感情とトピックに注目したメディアの報道姿勢の分析

中北　雄大† 　風間　一洋† 　吉田　光男†† 　土方　嘉徳†††

† 和歌山大学システム工学部　〒 640–8510 和歌山県和歌山市栄谷 930
†† 筑波大学ビジネスサイエンス系　〒 112–0012 東京都文京区大塚 3–29–1
††† 関西学院大学商学部　〒 662–8501 兵庫県西宮市上ケ原一番町 1–155
E-mail: †s236194@wakayama-u.ac.jp, kazama@ingrid.org, ††mitsuo@gssm.otsuka.tsukuba.ac.jp,
†††contact@soc-research.org

あらまし　メディアの報道はソーシャルメディアのユーザの意見に強い影響を与え，特に釣り記事やフェイクニュースは実世界の予期せぬ深刻な対立を招くことすらあるが，TV や新聞などのマスメディアに加えて，インターネット上に様々なオルタナティブメディアが乱立する現在，各メディアの報道姿勢を見分けることは容易ではない．本稿では，各メディアのニュース記事と，それに対するソーシャルメディアのユーザの反応を，感情的側面と意味的側面から比較・分析して，メディアの報道姿勢を報道側と読者側の双方から明らかにするとともに，将来的に保守・リベラルに限らない，多彩な分類が可能かを検証する．実際には，メディアのニュースとそれに言及したツイートを収集し，ツイートからユーザが自分の意見を表明している部分のテキストを抽出する．さらに，ML-Ask と LDA の処理結果から感情とトピックに関する特徴ベクトルを作成し，4 種類の方法でメディアをクラスタリングした結果を比較・分析する．

キーワード　ニュース，メディア，ML-Ask，LDA，クラスタリング

1 はじめに

ソーシャルメディア上の各種ニュースメディア（以後，メディア）の報道はソーシャルグラフ上ですみやかに拡散されるだけでなく，ユーザの意見や関連情報が付加されることで情報源としての価値が高まり，強い影響力を持つようになった．同類選択や選択的接触により構築されるソーシャルグラフは隣接ユーザと興味や信念，環境などが類似する傾向が高く，パーソナライズと同様にユーザの嗜好に適合した情報を選択する効果がある．反面，類似する隣接ユーザとの情報交換を繰り返すことで特定の偏った情報や信念が増幅されるエコーチェンバー現象を引き起こす問題も知られており，特に情報拡散力が高い釣り記事やフェイクニュースは実世界の予期せぬ深刻な対立を招くことすらある．この状況を踏まえて，情報受信者が公平な観点から情報を理解するために有効な手段の一つは，情報源であるメディアがどのような報道姿勢なのかを知ることである．しかし，TV や新聞に加えて様々なオルタナティブメディアが乱立する現在，各メディアの報道姿勢を見分けることは容易ではない．

本稿では，各メディアのニュース記事と，それに対するソーシャルメディアのユーザの反応を，感情的側面と意味的側面から比較・分析して，メディアの報道姿勢を報道側と読者側の双方から明らかにする．さらに，将来的に保守・リベラルに限らない，多彩な分類が可能かを検証する．実際には，メディアのニュースとそれに言及したツイートを収集し，ツイートからユーザが自分の意見を表明している部分のテキストを抽出する．さらに，ML-Ask と LDA の処理結果から感情とトピックに関する特徴ベクトルを作成し，4 種類の方法でメディアをクラスタリングした結果を比較・分析する．

2 関連研究

各メディアの報道はそれぞれの立場に基づいて報道する場合があり，そのような報道傾向を分析する研究がある．張らは，記事の書き方を「印象」という評価指標で分析することにより，それぞれの記事から受ける多様な印象を数値として推定し，さらにニュースサイトの報道傾向の時間的推移を視覚的に比較可能な分析手法を提案した [1]．小林らは，新聞読者にイデオロギー的極性化が起きているかを調べるために，13 個の調査データから求めた各新聞の購読者のイデオロギー得点の時系列的変化を分析した [2]．畑中らは，国会議事録における各党の議員の発言を教師データとして政党の判定器を作成して，新聞の社説に適用し，新聞社のイデオロギーを測定した [3]．

一方で，ソーシャルメディアの投稿を感情に注目して分析する研究が行われている．鳥海らは，ソーシャルメディア上の新型コロナウイルスに関する投稿を収集して，情報を発信するユーザの変化とその感情に着目して分析を行った [4]．笹原らは，感情辞書 ML-Ask と，彼らが開発した日本語版の心理学的カテゴリ辞書 J-LIWC，道徳基盤辞書 J-MFD を Twitter の大規模ソーシャルデータに適用し，コロナ禍における転売現象を対象として消費者心理・行動を定量化・分析した [5]．

また，ソーシャルメディアの投稿のコメント部分からメディアの政治的バイアスを分析する研究が存在する．久田らは，トピックモデルを用いて，ニュースに対するユーザのコメントからメディアの政治的バイアスを分析した [6]．その結果，政治的な話題のニュースに対するコメントの場合は，政治的バイアスによって，非政治的な話題のニュースに対するコメントの場合は，ユーザの反応やメディアのコンテンツによりメディアが

分離することを明らかにした．

3　メディアの報道姿勢の分析手法

3.1　感情とトピック

メディアの記者は，読者に伝えたいメッセージを込めてニュース記事を書く．また，ソーシャルメディアのユーザは，Web サイトや自分のタイムラインで各メディアのニュース記事を読み，時にはその記事を拡散または言及するなどの行動を起こす．つまり，メディアの報道姿勢は，読者側に影響を与えるはずであり，ソーシャルメディア上のニュース報道に関して，メディア側とユーザ側の両方に着目すれば，さらに詳細な分析ができると考えられる．

久田らは，ニュース記事に対する Twitter ユーザのコメント部分のトピックを LDA で求めて，メディアの政治的バイアスの分析に用いた [6]．本稿では，特定の話題に限定せずに分析すると共に，ニュース記事のタイトルにも適用する．これは，一般的にニュース記事のタイトルは内容の簡潔・適切な要約であると同時に，記者の意見を反映していると考えられるからである．

本稿では，さらにテキストに表れる感情にも注目する．記者が記事にどのような感情を込めて書いたかは，各メディアの報道姿勢に関係するはずである．例えば，中立・公正な報道を目指すメディアは記事全般で感情を抑えた表現を用いるだろうが，スポーツに関するメディアは試合の勝敗がはっきりわかるようなタイトルにするだろうし，記事の注目度を最優先するメディアは読者の目につきやすい記事のタイトルに誇張した表現を用いる，いわゆる「タイトル詐欺」を多用するかもしれない [7]．また，その記事を読んだユーザは，たとえ感情を抑えて書かれたとしても，記者の意図を理解したり，記事に直接書かれていない現在の状況や他のユーザの反応を理解した上で行動しているはずである．

そこで，各メディアが報道したニュース記事のタイトルと，それに言及したツイートのコメント部分から抽出した 2 種類のテキストデータから，ML-Ask で感情ベクトルを，LDA でトピックベクトルを抽出することで得られる 4 種類の特徴量を用いて，メディアの報道姿勢を分析する．

3.2　テキストの抽出

3.2.1　コメントテキスト

収集したニュース記事の URL を含む言及ツイートから，特にユーザの反応が現れているコメント部分のテキストのみを抽出し，分析に用いる．以降は，このテキストデータをコメントテキストと呼ぶ．

図 1 に，言及ツイートの例を示す．使用するツイートには収集の手がかりとした URL は必ず含まれるが，それ以外にも図 1 において緑の枠で囲われた部分のように，ニュースカテゴリ（【将棋】），タイトル（藤井聡太〜進出），メディア名（産経ニュース），メディアのユーザ名（@Sankei_news），ハッシュタグ（#将棋）などのニュース記事や言及したメディアに関する

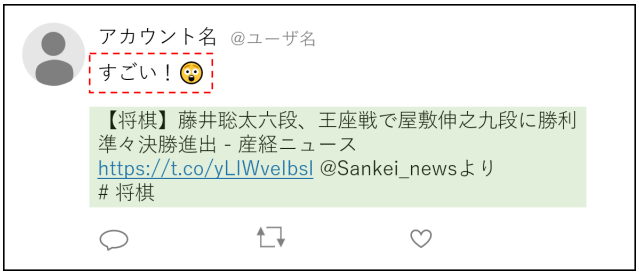


図 1: ニュースに言及しているツイートの例

情報が含まれることが多い．そこで，赤の点線で囲われたユーザのコメント部分だけを以下の手法で抽出する．

（1） テキストを neologdn[1]で正規化する．

（2） テキスト中の改行コードを空白に置換する．

（3） HTML エンコードされた文字（&，<，>，”）を復号化する．

（4） ツイートではニュース記事のタイトルの一部が省略されることがあるために，ツイートテキストとニュース記事のタイトルの最長共通部分文字列が l 文字以上であれば，タイトルとその直後のメディア名を日本語形態素の分割に影響を与えずに無効化するために，その開始位置から URL の直前までを代替文字として用いられている下駄記号「〓」に置換する．

（5） URL，ハッシュタグ，メディアのユーザ名，ニュースカテゴリを正規表現を用いて削除する．なおニュースカテゴリは，隅カッコ（【 】）や半角角括弧（[ ]）とそれで囲まれた文字列，メディアのユーザ名は URL 直後の “@” から始まる文字列（ただし，それに続く「から」や「より」なども含む）である．

3.2.2　ニュースタイトルテキスト

ツイートから言及されたニュース記事から，ニュース本文の注目部分を要約しているニュース記事のタイトルに以下の処理を行ってから分析に用いる．以降は，このテキストデータをニュースタイトルテキストと呼ぶ．

（1） テキストを neologdn で正規化する．

（2） テキスト中の HTML エンコードされた文字を復号化する．

3.3　特徴ベクトルの作成

3.3.1　感情ベクトル

抽出したニュースタイトルテキストとコメントテキストから，ML-Ask を用いて感情ベクトルを作成する．ML-Ask（eMotive eLement and Expression Ana-lysis system）[8,9] は，中村の 10 種類の感情分類 [10] に基づいた辞書を用いて，各感情においてマッチした表現の数から感情成分を推定するシステムであり，CVS（Contexual Valence Shifters）[11] に基づいて否定表現も考慮できる．

今回の分析には，新たに単語が追加された Python 版の実装である pymlask[2]を用いる．また，pymlask は内部で MeCab で日本語形態素解析を行うが，その辞書には mecab-ipadic-NEologd

1：https://github.com/ikegami-yukino/neologdn
2：https://github.com/ikegami-yukino/pymlask

を用いる．pymlask は，感情とその強さに加えて，感情極性，活性度なども推定する．

本稿では，各メディアごとに，ML-Ask で推定した感情の出現頻度から求めた 10 感情の割合（0~1）と，感情を表す傾向を考慮するために，感情が推定されたテキストの割合（感情推定率）を式 (1) で正規化した値（0~1）を用いて，11 次元の感情ベクトルを作成する．

$$x'_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{1}$$

ただし，x_i はメディア i の感情推定率を表す．

3.3.2 トピックベクトル

抽出したニュースタイトルテキストとコメントテキストから，トピックモデルを用いてトピックベクトルを作成する．テキストに LDA（Latent Dirichlet Allocation）[12] を用いて，各メディアのテキストに対するトピックの平均推定確率を特徴ベクトルとする．この手法は，久田らの手法 [6] に基づくが，本稿では特定の話題に制限しない点が異なる．

まず，テキストに対して，次の前処理を行う．

- 長音と中黒以外の記号を半角スペースに置換する．
- 絵文字を削除する．

次に，MeCab と mecab-ipadic-NEologd を用いて日本語形態素解析を行って，一般・固有名詞・サ変接続・形容動詞語幹の名詞の単語を抽出し，さらに出現文書数が 100 未満の単語とストップワードを除去し，単語数が 0 より大きいテキストだけを用いる．なお，日本語形態素辞書として mecab-ipadic-NEologd を用いる．ストップワードは，以下に示す単語である．

- ひらがなのみの単語
- 「笑」,「w」や「RT」
- 長音（ー）と中黒（・）の記号
- ニュースメディア名

次に，LDA を用いて k 個のトピックに分類する．トピック数 k は，トピック数を 10 から 20 まで変化させた時のトピック分類結果を目視で評価し，最適と思われる値を選択する．この評価はデータの収集期間に発生した出来事や事件のトピック分離の妥当性に注目して行う [3]．

各メディアごとにテキストに対するトピックの推定確率を求めた後に平均を求め，k 次元のトピックベクトルを作成する．

3.4 分析方法

本稿では，ニュースタイトルテキストとコメントテキストの感情推定結果とトピック抽出結果，そして感情ベクトルとトピックベクトルを用いたクラスタリング結果を分析する．

3.4.1 感情推定結果の分析方法

まず，各メディアが記事のタイトルにどの程度感情を込めるのか，またはユーザが記事に関してどの程度感情的なのかを，メディアごとに感情推定率を求め，それを運営会社の業種（新聞社，放送局，通信社，情報サービス，出版社，Web メディア）別に分類した後に，ニュースタイトルテキストの感情推定率の降順に並び変えて分析する．もちろん，ML-Ask は機械学習ではなく辞書ベースの手法であるために，必然的に感情推定率は低めになるはずだが，メディア間の相対的な評価は可能だと考えられる．また，業種別に分類することで，新聞社や雑誌社などの業種の傾向の差異を調べることができるはずである．

さらに，各メディアが記事にどのような感情を込めるのか，またはユーザが記事に対してどのような感情をもつのかを，ML-Ask で推定した 10 感情の割合を可視化することで比較する．なお，各メディアが後述するクラスタリングで得られたどのクラスタに所属するかを明らかにするために，クラスタと同じ色で着色する．

3.4.2 トピック抽出結果の分析方法

LDA でニュースタイトルテキストとコメントテキストからどのようなトピックが抽出されたのかを確認するために，各トピックの term-score [13] の上位 5 件の単語と，後述するクラスタリングで得られたクラスタの平均推定確率を用いて分析する．term-score は，TF-IDF のように特定のトピックに出現する単語ほど高く評価するための指標であり，トピック総数を K，トピック k の単語 w の出現確率を $\hat{\beta}_{k,w}$ として，式 (2) で計算する．

$$\text{term-score}_{k,w} = \hat{\beta}_{k,w} \log \left(\frac{\hat{\beta}_{k,w}}{\left(\prod_{j=1}^{K} \hat{\beta}_{j,w} \right)^{\frac{1}{K}}} \right) \tag{2}$$

3.4.3 クラスタリング結果の分析方法

各メディアのニュース記事とそれに対するソーシャルメディアのユーザの反応という観点に加えて，さらに感情的側面と意味的側面という観点から，どのようにメディアが分類されるかを比較・分析するために，ニュースタイトルテキストとコメントテキストに対して，それぞれ感情ベクトルとトピックベクトルを作成し，データ間の距離尺度としてユークリッド距離，クラスタ間の距離尺度として Ward 法を用いて階層的クラスタリングを行う．さらに，得られたデンドログラムに対して，人手で閾値を設定して，分割されたクラスタごとに異なる色で可視化する．

4 分　析

4.1 データセット

ニュースデータセットとして，統合型メタ検索エンジン Ceek.jp[4]が扱うニュース記事のうち，ツイートデータセットのツイートに含まれる URL のニュース記事 137,466 件を利用した．

ツイートデータセットとして，Twitter API で取得した 2018 年 5 月 1 日から 31 日の 1 か月間のツイートのうち，ニュース記事の URL を含む 326,906 件のツイートを利用した．

収集されたデータのうち，Twitter 上でニュース記事が言及されたメディア数は 491 であった．ただし，感情が推定できたツイート数が少ないメディアは感情推定結果の信頼性や重要性

3：https://www.nippon.com/ja/features/q201805/

4：http://www.ceek.jp/

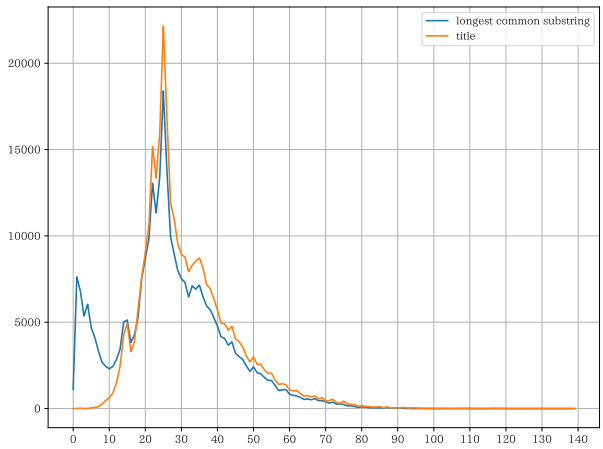


図 2: タイトル・最長共通部分文字列長とツイート数の分布

が低いと考えて，感情が推定できたツイート数が 100 件以上あった 43 のメディアを対象とした．

感情ベクトルの作成には，これらのメディアの 10,911 件のニュースタイトルテキストと，それらに言及した 21,214 件のコメントテキストを用いた．トピックベクトルの作成には，69,642 件のニュースタイトルテキストと 85,104 件のコメントテキストを用いた．

4.2 最長共通部分文字列長の閾値の決定

3.2.1 の（4）において，最長共通部分文字列長の閾値 l は，次のように決定した．図 2 に，ニュース記事のタイトルの長さと，タイトルとツイートテキストの最長共通部分文字列長の分布を示す．横軸が文字列長，縦軸がツイート数または記事タイトル数である．ただし，ニュースタイトルテキスト数はツイートから言及された数を重複カウントしている．最長共通部分文字列のツイート数（青色）は 1 文字から 10 文字までツイート数は減少し，この範囲では文字あるいは単語単位で一致しているだけで，タイトルが一致しているわけではないことがわかる．さらに，11 文字からタイトルと最長共通部分文字列のツイート数が共に増加傾向に転じて，24 文字でピークとなり減少傾向になることから，11 文字以上で正しく検出されているとみなして $l = 11$ とした．

4.3 感情推定率に関する分析

まず，メディアの感情推定率に関して分析した．各メディアのニュースタイトルテキストまたはコメントテキストの感情推定率を表 1 に示す．なお，運営会社の業種は，新聞（新聞社），放送（放送局），通信（通信社），情報（情報サービス），出版（出版社），Web（Web メディア）と表記して，業種ごとに記事の推定割合の降順に列挙した．

対象とするメディアの 81,461 件のニュースタイトルテキストに ML-Ask を適用した結果，感情が推定できた記事タイトル数は 10,911 件（13.4%）であった．各メディアのニュースタイトルテキストの推定率を見ると，新聞社では夕刊紙，スポーツ紙などの専門誌は高いが，一般紙は放送局や通信社と同様に低い傾向があった．これは，基本的に中立・公正な報道を目指しているからと思われる．出版社と Web メディアを見ると，総合ニュースの場合は高く，ファッションや IT などの専門ニュースの場合には低い傾向があった．なお，一番推定率が高かったリテラは月刊誌「噂の真相」の関係者が立ち上げた Web メディアであり，閲覧して調べたところ，「スキャンダル」というニュースカテゴリが存在し，多くの人の興味を引くようなニュースを積極的に報道する傾向があった．

次に，326,906 件のコメントテキストに ML-Ask を適用した結果，感情が推定できたツイート数は 26,070 件（8.0%）だった．つまり，ニュースタイトルテキストよりも推定率が低くなることがわかった．この理由として，ニュース記事への URL やタイトルは記入しても，自分自身のコメントを記入せずに投稿するユーザが多いことが考えられる．推定率が 0.13 より高いメディアは，新聞・放送・通信・情報では毎日新聞，朝日新聞，出版・Web ではプレジデントオンライン，ダイヤモンド・オンライン，弁護士ドットコム，ハフポストなどのリベラル系メディアであった．前者はニュースタイトルテキストの推定率が低い，つまりタイトルは中立でも内容でユーザの意見を引き出していると考えられる．なお，一番推定率が高かったプレジデントオンラインとハフポストを閲覧して調べたところ，前者のタイトルは人目を引くような誇張した表現で，後者はリベラル傾向が強い傾向にあった．

4.4 感情の割合に関する分析

次に，各メディアの感情の割合について分析した．各メディアのテキストによる感情の割合を図 3 と図 4 に示す．グラフの色は，それぞれ図 5(a) と図 5(b) のクラスタの色に一致させた．

ニュースタイトルテキストの感情の割合（図 3）を見ると，全般的に「厭」と「喜」の感情の割合が高いが，それらに加えてリテラや BBC のように「怒」の感情の割合が高いメディア，ナタリーや VOGUE JAPAN のように「好」の感情の割合が高いメディアが見られた．この結果から，メディア側は意図的にタイトルで感情を抑えたり，逆にさまざまな感情を込めるなど，いくつかのパターンがあることがわかる．

コメントテキストの感情の割合（図 4）を見ると，全般的に「厭」か「喜」の 2 つの感情に集中しており，朝日新聞や NHK のように「厭」の感情の割合の方が高いメディアや，ナタリーや VOGUE JAPAN のように「喜」の感情の割合の方が高いメディア，スポニチやサンスポのように「厭」と「喜」の感情の割合がある程度高いメディアが見られた．この結果から，ソーシャルメディアのユーザが記事の URL やタイトルに加えて，わざわざコメントを書くのは，「厭」か「喜」のように大きく感情を動かされた場合に限られることがわかる．

4.5 クラスタリング結果の分析

次に，ニュースタイトルテキストとコメントテキストから抽出した感情ベクトルとトピックベクトルを用いた，4 種類のクラスタリング結果について比較・分析した．それぞれのクラスタリング結果を図 5 に示す．閾値は，図 5(a) は 0.6，図 5(b) は 0.5，図 5(c) は 0.05，図 5(d) は 0.04 とし，得られるクラスタ数

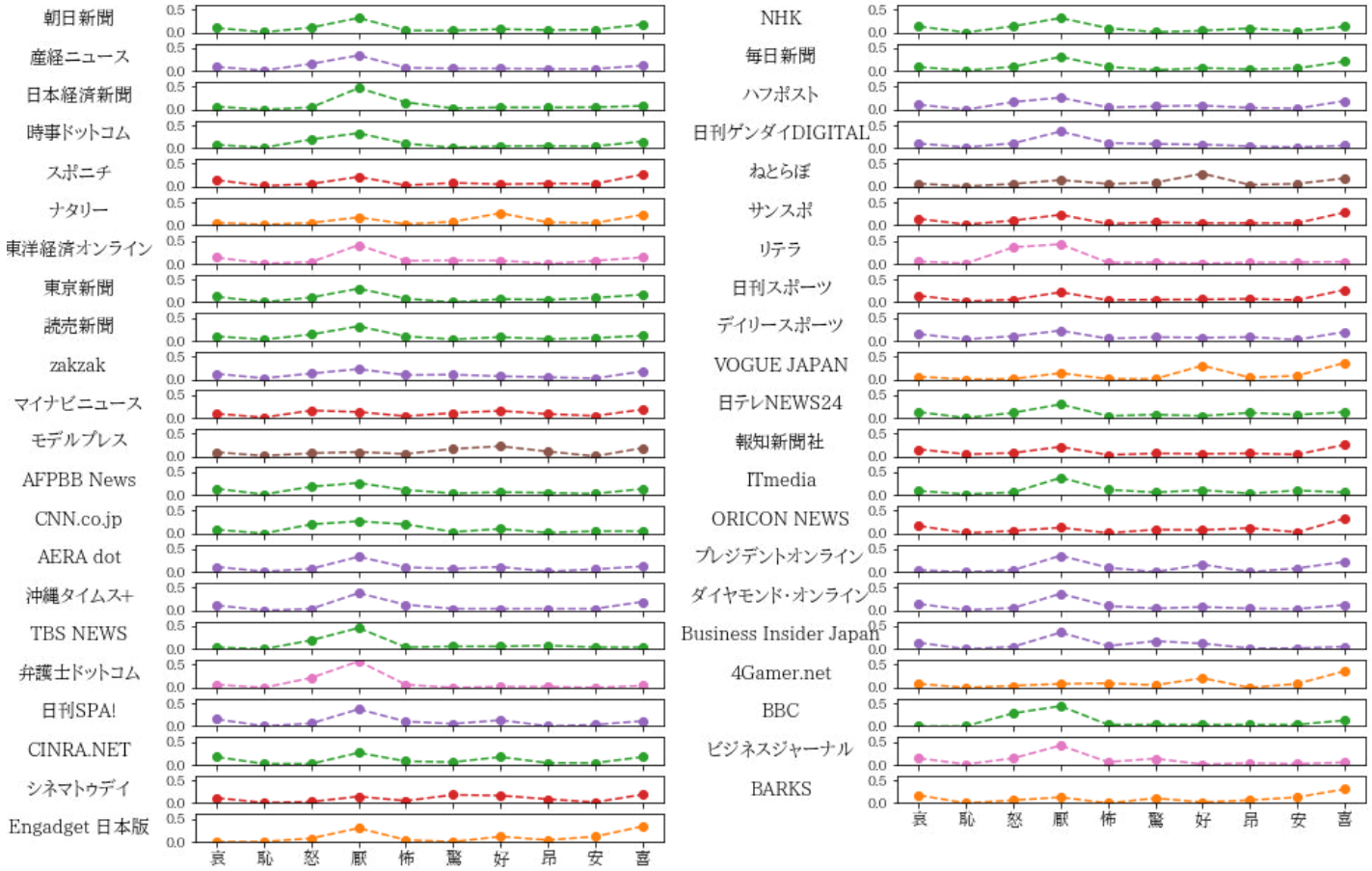


図 3: ニュースタイトルテキストの感情の割合

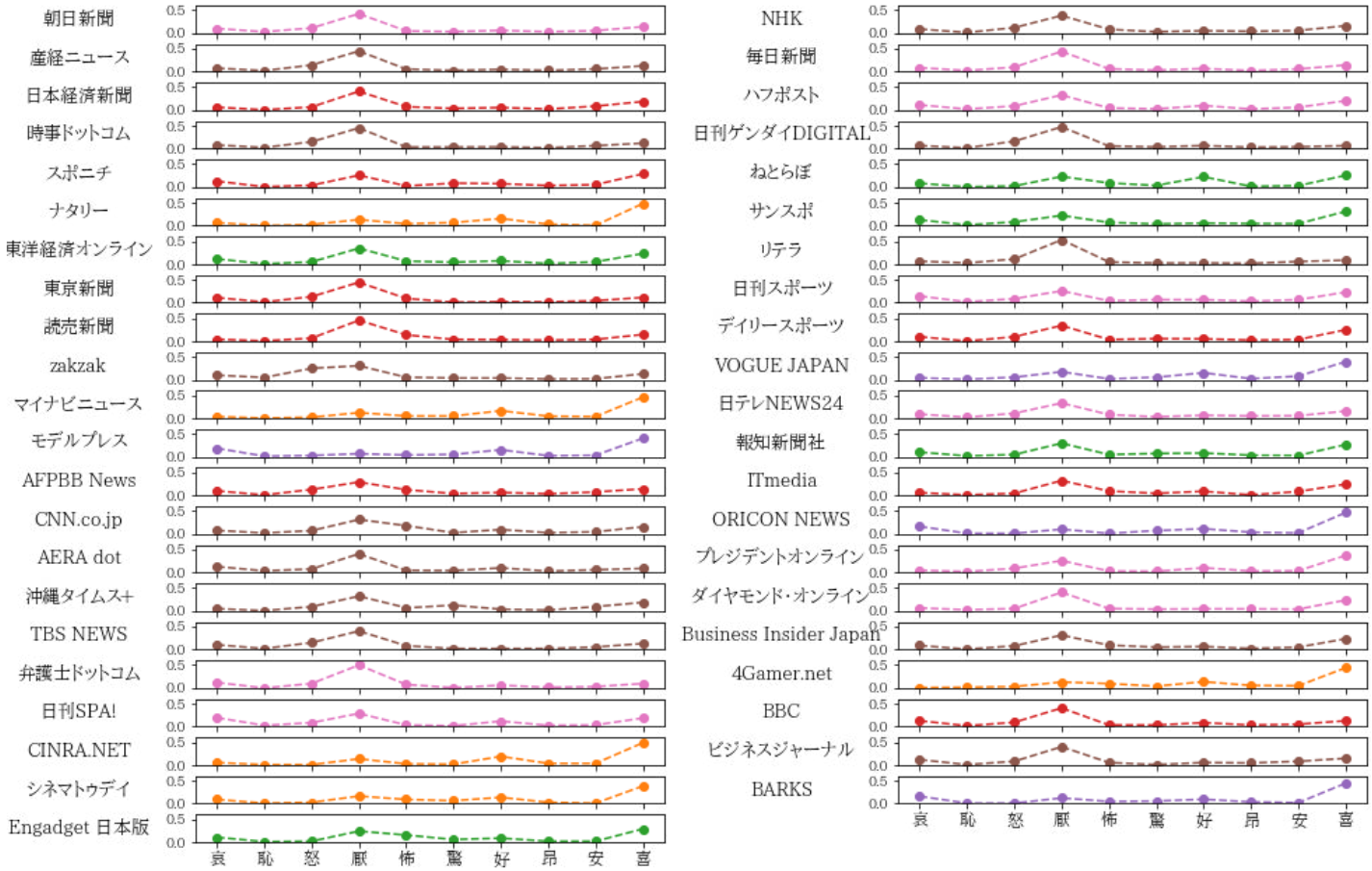


図 4: コメントテキストの感情の割合

表 1: 各メディアの感情推定率

メディア名	業種	タイトル	コメント
zakzak	新聞	0.213	0.094
デイリースポーツ	新聞	0.208	0.049
スポニチ	新聞	0.194	0.049
日刊スポーツ	新聞	0.165	0.114
サンスポ	新聞	0.158	0.084
産経ニュース	新聞	0.154	0.101
沖縄タイムス＋	新聞	0.154	0.099
報知新聞社	新聞	0.135	0.070
東京新聞	新聞	0.115	0.067
毎日新聞	新聞	0.112	0.136
朝日新聞	新聞	0.108	0.138
日本経済新聞	新聞	0.090	0.072
読売新聞	新聞	0.079	0.049
CNN.co.jp	放送	0.126	0.095
日テレ NEWS24	放送	0.097	0.125
BBC	放送	0.081	0.071
TBS NEWS	放送	0.078	0.112
NHK	放送	0.051	0.098
時事ドットコム	通信	0.117	0.112
AFPBB News	通信	0.113	0.047
ORICON NEWS	情報	0.194	0.117
東洋経済オンライン	出版	0.312	0.085
日刊 SPA!	出版	0.258	0.118
AERA dot	出版	0.230	0.097
プレジデントオンライン	出版	0.229	0.148
日刊ゲンダイ DIGITAL	出版	0.226	0.104
ダイヤモンド・オンライン	出版	0.171	0.139
VOGUE JAPAN	出版	0.131	0.124
リテラ	Web	0.351	0.090
モデルプレス	Web	0.317	0.089
ビジネスジャーナル	Web	0.316	0.092
ねとらぼ	Web	0.274	0.090
弁護士ドットコム	Web	0.240	0.133
ハフポスト	Web	0.215	0.148
マイナビニュース	Web	0.172	0.044
Business Insider Japan	Web	0.164	0.106
シネマトゥデイ	Web	0.149	0.051
ナタリー	Web	0.097	0.036
ITmedia	Web	0.090	0.045
BARKS	Web	0.082	0.095
Engadget 日本版	Web	0.068	0.075
CINRA.NET	Web	0.064	0.062
4Gamer.net	Web	0.056	0.048

が 5~7 となるように人手で設定した．

ニュースタイトルテキストの感情ベクトルによるクラスタリング結果（図 5(a)）では，業種でクラスタが分かれる傾向があった．例えば，桃・茶・橙色のクラスタは Web メディア，紫色は出版社と Web メディア，赤色は新聞社のスポーツ誌，緑色は新聞社の一般紙と放送局がほとんどを占めていた．この理由は，新聞社の一般紙や放送局は公正な報道を目出すために記事のタイトルに感情を出来る限り出さない傾向があるのに対して，スポーツ紙はタイトルで試合の勝敗などをアピールし，また出版社や Web メディアは多くの人に記事を読んでもらうようなタイトルにするなど，業種ごとのタイトルの作成方針の違いによりクラスタが形成されたと考えられる．

コメントテキストの感情ベクトルによるクラスタリング結果（図 5(b)）では，紫色のクラスタには音楽やファッション，橙色のクラスタにはゲームや映画などのエンターテイメント系のメディアが集まっているが，その他のクラスタでは新聞社が桃色，茶色，赤色，緑色のクラスタに分散するなど業種別に分かれる傾向がなくなり，明確な傾向が見られなかった．これは，ニュースタイトルテキストの感情の割合では多彩なパターンが見れたのに対して，コメントテキストでは「厭」か「喜」がほとんどだったことから，感情推定率と「厭」と「喜」の感情の極性が近いメディアでクラスタが形成されたと考えられる．

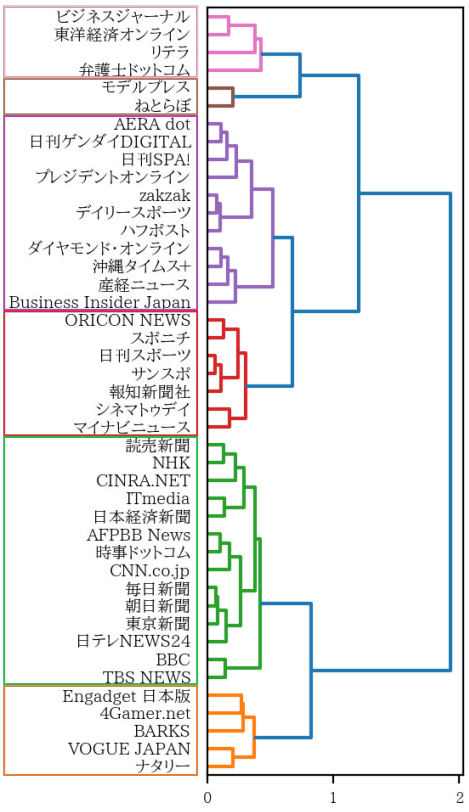
ニュースタイトルテキストのトピックベクトルによるクラスタリング結果（図 5(c)）では，橙・緑・赤色のクラスタは政治や事件，紫色のクラスタはビジネスや経済，茶色のクラスタはスポーツ，桃・灰色のクラスタのメディアはエンターテイメントなど，主に発信する記事の分野の類似性に基づいてクラスタが形成される傾向が見られた．タイトルには扱うニュースの内容を的確に示す特徴語が含まれる確率が高いことから，そのトピックでクラスタリングすればこのような結果になるのは妥当だと考えられる．

コメントテキストのトピックベクトルの結果（図 5(d)）では，新聞社のスポーツ紙が橙色のクラスタが形成されているが，新聞社の一般紙は茶・紫・赤・緑色のクラスタに分散し，出版社や Web メディアと混在しており，これだけではどのような根拠で分類されたのかが明確ではない．

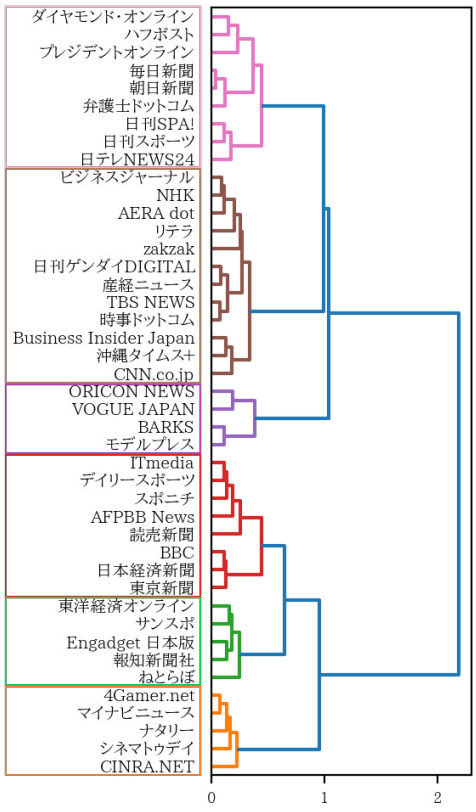
4.6 トピックとクラスタとの関係の分析

最後に，ニュースタイトルテキストとコメントテキストから抽出した各トピックの代表的な単語と，クラスタごとの平均推定確率から，トピックベクトルを用いた場合のクラスタリングの特徴について分析した．トピック数 k は，各トピックの内容が明確になり，トピックが過度に分割されすぎないように人手で選択した．

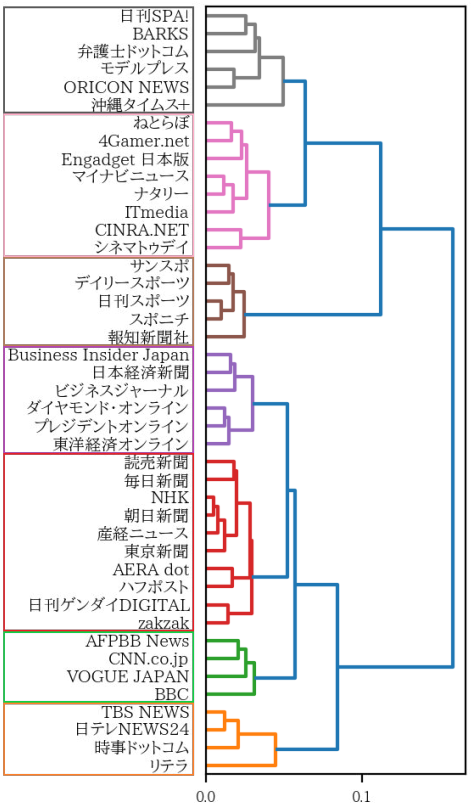
表 2 はニュースタイトルテキストから抽出したトピックに関する単語と，クラスタごとの平均推定確率を示す．トピック数は $k = 12$ とした．各トピックの内容を見ると，記事の分野によって分かれている傾向があり，トピック 4 は企業犯罪，トピック 5 と 8 は国内スポーツ，トピック 7 は芸能，トピック 9 は外交，トピック 12 は海外スポーツに関する内容であり，それぞれ



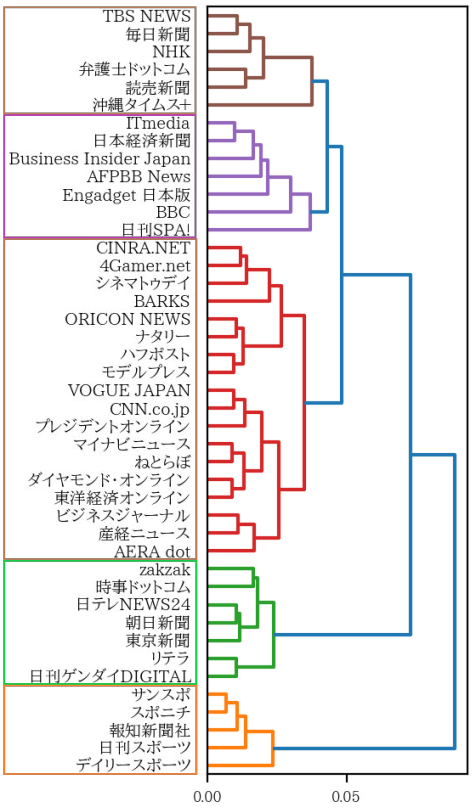
(a) 感情ベクトル（ニュースタイトルテキスト）



(b) 感情ベクトル（コメントテキスト）



(c) トピックベクトル（ニュースタイトルテキスト）



(d) トピックベクトル（コメントテキスト）

図 5: メディアのクラスタリング結果

表 2: ニュースタイトルテキストのトピックとクラスタの関係

	単語	クラスタ						
		橙	緑	赤	紫	茶	桃	灰
1	広島 開始 プロ 超特急 大阪	7.8	7.9	8.3	8.5	7.8	8.6	8.6
2	発表 W 杯 代表 人気 批判	6.9	6.8	6.9	7.1	7.3	7.0	7.5
3	死亡 2 人 記念 女子 声	6.3	7.1	6.5	6.4	6.3	6.2	6.5
4	逮捕 男 決定 社長 容疑	9.5	8.6	9.2	10.0	8.3	8.9	9.3
5	巨人 阪神 出演 日本ハム V	7.7	8.3	8.8	8.4	12.2	9.4	9.2
6	発売 連続 安打 ライブ 写真	8.2	8.9	8.8	8.8	9.0	10.0	8.9
7	挑戦 舞台 企業 TOKIO 提訴	6.3	7.0	6.8	7.2	6.8	7.4	7.3
8	監督 日本 日大 選手 会見	8.4	7.3	8.7	8.2	10.2	8.1	7.8
9	世界 北朝鮮 首相 韓国 安倍首相	13.6	10.9	11.2	9.9	8.2	8.4	8.7
10	公開 中国 動画 イベント 人	7.9	8.7	8.2	8.4	8.1	9.3	9.1
11	女性 男性 調査 疑い セクハラ	8.5	8.2	8.1	7.7	7.6	7.4	8.9
12	米 開催 登場 大谷 映画	9.0	10.2	8.6	9.3	8.1	9.4	8.1

表 3: コメントテキストのトピックとクラスタの関係

	単語	クラスタ				
		橙	緑	赤	紫	茶
1	自分 森友 言葉 内容 知事	7.1	7.5	7.5	7.4	7.1
2	発言 仕事 写真 名前 動画	5.7	5.6	5.8	6.0	5.4
3	記事 男性 逮捕 掲載 会社	7.0	6.8	7.8	7.4	7.2
4	北朝鮮 国 期待 アベ 公開	6.9	7.1	7.3	7.4	6.9
5	国民 文書 安倍首相 批判 財務省	5.9	7.0	6.2	6.0	6.2
6	女性 中国 セクハラ 確認 感じ	6.4	6.1	6.3	6.3	6.6
7	会見 明らか 社会 指示 判断	5.5	5.6	5.4	6.0	5.8
8	日大 選手 監督 話 アメフット	11.4	8.2	8.8	8.2	9.0
9	世界 アメリカ 可能性 理事長 米国	6.6	7.1	7.2	7.3	7.1
10	ニュース 安倍 米 大学 対応	7.7	7.7	7.7	8.3	8.1
11	発表 会談 男 韓国 コーチ	6.6	6.2	6.4	6.9	6.6
12	野党 自民党 政府 議員 国会	7.2	8.8	7.5	7.4	8.4
13	必要 加計 嘘 首相 加計学園	7.6	8.6	7.7	7.4	7.8
14	日本 人 事件 情報 記者	8.5	7.8	8.5	8.0	7.8

紫色，茶色，桃色，橙色，緑色のクラスタとの関係が強く，記事の分野の類似性に基づいてクラスタリングされることが確認できた．

表3はコメントテキストから抽出したトピックに関する単語と，クラスタごとの平均推定確率を示す．トピック数は $k = 14$ とした．各トピックの内容を見ると，コメントテキストよりもトピックの範囲が限られるとともに，トピック 1 は森友学園問題，トピック 5 は財務省文書改竄問題，トピック 6 は財務事務次官セクハラ発言問題，トピック 8 は日大アメフト部反則タックル問題，トピック 13 は加計学園問題のように，犯罪や事件が推測されるトピックが抽出されていると思われる．これは，ツイートにニュース記事の URL やタイトルなどの情報以外に，自分の意見を示すコメントを書く場合には，犯罪や事件のように強く心が動かされる場合に限られるからだと考えられる．それゆえに，コメントテキストから抽出したトピックベクトルでクラスタリングした場合には，業種はあまり関係なく，報道する犯罪や事件の傾向が類似しているメディアでクラスタが形成されたと考えられる．さらに，茶色のクラスタでは，保守派の読売新聞，中道派の NHK，リベラル派の毎日新聞と沖縄タイムス＋が混在していることから，政治的なバイアスで分かれているわけではないことがわかった．

5 おわりに

本稿では，メディアの報道姿勢の違いを，各メディアのニュース記事のタイトルとそれに言及したユーザのコメントを ML-Ask と LDA で処理してそれぞれに感情ベクトルとトピックベクトルを作成し，それらの 4 種類の組み合わせで分析した．その結果，感情の場合は，タイトルにはメディアの業種別の報道姿勢の違いが，コメントにはメディアのメッセージの極性（「厭」と「喜」）の違いと大きさが反映され，トピックの場合は，タイトルにはメディアのトピックの違いが，コメントにはメディアが注目する犯罪や問題が反映されることがわかった．

今後は，今回用いた 4 種類の手法を統合してメディアの政治的なバイアスの推定を試みると共に，特定のニュースの話題に限定してメディアを分析する予定である．

謝　辞

本研究は JSPS 科研費 21H03557 の助成を受けた．

文　献

[1] 張建偉, 河合由起子, 熊本忠彦, 白石優旗, 田中克己. 多様な印象に基づくニュースサイト報道傾向分析システム. 知能と情報, Vol. 25, No. 1, pp. 568–582, 2013.

[2] 小林哲郎, 竹本圭祐. 新聞読者は極性化しているか. 日本世論調査協会報「よろん」, Vol. 117, pp. 22–26, 2016.

[3] 畑中允宏, 村田真樹, 掛谷英紀. 新聞社説・国会議事録に基づく言論のイデオロギー別分類. 言語処理学会第 15 回年次大会 (NLP2009), pp. 408–411. 言語処理学会, 2009.

[4] 鳥海不二夫, 榊剛史, 吉田光男. ソーシャルメディアを用いた新型コロナ禍における感情変化の分析. 人工知能学会論文誌, Vol. 35, No. 4, pp. F–K45_1–7, 2020.

[5] 笹原和俊, 奥田慎平, 五十嵐祐. テキストマイニングによるコロナ禍の消費者心理・行動の定量化. 第 35 回人工知能学会全国大会 (JSAI2021), 1D3-OS-3b-04, 2021.

[6] 久田祥平, 村山太一, 矢田竣太郎, 若宮翔子, 荒牧英治. SNS コメントを用いたニュースメディアバイアスの分析. 第 35 回人工知能学会全国大会 (JSAI2021), 1D2-OS-3a-04, 2021.

[7] 宮田健.「タイトル詐欺」なぜ横行？ 煽り記事に“釣られない”ための心構え. `https://www.itmedia.co.jp/news/articles/1801/20/news012.html`, 2018.

[8] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka, and Kenji Araki. A system for affect analysis of utterances in Japanese supported with web mining. 知能と情報, Vol. 21, No. 2, pp. 194–213, 2009.

[9] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Fumito Masui. ML-Ask: Open source affect analysis software for textual input in Japanese. *Journal of Open Research Software, Journal of Open Research Software*, Vol. 5, , 2017.

[10] 中村明. 感情表現辞典. 東京堂出版, 1993.

[11] Livia Polanyi and Annie Zaenen. Contextual valence shifters. In *Computing attitude and affect in text: theory and applications*, pp. 1–10. Springer, 2006.

[12] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[13] A. Srivastava and M. Sahami, editors. *Text Mining: Theory and Applications*, chapter TOPIC MODELS. Taylor and Francis, 2009.