高評価数の全てを使用した場合である.

まずは各コメントの感情スコアのみを利用した場合の実験結果を表 1 に示す．各コメントの感情スコアに加え，動画のメタデータとコメントの高評価数も利用した場合の結果を表 2 に示す.

F-measure を見ると，全ての情報を特徴量として使用し，LightGBM を用いた場合が最も高い結果となった.

表 1: コメントの感情スコアのみを利用した場合の実験結果

手法	Accuracy	Precision	Recall	F-measure
MLP	0.722	0.780	0.603	0.681
LightGBM	0.680	0.696	0.651	0.672

表 2: 動画のメタデータとコメントの評価数も利用した場合の実験結果

手法	Accuracy	Precision	Recall	F-measure
MLP	0.722	0.714	0.789	0.750
LightGBM	0.814	0.863	0.791	0.826

7　炎上動画コメントの特徴抽出

最後に YouTube における炎上動画のコメントを分析し，その特徴抽出を試みる．本節ではユーザーをアイコン画像によって分類し，各ユーザーのコメントが他と区別できる特徴を持つか否かを検証する実験の検討内容を報告する.

SVM は多量の属性で構成されるデータの分類に優れている。正例と負例データによる学習で SVM 分類モデルを作成すると，属性の重みも出力される。正の重みの大きな属性は正例を特徴付けるもので，負の重みの絶対値が大きな属性は負例を特徴付けるものとなる [Sakai 12]。この手法を使うことで，ある集合のコメント群と，それ以外のコメントとの違いを，単語等の属性で特徴づけることが出来る。

インターネット上のサービスにおけるアイコン画像は，そのサービス上でのユーザーのアイデンティティの 1 つであり，個人の嗜好が強く反映されると考えている．すなわちアイコン画像によって他者からの印象も変化し得る．筆者が注目したのは，アニメやゲームの画像をアイコンに設定しているユーザーに対する別ユーザーからの印象である．近年インターネット上で

「YouTube のコメントを，アニメの画像をアイコンにしているユーザーばかりが荒らしている」という趣旨の言説を複数見かけることがある．各発言者が言及している動画やユーザーの特定を行ってはいないが，アニメ等の画像がアイコンのユーザー全てが他者にとって不快なコメントを残すとは考えられない．そのため，コメントが荒れている動画を対象として，その動画へのコメントをユーザーのアイコンごとに分類し，それぞれに他と異なる特徴があるか検証したいと考えた．

第 2 節の最後に示した関連研究では，Twitter ユーザーの行動にアイコン毎の違いがあるかを検証している [冨永 14]．上記の研究で違いがあるとされていた，「オタク」アイコンと標準アイコンのユーザーを本研究では対象としたい.「オタク」アイコンとは，上記の研究によると,「アニメやゲームの画像を使用したアイコン」である．標準アイコンとは，サービスにより初めから設定されているアイコンである．

本研究における実験手順は大きく分けて 2 つである．まず YouTube のコメントから「オタク」アイコンと標準アイコンを設定しているユーザーによるコメントを目視で選定する．その後 SVM や BERT 等を用いて分類器を作成し，オタクアイコンとそうでないアイコンのユーザーによるコメント，また標準アイコンとそうでないアイコンのユーザーによるコメントの 2 値分類タスクをそれぞれ解く．作成したモデルの分類結果などから，それぞれのアイコンのユーザーによるコメントにおける特徴語を求める．以上の実験から，異なるアイコンのユーザーによるコメントに差異があるか，また差異があればそれぞれどのような特徴があるのかを検出できると考えている．

8　おわりに

本研究では YouTube において視聴者コメントが荒れている炎上動画を検出する分類器を作成した．使用したメタデータは YouTube Data API を用いて取得した．学習データに使用する正例と負例は絞り込みを行い選定した．動画をベクトル化する際は，動画の再生回数，コメント数，及び各コメントのセンチメント分析結果と高評価数を用いた．LightGBM を使用して分類器を作成し実験を行った結果，0.826 の F-measure で炎上動画を分類できた．

今後は特徴量の表現を工夫したいと考えている．本稿で紹介した手法では動画の投稿日時など，未使用のメタデータがあるためである．またコメントについても，トップレベルコメントのみを使用しておりそれに対する返信コメントは未使用である．今回対象とした，ゲームカテゴリ以外のカテゴリについても実験を行いたい．

また炎上動画におけるコメントの特徴抽出として，ユーザーが設定したアイコンによるコメントの違いを検証する実験を提案した．今後実験を行い，方法の評価や結果を考察し報告する予定である．

参考文献

[abhishek 21] abhishek, : autonlp-japanese-sentiment-59363, https://huggingface.co/abhishek/autonlp-japanese-sentiment-59363 (2021)

[Ke 17] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. eds., *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc. (2017)

[Mori 19] Mori, K. and Takeda, F.: Corporate Responses to Internet Flaming: Evidence from Japan, in *2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 359–363IEEE (2019)

[Rajapaksha 19] Rajapaksha, P., Farahbakhsh, R., Crespi, N., and Defude, B.: Uncovering flaming events on news media in social media, in *2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)*, pp. 1–8IEEE (2019)

[Sakai 12] Sakai, T. and Hirokawa, S.: Feature words that classify problem sentence in scientific article, in *The 14th International Conference on Information Integration and Web-Based Applications and Services*, pp. 360–367 (2012)

[Salawu 17] Salawu, S., He, Y., and Lumsden, J.: Approaches to automated detection of cyberbullying: A survey, *IEEE Transactions on Affective Computing*, Vol. 11, No. 1, pp. 3–24 (2017)

[Wikipedia 20] Wikipedia, : 木村花 (May 27, 2021, 05:18 UTC), Retrieved from https://ja.wikipedia.org/wiki/%E6%9C%A8%E6%9D%91%E8%8A%B1 (2020)

[竹内 21] 竹内幹太, 伊東栄典：文書分類手法による炎上動画検出手法の検討, 火の国情報シンポジウム 2021, pp. B3–3, 情報処理学会 (2021)

[中野 18] 中野信子：シャーデンフロイデ, 第 4 巻, 幻冬舎新書 (2018)

[冨永 14] 冨永 登夢, 土方 嘉徳, 西田 正吾：アイコン画像に注目した Twitter 研究の提案, 人工知能学会全国大会論文集, Vol. JSAI2014, pp. 3M44in–3M44in (2014)

[李子 16] 李子怡, 川本淳平, フォン・ヤオカイ, 櫻井幸一：コメントの親子関係を利用したネットいじめコメントの検出, コンピュータセキュリティシンポジウム 2016 論文集, Vol. 2016, No. 2, pp. 1161–1168 (2016)