Vol.2021-ICS-203 No.9
2021/7/2

コメントを利用した炎上動画検出に関する検討

堺 雄之介[1,a)]　竹内 幹太[1,b)]　伊東 栄典[2,c)]

概要：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．本研究では，機械学習による動画サイトにおけるコメントが誹謗中傷状況にあることの検出を目指す．対象とする動画サイトは YouTube を想定している．コメントが荒れている動画を人力で見つけ，そのコメントデータを正例とする．また荒れていない一般動画のコメントを負例とする．これらのデータを機械学習に適用して炎上状態発見器を作る．機械学習を適用するには，対象のベクトル化と，判別アルゴリズムが重要である．本論文では，コメント収集手法，コメントの炎上判定，ベクトル化，アルゴリズムについて，検討内容を報告する．

キーワード：炎上, ネットいじめ, コメント分析, YouTube, 機械学習, 2 クラス分類

Yunosuke Sakai[1,a)]　Kanta Takeuchi[1,b)]　Eisuke Ito[2,c)]

Abstract: In recent years, cyber slander, cyberbullying and comments flaming have become serious problems on SNS and video services such as YouTube. There are few cases of suicide caused by them. In this paper, we study flaming detection methods using document classification and machine learning. The target video service is YouTube. At first, we will manually find flaming comment threads in YouTube, and use them as positive training data. We also select comments thread which isn't flaming, and use them as negative training data. Apply these data to machine learning to create a flaming detector. In order to apply machine learning, it is important to vectorize target data, and to find appropriate classification machine learning algorithms. In this paper, we report how to collect YouTube comments, how to select comments flaming, and vectorization of comments.

Keywords: Flaming, Cyberbulling, comments analysis, YouTube, Machine Learning, Binally classification

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 [1]．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や誹謗中傷コメントの発見には意味がある．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 [2]．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，二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 [2]．YouTube には映像の上を流れる弾幕

[1] 九州大学大学院システム情報科学研究院
[2] 九州大学情報基盤研究開発センター
[a)] y.sakai.a96@s.kyushu-u.ac.jp
[b)] takeuchi.kanta.881@s.kyushu-u.ac.jp
[c)] ito.eisuke.523@m.kyushu-u.ac.jp

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す．本論文では主に，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．

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 3.2 章では学習用の炎上動画選定を説明する．第 4 章では動画の数値ベクトル変換を述べる．第 5 では教師あり機械学習 SVM, 決定木,MLP を用いた炎上分類器作成を説明する．最後に 6 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 (Schadenfreude) 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 [3]」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 らは文献 [4] で，ネットいじめ (Cyberbulling) の自動検出手法について報告している．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．

李らは文献 [5] で，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．李らの手法は，本研究の目的と近いため，参考にする部分が多い．

Mori らは文献 [6] で，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．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 らは文献 [7] で，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．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．

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．この結果得られた約 27,000 件の動画 ID を本研究の対象とする．

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，動画投稿日時，動画長，高評価/低評価の数が含まれる．

動画に付随する視聴者からのコメント群は，文献 [5] で李らが記載しているように，木構造になっている．図 1 の右側に示すように，１次コメントと，１次コメントへの返信である２次コメントの２層構造で構成される．図 1 の右側では，１次コメントを $C_1, C_2, ..., C_n$ とし，１次コメント C_i への返信である２次コメントを $R_i1, R_i2, ...$ としている．また，各コメントには高評価/低評価のスコアが有る．

3.2 炎上動画の選出

炎上動画か否かの判定は，基本的な２値分類（２クラス分類）問題になる．教師あり機械学習で分類する場合，正例・負例の訓練データの収集，対象の数値ベクトル化手法，機械学習による分類手法，の 3 つが問題になる．ここでは正例・負例の訓練データの収集について述べる．

教師あり機械学習では，正例と負例の学習用データが多い方が良い．本研究の目的はコメントが炎上している動画の検出であるため，炎上している動画を正例，そうでない動画を負例とする．膨大な数の動画で，殆どのコメントは炎上していないのに対し，正例となる炎上動画はごく少数である．そのため正例の選定は難しい．

膨大な動画データから炎上動画を見つけるため，候補を大雑把でも絞り込むことにする．その手法として，３つの指標を用いる．

１つ目の絞り込み指標は，動画に対する視聴者からの高評価/低評価の数である．炎上する動画は，低評価の数が多めであろう．低評価のスコアが高い動画を機械的に抽出する．

２つ目の絞り込み指標は，コメント数である．炎上する動画では，ある程度の数の視聴者が，コメントを多数投稿していると考えられる．コメント数が多い動画を，炎上動画の候補とする．

３つ目の絞り込みは，コメントの感情分析 (sentiment analysis) である．コメントが炎上している場合，投稿コメントはネガティブな文章が多いと思われる．動画に投稿された全コメントを感情分析し，動画への全コメントについてネガティブ/ポジティブ度を算出する．コメント群の感情が，ある値以下のネガティブ度であれば，炎上動画の候補とする．

最後に，炎上動画候補に対し，動画とコメント群を人間が確認して，炎上動画か否かを決める．３つの絞り込み指標について，現在の所，適切な値を設定できていない．多数の動画を調べることで，手頃な絞り込み値を設定できると考えている．

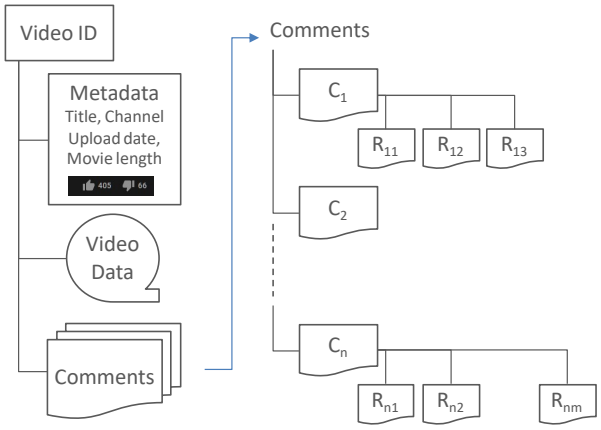

図 1 YouTube におけるデータの構造

4. 動画のベクトル化

本研究では再生される動画本体は扱わず，動画に付随するテキストデータを扱う．図 1 に示すように，動画に付随するテキストデータには，動画メタデータ（投稿日・タイトル・作者・高評価/低評価の数）と動画コメントの２つが在る．炎上動画ではコメントが荒れているため，コメントから数値ベクトルを作る．本研究で検討したベクトル化手法の概要を図 2 に示す．

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

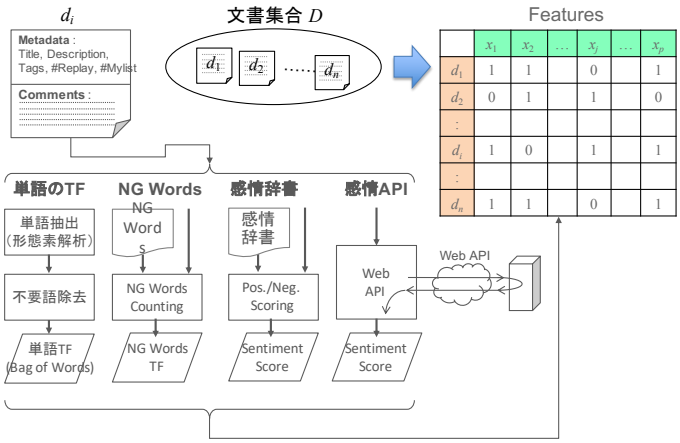

図 2 動画コメントのベクトル化

文書の数値ベクトル化で，古くから使われる手法は単語ごとの出現頻度である [8]．文書内の文を形態素解析機で単語に分割し，単語の出現回数を調べる．文書に出現する単語の頻度を数え上げれば，文書を数値ベクトルで表現できる．単語頻度を用いる場合，単語の切り分け手法の検討が必要である．YouTube 動画で投稿されるコメントの多くは，短い文章であることが多い．顔文字のような辞書に記載のないものも多い，そのため普通の形態素解析は適していないかもしれない．

辞書ベースの手法も援用できる．誹謗中傷で多様される単語の辞書があれば，その単語の頻度を別データとすることで，ベクトルに新たな数値を追加できる．誹謗中傷関連用語集[*2]では，炎上状態を説明する「ネットいじめ」「炎上」「ネットイナゴ」などの単語について，その単語を説明している．しかしながら，誹謗中傷コメントに多用される単語が網羅されているわけではない．ニコニコ動画の「NG 推奨ワード一覧 [9]」は誹謗中傷関連用語集に利用できるかもしれない．用いる場合は YouTube の日本語動画でも適用可能かを調査する必要がある．

文書のセンチメント分析も利用きる．高村は自身の Web サイトで感情辞書を公開している [10][11]．この辞書では，感情的な単語について -1 から 1 の範囲で感情数値を付与している．ポジティブな単語には正の値が，ネガティブな単語は負の値が付与されている．感情辞書に記載のある単語をコメントから拾い上げ，感情辞書のスコアでコメントのポジティブ度・ネガティブ度をスコアにする．感情辞書の援用には２つの問題がある．１つは日本語の感情辞書に登録された単語はネガティブ単語が多いことで，もう２つ目の問題はポジティブ・ネガティブの値しかないことである．

日本語文のセンチメント分析のために，abhishek は機械学習によるセンチメント分析モデルを公開している [12]．このモデルは，入力した日本語文章が Positive か Negative かを数値で出力する．最もポジティブが 1, 最もネガティブが -1 である．コメント文の感情値を利用可能であろう．

単語の分散表現を用いたコメント文の分散表現も利用できる．単語をベクトルで表現する分散表現を得る手法に Word2Vec と fastText がある．Word2Vec は Tomas Mikolov らが開発した分散表現生成法およびツールである [13]．Word2Vec では文書中に含まれる単語の出現数を利用する Countinuous Bag-of-Words (CBOW) モデルと，文章中に含まれる単語の並びから単語の出現確率を利用する Skip-gram モデルの両方の学習モデルを用いて，Hierarchical Softmax 及び Negative Sampling で処理を高速化している．fastText は，Facebook AI Research が 2016 年に開発した自然言語処理向けアルゴリズムおよびツールである [14][15]．fastText は単語の分散表現に加え，テキスト分類も可能である．

コメント文全体をベクトル化する場合，Doc2vec も利用できる [16]．Doc2Vec も Mikolov らにより開発されたアルゴリズムおよびツールである．

炎上動画判定では，各コメントをバラバラに用いて多数のベクトルを検出器（文書分類器）の入力とする方法もある．またコメント群を１つのベクトルに変換し，検出器（文書分類器）の入力にする方法も考えられる．ここで述べた多数のベクトル化手法について，どの手法が適しているかは，今後の実験で明らかにする予定である．

5. 機械学習による炎上動画分類

分類問題を解く教師あり機械学習手法にはいくつかのある [17]．本研究では Python 言語用の scikit-learn モジュール [18] に付随する分類器の利用を検討する．具体的には，SVM (Support Vector Machine), 決定木, MLP (Multi Layer Perceptron) で学習を行い，出来たモデルを炎上動画分類器とする．作成した分類器の性能評価は４つの指標 (Precision, Recall, Accuracy, F-measure) で判断する予定である．

6. おわりに

本研究では YouTube 動画を対象に，教師あり機械学習である文書分類器を用いて，誹謗中傷コメントで炎上している動画の検出について検討した．

動画のメタデータおよびコメントデータの取得については，動画 API を用いる．API の利用方法について説明した．次に教師あり機械学習に用いるための，正例である炎上動画の選出について説明した．YouTube の動画数は膨大であるため，炎上動画の候補を絞り込むための３つの指標を説明した．動画コメントのベクトル化では，単語頻度，単語の感情度，コメント感情度，NG ワード，および単語の分散表現を用いたコメント文のベクトル化を提案した．学習用の動画のテキストデータと，ベクトル化手法が有れば，

[*2] http://guardman-pro.net/word.html

SVM・決定木・MLP を用いて分類器を作成できる．今後は実際に学習用の正例データの選出を行う予定である．負例候補も適切に選出し，ベクトル化および分類器の性能を網羅的に調べる予定である．最終的には，学習データで作成した炎上検出器を多数の動画データに適用し，埋もれた炎上動画の検出を目指したい．

参考文献

[1] 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).
[2] 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，情報処理学会，pp. B3–3 (2021).
[3] 中野信子：シャーデンフロイデ, Vol. 4, 幻冬舎新書 (2018).
[4] 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).
[5] 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).
[6] Mori, K. and Takeda, F.: Corporate Responses to Internet Flaming: Evidence from Japan, *2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, pp. 359–363 (2019).
[7] Rajapaksha, P., Farahbakhsh, R., Crespi, N. and Defude, B.: Uncovering flaming events on news media in social media, *2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)*, IEEE, pp. 1–8 (2019).
[8] 北研二，津田和彦，獅々堀正幹：情報検索アルゴリズム，共立出版 (2002).
[9] ニコニコ大百科：NG 推奨ワードの一覧，https://dic.nicovideo.jp/a/ng%E6%8E%A8%E5%A5%A8%E3%83%AF%E3%83%BC%E3%83%89%E3%81%AE%E4%B8%80%E8%A6%A7.
[10] Takamura, H., Inui, T. and Okumura, M.: Extracting Semantic Orientations of Words using Spin Model, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)*, pp. 133–140 (2005).
[11] 高村大也：単語感情極性対応表，http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html.
[12] abhishek: autonlp-japanese-sentiment-59363, https://huggingface.co/abhishek/autonlp-japanese-sentiment-59363 (2021).
[13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality, *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, Vol. 2, Curran Associates Inc., pp. 3111–3119 (2013).
[14] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146 (2017).
[15] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C. and Joulin, A.: Advances in pre-training distributed word representations, *arXiv preprint arXiv:1712.09405* (2017).
[16] Le, Q. and Mikolov, T.: Distributed Representations of Sentences and Documents, *Proceedings of the 31st International Conference on Machine*, pp. 1188–1196 (2014).
[17] 秋庭伸也，杉山阿聖， 寺田学，加藤公一：見て試してわかる機械学習アルゴリズムの仕組み機械学習図鑑，翔泳社 (2019).
[18] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830 (2011).