Vol.2022-DBS-175 No.36
Vol.2022-IFAT-148 No.36
2022/9/10

Sentence-BERTを利用したSNS上のリプライに基づいたニュース特徴の抽出と可視化

高橋　陸[1,a)]　牛尼　剛聡[2,b)]

概要：近年 SNS の普及に伴い，ユーザが興味のあるトピックに関する情報収集を行う際に SNS を用いることが一般化した．SNS を用いた情報収集では，リアルタイムで更新される情報や投稿者から直接意見が得られるなどの利点がある．一般的に SNS を用いて情報収集する際，対象とするトピックに関連する単語をキーワードとして検索を行うキーワード検索を用いる．しかし，SNS はユーザが個別に投稿を行うため情報の整理が行われておらず，同一のトピックに関する投稿であっても含まれるキーワードが一致していない場合が多い，そのため，SNS から特定のトピックに関する意見や反応を網羅的に検索するのは困難である．本研究では，Twitter 上に投稿されたニュース記事とそれに対するリプライに着目し，機械学習を用いてニュースとリプライの関係の学習を行うことで，ニュースに対するリプライをもとにした社会的反応の特徴を抽出する手法を提案する．本手法では，ニュース記事とリプライの文章ベクトルから，同一ニュースに関する投稿の文章ベクトルの類似度が高くなるよう学習を行うことで，投稿から類似度の高いニュースを求め，ユーザが興味を持ったニュースに関連する情報を提示することで情報収集を支援する．

キーワード：Sentence-BERT，SNS，ニュース，ユーザ支援

1. はじめに

近年，個人が手軽に情報を発信できる SNS(Social Networking Service) の普及に伴い，Twitter や Instagram など，多くの SNS で様々なトピックに関する投稿が多く行われている．SNS 上にはユーザがいつでも気軽に文章を投稿できることから，個人の意見や反応が多く存在する．また，SNS 上にはテレビ局や新聞社などの報道機関が，取材などで得た情報をリアルタイムにニュース記事として投稿している．こうした背景から，SNS は情報収集の手段として用いられることも多くなってきている．

新聞通信調査会による「第 13 回メディアに関する全国世論調査」[1] では，インターネットは情報源としての信頼度は新聞やテレビに劣る一方で，情報源として最も欠かせないメディアとして支持されている．また，NTT ドコモのモバイル社会研究所による「週１回以上アクセスして日常的に生活情報を得ているメディア」に関する調査 [2] では，10 代 20 代においてソーシャルメディアの情報収集としての利用率ははテレビや新聞による情報収集の割合を上回っており，特に若い世代を中心に SNS が情報源として広く利用されていることが分かる．

SNS を情報収集に用いることで，ユーザは鮮度の高い情報や多様な視点からの反応や意見を得ることができる．一方で，SNS 上に存在する情報はトピックごとや関連する内容に関して整理されておらず，様々な投稿を網羅的に取得することは困難である．現在，ユーザが SNS を用いてニュースに関する情報を探索する際に，一般的に用いられている方法は，対象とするニュースに関する代表的なキーワードを用いたキーワード検索である．キーワード検索では，対象とするニュースに対してユーザが関係があると考える単語や，知りたい内容についてのキーワードをクエリとして投稿を検索する．しかし，一般に対象とするニュースに関して，そのニュースに対する SNS の投稿全てに含まれるような単語は存在しない．したがって，単一のキーワードを用いた検索では，ユーザが欲しい情報が十分に得られないことが多く，情報を網羅的に集めるにはキーワードを変えて何度も検索を行う必要がある．以上の問題点から，現状では SNS を情報源として有効に活用できているとは言い難い．

本研究では、代表的な SNS の一つである Twitter を用

[1] 九州大学大学院芸術工学府
Graduate School of Design, Kyushu University, Fukuoka 815–8540 Japan
[2] 九州大学大学院芸術工学研究院
Faculty of Design, Kyushu University, Fukuoka 815–8540 Japan
[a)] takahashi.riku.110@s.kyushu-u.ac.jp
[b)] ushiama@design.kyushu-u.ac.jp

いて，Twitter 上に投稿されたニュース記事とそれに対するリプライの関係に着目し，機械学習を用いてニュースとニュースに対する SNS 上の反応の関係を学習する．その後，学習したモデルを用いてニュースやリプライの文書ベクトルを生成し，各ニュースや投稿の関係性を可視化することでユーザが情報収集を行う際の支援とすることを目的とする．

本研究の貢献は，以下の通りである．

- ニュースの内容ではなく，SNS を用いてニュースに対する反応という観点からニュースの特徴を抽出することで，SNS による情報収集を支援する．
- 2 文間の関係を学習することにより，ニュースや SNS 投稿の反応を考慮した文書ベクトルを生成する．

本論文の構成は以下のようになっている．第 2 章では，関連研究の紹介を行う．第 3 章では，提案手法について述べる．第 4 章では，提案手法の有効性を評価するための実験内容について述べ，実験結果について述べる．第 6 章では，まとめと今後の課題について述べる．

2. 関連研究

これまでにも，文章の特徴から，機械学習を用いて関連する情報を抽出する手法が提案されている．ここでは，それらの手法のうち，提案手法に関連する手法について述べる．

2.1 ニュース理解支援に関する研究

神谷 [3] らは，ウェブ上から社会問題に関連する情報を抽出する手法を提案している．この手法では，対象とするウェブ上のテキストに対して，それが社会問題に関連する内容であるかを判断し，関連する社会問題に分類している．この手法では，日本語 wikipedia 中の，社会問題に関する記事を教師データとして，自然言語処理の代表的な機械学習モデルの一つである BERT[4] を用いたテキスト分類モデルを構築している．この手法では，対象とするテキストが社会問題に関連するか否かを判定した後に，どの社会問題に関連するかを分類している．評価実験では，社会問題であるか否かの判定に関しては良い結果が得られなかった一方で，社会問題であると判定された記事を分類するタスクでは高い精度であった．

この手法では，社会問題との関係が明示的に示されていないテキストを，社会問題に分類するという点が本研究の提案手法と類似している．しかし，本論文の提案手法では SNS 上の投稿を学習データとして用いるため，Wikipedia に掲載されていないニュースに関しても対応可能である．

2.2 文書の特徴抽出に関する研究

大倉 [5] らは，ニュース記事に対する関連記事を判定するためのキーフレーズ抽出方法を提案している．この手法では，記事内容の特徴を捉え，かつ同一トピック内で広く使用されるフレーズを抽出するために，記事内での単語の出現頻度と類似記事における出現数からフレーズを抽出し，RNN[6] を用いて学習を行うことで関連記事を判定するためのキーフレーズを抽出している．

この手法では，文章中に現れるフレーズを用いて機械学習モデルに類似記事の学習を行っている．関連する記事を得るという点で本研究の目的を類似している．しかし，本研究では SNS 投稿を用いて間接的にニュース記事間の関係性を求めるため，記事中に含まれるフレーズに関係なくニュース記事の特徴を抽出するという点で異なる．

3. 提案手法

3.1 概要

本研究では，入力として用いる文章間の関係性を学習するため，代表的な自然言語処理モデルである BERT[4] ではなく，文書ベクトルの生成に関して改良したモデルである Sentence-BERT[7] を用いる．Sentence-BERT は図 1 に示す様に，入力として与えられた二つの文章をそれぞれベクトル化し，文書ベクトルの類似性や関係性を学習することで，文章間の関係を文書ベクトルに反映するモデルである．Sentence-BERT は通常，2 つの類似した内容の文章を入力として用い，それぞれの文章のベクトルがベクトル空間上で近い位置になるよう学習を行うことで，類似文章検索などのタスクに用いられる．本研究では，入力文として SNS に投稿されたニュース記事の見出し部分と，それに対して行われたリプライを用いる．関係のある二つの文章のベクトルがベクトル空間上で近くなるよう学習を行うことで，ニュースと SNS 投稿の関係性や投稿同士の関係性からニュースに対する SNS 上の反応の特徴を抽出する．

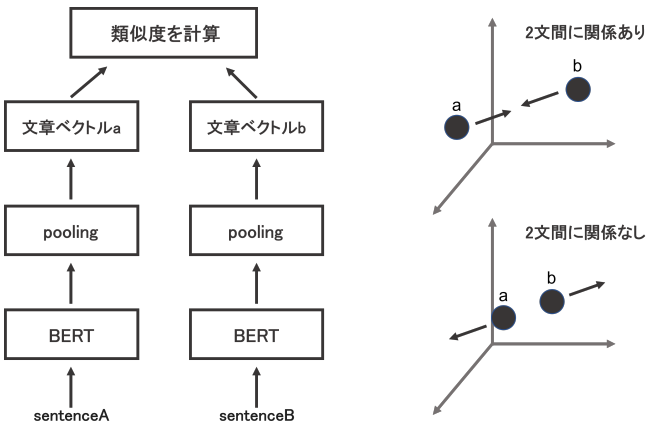

図 1 Sentence-BERT の学習

本研究では入力に用いる文章を変えることで 2 種類のモ

デルの学習を行う．SNS に投稿されたニュース記事の見出し文章とそれに対するリプライをモデルの入力文章として用いた場合と，ニュース記事に対するリプライ同士を用いて関係の学習を行った場合の２つの手法を提案し，それぞれを対象とした実験を行う．それぞれにおいて，学習したモデルを用いてリプライとニュースを表す文章の文章ベクトルを生成し，二つの文章の類似度を求めることで，リプライから関係するニュースの予測を行う．また，学習したモデルによって生成された文章ベクトルを t-SNE[8] を用いて二次元に圧縮することで可視化し分析を行う．

提案手法に対する比較手法として，BERT を用いてリプライを入力とし，正解ラベルをニュースとする分類を行う．

3.2 ニュース-SNS 投稿の学習

ここでは Twitter 上に投稿されたニュース記事の見出し文章とそれに対するリプライを入力として用いて Sentence-BERT の学習を行った場合について説明する．ニュースとリプライの関係の学習を行うため，Sentence-BERT の入力として Twitter に投稿されたニュース記事の見出し部分と，それに対するリプライをペアとして文章間の関係の学習を行う．Twitter 上で投稿間の関係がリプライとして明示されている二つの文章の文章ベクトルがベクトル空間上で近くなるよう学習を行うことで，ニュースの内容を考慮したリプライの文書ベクトル生成を行い，リプライと各ニュースの類似度を計算することでニュース間の関係性について調査を行う．

3.3 コメント同士の学習

ニュースに対するリプライ同士の関係の学習を行うため，Sentence-BERT の入力として二つのリプライを用いて文章間の関係の学習を行う．同一ニュースに対して行われたリプライ同士には関係があるとし，リプライ元のニュースが異なるリプライ間には関係がないとして，リプライの関係性の学習を行う．学習したモデルを用いて文書ベクトルの生成を行い，リプライと各ニュースの類似度を計算することでニュースとリプライの分類を行う．この手法では，各ニュースの社会的な反応という観点からニュースの関係性について調査を行う．

4. 評価実験

4.1 評価実験の概要

提案手法を用いて，入力に用いた 2 文章間の関係性の訓練を行うことでモデルを構築した．学習したモデルを用いてニュース記事タイトルとリプライの文書ベクトルを生成し，リプライと各ニュース間の類似度を計算することで分類を行った．分類では類似度に対して閾値を指定し，二文間の関係の有無について判定を行った．また，リプライに対して各ニュースとの類似度を求め，類似度の Top_k 内に正解ニュースが含まれるか否かを求めることで提案手法によるリプライの分類精度を調査した．Sentence-BERT によって生成された 768 次元の文章ベクトルを，t- SNE を用いて次元削減を行うことで二次元の散布図に可視化し，各ニュース間の関係について分析した．

また，提案手法との比較として BERT を用いた分類を行った．リプライを入力とし，投稿の内容から関係するニュースの予測を行うことで，提案手法との比較を行った．

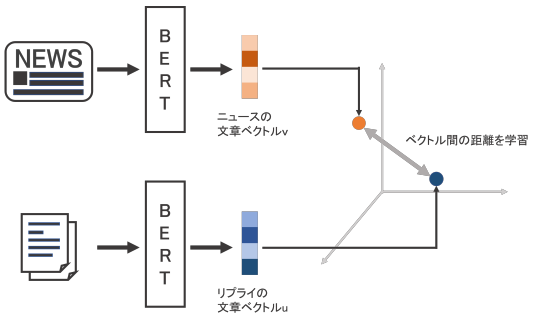

図 2 提案手法の概要（ニュース-リプライの学習）

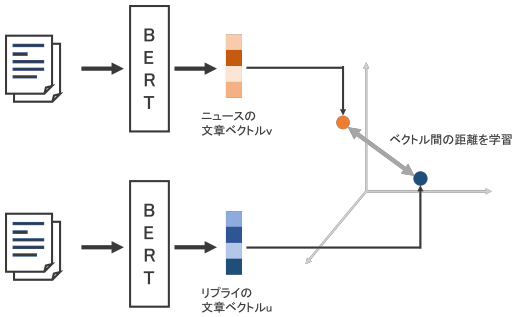

図 3 提案手法の概要（リプライ同士の学習）

4.2 データセット

本研究では代表的な SNS の一つである Twitter を用いた．Twitter API を利用して，投稿されたニュース記事 16 件と各ニュースに対するリプライの取得を行った．図 4，5 に示すように，取得したニュース記事とリプライについて，Twitter 上でリプライとして関係が明示されたニュースとリプライには関係が存在するとし，リプライとして関係が明示されていない投稿間には関係が存在しないとしてラベル付けを行い，データセットを作成した．データセットの正例と負例の数は 1:1 とし，合計 14282 件のデータセットを作成した．

4.3 ニュース-リプライ

各ニュースとそれに対するリプライについて 2 文間に関係が存在し，ニュースと他のニュースに対するリプライには関係が存在しないとして，Sentence-BERT の入力として用いるデータセットを作成した．このデータセットを用い

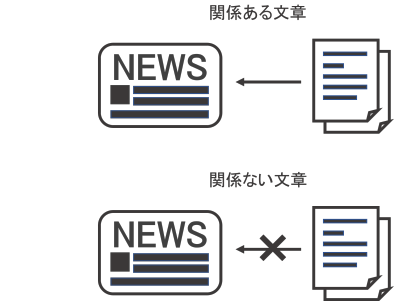

図 4 データセットの作り方（ニュース-リプライ）

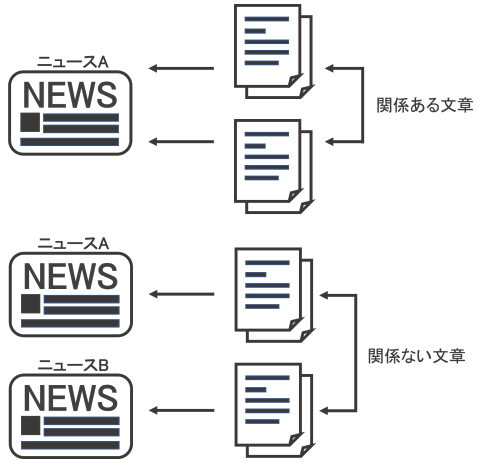

図 5 データセットの作り方（リプライ-リプライ）

表 1 データセットに用いたニュース一覧

ニュース記事見出し
岸田首相は当初は積極的に取材に応じる姿勢を示していたが...
先進 7 カ国 7 は 49 日オンラインで首脳会議を開きロシア産石油...
日経平均株価は先週末の終値より 684 円安い 2 万 6319 円 34 銭で...
新型コロナウイルスの水際対策について緩和すべきと思う人...
財務省は 10 日税収で将来返済する必要がある国の借金長期債...
新型コロナウイルスについて東京都は 8 日新たに 4711 人の感...
東京都医師会は 101 日午後に行われた定例会見で段階的に感染...
小池氏、今度は「神宮外苑再開発」で樹齢 100 年の樹木を約 1000...
ロシアの国連代表部に勤務してきた露外交官、ボリス...
近年、「体育座り」は子供の体に負担がかかるとの指摘があり...
マイナンバーカードと健康保険証の機能を併せ持つ「マイナ保険...
国民審査 “在外の日本人が投票できないのは違憲” 最高裁判決
東京都は、一戸建て住宅を含む都内の新築建物に、太陽光パネル...
「他意はなく、不用意な発言だった。心からおわびしたい」と...
性行為伴うＡＶ禁止する法制定を別途検討　立憲民主党が方針
【誤送金問題】「約 4299 万円を法的に確保した」阿武町長が会見...

てニュース記事とリプライの関係について学習を行い，学習したモデルを用いてリプライとニュース記事タイトルの文書ベクトルを生成した．リプライと各ニュース記事タイトルとの類似度を計算し，2 文間の関係性の有無の判定とリプライとニュースの分類を行い，t-SNE を用いて文書ベクトルを 2 次元にすることで可視化を行った．

4.4 リプライ同士

同一のニュース記事に対するリプライ間には関係があるとし、異なるニュースに対するリプライ間には関係が存在しないとしてデータセットを作成した．作成したデータセットを用いて各ニュースに対する Twitter 上の反応に関してモデルの学習を行い，学習したモデルを用いてリプライとニュース記事タイトルの文書ベクトルを生成した．リプライと各ニュース記事タイトルとの類似度を計算し，2 文間の関係性の有無の判定とリプライとニュースの分類を行い，t- SNE を用いて文書ベクトルを 2 次元にすることで可視化を行った．

4.5 結果

結果を表 2，図 6，7 に示す．文章間の関係性の判定では，求めた 2 文間の類似度に対して閾値を設定し，2 文間の関係の有無を判定した．結果から，ニュース記事の見出しとリプライの関係を学習したモデルが最も精度が高く，リプライ同士の関係を学習したモデルにおいても一定の精度が得られたことが分かる．

リプライに関係するニュースの予測では，リプライに対して各ニュースとの類似度を求め，その上位数件に正解のニュースが存在するかを調査した．図 7 を見ると，ニュース-リプライの学習を行ったモデルでは正解ニュースとの類似度が高く予測できていることが分かる．一方で，リプライ同士の学習を行ったモデルではそれほど正確な予測が出来ていない．

続いて，入力に用いた文章間の関係を学習したモデルによる文書ベクトルを散布図に可視化した図について確認する．Sentence-BERT 及び BERT による文書ベクトルは 768 次元で表現されており，これを散布図にプロットするにあたって t-SNE を用いて 2 次元にベクトルを変換した．また，リプライとニュースの分類を行った BERT による文書ベクトルは最終層の CLS トークン部分を用いた．

図 8〜10 は各ニュースに対するリプライの文書ベクトルを t-SNE を用いて 2 次元の散布図にしたものである．図 8 から，ニュース記事の見出し文章とそれに対するリプライの関係を学習したモデルでは，内容の類似したニュースに対するリプライの文書ベクトルが近くに現れており，投稿の文書ベクトルにニュースの内容を含んでいるといえる．一方で，図 9 を見ると，リプライ同士の関係を学習したモデルによる文書ベクトルでは，ニュースの内容は考慮

されておらず，ニュースに対する純粋な SNS 上の反応の類似性を表している．図 10 の BERT による文書ベクトルでは，Sentence-BERT を用いた場合と比べ，ニュースごとに SNS 投稿の文書ベクトルの分布が固まっていることが分かる一方，ニュース間の関係やコメントの類似性は考慮されていないといえる．

次に，ニュース間の関係の強弱を調べるため，各ニュースに対するリプライの文書ベクトルからそれぞれのニュースに対するリプライの重心を求めた．それぞれのニュースの重心間のコサイン類似度を求めることで，ニュース間の関係性の強弱を調査した．表 3，4 はニュース-リプライの学習を行ったモデルとリプライ同士の学習を行ったモデルのニュース間の関係を求めた結果の例である．

結果から，それぞれのニュースに対するリプライの文書ベクトルの重心によるニュース間の関係は似た内容となっていることが分かる．

4.6 考察

実験結果より，提案手法によりニュースの内容を考慮した文書ベクトルとニュースに対する SNS 上の反応を特徴を表す文書ベクトルを生成できたと考える．一方で，リプライの文書ベクトルからニュースを予測するタスクでは，リプライ同士の学習を行ったモデルでは精度が低い結果となった．これは，リプライにはニュースの内容が含まれているとは限らず，生成した文書ベクトルにはニュースの内容が含まれていないためだと考えられる．また，表 3，4 に示したように，ニュースの予測ではなく関係性の強弱を求める場合では 2 つのモデルで似た結果となっていることから，それぞれのモデルにおいてニュースごとの特徴を学習できていると考えられる．

表 2 BERT によるリプライのニュース分類

	val acc	test acc
BERT	0.734375	0.7058

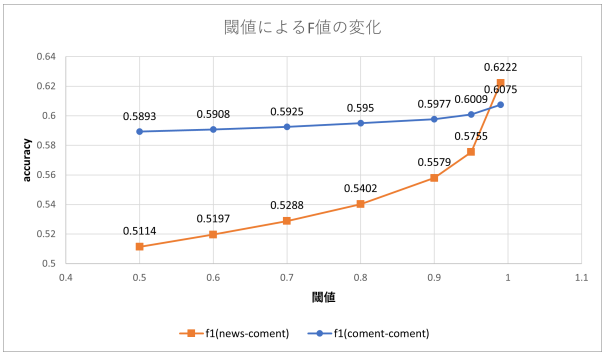

図 6 二つの文章間の類似度による関係性有無の判定

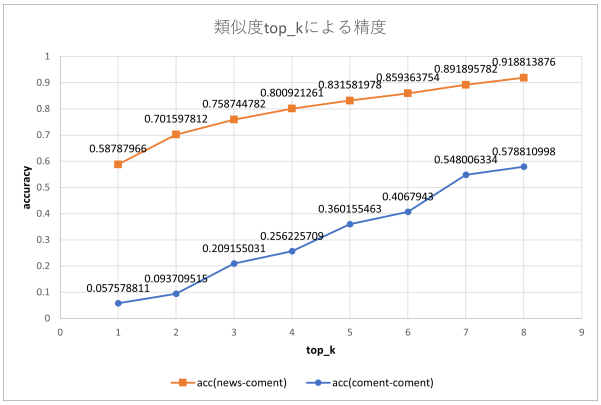

図 7 リプライとニュースの予測精度

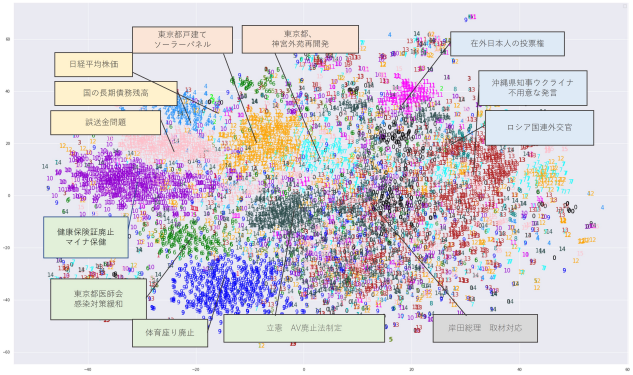

図 8 文章ベクトルの可視化（ニュース-リプライ）

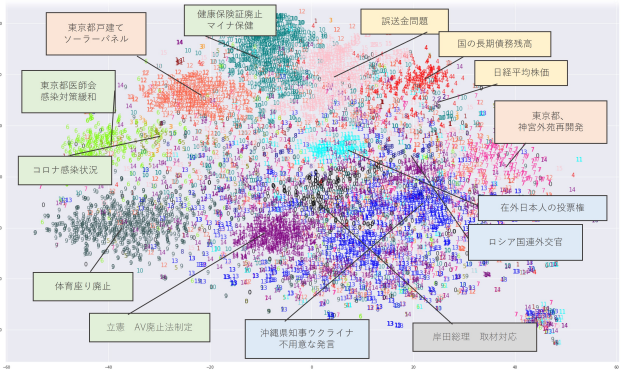

図 9 文章ベクトルの可視化（リプライ-リプライ）

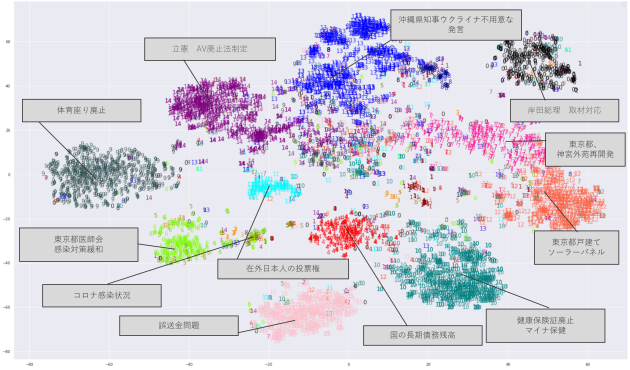

図 10 文章ベクトルの可視化（BERT）

表 3 類似度の最も高いニュースの例（ニュース-リプライ）

ニュース	類似度の最も高いニュース
岸田首相は当初は積極的に取材に応じる姿勢を示していたが...	ロシアの国連代表部に勤務してきた露外交官、ボリス...
先進 7 カ国 7 は 49 日オンラインで首脳会議を開きロシア産石油...	財務省は 10 日税収で将来返済する必要がある国の借金長期債...
日経平均株価は先週末の終値より 684 円安い 2 万 6319 円 34 銭...	財務省は 10 日税収で将来返済する必要がある国の借金長期債...
新型コロナウイルスの水際対策について緩和すべきと思う人...	東京都は、一戸建て住宅を含む都内の新築建物に、太陽光パネ...
新型コロナウイルスについて東京都は 8 日新たに 4711 人の感...	東京都医師会は 101 日午後に行われた定例会見で段階的に感染...
小池氏、今度は「神宮外苑再開発」で樹齢 100 年の樹木を...	ロシアの国連代表部に勤務してきた露外交官、ボリス...
近年、「体育座り」は子供の体に負担がかかるとの指摘があり...	新型コロナウイルスの水際対策について緩和すべきと思う人...
マイナンバーカードと健康保険証の機能を併せ持つ「マイナ保険...	【誤送金問題】「約 4299 万円を法的に確保した」阿武町長が会...
国民審査 “在外の日本人が投票できないのは違憲” 最高裁判決	小池氏、今度は「神宮外苑再開発」で樹齢 100 年の樹木を...
「他意はなく、不用意な発言だった。心からおわびしたい」と...	性行為伴うＡＶ禁止する法制定を別途検討　立憲民主党が方針

表 4 類似度の最も高いニュースの例（リプライ-リプライ）

ニュース	類似度の最も高いニュース
岸田首相は当初は積極的に取材に応じる姿勢を示していたが...	「他意はなく、不用意な発言だった。心からおわびしたい」と...
先進 7 カ国 7 は 49 日オンラインで首脳会議を開きロシア産石油...	日経平均株価は先週末の終値より 684 円安い 2 万 6319 円 34 銭...
新型コロナウイルスの水際対策について緩和すべきと思う人...	マイナンバーカードと健康保険証の機能を併せ持つ「マイナ保険...
財務省は 10 日税収で将来返済する必要がある国の借金長期債...	東京都は、一戸建て住宅を含む都内の新築建物に、太陽光パネ...
新型コロナウイルスについて東京都は 8 日新たに 4711 人の感...	東京都医師会は 101 日午後に行われた定例会見で段階的に感染...
小池氏、今度は「神宮外苑再開発」で樹齢 100 年の樹木を...	ロシアの国連代表部に勤務してきた露外交官、ボリス...
近年、「体育座り」は子供の体に負担がかかるとの指摘があり...	新型コロナウイルスについて東京都は 8 日新たに 4711 人の感...
マイナンバーカードと健康保険証の機能を併せ持つ「マイナ保険...	【誤送金問題】「約 4299 万円を法的に確保した」阿武町長が会...
国民審査 “在外の日本人が投票できないのは違憲” 最高裁判決	小池氏、今度は「神宮外苑再開発」で樹齢 100 年の樹木を...
「他意はなく、不用意な発言だった。心からおわびしたい」と...	性行為伴うＡＶ禁止する法制定を別途検討　立憲民主党が方針
【誤送金問題】「約 4299 万円を法的に確保した」阿武町長が見...	財務省は 10 日税収で将来返済する必要がある国の借金長期債...

5. おわりに

本論文では，ニュース記事の内容を考慮した文書ベクトルの生成と，ニュースに対する SNS 上の反応の特徴を埋め込んだ文書ベクトルを生成する手法を提案した．実験の結果から，Sentence-BERT を用いて，ニュースと SNS 投稿の関係性を学習することで，SNS 上に投稿されたニュース記事の特徴を，SNS 投稿を用いて間接的に抽出可能であることが示された．今後の課題として，生成された文書ベクトルから，各ニュース同士の境界における SNS 投稿の内容の確認や，代表的な SNS 投稿の抽出を行うことで，より良いニュースの理解支援の実現を目指す．また，Sentence-BERT を用いた手法では，入力された文章のベクトルから，類似度を計算することで関連性の有無の判定を行う．そのため，学習に用いていないニュースに対しても対応できる可能性があると考えられる．今後は，より大規模な学習を行った際の，未知のニュースに対する文書ベクトルの生成と分類精度などの調査を行いたいと考えている。

謝辞

本研究は JSPS 科研費 19H04219 の助成を受けたものです.

参考文献

[1] 新聞通信調査会：第 13 回メディアに関する全国世論調査.
[2] NTT ドコモモバイル社会研究所：2021 年一般向けモバイル動向調査.
[3] 白松 俊神谷 晃：市民協働のための Web 記事上の社会問題の自動タグ付けと関連事例抽出手法，人工知能学会全国大会論文集 JSAI2020.
[4] Devlin Jacob, Chang Ming-Wei, L. K. T. K.: Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
[5] 小野 真吾大倉 俊平：関連記事判定のためのニュース記事キーフレーズ抽出.
[6] Williams Ronald J, Z. D.: A learning algorithm for continually running fully recurrent neural networks, *Neural computation*, Vol. 1, No. 2, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, pp. 270–280 (1989).
[7] Nils Reimers, I. G.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
[8] Laurens van der Maaten, G. H.: Viualizing data using t-SNE, *Journal of Machine Learning Research 9(2605)*, pp. 2579–2605 (2008).