DEIM Forum 2022 B24-1

ニュースの理解支援のための「エアリプ」ツイート抽出

高橋　陸† 牛尼　剛聡††

† 九州大学芸術工学部芸術情報設計学科　〒 815-8540 福岡県福岡市南区塩原４丁目９－１
†† 九州大学大学院芸術工学研究院　〒 815-8540 福岡県福岡市南区塩原４丁目９－１
E-mail: †takahashi.riku.110@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　近年，SNS 利用者の増加に伴い，SNS を情報収集の場として活用する人が増えている．一般的に，ユーザが SNS を用いて報道されたニュースに関する情報を収集する場合，対象とするニュース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ことが一般的である．キーワード検索では，検索に用いた単語を含む SNS 上の投稿を取得可能であるが，対象とするニュースに関する代表的なキーワードは複数存在することが多い．そのため，ユーザがキーワード検索のみで，ニュース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，代表的な SNS の一つである Twitter におけるリプライ機能に注目し，投稿されたニュース記事と，それに対するリプライを教師データとして，ニュースに関する意見や反応を表すツイートを発見するための機械学習モデルを構築するための手法を提案する．本手法では，ニュース記事とそれに対するリプライという，必ずしもニュースに関する代表的なキーワードを含まない投稿とニュースの関係を用いて機械学習モデルを訓練する．このモデルを用いることで，特定のニュースに対する直接的な反応ではない，ニュースに対する暗黙的なリプライツイートを抽出し，ニュースの理解支援を実現する．

キーワード　SNS, Twitter, 自然言語処理, 機械学習, 反応ツイート

1　はじめに

1.1　SNS 利用について

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，ツイッターやインスタグラムなど，多くの SNS 上で様々なトピックに関しての多くの投稿が行われている．SNS 上では，登録したユーザがいつでも気軽に文章を投稿でき，投稿には，ユーザ個人の考えが反映されることが多い．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．

一方，SNS は情報収集の場としても活用されている．総務省による「令和 2 年度情報通信メディアの利用時間と情報行動に関する調査報告書」[1] によると，各年代における情報源としての重要度の調査でも 10 代から 60 代の 77.3 %がインターネットが情報源として重要と回答している．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」[2] 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．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．

また，総務省による「ICT によるインクルージョンの実現に関する調査研究」[3] によると，SNS 別利用状況では，LINE を除いた代表的な SNS において，自ら発信を行うよりも，他者の発言や書き込みの閲覧を好んで行っている．以上のことから，現在 SNS は重要な情報源として広く利用されている．

1.2　SNS を用いた情報収集における問題点

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，多くの情報を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．

SNS では，ユーザが興味を持つトピックをピックアップし，ユーザに提示するトレンド機能と呼ばれる機能が存在する．トレンド機能では，ユーザが SNS を利用している時点において，多くのユーザが注目しているトピックを提示することで，ユーザはリアルタイムに話題となっているトピックに関する投稿を見ることが出来る．しかし、トレンド機能によってユーザに提示されるトピックはその時点で話題になっているごく一部のものに限られる．そのため，トレンド機能でピックアップされない大多数のトピックに関して，ユーザが情報収集を行うことは難しい．

他にも，ユーザが SNS を用いて情報収集を行う際に十分な情報を得られない理由として,「エアリプ」の存在があげられる．図 1 に示すように,「エアリプ」とは一般的には特定の相手に対して返信する意図がありながら，相手に対するリプライではな

くシステム上つながりのない通常の投稿を行う行為を指す．主に，個人間のやり取りの中で，意図して返信ではなく「エアリプ」を用いる場合が多い．話題になっているニュースに関する投稿では，話題としている内容に関する単語の省略などが行われることが多い．そのため検索に用いた単語を含む投稿を抽出するキーワード検索では，こうしたニュースに対する「エアリプ」を抽出することが難しい．

図 1　リプライとエアリプ

1.3　研 究 目 的

英語などの言語と異なり，日本語では文脈から話題にしている対象が明らかである場合には，文中の主語や目的語を省略することが頻繁に行われる [4]．そのため，日本語は文脈の流れで言いたい内容を慮ることが多く，ハイコンテクストな言語とされる．したがって，日本語で書かれた SNS 投稿は，キーワード検索では見つけることが困難な，所謂「エアリプ」的な投稿が多く存在する．また,「エアリプ」的な投稿は口語的な文体で投稿されることが多く，投稿者の率直な意見が述べられていることも多い．本論文では，キーワード検索では抽出することが難しい，ニュースに対する反応を表す投稿を，ニュースと関連する情報として抽出することで，ニュースに対する理解支援とすることを目的とする．

本研究の貢献は，SNS における「エアリプ」の役割に注目し，機械学習を用いて明示的に関係が示されていない投稿と話題を関連付け，ユーザが利用可能とする手法を提案し，その有効性を実験により評価したことである．

1.4　本論文の構成

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，本研究で利用した Twitter におけるデータ構造と特徴について述べる．第 4 章では，提案手法について述べる．第 5 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，実験結果と考察について述べる．第 6 章では，まとめと今後の研究予定について述べる．

2　関 連 研 究

これまでにも，ニュースや SNS を対象として，単純な検索では取得が困難な情報を, 機械学習を利用して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　ニュースに関連する情報の抽出

神谷 [5] 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 [6] 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とが示されている．

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，本研究の提案手法では 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　SNS からニュースに関する情報の抽出

牧野 [7] [8] らは，Twitter からニュースに関連する情報を抽出するための手法を提案している．この手法では，ランダムに取得したツイートに対して LSTM [9] を用いてニュース性の有無を判別し，その内容がニュースの第一報となる非既出のツイートか，ニュースの続報や意見などを述べた既出ツイートであるかを分類をする．この研究では，ニュースに対する反応ではなく，ニュース性を有するツイート抽出することを目的としている点が本研究と異なる．また，この手法では，報道現場で抽出されたツイートを教師データとして訓練を行うため，少量のデータでの分類は考慮されていない．

3　SNS におけるニュースとその反応

3.1　Twitter の構造

本研究では代表的な SNS である Twitter を対象とする．Twitter では，ユーザは全角 140 文字，半角 280 文字以内の文章を投稿できる．Twitter では，ユーザが投稿した文章をツイートと呼び，ある投稿に対して返信としてツイートされた投稿をリプライと呼ぶ．図 1 に示したように，リプライは対象となる投稿にぶら下がる構造となる．そのため，リプライの対象となった投稿を明確に取得可能である．

また，Twitter にはリツイートと呼ばれる他者に特定のツイートを拡散するための機能が存在する．リツイートでは，拡散したい投稿のみをリツイートする場合と，引用リツイートと呼ばれる拡散したいツイートにユーザがコメントをつけて投稿する方法の 2 つが存在する．

3.2　Twitter におけるニュース記事

Twitter では，多くの報道機関がニュース記事の投稿を行っている．ニュース記事の投稿は，それぞれの報道機関が独立し

て行っているため，同じ内容のニュースが複数のニュース記事として投稿される場合が多い．そのため，ユーザが，あるニュースに対する意見や反応などの情報を探す際，一つのニュース記事投稿では十分な情報を得ることが困難であり，同じ内容のニュース記事を複数調べる必要がある．

3.3 ニュース記事への反応

Twitter に投稿されたニュース記事に対するユーザの反応は，主にニュース記事へのリプライという形で存在する．ニュース記事は同一の内容であっても複数存在することが多いため，ニュース記事に対するリプライも分散し，一つのニュース記事投稿に対するリプライ数というのはそれほど多くない．

また，ニュース記事に対する直接的なリプライではなく，リツイート後に通常のツイートでニュースの内容に対する投稿を行う場合や，引用リツイートを用いてニュースに対してコメントを行う場合もある．このようなニュースへの反応はキーワード検索やニュース記事に対するリプライでは発見することができない．

本研究で利用する 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関して，その投稿数を調査した結果を図 2 に示す．この結果から，投稿直後のリプライが最も多く，時間の経過に伴って徐々に減少していき，約 6 時間程度でほぼ 0 となることがわかる．その後，数時間おきに数件のリプライが行われるものの，ニュース記事が話題となるのは投稿直後から数時間の短い時間である．このことから，一般的なニュースにおいて話題となる時間は短く，ニュース記事に対するリプライを多く得ることは難しいことがわかる．

3.4 ニュース記事に対する暗黙的なリプライ

Twitter で用いられる用語として「エアリプ」という用語がある．エアリプとは，エアリプライを短縮したものであり，リプライのように特定の投稿や話題に対する反応を含む投稿であるが，リプライのように元となる投稿との間に明示的な関係を示しておらず，一般的なツイートして投稿されたものを指す.「エアリプ」は一般的に個人に対する返信などで用いられることが多いが、ニュース記事に対する暗黙的なリプライに関してもエアリプと同様のものであると考えられる．本論文では，ニュースに対する暗黙的なリプライを「エアリプ」と捉える．「エアリプ」では何かしらの投稿に対して反応しているため，文章内で主語や目的語などの単語が省略されることも多く，また明示的なリプライのように元の投稿との関係性が明示されてないことから人手によるキーワード検索を用いた網羅的な発見は困難である．

4 提案手法

4.1 概　　要

本研究で提案する手法の概念図を図 3 に示す．本研究では，Twitter において，キーワード検索では抽出困難なニュースに対する暗黙的なリプライを抽出するための手法として，機械学習を用いる．ニュースとそれに反応したツイートの関係を教師あり学習を行い機械学習モデルに訓練することで，ニュースに対する暗黙的なリプライを抽出する．

ニュースに対する暗黙的なリプライを機械学習を用いて抽出するため，本手法ではツイッターの明示的なリプライに注目する．明示的なリプライでは，返信先となるツイートが存在する．そのため，ニュース記事に対するリプライでは，話題にしているニュース記事が明確に示されている．明示的なリプライと暗黙的なリプライでは，返信先を明示的に示すか否かという点以外の特徴は類似している．この性質を利用し，ツイッター上に多数存在するニュース記事の投稿と，それに対するリプライの特徴を利用することで，ニュースに関連する代表的な単語を含まないが，対象とするニュースに関係する内容のツイートを機械学習によって発見する．

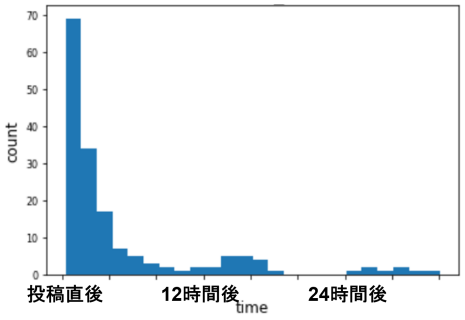

図 2　投稿されたニュース記事に対するリプライ数の推移の一例（縦軸：tweet 数、横軸：時刻）

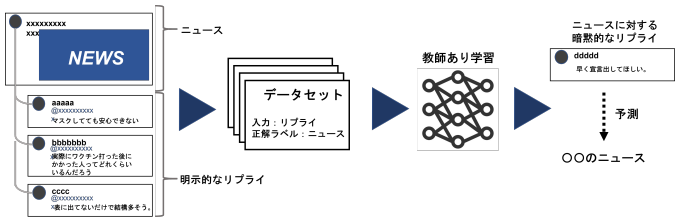

図 3　提案手法の概念図

4.2 提 案 手 法

本研究での提案手法を説明する．

SNS 上に投稿されたニュースの集合とツイートの集合を，それぞれ

$$NW = \{nw_1, nw_2, \cdots, nw_{|NW|}\} \tag{1}$$

$$TW = \{tw_1, tw_2, \cdots, tw_{|TW|}\} \tag{2}$$

とする．このとき，ツイート tw が，参照しているニュースを $news(tw)$ と表記し，ツイート tw が別のツイートに対して明示的に行われたリプライであるとき，ツイート tw のリプライ先のツイートを $reply(tw)$ と表記する．

いま，ツイート tw が与えられたとき，そのツイートがニュース nw に関連した内容である場合，その確率を $P(nw|tw)$ と表すものとする．ここで，確率 $P(nw|tw)$ は以下を満たすものとする．

$$\sum_{nw \in NW} P(nw|tw) = 1 \tag{3}$$

また，ニュースの集合 NW の要素には，どのニュースも表さないニュース NaN があるとする．ここで，確率 $P(\mathrm{NaN}|tw)$ はツイート tw がどのニュースにも関連しないと考えられる確率である．

本研究では，ツイート tw がニュース nw に関連する確率 $P(nw|tw)$ を予測するための手法として機械学習モデル M を利用し教師あり学習を行う．いま，機械学習モデル M を訓練するための教師データを (x, y) と表す．ここで，x は機械学習モデル M への入力を表し，y は入力 x に対する正解ラベル表す．

本研究では，教師データセット TS を以下のように定義する．

$$TS = \{(tw, news(reply(tw)))|tw \in TW\} \tag{4}$$

ここで，$news(reply(tw))$ はツイート tw がリプライした先のツイートが表すニュースを指す．このデータセット TS は，機械学習モデル M が，ツイートを入力としてそのツイートが言及するニュースを予測するため訓練に利用する教師データである．

5 評 価 実 験

本研究では，提案手法の有効性を評価するため，以下の実験を行った．

- ニュースに対するエアリプの分類
- データ数による暗黙的なリプライの分類精度の比較

実験において，機械学習モデルの訓練に用いるデータは，Twitter API を用いて収集をした．収集したデータは，日本の報道機関がツイッター上に投稿したニュース記事のツイートから，そのツイートに対して行われたリプライを取得した．また，多くのデータを教師データとして収集するため，Twitter API を用いてデータを取得した時点において，Twitter のトレンド機能によってピックアップされた，その時点で日本で話題となっているニュースを対象とした．

5.1 ニュースに対するエアリプの分類、モデルごとの精度

この実験では，提案手法によって，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を教師データとして用いて，ニュースに関する明示的な参照を有さないツイートをニュースと関連付けられるかを評価する．

本実験では，文章のベクトル化に関して，文中に出現する単語の有無を表現する one-hot ベクトルを用いた場合と単語分散表現の一手法である Word2vec [10] を用いた場合の分類精度を比較する．また，複数の機械学習モデルによる分類精度を比較し，ニュース記事に対する「エアリプ」の抽出に適した機械学習モデルを明らかにすることを目的とする．

本実験において，機械学習モデルの入力として用いる教師データは，スポーツ，政治，感染症などジャンルの異なるニュース記事とそれに対するリプライを使用した．

この実験の流れを図 4 に示す．Twitter に投稿されたニュース記事と，それに対するリプライを入力データとして収集を行った．各モデルで同じ入力データを用いて，ニュース記事を正解ラベルとし，ニュース記事に対して行われたリプライを入力データとして学習を行い，CNN [11]，RNN [12]，LSTM [9]，BERT [6] という 4 種類の機械学習を用いてツイートの分類を行った．そして，各モデルごとに，ニュースに対するツイートのニュース記事をカテゴリとした分類精度の比較を行った．

5.1.1 データの前処理

データの前処理として，入力データとして用いるリプライツイートに対して，メンション，ハッシュタグ，URL の除去を行い，ツイート本文のみとした．次に，ツイート本文に対して，分かち書きを行い，文を構成要素に分割した．

次に，機械学習モデルに入力するため，分かち書きを行った文をベクトル表現に変換した．CNN，RNN，LSTM を用いた分類では，単語を one-hot ベクトルに変換した．また，LSTM では学習済み単語分散表現 Word2Vec を用いたものも対象とした．本実験で使用した Word2vec では，コーパスとして日本語版 Wikipedia を使用した．それぞれの機械学習モデルに，全結合層にソフトマックス関数を組み合わせることで各クラスの分類確率を予測した．

BERT では東北大学の乾研究室が公開している事前学習済みのモデル [13] を使用した．BERT の事前学習には日本語 Wikipedia のデータを使用した．モデルへの入力として，他のモデルと同様の前処理を行った後に入力文章のベクトル化を行った．

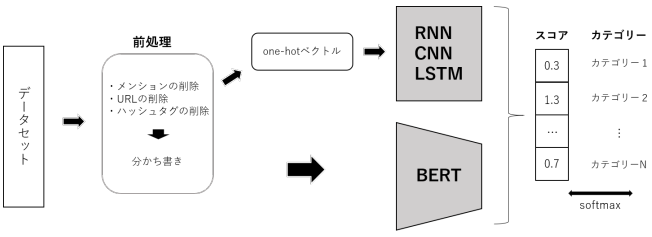

図 4 ニュースに対するエアリプの分類の流れ

5.1.2 データセット

使用したデータの概要を表 1 に示す．教師データとして，正解ラベルをニュース記事として各ラベル 150 件，計 750 件のリプライデータの収集を行った．また，ニュースに関係ないツイートの判別も行うために，ニュースとの明確な関連性のないツイートを，収集を行った際の投稿順に 450 件取得した．表 1 にあるように，分類に用いたニュース記事は環境問題，スポーツ，政治などのジャンルが異なるニュースを用いてツイートの収集を行った．学習の際．データの順番をランダムにシャッフルした後，データセットを訓練データと検証データが 8:2 になるように分割を行い実験した．

5.1.3 実 験 結 果

実験結果を表 2 に示す．テストデータを用いた予測を行う際には，検証データの損失が最も低かった状態のモデルを用いた．

実験結果から，BERT による分類の精度が最も高く，続いて Word2Vec を用いた LSTM の精度が高かった．RNN を用いたニュースに対する暗黙的なリプライの分類では，分類精度が 0.6218 であり，あまり精度が良くない．また，その他のモデルである Word2vec を用いていない LSTM，CNN ともに精度は 0.8 前半であり，BERT による分類の精度 0.91 が他の手法よりも大幅に精度が良かった．

BERT による入力文章とニュースの予測の結果を表 3 に示す．入力文章と，正解ラベルのニュースに対する予測スコアから，ワクチンや給付金といったニュースに関連する代表的なキーワードを含まないツイートに対しても高精度で予測できていることがわかる．

5.1.4 考　　察

本実験で使用したデータセットでは，スポーツ，政治，感染症など，正解ラベルとしたニュース記事の内容が明確なものを使用した．そのため異なるニュース間で類似した単語や内容が少なく，ラベルの特徴が分かりやすいため分類が容易であったと考えられる．その結果，ニュースを代表するような単語を含まないツイートであっても，明らかに内容が異なるニュースに関しては，他のニュースと混同することなく，ニュース記事とそれに対するリプライの関係から高い精度で分類出来たと考えられる．また，過去の入力から重要な情報を記憶することで予測に用いる LSTM，語順を考慮した分類を行う CNN と比較して，BERT は高い精度を出している．このことから，ニュースに対する暗黙的なリプライの分類において，文脈を考慮できるという BERT の特徴は有効であると考えられる．

表 1　使用したデータセット

ラベル	件数
ニュースと関連のないツイート	450
ワクチン	150
寄付金	150
スポーツ	150
水際対策	150
環境問題	150

表 2 「エアリプ」ツイートの分類

model	Accuracy
RNN	0.6218
CNN	0.8384
LSTM	0.8289
LSTM(Word2vec)	0.8446
BERT	**0.9173**

5.2 データ数毎におけるニュースに対するエアリプの分類

5.2.1 実験の概要

本実験の流れを図 5 に示す．本実験では，学習に用いる教師データのデータ数を変化させることで，少量のデータを入力とした場合のテキスト分類の精度を比較する．

この実験の背景には，ニュースが話題になる期間が関係している．一般的に，機械学習モデルを用いたテキスト分類タスクでは，大量のデータが教師データとして必要となる．しかし，Twitter 上に投稿されたニュース記事に対するリプライでは，ニュース記事の投稿直後の反応が最も多く，時間の経過とともに減少するため，ニュースが話題となる期間は短い．そのため，分類を行うための教師データを集めることが難しいという問題がある．そこで，十分に教師データを集めることが困難な状況における，分類性能を調査するため，訓練に用いる教師データ数を変化させ，異なるデータ数での分類の精度を調べた．使用する機械学習モデルは，5.1 に示したニュースに対するエアリプの分類の結果から，精度の良かった BERT 及び Word2Vec を用いた LSTM を使用した．

また、内容が類似したニュースに対するエアリプの分類が可能かどうかを調査するため，教師データとして類似した単語が出現するニュース記事とそれに対するリプライを用いた．これにより，SNS 上に多く存在する，スポーツや政治などのジャンル内での内容の類似したニュース記事に対する分類についても調査する．

データの前処理は実験 1 と同様に行い，ツイートに含まれるメンション，ハッシュタグ，URL の除去を行いツイートの本文のみを使用した．

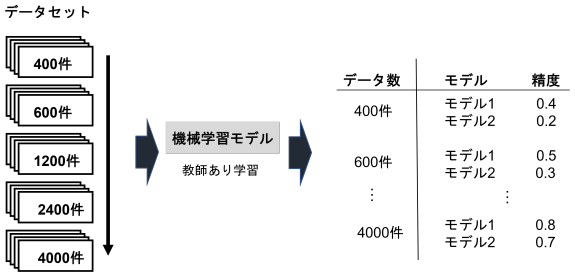

図 5　データ数を変えた暗黙的なリプライ分類の流れ

表 3　BERT による分類結果

正解ニュース	予測スコア	ツイート
ニュースと	0.998917	よっしゃ仲間出来たバイト中かい?早く夜ご飯たらふく食べようぜ
関連のない	0.998915	今日は 1 人ぼっち、常にボッチやった
ツイート	0.998832	全然話を聞けない、立ち歩き、提出物そろわない... 疲れた。
	0.998906	今日は冷たいカフェオレ押したらちゃんと冷たいのがでた.
	0.99884	リゾットを初めて食べたおいしい
ワクチン	0.995248	厚労相は効果があれば、230 万人以上の健康被害は知ったこっちゃないとおっしゃっているんですか?
	0.993073	接種したいけど、不安障害を持っているので、その場に自分がいる事を想像するだけで心がおかしくなってしまう
	0.972196	「有効率 95 %」のカラクリ (打たなくても 99.2 %%発症しない、打ったらちょっと発症しにくくなるけど 8 割に副反応)
給付金	0.998146	とりあえず全員に配って、金持ちからは税金で回収すればいいだろ。
	0.998222	生活困窮者の基準は?世帯年収でも個人年収でもだいぶ変わってくるのでは???そろそろコロナ給付金詐欺を摘発して詐欺すると捕まるぞ ってのを見せつけてよ。もうやってたらすみません。
	0.998188	一律給付か給付しないかの二択しかない。他は全て線引きで揉める。
	0.998293	一律給付なんてしなくていい。絶対反対!国民が堕落するだけ。18 歳以下の給付でよし!
スポーツ	0.994186	FA 権持ちの選手を CS でなんとも微妙な起用法をして FA 権行使までいったらまんま大和選手とおんなじ流れになってしまうんじゃないのかな
	0.994865	宣言したら欲しい球団多そうだな。
	0.994633	まぁ使われないんじゃ仕方ない
	0.994576	関係ないかもだけど FA 取れるこのタイミングでスタメン減ればそりゃFA 使いたくもなるよね
水際対策	0.994741	此処で緩和?今が一番肝心な時期なのに!
	0.995966	しかしまぁ、ほんとに国民の意見を全く聞かない政府だよな....
	0.995611	どうしても 6 波が欲しいんだね... いい加減にしてもらいたい
	0.996465	大丈夫なのかなぁ?水際対策ができてたから感染拡大を抑えれてたと思うのに。諸事情はあるとしてももう少しだけガマンした方が良いと思うけどなぁ。
環境問題	0.994657	何も知らないのになんか良さげ?的なファッション。理系の若者なら終わってる。
	0.934595	こういう反対だけの集まり大嫌い代案だせや。能力がないから、代案出せないんだろう。

5.2.2　使用したデータセット

本実験で使用したデータセットの概要を表 4 に示す．内容的が類似したニュースに対する暗黙的なリプライツイートの分類精度を調べるため，コロナウイルスに関係するニュース中心に，同時期に話題となった，ツイート内に類似した単語が出現する 8 つのニュースに対するリプライを収集した．各正解ラベルにつき入力として用いるツイートが 500 件の計 4000 件の教師ありデータセットを作成し，このデータセットを用いてデータ数と分類精度の関係を調べた．モデルの学習を行う際，データの順番をランダムにシャッフルした後，訓練データと検証データは 8:2 に分割した．また，データ数毎に学習したモデルに対して，同じニュース記事から別に収集した 800 件のテストデータを用いて分類を行うことで同じ条件で比較を行った．

表 4　データセット

ラベル	件数
石油問題	500
マスク	500
給付金	500
オミクロン株	500
ワクチン	500
五輪外交	500
マスクに関する傷害事件	500
入国制限	500

5.2.3　実 験 結 果

実験結果を，表 5 に示す．5.1 の実験と同様に，訓練に用いた教師データ数を 1200 件として学習を行い，ニュースに対するエアリプの分類精度を比較した．Word2vec を用いた LSTM, BERT を用いた，類似した内容のニュースにおけるエアリプの分類では，先に行ったニュースに対するエアリプの分類結果と比較して両モデルとも精度が低かった．

表 6 に教師データの数を変化させた場合のテストデータの分類精度を示す．学習に用いるデータ数を 400 件から 4000 件までの 5 つの異なるデータ数に変化させた場合の分類精度を確認した．BERT はデータ数が少ない場合においても検証データに対する分類精度が大きく低下せず，80 %以上の精度であった．また，テストデータにおける分類では，精度はデータ数の減少に伴い少しずつ下がっていくという結果となった．

一方で，Word2Vec を用いた LSTM を使用した分類では，検証データの精度では一番高い 4000 件の場合でも 0.68 という結果となった．BERT と比較しても，検証データの精度はデータ数の減少に伴い明らかに低下した．また，テストデータの精度は 4000 件では高い精度であるが，2400 件では 0.148 と 75 %以上精度が悪化した．

また、表 7 に BERT による分類において予測と正解が異なる入力文章を示す．類似した内容なニュースに対する「エアリプ」の分類において，正解と異なる予測をした入力文章では，「本物だ」,「賛成です」といった非常に短く，様々な話題で出現

する投稿があった．また，オミクロン株のニュース記事に対してワクチンに関するリプライや，入国制限に関するリプライなど，関連する類似した他のニュースへの意見が投稿されていることが分かった．

表 5　類似したニュースに対するエアリプ分類の比較

model	Accuracy
ジャンルの異なるニュース	
LSTM(Word2vec)	0.8446
BERT	**0.9173**
類似したニュース	
LSTM(Word2vec)	0.14
BERT	**0.44**

表 6　データ数毎の分類精度

model	400	600	1200	2400	4000(データ数)
LSTM(Word2vec)	0.128	0.15	0.14	0.148	**0.625**
BERT	**0.37**	**0.39**	**0.44**	**0.45**	0.47

5.2.4 考　　察

BERT，Word2vec を用いた LSTM ともに分類精度の悪化について考察する．内容が類似したニュースに対するリプライでは，ニュース記事の内容に対して異なる，別の類似したニュースへの意見が投稿されている場合がある．こうした例では，投稿者が適切ではないニュースへ投稿しているため，正解のニュースよりも予測したニュースへ分類したほうが正しいものも存在した．

BERT と Word2Vec を用いた LSTM のテストデータの精度を比較する．実験結果から，データ数 4000 件の場合では LSTM を用いた分類の精度が BERT を用いた場合より優れているが，データ数 2400 件以下では明らかに BERT の精度が LSTM を上回っていることがわかる．また，BERT を用いた分類では，入力データ数の減少に伴い緩やかに精度が悪化しているものの，LSTM と比較して悪化の程度が緩やかであることから，BERT は少量のデータの場合でも安定した精度で分類出来ると言える．

一般的に，Twitter に投稿された一つのニュース記事に対するリプライ数はそれほど多くない．今回の実験で収集したデータでは，一つのニュース記事に対する 100 件以上のリプライの取得が困難であった．データ数を変化させた場合の分類において，入力データ 400 件の場合における各ラベルのデータ数は 50 件であることから，BERT を用いてデータ数 400 件での精度を向上させることで，ニュースの話題の変化，データ数の少なさに対応できると考えられる．

6 ま と め

6.1 実験のまとめ

本論文では，SNS 上に多く存在するニュースに関連する代表的なキーワードを含まないニュースに対するエアリプを抽出する手法を提案した．本手法では，Twitter のリプライ機能に注目し，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から，機械学習を用いることで, ニュースに対するエアリプの分類，抽出を行った．4 つの機械学習モデルを用いて精度の比較を行った結果，BERT を用いることで，高い精度で比較的内容が異なるニュースと，それに対するエアリプの類を行うことができた．

次に，話題の変化の激しいニュースに対応するためより少量のデータを用いた分類精度の調査では，BERT を用いることで，各ラベル 50 件のデータであっても，word2Vec を用いた LSTM と比較しても安定した精度を出せるという結果が得られた．これにより，入力データの数に分類精度が左右されづらい BERT を用いることで，より少量のデータを用いたニュースに対するエアリプの分類が可能であると考えられる．

6.2 今後の課題

本研究の実験結果から，ツイッターのリプライ機能に着目し，ニュース記事とそれに対するリプライの関係を機械学習モデルを用いて教師あり学習を行うことで，SNS 上に数多く存在するキーワード検索では見つけることが困難な，ニュースに対するエアリプを抽出するという本手法の試みは成功したと言える．しかし、少量の教師データにおける分類や，より内容が類似した，近しいニュースに対する分類では精度が低く，課題が残る．

今後，少量の教師データによる分類精度の向上を目指していく．本研究では，入力データにはニュース記事に対するリプライとして投稿されたすべてのツイートを用いているが，こうしたリプライの中には，SNS 上の数多くの投稿で見られる感情を述べた投稿や，ニュース記事の内容に対してあまり関係のない投稿が存在する．こうした投稿をニュースに対する関連度の低いものとして取り除くことで，近しい内容のニュースであってもより精度が高く分類できるのではないかと考えている．また，少量のデータから効率的に学習を行う手法として，Few-shot 学習に注目している．事前にニュースに関するデータを学習させておくことで，新しいニュースを対象とした分類を行う際に，過去に学習したモデルのパラメータから少ないデータで対応できるシステムの構築を目指す．

謝　　辞

本研究は JSPS 科研費 19H04219 の助成を受けたものです．

文　　献

[1] 総務省. 令和 2 年度情報通信メディアの利用時間と情報行動に関する調査報告書＜概要＞. 2021. https://www.soumu.go.jp/

表 7　BERT による分類結果 予測と正解が異なる入力文章

正解ニュース	予測ニュース	ツイート
石油問題	ワクチン	本物だ
石油問題	五輪外交	馬鹿だ馬鹿だ
マスク	マスク傷害事件	日本の法律ってアホな条件加えるからなその内公の場でマスクしてない場合公然罪に該当するとか言うんやで
マスク	マスク傷害事件	ななんだとーやっぱり若者は全員タチの悪い犯罪者じゃないか
給付金	オミクロン株	一周回ってまた戻ってきた感じそれが時間と経費の無駄
給付金	マスク傷害事件	そんないらないから打撃受けてる飲食店とかに回した方が一番と思う
オミクロン株	ワクチン	少しずつ国民がワクチンのポンコツっぷりに気づき始めたようです 3 回目で投与量 500 μに増やして一気に刺しましょう
オミクロン株	入国制限	鎖国だな攘夷だな
オミクロン株	入国制限	国内に感染者入れちゃってから入国させないって言ってももう遅せぇよ
オミクロン株	入国制限	賛成です
オミクロン株	ワクチン	医療従事者だけでも 6ヶ月での接種にしてほしい それで結構な人を守られると思う
ワクチン	五輪外交	渡航者法じゃ俺たちは縛れねぇ
オミクロン株	入国制限	一年前の専門家こびナビ発言のレビューをすべきですねゴールポストずらしすぎ
五輪外交	マスク	ホンマたまーにまともな事言うよなぁ
五輪外交	入国制限	とても正当な主張です素晴らしいですでもくれぐれもくれぐれも昔に戻らないでください それがすごく心配です
入国制限	ワクチン	まだワクチン打たせたいのかほんとに誰かさんの犬みたいだな

main_content/000765135.pdf.

[2] 新聞通信調査会. 第 13 回メディアに関する全国世論調査. 2021. https://www.chosakai.gr.jp/oshirase/●第 13 回メディアに関する全国世論調査（2020 年）報告書.pdf.

[3] 総務省. Ict によるインクルージョンの実現に関する調査研究（2018）. https://www.soumu.go.jp/johotsusintokei/linkdata/h30_03_houkoku.pdf.

[4] 加藤薫. 日本語の構文的特徴から見えてくるもの:「主体・客体」と「自分・相手」. 文化学園大学紀要. Journal of Bunka Gakuen University. 人文・社会科学研究, 第 20 巻, pp. 1–13. 文化学園大学, 2012.

[5] 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, pp. 1C3OS6a03–1C3OS6a03, 2020.

[6] Lee Kenton Toutanova Kristina Devlin Jacob, Chang Ming-Wei. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018.

[7] 宮崎 太郎 後藤 淳牧野 仁宣. ニューラルネットワークを用いた既出 tweet 分類. 言語処理学会　第 24 回年次大会発表論文集 (2018-3), 2018.

[8] 武井 友香 山田 一郎 後藤 淳宮崎 太郎. Twitter からの有用情報抽出のための学習データのマルチクラス化. 情報処理学会研究報告　, 2017.

[9] Schmidhuber Jürgen Hochreiter Sepp. Long short-term memory. In *Neural computation*, Vol. 9, pp. 1735–1780. MIT Press, 1997.

[10] Greg Corrado Jeffrey Dean Tomas Mikolov, Kai Chen. Efficient estimation of word representations in vector space. 2013. https://arxiv.org/pdf/1301.3781v3.pdf.

[11] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, p. 1746–1751., 2014.

[12] Zipser David Williams Ronald J. A learning algorithm for continually running fully recurrent neural networks. In *Neural computation*, Vol. 1, pp. 270–280. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …, 1989.

[13] 東北大学乾研究室. Pretrained japanese bert models / 日本語 bert 訓練済みモデル. https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/.