2S-08

新型コロナ禍に関するツイートの自動カテゴリ分類

木村 優平[♭]　西村 百之輔[♭]　片倉 多智[♭]　戎 涛[♭]　早坂 絵央[♭]　延澤 志保[*]

[♭]東京都市大学知識工学部　[*]東京都市大学情報工学部

1　研究背景

2019年に発生が確認された新型コロナウイルス感染症は全世界に多大な社会的，経済的な影響を及ぼしている．ソーシャルメディアなどで拡散する大量の情報が実社会に影響を及ぼす現象はインフォデミックと呼ばれ，実生活に影響を及ぼすことが懸念される [1, 2, 3]．鳥海らは新型コロナ禍の世情を俯瞰的に捉えることを目的として，ツイートの感情分析を行い可視化した [4]．本研究ではコロナ禍のツイートに着目し，既知のカテゴリに分類することで興味関心の推移を俯瞰的に捉えることを目指す．

2　ツイートのカテゴリ分析

興味関心を示すカテゴリとしては，医療，生活，政治，経済，スポーツ，エンタメの6カテゴリ分類が提案されている [5]．本研究ではスポーツとエンタメを合わせて娯楽カテゴリとし，5カテゴリに分類する．

本研究では実験コーパスとして新型コロナウイルスに関連するツイート[1]を用いる．リツイートは分析対象に含めない [6]．リプライツイートは，50文字以上の場合に限り，返信先のツイートとまとめて1ツイートとして扱う．100文字未満のツイートは短過ぎるため対象から外す．

2020年1月から9月のコロナウイルスに関するツイートを月ごとに人手でカテゴリ分類した結果をツイート正解データセットとする (図 1)．図1の縦軸はツイート数

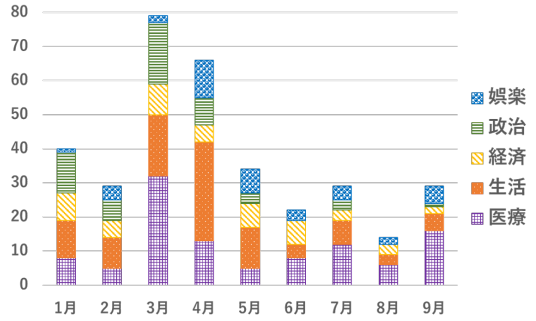

図 1: ツイート正解データセットのカテゴリ分類結果

を示す．図1より，3月から5月は生活カテゴリツイートが他の月に比べて多いことがわかる．これは緊急事態宣言による生活環境の変化が原因であると考えられ，ツイートのカテゴリ分類によってコロナ禍の影響の推定が可能と期待できる．

3　ニュース記事を用いたツイートの分類手法

ツイートはもともとカテゴリ情報を持たず，ツイート群から分類学習を行うことは困難である．そのため本研究では，ニュース記事を学習のベースに用いてツイートのカテゴリ分類を行う手法を提案する．学習にはNHK新型コロナウイルス特設サイト[2]，日経新聞[3]，読売新聞[4]の新型コロナウイルスに関係するニュースを用いる．

ニュース記事とツイートでは時系列的に相関関係がある [5]．そこで本研究では1カ月毎に学習データを増やしながら分類器を更新する学習方法を提案する (図 2)．図

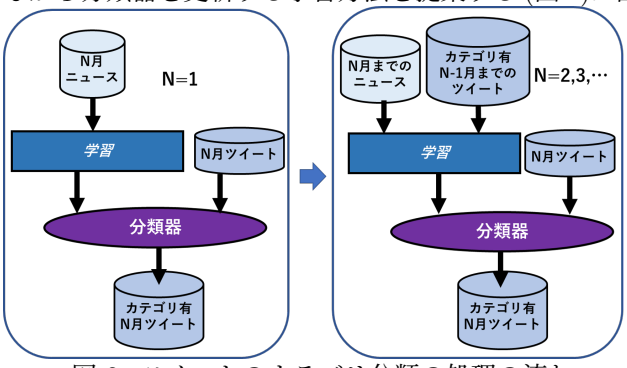

図 2: ツイートのカテゴリ分類の処理の流れ

2に示すとおり，n月のツイートのカテゴリ分類を$n+1$月に行うことを想定し，学習コーパスとして1月からn月までのニュース記事と1月から$n-1$月までの分類済みツイートを用いる．1月のツイートに対しては1月のニュース記事のみを学習コーパスとする．

本研究では，各カテゴリ毎の学習データ群における名詞を抽出し，閾値を超える頻度を示す形態素を素性として用いる．ニュース記事の見出しと本文，は分けて集計する．ツイートについてはハッシュタグを該当ツイートの見出しとして扱う．表1に各カテゴリについて見出しから得られた素性と本文から得られた素性の7月時点での上位3件ずつを示す．表1から各カテゴリにおいて，

表 1: 7月時点の各カテゴリの素性

カテゴリ	見出し			本文		
医療	拡散	希望	事態	検査	情報	医療
生活	事態	宣言	マスク	マスク	中国	可能
政治	中国	ニュース	韓国	影響	日本	経済
経済	肺炎	武漢	ニュース	対策	首相	肺炎
娯楽	時間	好き	肺炎	時間	今日	情報

該当カテゴリを連想可能な単語が抽出できていることがわかる．

Automatic Categorization of Tweets on the COVID-19 Calamity.

Yuhei Kimura[♭], Momonosuke Nishimura[♭], Taichi Katakura[♭], Bo Rong (戎 ボツ)[♭], Kaio Hayasaka[♭], and Shiho Hoshi Nobesawa[*].
[♭] Faculty of Knowledge Engineering, Tokyo City University
[*] Faculty of Information Technology, Tokyo City University

[1]COVID-19-TweetIDs, https://github.com/echen102/COVID-19-TweetIDs.

[2]NHK 新型コロナウイルス特設サイト，https://www3.nhk.or.jp/news/special/coronavirus/latest-news/.

[3]日経新聞，https://www.nikkei.com/.

[4]読売新聞，https://www.yomiuri.co.jp/.

4　ツイート分類実験

本研究では分類器としてサポートベクターマシン (SVM) を用いる．学習データは見出しと本文をそれぞれの素性を用いてベクトル化しマージして利用した．分類器の評価方法として，ニュース記事のみで学習したモデル，ツイート正解データセットとニュースで学習したモデル，出力結果とニュース記事で学習したモデルの 3 種類について，ツイート正解データセットをテストデータとして分類実験を行った結果を表 2 に示す．各カテゴリの分類

表 2: ツイート分類実験結果 (9 月時点)

カテゴリ	学習	テスト	モデル (F1 値)		
	記事	ツイート	記事のみ	正解追加	出力追加
医療	9,079	402	0.78	0.97	0.63
生活	1,924	284	0.75	0.98	0.45
政治	3,150	143	0.77	0.96	0.64
経済	7,556	94	0.75	0.97	0.60
娯楽	1,486	127	0.77	0.98	0.62
平均	4,639	210	0.76	0.97	0.59

器を学習する際の学習データは，正例と負例が 1 対 1 になるようデータ数を調整した．各モデルの 9 月時点の平均 F1 値を見ると，記事のみのモデルに対して正解ツイートを用いたモデルの方が約 0.2 ポイント高く，適切に分類されたツイートを学習データに加えることが有効であるとわかる (表 2)．しかし，出力結果を用いた場合には，記事のみのモデルに比べ約 0.1 ポイント減少している．

図 3 にニュース記事のみを用いて分類した場合の月毎の各カテゴリの F1 値を示す．各カテゴリの F1 は 0.7 前

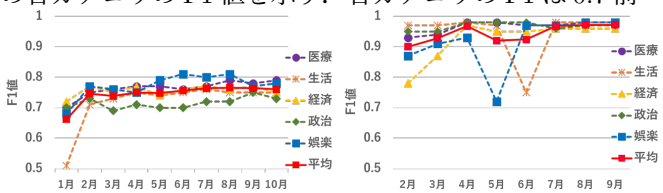

図 3: 分類結果 (左: 記事のみ，右: 正解追加)

後から 0.8 前後へ緩やかに上昇しており，1 月時点の F1 値より 10 月時点の F1 値の方が全てのカテゴリにおいて高い (図 3 左)．このことから，ニュース記事のみを用いたツイートのカテゴリ分類は十分な精度とは言えない．

ニュース記事とツイート正解データセットを合わせて学習したモデルの 2 月から 9 月の実験結果を図 3(右) に示す．このモデルでも 1 月時点ではニュース記事のみで学習している．図 3(右) は図 3(左) と比べて F1 値が最大約 0.4 ポイント高く，正しくカテゴリが付与されたツイートの利用の有効性が示された．図 3(右) では 6 月以降の F1 値がほぼ 1 であり，ツイートを適切に分類できていることがわかる．このことから，正しく分類されたツイートを用いて学習を行うことで，ニュース記事のみで学習する場合に比べ精度が向上することがわかる．表 2 出力追加モデルでは 1 月時点での分類器はニュース記事と正解ツイートとで学習しているため，図 3 の 1 月の F1 値に見られるように学習データに加えるツイートの分類精度が低いことが連鎖的に精度悪化を引き起こしている．

図 4 に各カテゴリの分類結果の F1 値を示す．図 4 で

2月	医療	生活	経済	政治	娯楽
医療	0.52	0.56	0.26	0.15	0.18
生活	0.44	0.56	0.34	0.06	0.22
経済	0.34	0.50	0.50	0.09	0.26
政治	0.35	0.56	0.23	0.38	0.19
娯楽	0.45	0.38	0.14	0.01	0.41

10月	医療	生活	経済	政治	娯楽
医療	0.42	0.31	0.08	0.23	0.28
生活	0.21	0.50	0.16	0.16	0.35
経済	0.17	0.29	0.49	0.28	0.29
政治	0.23	0.25	0.14	0.56	0.21
娯楽	0.17	0.35	0.12	0.08	0.60

図 4: ツイート分類結果のカテゴリ重複状況

は，例えば 2 月の医療カテゴリのツイートを医療カテゴリと分類する場合の F1 値 0.52 に対して生活カテゴリと分類する場合の F1 値は 0.56 となるなど，2 月時点ではカテゴリの混乱が見られるのに対して，10 月の時点ではカテゴリの分離が進んでいることがわかる．

5　まとめ

新型コロナウイルス感染症は，発生以来 1 年経ってもなお全世界に多大な社会的，経済的な影響を及ぼしている．こういった社会現象や自然災害に対し，俯瞰的に状況を捉えることができる手法として，ソーシャルメディアの分析が有用とされている．コロナ禍での興味関心の推移を捉えるため，本研究では，ニュース記事のカテゴリに合わせたツイートの自動分類を提案した．

本研究ではまずニュースとツイートの時系列的な内容に相関関係があることを確認した．その上で，ニュース記事のみで学習したモデルと分類対象のツイートを適宜学習データに加え分類器を更新したモデルを提案し，人手により正解カテゴリを付与したツイート群を用いて分類実験を行った．実験結果から，正しく分類されたツイートを用いて学習することで，ニュースのみの場合と比べ精度が向上した．このことより，ニュース記事での学習を用いたツイート分類は十分に実現可能と考えられる．

参考文献

[1] John Zarocostas, “How to Fight an Infodemic,” The Lancet World Report, Vol.395, No.10225, p.676, 2020.

[2] Deborah Bunker, “Who do you trust? The Digital Destruction of Shared Situational Awareness and the COVID-19 Infodemic,” International Journal of Information Management, 2020.

[3] 笹原和俊, “ウェブの功罪,” 情報の科学と技術, vol.70, no.6, p.309–314, 2020.

[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.1–7, 2020.

[5] 木村優平, 戎淳, 王羽, 片倉多智, 中山碧士, 西村百之輔, 早坂絵央, 真鍋大雅, 吉原圭祐, 藤田和成, 延澤志保, “ツイートに基づく新型コロナ禍に関する興味関心の推移の可視化,” NLP 若手の会第 15 回シンポジウム, 2020.

[6] 浅沼爽汰, 藤田和成, 田村亮介, 白石絵里奈, 白井聡一, 町田翔, 延澤志保, “災害時避難支援のための twitter からの現在地周辺情報の抽出,” 情報処理学会研究報告, no.2018-NL-238-8, pp.1–4, 2018.