ソーシャルメディアで言及されたニュースのタイトルに関する分析

関本　健臣† 　関　 喜史†† 　吉田　光男† 　梅村　恭司†

† 豊橋技術科学大学 情報・知能工学科　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1
†† 株式会社 Gunosy　〒 106–6125 東京都港区六本木六丁目 10 番 1 号六本木ヒルズ森タワー
E-mail: †{sekimoto.kenshin.lo,umemura}@tut.jp, ††yoshifumi.seki@gunosy.com, †††yoshida@cs.tut.ac.jp

あらまし　インターネット及び情報端末が普及し，ソーシャルメディア上で自身の生活に関わる情報のほかに，ニュースなどの情報を「シェア」する行為も一般的になりつつある．本稿では，Twitter においてリツイートされるニュースは「重要」なもの，いいねされるニュースは「おもしろい」もののように，それぞれ異なる特徴をもつものであると仮定した上で，リツイートされやすいニュースといいねされやすいニュースに分かれる要因を分析した結果を報告する．

キーワード　SNS, Twitter, ソーシャルメディア, ニュース, ニュースタイトル, 特徴語

1 はじめに

インターネット及び情報端末の普及した現代において，ソーシャルメディアもまた普及が進んでいる．これに付随し，Twitter[1]，Facebook [2]，Instagram[3] をはじめとするソーシャルメディア上での情報の「シェア」行為はソーシャルメディアユーザの多くが体験しており [1]，今や自身の生活に関わる情報のほかに，ニュースなどの情報を「シェア」する行為も一般的になりつつある．代表的なソーシャルメディアである Twitter には機能として，あるユーザが投稿した書き込み（ツイート）に対してのリツイート（以下 RT）といいね（以下 fav）が存在し，この機能によりユーザは情報のシェア行為を簡便に行うことが可能である．まず RT とは，あるツイートを自身のフォロワーに対して共有する機能である．これはフォロワーのタイムライン上に，共有したいツイートをコピーして表示させることとほぼ同義であり，そのツイートに対する強い拡散・共有行為であると言える．他方 fav とは，ツイートに対する好意的な気持ちを示す機能である．fav したツイートは記録され， fav したユーザのアカウントのいいね欄にて全て確認できる．このため，fav はツイートの公開お気に入り機能であると言える．これらの 2 つの機能が存在するため，平常時ではユーザが主に自身の状況や雑記などをつぶやくマイクロブログとして使用されている Twitter が，重大な出来事が発生した場合には情報拡散用のツールとしても使われる [2]．このことから，ツイートが RT された回数（RT 数）と fav された回数（fav 数）の 2 つの数値は，ツイートの拡散具合を示す指標としてよく用いられる．

ソーシャルメディア上での投稿に対するシェア行為について，投稿の要素に着目した研究では，Twitter への投稿（ツイート）内に URL やそのツイートをラベリングするために使用される「ハッシュタグ」が含まれていると，ユーザに RT されやすく，対して，ツイートしたユーザの過去のツイート数は RT のされやすさにほとんど影響を与えていないことが明らかになっている [3]．また，投稿したユーザとその投稿をシェアしたユーザとの，コミュニティの違いに着目して拡散・共有行為の分析した研究 [4] では，[3] について明らかになった要素と比較し，コミュニティの違いが拡散規模の予測に有用であることを明らかにしている．その他にも，情報価値の低いツイートが RT される経路についての分析 [5] や，ツイート本文がネガティブかポジティブであるかについて着目した RT の分析 [6] がなされている．しかし，これらの研究ではツイートのみの分析に留まっており，リンク先のコンテンツの影響は明らかになっていない．我々はニュースを閲覧した後のユーザに付随する行動に興味を持っており，本研究では，ツイートのリンク先のコンテンツとしてニュースに着目し，ニュースのどのような要素が影響してユーザによるシェアに繋がるのかを解明したいと考えている．

ソーシャルメディア上のニュースに関する投稿について，コメント数やお気に入り数に着目し，これらに影響を与えているニュースのコンテンツについて，複数のソーシャルメディアを横断した分析 [7] [8] が行われている．また Twitter において，どのようなユーザがどのようなニュースを共有･拡散するのかについての分析もある [9]．この他にも，ユーザの目を引くニュースの説明文の自動生成を目指した研究があり，ツイート内の画像の有無，投稿時間やいくつかのキーワードの記述が有効である可能性が明らかになっている [10]．しかし，これらの研究では fav などのお気に入り行為または RT などの共有拡散行為のどちらかにしか焦点を当てない，もしくはその 2 つを区別せずに分析している，という問題がある．本研究では，Twitter において RT されるニュースは「重要」なもの，fav されるニュースは「おもしろい」もののように，それぞれ異なる特徴をもつものであると仮定した上で，RT と fav との両方に着目し，RT されやすいニュースと fav されやすいニュースに分かれる要因を分析する．なお， 2017 年 3 月に仕様が変わり，一部の fav については，RT 同様にフォロワーのタイムラインに表示されるようになった．しかし，RT とは異なり，全てがフォロワーのタイムラインに表示されるわけではなく，さらに，これまでと同様に自身のアカウントのいいね欄にて一覧で確認できるた

1 : https://twitter.com/ (accessed 2019–12–24)
2 : https://ja-jp.facebook.com/ (accessed 2019–12–24)
3 : https://www.instagram.com/ (accessed 2019–12–24)

め，仕様は変更されたものの，ユーザの使用目的にほとんど変化はないものとして，共有・拡散行為の RT とは別のものと考えて，分析を進める．

本論文では，RT や fav に影響を与える要因として，ニュースを閲覧する際に最初に見るであろうニュースのタイトルに着目して，カテゴリ毎にニュースタイトルに現れる単語を分析する．まず，そもそもツイート本文を分析した方がいいのではないか，実際にカテゴリ毎にニュースの RT や fav のされやすさに違いがあるのかといった疑問に答えるために，事前分析を行う．事前分析の結果，ニュース記事の URL を含むツイートの内容について，大半のツイートが，言及しているニュースタイトルに現れる単語を 8 割以上カバーしていること，また，カテゴリ毎に RT・fav のされやすさが異なる傾向があることが確認できた．これらの結果をもとに，次の 3 つに焦点を当て，分析した．

- 公式アカウントとそれ以外のアカウントとの RT・fav されやすさに差はあるか？
- カテゴリによってニュースタイトルは異なるか？
- fav に比べ RT が多い，RT に比べ fav が多いニュース間においてニュースタイトルに現れる特徴語に差はあるか？

分析の結果，まず，公式アカウントとそれ以外のアカウントの RT や fav のされやすさにほとんど違いがないこと，ニュースタイトルに現れる単語の出現傾向がカテゴリによって異なることが明らかになった．そして，fav に比べ RT が多いニュースと RT に比べ fav が多いニュースを比較すると，ニュースタイトルに現れる特徴語に差があり，さらにカテゴリによってもそれぞれ異なることが示された．

第 2 章では使用したニュースデータ及びツイートデータについての説明と，これら 2 つのデータの対応付けについて述べる．第 3 章では事前分析として，ツイート本文でのニュースタイトルのカバー率と，ニュースの RT・fav されやすさがカテゴリ毎に異なるかについて分析し，その結果について述べる．第 4 章では 3 章での分析結果を踏まえ，先述した 3 つの焦点について分析し，その結果について述べる．第 5 章では本研究のまとめについて述べる．

2 使用データについて

2.1 ニュースデータについて

本研究で使用するニュースデータは，ニュースポータルサイト Ceek.jp News[4] にて 2017 年 1 月 1 日から 2017 年 12 月 31 日の 12ヵ月間に収集された 355,086 件のデータである．ニュースデータからはニュースカテゴリや日時，ニュースタイトルなどの情報が取得できる．本研究では，ニュース URL，ニュースタイトル，ニュースカテゴリの 3 つの情報を使用する．ニュースカテゴリは経済，エンタメ，その他，IT，地方，社会，訃報，政治，科学，スポーツ，中韓，国際の 12 種であり，あるニュースは 1 つのカテゴリにのみ属する．

4 : http://news.ceek.jp/ (accessed 2019–12–24)

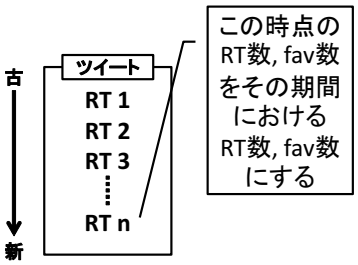

図 1: RT 数，fav 数の決定手法

2.2 ツイートデータについて

本研究で使用するツイートデータは，Twitter Search API[5] を用いて 2017 年 1 月 1 日から 2017 年 12 月 31 日の 12ヵ月間に RT されたツイートを収集した RT データである．各ツイートデータは属性値として様々な値を有しており [6] ，その中でもツイート ID（*id*），ユーザ名（*screen_name*），ツイート日時（*created_at*），RT 数（*retweet_count*），fav 数（*favorite_count*），ツイート本文（*text*），ツイート内 URL（*expand_url*）を抽出し，分析に使用する．

RT データはあるツイートが RT された瞬間に記録され，収集される．このため，同じツイートのツイートデータは最初に RT された時から，最後に RT された時までのデータが残っていることになる．そこで，各ツイートについて図 1 に示すように，集計期間内で最後に RT された時点での RT 数，fav 数をそのツイートの期間内における最新の RT 数，fav 数とし，それ以前の RT データは除去する．また，本研究ではリンク先のニュースに着目するため，そもそもツイート内に URL が含まれてないものも除去する．これらの処理により，全ツイートデータ 4,552,513,378 件のうち，135,264,733 件を使用することとなった．

2.3 言及ツイートの抽出

本研究では，あるニュースへの言及ツイートに対するシェア行為が，ニュースタイトルによって変化するかについて分析する．そのため，先述した 2 種類のデータ（ニュースデータ，ツイートデータ）について，対応付けを行い，ニュース以外の URL を持つツイートを除去する．2.2 節にて述べた 135,264,733 件のツイートについて，ツイート内 URL とニュース URL を照合し，一致したものを対応付ける．ニュースによっては複数のツイートで言及されている場合もあるため，1 つのニュースが複数のツイートと対応している場合がある．また，1 つのツイートで複数のニュースについて言及している場合もあり，この場合は各ニュースにそのツイートを対応付ける．これらの一例を図 2 に示す．対応付けの結果，対応データは 2,615,563 件となった．1 つのツイートで複数のニュースに言及する場合があることから，対応データのツイートには重複が存在する．

5 : https://developer.twitter.com/ja/docs/ads/general/api-reference (accessed 2019–12–24)

6 : https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object (accessed 2019–12–24)

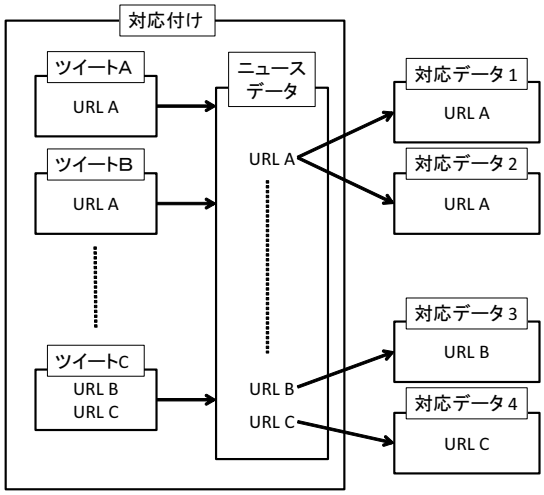

図 2: 対応データの一例

表 1: カテゴリ毎の言及ツイート数

カテゴリ	言及ツイート数 [件]
経済	21462
エンタメ	116604
その他	68923
IT	104181
地方	42750
社会	65992
訃報	259
政治	53711
科学	7987
スポーツ	28682
中韓	3203
国際	53654
合計	567408

RT 数が 1，fav 数が 0 のような殆どシェアされていないデータを除くため，本研究では RT 数・fav 数がともに 10 以上のものに限定することとし，結果的に，567,408 件の対応データを分析することとした．この 567,408 件の対応データを本論文では「言及ツイート」と呼ぶ．言及ツイートのカテゴリ毎のデータ数を表 1 に示す．

3 事前分析

3.1 言及ツイートにおけるニュースのタイトルに現れる単語のカバー率

本研究ではニュースのタイトルに着目するが，ツイート本文に着目する方が良いという立場もあるため，事前分析として，タイトルと本文との関係を調べる．ここではニュースタイトルに現れる単語が，そのニュースに対する言及ツイート内でどれだけカバーされているのか全言及ツイートを対象に調査した．t はある 1 つの言及ツイート，n はある 1 つのニュースとし，ある言及ツイート t の単語集合を W_t，その言及ツイートと対応付けられるニュース n のタイトルの単語集合を W_n とする．カバー率は (1) で求められる．

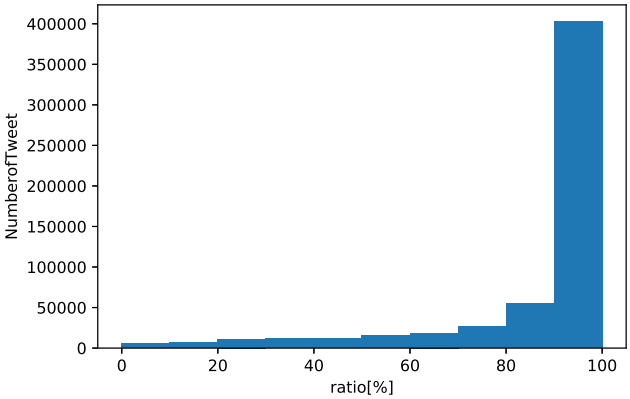

図 3: 言及ツイートにおけるニュースタイトルに現れる単語のカバー率

$$\mathrm{ratio}(t,n) = \frac{|W_n \cap W_t|}{|W_n|} \tag{1}$$

単語集合 W_t ,W_n を求めるためのテキストの分かち書きは，形態素解析エンジン MeCab[7] を用いた．辞書としては，Web 上から得た新語に対応しており，毎週更新されている mecab-ipadic-NEologdNews[8] を用いた．

それぞれの言及ツイートの，カバー率を求め，10%刻みのヒストグラムにしたものを図 3 に示す．横軸はカバー率を，縦軸はその区分に含まれる言及ツイートの数を示す．この結果から，言及ツイートの 8 割以上が，言及しているニュースタイトルに現れる単語を 8 割以上カバーしていることがわかった．このため，ニュースタイトルを用いても，ツイート本文の特性を反映した分析結果が得られると考えられる．

3.2 各カテゴリの RT・fav されやすさ

ニュースページでの滞在時間などに着目した場合，カテゴリ毎にニュースの閲覧行動に差があることが明らかになっている [11]．このことから，Twitter での RT・fav されやすさにも，カテゴリ毎に差があるのではないかと考えた．そこで，あるツイートの RT・fav されやすさ（RT_fav）を，RT の中央値（$median\_RT$）と fav の中央値（$median\_fav$）の比率とし，2.3 節にて述べている対応付けデータを用いて分析した．RT_fav は，1.0 よりも小さければ fav されやすく，1.0 よりも大きければ RT されやすいことを表す．

$$\mathrm{RT\_fav} = \frac{median\_RT}{median\_fav} \tag{2}$$

カテゴリ毎に RT_fav を計測した結果を表 2 に示す．表 2 より，RT されやすいカテゴリとして,「社会」,「訃報」,「政治」,「国際」が，fav されやすいカテゴリとして,「エンタメ」,「IT」があることがわかる．よって，カテゴリ毎に RT・fav されやすさに違いがあることが確認できた．本章での結果を踏まえ，次章では次の 3 つに焦点を当て分析する．

7：使用バージョン:0.996 https://taku910.github.io/mecab/ (accessed 2019–12–24)

8：使用バージョン:102(2019/9/30 更新) https://github.com/neologd/mecab-ipadic-neologd (accessed 2019–09–30)

表 2: カテゴリ毎の RT・fav されやすさの比較

Category	median_RT	median_fav	RT_fav
経済	29	26	1.115
エンタメ	43	90	0.478
その他	40	36	1.111
IT	34	39	0.872
地方	28	25	1.120
社会	42	29	1.448
訃報	32	22	1.455
政治	41	30	1.367
科学	41	39	1.051
スポーツ	24	34	0.706
中韓	26	23	1.130
国際	35	26	1.346

- 公式アカウントとそれ以外のアカウントとの RT・fav されやすさに差はあるか？
- カテゴリによってニュースタイトルは異なるか？
- fav に比べ RT が多い，RT に比べ fav が多いニュース間においてニュースタイトルに現れる特徴語に差はあるか？

4 分析実験

4.1 公式とそれ以外のアカウントの RT・fav されやすさ

一般的に，公式アカウントはそのニュースサイトを好むフォロワーを多数持ち，そのニュースサイトのニュースにしか言及しないが，他方その他のアカウントのフォロワーは特定のニュースサイトを好むユーザに限定されず，様々なニュースサイトのニュースに言及する．このことから，言及するユーザの性質によって RT・fav のされやすさに差が生じる可能性があるため，言及ツイートに対して，全アカウントをまとめて分析することの妥当性を検証する必要があると考えた．

本節では，ニュースメディアの公式アカウントによる言及ツイート群と，公式アカウントを含むすべての言及ツイート群とを対象とし，RT・fav のされやすさを分析する．後者の言及ツイート群は，対象とするニュースメディアのニュースと対応付けられる言及ツイートに限定される．分析するニュースメディアとして，ニュースの言及ツイート数が多い，産経ニュース （Sankei）[9]，NHK ニュース （NHK）[10]，ナタリー （Natalie）[11]，ねとらぼ （Nlab）[12]，AFPBB News （AFPBB）[13] の 5 つを選んだ．

表 3 は分析対象メディアについて，公式ツイートのみの場合とそれ以外の言及ツイートも含んだ場合の両方の RT_fav（式 (2)）を計測した結果である．この結果から，5 つのメディアのどれにおいても，双方のツイート群の RT_fav に大きな差がないことがわかる．このことから，言及ツイートに対して，公式以外のアカウントによる影響は小さく，全アカウントをまとめて分析することは妥当と考えられる．

表 3: 公式とそれ以外のアカウントにおける RT・fav されやすさの比較

ニュースメディア	ツイート群の種別	ツイート数	median_RT	median_fav	RT_fav
Sankei	公式アカウントのみ	37305	45	37	1.216
	すべてのアカウント	79562	37	30	1.233
NHK	公式アカウントのみ	41523	45	37	1.216
	すべてのアカウント	69907	44	35	1.257
Natalie	公式アカウントのみ	18232	108	245	0.441
	すべてのアカウント	39501	80	166	0.482
Nlab	公式アカウントのみ	20599	66	55	1.200
	すべてのアカウント	24714	56	50	1.120
AFPBB	公式アカウントのみ	16954	40	26	1.538
	すべてのアカウント	18974	39	26	1.500

4.2 ニュースタイトルのカテゴリ間相関

本研究では，3.2 節にてカテゴリ毎にシェア行為に差があることを確認した．この差をもたらした一因として，我々はニュースタイトルに着目した．しかし，カテゴリ全体が似たニュースを扱っていた場合，カテゴリ別に分析する意義が薄れるため，カテゴリ間のニュースタイトルが異なるか検討する必要があると考える．よって，ニュースタイトルのカテゴリ間の相関関係を調査する．

まず，カテゴリ毎の全てのニュースタイトルを分かち書きし，各単語の出現回数をカテゴリ毎に集計する．そして，各カテゴリの各単語に対して計測した相対頻度（$relative\_count$）の大きさでランク付けしたものを順位データとし，この順位データを用いてスピアマンの順位相関係数を算出し，カテゴリ間の相関関係を見る．$T(c,w)$ をあるカテゴリ c に属する単語 w の出現回数，$T(w)$ をある単語 w の総出現回数とした場合，相対頻度は式 (3) で求められる．

$$\mathrm{relative\_count(c,w)} = \frac{T(c,w)}{T(w)} \tag{3}$$

スピアマンの順位相関係数は式 (3) にて求められる．N はニュース全体に現れる単語の種類数，R_x と R_y はある単語のカテゴリ x, y での出現頻度ランク，t_i,t_j はカテゴリ x, y での同順位の組それぞれの個数（$i = 1, 2...n_x, j = 1, 2...n_y$）を表す．

$$\text{相関係数}\rho = \frac{T_x + T_y - \sum(R_x - R_y)^2}{N^3 - N} \tag{4}$$

$$\mathrm{T_x} = \frac{N^3 - N - \sum(t_i^3 - t_i)}{12} \tag{5}$$

$$\mathrm{T_y} = \frac{N^3 - N - \sum(t_j^3 - t_j)}{12} \tag{6}$$

結果を表 4 に示す．大きな値として政治カテゴリと国際カテゴリの 0.3694 や中韓カテゴリと国際カテゴリの 0.3272 などが見られる．これらはニュースのジャンル上，国名などの似通っ

9 : https://www.sankei.com/ (accessed 2019–12–24)
10 : https://www3.nhk.or.jp/news/ (accessed 2019–12–24)
11 : https://natalie.mu/ (accessed 2019–12–24)
12 : https://nlab.itmedia.co.jp/ (accessed 2019–12–24)
13 : https://www.afpbb.com/ (accessed 2019–12–24)

表 4: カテゴリ間の相関係数（ニュースタイトル）

	経済	エンタメ	その他	IT	地方	社会	訃報	政治	科学	スポーツ	中韓	国際
経済	1											
エンタメ	-0.0160	1										
その他	0.2138	-0.0870	1									
IT	0.0262	-0.2220	-0.0760	1								
地方	0.2055	-0.1260	0.1082	-0.1510	1							
社会	0.2995	-0.0630	0.1980	-0.0750	0.2931	1						
訃報	0.0638	0.0116	0.0381	0.0002	0.0395	0.0574	1					
政治	0.3464	-0.0220	0.1897	-0.0210	0.1962	0.3242	0.0655	1				
科学	0.2663	0.0290	0.1702	0.0396	0.1774	0.2669	0.0910	0.2470	1			
スポーツ	0.2152	-0.0080	0.1141	-0.0320	0.1648	0.1868	0.0520	0.2359	0.1892	1		
中韓	0.2904	0.0242	0.1529	0.0304	0.1776	0.2441	0.0795	0.2977	0.2828	0.2151	1	
国際	0.3466	-0.0400	0.2006	-0.0320	0.1812	0.3031	0.0623	0.3694	0.2807	0.2076	0.3272	1

た単語が現れるカテゴリ間であるからと考えられる．そのほかのカテゴリ間においては比較的低い値が見られるため，ニュースタイトルに現れる単語の出現傾向がカテゴリによって異なると考えても良いと捉える．

4.3 RT 優位・fav 優位なニュースのタイトルに現れる特徴的な単語

4.2 節にて，ニュースタイトルに現れる単語の出現傾向がカテゴリによって異なることを確認した．よって本節では，あるニュースに対する言及ツイートの総 RT 数が総 fav 数よりも多いニュースを RT 優位のニュース，総 fav 数が総 RT 数よりも多いニュースを fav 優位のニュースとし，この二種類のニュース間においてニュースタイトルを構成する特徴語が異なるか分析する．

カテゴリ毎のニュースタイトルの特徴語を求める手法を説明する．まず，ニュースタイトルを全て分かち書きし，カテゴリ毎にそれぞれの単語の出現回数を数える．この回数を実測値とする．そして，特徴語を決定する指標として，式 (7) に示す期待度数を用いた式 (8) のスコアを求める．この指標は実測値が期待度数からどれほど離れているのかを表す．この指標が大きければ大きいほど，他カテゴリに比べてその単語が特定のカテゴリに多く出現していることを表す．

$$\mathrm{E(c,w)} = \frac{\sum_{c' \in C} T(c^{'}, w) \times \sum_{w' \in W} T(c, w^{'})}{\sum_{c' \in C, w' \in W} T(c^{'}, w^{'})} \qquad (7)$$

$$\mathrm{score(c,w)} = \frac{(T(c,w) - E(c,w))^2}{E(c,w)} \qquad (8)$$

前述した方法で求めた各カテゴリの特徴語上位 10 件を降順に並べた結果を表 5, 表 6 に示す．本研究では，表 1 に示した通り，訃報カテゴリについてはニュースの数がそもそも少なかったため，結果として見るには不十分であると考え，このカテゴリの考察は控える．

はじめに RT 優位・fav 優位なニュース間で現れる特徴語がわかりやすく異なる地方，社会，科学カテゴリを見る．地方カテゴリでは RT 優位なニュースのタイトルの特徴語として「大阪県警」「逮捕」「容疑」といった犯罪に関係する単語や「辺野古」「沖縄」といった軍事基地関連の単語が，fav 優位な方では「ベガルダ」「高校野球」といった地方のスポーツに関連する単語や「うめきたガーデン」「富士山」といった地方の観光地に関連する単語が上位に現れている．社会カテゴリでは RT 優位な方に「避難勧告」,「氾濫危険水域」などの災害情報や「運転見合わせ」,「運転再開」などの交通情報に関連する単語が，fav 優位なニュースでは,「皇太子」,「真子さま」などの皇族関連の単語が多く上位に現れている．科学カテゴリでは RT 優位な方に「福島第一原発」,「被ばく事故」などの原発関連の単語が，fav 優位なニュースでは,「新種」,「化石」,「発見」などの進展や調査関連の単語が多く上位に現れている．これら 3 つのカテゴリの特徴語に共通することとして，生活に関わってくる重要な単語や，生活の不安に繋がる単語が RT 優位なニュースのタイトルに現れていることである．また，fav 優位なニュースのタイトルには，個人の興味が反映される「おもしろい」単語が多く現れているのではないかと考えられる．

次に特徴語の違いがわかりにくく，また比較的専門的なカテゴリである経済，エンタメ，IT カテゴリについて見る．RT 優位な特徴語として経済カテゴリでは「東芝」,「神戸製鋼」,「損害」などが上位に来ており，これは 2017 年に東芝が半導体事業を売却したことや神戸製鋼の不正が発覚したためだと考えられる．また，エンタメカテゴリでは「テレビアニメ」,「マンガ」などの単語の中,「死去」,「役」が特徴語の上位に出現している．これは国民的アニメにも出演していたベテランの声優が 2017 年に亡くなったことが関係していると思われる．これらから，ニュースカテゴリにおける重大なニュースの RT が多くなる傾向にあると考えられる．これを踏まえてみると，IT カテゴリでは家庭用コンシューマ機などのゲームに関連するニュースは重要であり RT されやすく，他方アニメや先行情報に関連するニュースは fav されやすいと考えられる．この他，エンタメカテゴリの他の特徴語を見ると，アニメや漫画関連の単語が RT 優位なニュースのタイトルに,「MV」,「ツアー」,「アルバム」のような音楽関連の単語が特徴語の上位に現れている．このことから同じサブカルチャーの中でもシェア行為に違いがあることがわかる．特に，IT カテゴリでは fav 優位なニュースのタイトルに現れていたアニメ関連の単語が，エンタメカテゴリでは RT 優位なニュースの特徴語として現れていることから，同じ単語であってもカテゴリ間にて重要度が異なる可能性が示されている．

その他カテゴリについては，RT 優位な方に「ニュースの深層」,「安倍政権」,「産経抄」などの政治カテゴリや経済カテゴ

表 5: 各カテゴリの特徴語（1/2）

経済		エンタメ		その他		IT		地方		社会	
RT 優位	Fav 優位	RT 優位	Fav 優位	RT 優位	Fav 優位	RT 優位	Fav 優位	RT 優位	Fav 優位	RT 優位	Fav 優位
東芝	株価	テレビアニメ	MV	ニュースの深層	猫ちゃん	PS4	さん	大阪府警	ベガルタ	逮捕	皇太子
売却	値上がり	死去	映画	安倍	かわいい	版	TV アニメ	逮捕	弘前	避難勧告	さま
株価	円相場	連載	新曲	安倍政権	ビデオ	配信	カット	容疑	富士山	運転見合わせ	陛下
値上がり	投資	マンガ	主演	産経抄	ご	PC	より	兵庫県警	高校野球	容疑	段
半導体	値下がり	次号	ツアー	て	ワンコ	Nintendo Switch	先行	辺野古	うめきたガーデン	氾濫危険水位	パンダ
値上げ	リーダーシップ	アニメ	ライブ	聞い	た	開始	到着	道内	見頃	死亡	眞子さま
円相場	教養	舞台化	アルバム	た	て	向け	場面	沖縄	J1	疑い	皇后
神戸製鋼	終値	役	ドラマ	インサイド	新型	対応	登場	男	関西	運転再開	藤井
損失	小幅	放送	ら	Twitter	モーターショー	スマホ	レポート	関西	県	避難指示	津波
国内政治	ビットコイン	NHK	公演	は	そう	VR	公開	大阪	盛岡	男性	雪崩

表 6: 各カテゴリの特徴語（2/2）

訃報		政治		科学		スポーツ		中韓		国際	
RT 優位	Fav 優位	RT 優位	Fav 優位	RT 優位	Fav 優位	RT 優位	Fav 優位	RT 優位	Fav 優位	RT 優位	Fav 優位
死去	術語	民進	安倍晋三首相	福島第一原発	新種	ロッテ	阪神	台湾	台湾	米	米
訃報	笑福亭仁勇	自民	衆院選	打ち上げ	研究	ヤクルト	DeNA	中国大陸	日本統治時代	北朝鮮	トランプ大統領
はしだ のりひこ	辰濃	氏	自民	研究	発見	WBC	西武	台北	台北	中国	トランプ氏
天声人語	元吉	首相	官房長官	柏崎刈羽原発	金井	巨人	ロッテ	中国メディア	台南	トランプ氏	北朝鮮
松本俊夫	アビルドセン	衆院選	外相	3 号	化石	抹消	巨人	総統	総統	シリア	トランプ政権
筑豊	天声人語	野党	菅義偉	高浜原発	絶滅	日本ハム	羽生	韓国	高雄	ロシア	大統領
赤塚	倉嶋厚	共産	氏	4 号機	宇宙飛行士	プロ野球	大谷	空港線	屏東	IS	中国
書風	したたる	衆院	幹事長	探査機	丸のみ	西武	藤浪	メトロ	台東	死亡	英
元吉	九條	幹事長	河野太郎	被ばく事故	太古	2 軍	優勝	台湾鉄道	蔡	北	ロシア
路加	中井治	安倍首相	代表	大気汚染	探査	中日	先発	桃園	阿里山	大統領	トランプ

リに現れそうな単語が，fav 優位なニュースでは,「猫ちゃん」,「かわいい」,「ワンコ」などの他のカテゴリには分類されなさそうな単語が多く上位に現れている．その他カテゴリには他の 11 カテゴリに当てはまらないコラムなどのニュースが分類されているためニュースの幅が広い．よって，その中でも専門的な単語が RT 優位なニュースのタイトルに現れ，さらに雑多な単語の中でも特徴的な単語が fav 優位なニュースのタイトルの特徴語として上位に現れたのだと考えられる．

最後に，より専門的な政治，中韓，スポーツ，国際カテゴリについて見る．この 3 つのカテゴリでは，これまでと異なり，RT 優位・fav 優位なニュース間の特徴語に目立った違いが見受けられなかった．この原因として，これらのカテゴリは専門性が高いかつニュースの幅も狭いため，RT・fav されるニュースのタイトルに大きな違いが無かったのではないかと考えられる．

4.4 各カテゴリの特徴語の類似度

4.3 節では RT 優位・fav 優位なニュースのタイトルに現れる特徴語について分析し，考察した．しかし，その考察は主観的なものであるため，RT 優位・fav 優位なニュースのタイトルに現れる特徴語の上位 50 件を抽出し，両ニュース間の同一カテゴリで特徴語の類似度を測り，考察の信頼性を調査した．本研究では，類似尺度として Dice 係数を用いる．Dice 係数は 2 つの集合の平均要素数と共通要素数の割合を表す値であり，0 から 1 の間の値をとる．あるカテゴリ c の RT 優位なニュースの特徴語群を R_c，あるカテゴリ c の fav 優位なニュースの特徴語群を F_c とすると，Dice 係数は式 (9) で求められる．

$$\mathrm{Dice}(\mathrm{R_c}, \mathrm{F_c}) = \frac{2|R_c \cap F_c|}{|R_c| + |F_c|} \tag{9}$$

RT 優位・fav 優位なニュース間の同カテゴリにおいて算出された上位 50 件の特徴語の Dice 係数を表 7 に示す．この結果を見たところ，4.3 節にて大きな違いが見られなかった中韓，スポーツ，国際カテゴリの特徴語が RT 優位・fav 優位なニュース間にて，他カテゴリに比べて Dice 係数が大きいことがわかる．また，特徴語に違いが見られた科学，その他，地方カテゴリの Dice 係数は小さく，少し違いが見られた経済カテゴリは，科学カテゴリなどの Dice 係数に比べると大きく，政治カテゴリなどの Dice 係数に比べると小さいことがわかる．つまり，それぞれのカテゴリにおいて，他カテゴリには現れない専門的な単語が多くタイトルに出現する程，RT 優位なニュース・fav 優位なニュースの特徴語に違いが無くなると考えられる．この結果は 4.3 節での，考察を支持する結果になっていると考える．

表 7: カテゴリ毎の RT・fav されやすさの比較

カテゴリ	Dice 係数
経済	0.32
エンタメ	0.16
その他	0.14
IT	0.16
地方	0.16
社会	0.08
訃報	0.1
政治	0.46
科学	0.16
スポーツ	0.46
中韓	0.46
国際	0.48

5 おわりに

本研究の目的はソーシャルメディアで言及されたニュースの

タイトルに関する分析をすることである．そこで，ソーシャルメディアの 1 つである Twitter での言及ツイートを対象に，ニュースタイトルに着目してシェア行為の分析を行った．まず，ニュースデータとツイートデータに対して，ツイート内 URL とニュース URL を照合し，一致したものを対応付けた．対応付けたデータを用いて，言及ツイートにおけるニュースタイトルのカバー率及び，ニュースカテゴリ毎の RT・fav されやすさを事前に分析した．この分析結果より次の 3 つに焦点を当て，分析を行った．

- 公式アカウントとそれ以外のアカウントとの RT・fav されやすさに差はあるか？
- カテゴリによってニュースタイトルは異なるか？
- fav に比べ RT が多い，RT に比べ fav が多いニュース間においてニュースタイトルに現れる特徴語に差はあるか？

分析の結果，公式アカウントのみの RT・fav されやすさと，それ以外のアカウントのツイートを含む場合の RT・fav されやすさの間に大きな差は見受けられないため，言及ツイートに対して，全アカウントをまとめて分析するのが妥当とわかった．また，ニュースタイトルに現れる単語について，カテゴリ間の相関関係を見たところ，ニュースタイトルに現れる単語の出現傾向がカテゴリによって異なると考えても良いことが確認できた．さらに，RT 優位・fav 優位な大多数のニュース間において，ニュースタイトルの特徴語を分析し比較した所，ニュースの幅が狭く専門性が比較的低いニュースカテゴリにおいては，RT 優位・fav 優位なニュースのタイトルに現れる特徴語に違いがあることを確認した．

文　　献

[1] 総務省．社会課題解決のための新たな ict サービス・技術への人々の意識に関する調査研究．https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/pdf/n4200000.pdf, 2015.

[2] 塩田茂雄 and 中島圭佑．Twitter データに見られる特徴と人間のリツイート行動．In 人工知能学会全国大会論文集 一般社団法人人工知能学会, pages 2E5J601–2E5J601. 一般社団法人人工知能学会, 2019.

[3] Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In 2010 IEEE Second International Conference on Social Computing, pages 177–184. IEEE, 2010.

[4] 津川 翔．ソーシャルネットワークのコミュニティ構造がソーシャルメディア上の投稿の拡散規模に与える影響の分析．人工知能学会全国大会論文集, 2018:2C204–2C204, 2018.

[5] 山下 玲子 and 三浦 麻子．おもしろツイートはいかに広まったか：事例研究による「じわる」プロセスの解明．メディア・情報・コミュニケーション研究, (3):1–18, mar 2018.

[6] Sho Tsugawa and Hiroyuki Ohsaki. Negative messages spread rapidly and widely on social media. In Proceedings of the 2015 ACM on Conference on Online Social Networks, pages 151–160. ACM, 2015.

[7] Kholoud Khalil Aldous, Jisun An, and Bernard J Jansen. Predicting audience engagement across social media platforms in the news domain. In International Conference on Social Informatics, pages 173–187. Springer, 2019.

[8] Kholoud Khalil Aldous, Jisun An, and Bernard J Jansen. Stylistic features usage: Similarities and differences using multiple social networks. In International Conference on Social Informatics, pages 309–318. Springer, 2019.

[9] 李光鎬. ツイッター上におけるニュースの普及: どのようなニュースを誰がリツイートするのか．メディア・コミュニケーション: 慶応義塾大学メディア・コミュニケーション研究所紀要, (65):63–75, 2015.

[10] 興梠, 木村, 藤代, and 西川．Sns 上での拡散を誘発する web ニュース説明文の調査と自動選択．電子情報通信学会論文誌, 2016.

[11] Yoshifumi Seki and Mitsuo Yoshida. Analysis of user dwell time by category in news application. In 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pages 732–735. IEEE, 2018.