DEIM2022 E24-5(day2 p29)

ニュースコメントの閲覧支援のための ニュースへの反応に基づくユーザ埋め込み表現の生成

中原　輝樹† 　牛尼　剛聡†

† 九州大学芸術工学部　〒 815-8540 福岡市南区塩原 4-9-1
† 九州大学大学院芸術工学研究院　〒 815-8540 福岡市南区塩原 4-9-1
E-mail: †nakahara.teruki.528@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　SNS やニュースサイトを利用してニュース記事を読む際，ユーザはニュースに対する他者のコメントを読むことができる．ユーザは他者のコメントを読むことにより，ニュースに対する世論を理解でき，ニュースの全体像の把握に役立つことが多い．しかし，コメントの読者は表示されるコメントがどのような特徴を持つユーザから投稿されたのかを把握できないという問題点がある．本論文では，コメントを投稿したユーザの特徴を考慮したコメントの閲覧支援を行うために，過去のニュースへの反応に注目し，コメントを行ったユーザの類似性を予測する手法を提案する．さらに，予測されたユーザの類似性に基づいた SNS 上でのニュースコメント閲覧支援手法を提案する．提案手法では，過去のニュースへの反応からユーザ予測を行う機械学習モデルを利用して，ユーザの埋め込み表現を生成する．評価実験の結果，提案するユーザ埋め込み表現が複数の観点からユーザの特徴を表していることを明らかにした．

キーワード　ユーザ埋め込み表現，ニュースコメント，SNS, Attention，自然言語処理

1　はじめに

近年，インターネット上でニュースを読むことが一般化した．新聞通信調査会によるメディアに関する全国世論調査 [1] では，週に 1 日以上インターネット上でニュースを読むと回答したのは，全世代の 73.1 %に及び，40 代以下では 90 %を超えている．ポータルサイトや，テレビ局・新聞社が運営するニュースサイトだけでなく，SNS(ソーシャル・ネットワーキング・サービス) 上でニュースを読む人々も増えている．特に，10 代・20 代の若い世代では，インターネット上でニュースを読む際にアクセスするサイトとして，SNS が最も多く利用されている [1]．

日本における代表的なニュースサイトの一つである Yahoo! ニュース [1]では，ニュース記事毎に，ニュースに対するコメントを投稿することができるコメント欄が設けられている．また，代表的な SNS の一つである Twitter[2]では，テレビ局・新聞社などのアカウントが，ニュースやニュース記事に対するリンクをツイートとして投稿している．ニュースを読むユーザは，ニュースサイトにおけるコメント欄のように，Twitter においてもニュースに対するコメントをツイートへの返信として投稿することができる．このため，インターネット上でニュースを読む際，ユーザはニュースだけではなく，ニュースに対する他者のコメントを読むことができる．他者のコメントを読むことで，ニュースに対する世論を理解でき，ニュースの全体像を捉えることができる．これにより，ユーザはニュースに対する理解を深めることができる．

Stroud ら [2] の調査では，インターネット上でニュースを読む人の 49 %がニュースに対するコメントを読むと回答しており，コメントの重要度は高いと考えられる．一方，ニュースのコメントを読む際，コメントの読者は表示されるコメントがどのような特徴を持つユーザから投稿されたのかを把握できないという問題点がある．例えば，あるニュースに対して批判的なコメントが投稿された際に，そのコメントは普段からニュースに対して建設的な意見を投稿するユーザによるコメントなのか，ニュースに対して批判的なコメントばかりを投稿するユーザによるコメントなのか区別することができない．内容が類似するコメントであっても，コメントを投稿したユーザの特徴によってコメントの持つ意味や価値は異なるため，ユーザの特徴を考慮したニュースコメントの閲覧支援が重要となる．

本論文では，ニュースに対するコメントを投稿したユーザの埋め込み表現を取得することで，コメントを投稿したユーザの特徴を考慮したコメントの閲覧支援手法を提案する．

本研究ではインターネット上のニュースとして Twitter を対象とし，テレビ局・新聞社などのニュースを配信するアカウントによるツイートをニュースとして収集する．さらに，ニュースに対するユーザの返信をコメントとして収集する．提案手法では，コメントを投稿したユーザの埋め込み表現を取得するために，ユーザの過去のニュースへの反応に注目する．ユーザの過去のニュースへの反応は，ニュースに対してコメントを行ったユーザを特徴づける重要な情報になると考えられる．そこで本手法では，ニュースに対して同様な反応をするユーザ同士は，類似した特徴を持つという仮説に基づいて，ニュースとコメントのペアからユーザを予測する機械学習モデルにより，ユーザ埋め込み表現を生成する．提案手法によって得られたユーザ埋め込み表現を基にニュースへのコメントに対してクラスタリングを行い，各クラスタから代表的なコメントを抽出してユーザ

1：https://news.yahoo.co.jp/
2：https://twitter.com/

に提示することで，コメントを投稿したユーザの特徴を考慮したニュースコメントの閲覧支援手法を提案する．

本論文の貢献は，ユーザが過去にコメントを行ったニュースと，そのコメントのテキストから，ユーザの特徴を表現したユーザ埋め込み表現を生成する手法を提案することである．また，取得したユーザ埋め込み表現がどのような観点からユーザの特徴を表現しているのかについて被験者からの主観評価により検証し，取得したユーザ埋め込み表現を用いることで，コメントを投稿したユーザの特徴を基にニュースコメントのクラスタリングと抽出を行うことが可能になることを示したことである．

本論文の構成は次の通りである．2 章で関連研究について述べる．3 章では提案するユーザ予測モデルによるユーザ埋め込み表現の取得について説明する．4 章で実験結果及び考察について述べる．5 章ではユーザの特徴を考慮したニュースコメントの閲覧支援手法について説明する．6 章でまとめと今後の課題を述べる．

2 関 連 研 究

2.1 ユーザ埋め込み表現の取得

本研究では，ユーザのニュースへの反応を利用してユーザ埋め込み表現を取得し，投稿者の特徴を考慮したニュースコメントのクラスタリングと抽出を行う．これまでにもユーザの過去の投稿を利用したユーザ埋め込み表現の取得に関して，いくつかの研究が行われている．

Hallac ら [3] は，文章の埋め込み表現を獲得する手法の一つである doc2vec [4] を利用してユーザの過去の投稿をベクトル化したものを平均化することで，ユーザ埋め込み表現を取得する手法を提案している．Wu ら [5] は，ユーザの過去の投稿を，自然言語処理モデルである BERT [6] と再帰型ニューラルネットワークである GRU [7] を利用してベクトル化し，得られたベクトルを用いてユーザを分類する事前学習を行うことで，ユーザ埋め込み表現を取得する手法を提案している．一般にユーザの過去の投稿には，ユーザの趣味や日常に関する投稿が含まれる．このため，ユーザの過去の投稿を用いてユーザ埋め込み表現を取得する際には，趣味や日常に関する情報によりユーザが特徴づけられることが想定される．一方，本手法ではニュースに対してコメントを行ったユーザを対象とし，ユーザの過去のニュースへの反応がユーザを特徴づける重要な情報になると考える．したがって，本手法ではユーザの過去の投稿の一部であるニュースへの反応に注目する．ニュースに対して同じような反応をするユーザ同士は似た特徴を持つと考えられるため，ユーザの過去のニュースへの反応を利用することでユーザ埋め込み表現を取得する．

2.2 ニュースコメントの閲覧支援

ニュースへのコメントの全体像を把握するために，ニュースへのコメントに対してクラスタリングを行う研究が行われている．Ma ら [8] は，トピックモデルの 1 つである LDA(Latent Dirichlet Allocation) [9] を用いてコメントのトピックを推定することで，ニュースへのコメントのクラスタリングを行っている．Aker ら [10] は，コメント間の類似度をグラスベースの手法でモデル化することで，ニュースへのコメントのクラスタリングを行っている．本論文では，ニュースに対してコメントを行ったユーザの特徴を基にコメントのクラスタリングと抽出を行い，ユーザの特徴ごとにコメントを表示することで，コメントを投稿したユーザの特徴を考慮したコメントの閲覧支援を提案する．

3 提 案 手 法

3.1 提案手法の概要

本論文で提案する手法では，ユーザが過去にコメントを行ったニュースと，そのコメントを入力とし，ユーザを予測するモデルを利用してユーザ埋め込み表現を求める．このモデルを利用することで，ニュースへの反応に基づいてユーザの特徴を取得する．つまり，ニュースに対して同じような反応をするユーザ同士は似た特徴を持つと考え，モデルの出力値をニュースへの反応に基づくユーザ埋め込み表現とする．

ニュースへの反応に基づいてユーザ埋め込み表現を求める提案手法の概要を図 1 に示す．本手法で利用する機械学習モデルは，はじめにニュースとそれに対するコメントのテキストを BERT を用いてベクトルに変換する．次に，このベクトルを用いてそのコメントを投稿したユーザを予測する．このモデルを利用して，ユーザの過去のニュースへの反応から，ユーザの特徴を表すベクトル表現を取得する．

3.2 ユーザ予測手法の定義

ニュースツイートの集合を $NT = \{nt_1, \ldots, nt_{|NT|}\}$，コメントツイートの集合を $CT = \{ct_1, \ldots, ct_{|CT|}\}$，Twitter のユーザ集合を $U = \{u_1, \ldots, u_{|U|}\}$ と表記する．コメントツイートとは，ニュースツイートに対するユーザの返信である．コメントツイート ct が対象とするニュースツイートを $\mathrm{news}(ct)$ と表記する．ここで，$\mathrm{news}(ct) \in NT$ である．また，コメントツイート ct を投稿したユーザを $\mathrm{user}(ct)$ と表記する．ここで，

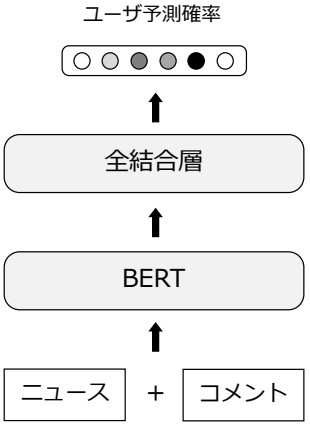

図 1　提案手法の概要

user(ct) $\in U$ である．

ニュースツイート nt 及びコメントツイート ct が与えられたときに，そのユーザを予測することを考える．このとき，そのユーザが u である確率を $\mathrm{P}(u \mid nt, ct)$ と表記する．ここで，$\sum_{u \in U} P(u \mid nt, ct) = 1$ を満たすものとする．本論文では，$\mathrm{P}(u \mid nt, ct)$ を予測する機械学習モデルを提案し，そのモデルを利用して，ユーザの埋め込み表現を取得する．上記の予測を行う機械学習モデルを M とする．このモデルの訓練を行うための教師データを (x, y) と表記する．ここで，x は入力データであり，y はそのデータに対する正解ラベルである．本手法では，M を訓練するためのデータセット TS を以下の式 (1) で定義する．

$$TS = \{((nt, ct), u) \mid nt \in NT, ct \in CT, u \in U, \\ \mathrm{news}(ct) = nt, \mathrm{user}(ct) = u\} \quad (1)$$

(nt, ct) はニュースツイートとコメントツイートのペアであり，入力データとなる．教師データでは $nt = \mathrm{news}(ct)$ を満足するため，コメントツイートとその対象となるニュースツイートが入力となる．正解ラベルは，そのコメントツイート ct を投稿したユーザである．

3.3 ユーザ予測モデルへの入力データ形式

提案するユーザ予測モデルでは，BERT を利用する．BERT は単語同士の関係を考慮することで，入力文を文脈に応じたベクトルに変換することができ，文章分類などの様々な自然言語処理のタスクを高い精度で処理可能とする．BERT への入力データの形式を図 2 に示す．提案モデルでは，ニュースとコメントの文章を以下の形式で結合し，入力を行う．

[CLS] ニュース文 [SEP] コメント文 [SEP]

ここで，[CLS] は入力文の先頭を表す特殊トークンであり，[SEP] は文のペアの境界や，入力文の終わりを表す特殊トークンである．文章をトークンに分割した後，各トークンがニュースの文章によるものか，コメントの文章によるものかを表す Segment Embeddings を各トークンに埋め込む．これにより，BERT においてニュースとコメントの文章間の対応関係を考慮することができる．ニュースとコメントの対応関係を考慮することにより，ユーザのニュースに対する捉え方を表す埋め込み表現を取得することが期待される．

3.4 BERT によるベクトル化

提案モデルで用いる BERT では，Transformer [11] の Encoder における Self-Attention 機構により，入力データ内の単語同士の関係を考慮したベクトルを出力する．提案モデルで用いる BERT の構造を図 3 に示す．E は入力埋め込み，Trm は Transformer Encoder，T は最終層の Transformer Encoder による出力を表す．図 2 に示すように各トークンごとに埋め込みが行われ，12 層の Transformer Encoder による処理の後，最終層の Transformer Encoder の [CLS] トークンの出力を，次の全結合層へ出力する．

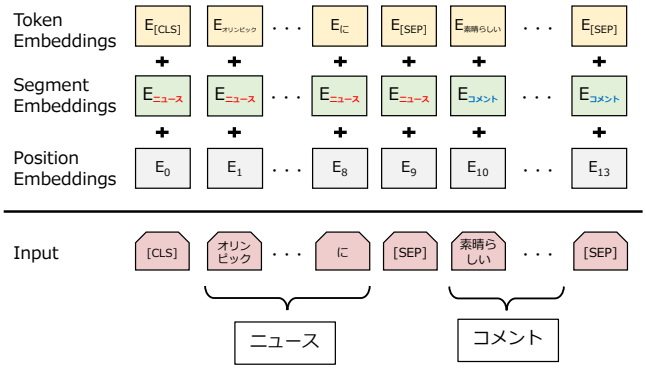

図 2　BERT への入力データ形式

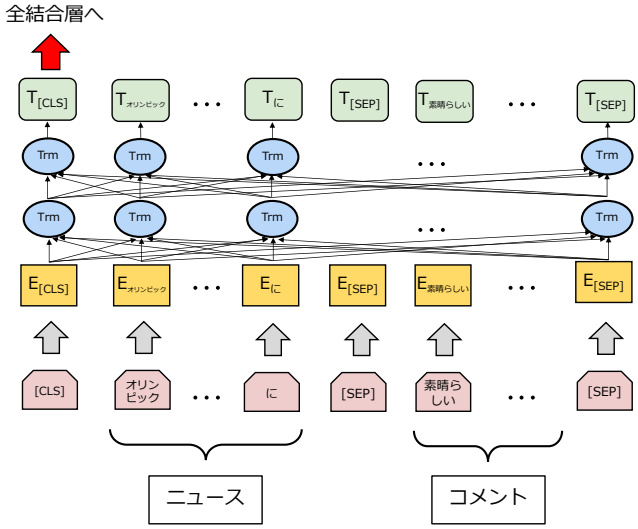

図 3　BERT の構造

BERT において，n 個のトークンからなる入力 $X = (x_1, x_2, \ldots, x_n)$ が与えられた場合，各層の Transformer Encoder では，入力 X に対し，クエリ Q, キー K, バリュー V を行列 W_Q, W_K, W_V の線形変換により以下のように求める．

$$Q = XW_Q \quad (2)$$

$$K = XW_K \quad (3)$$

$$V = XW_V \quad (4)$$

Self-Attention は，クエリ Q，キー K，バリュー V と，Q，K の次元数 d より，以下の式 (5) により求められる．

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \quad (5)$$

Self-Attention では各トークンのクエリとキーの内積を計算することで，各トークンの情報を考慮した値を出力する．BERT への入力には，各トークンがニュースの文章によるものか，コメントの文章によるものかを表す Segment Embeddings が埋め込まれているため，ニュースとコメントの関係性を考慮した出力を各層の Transformer Encoder から得ることができる．

3.5 ユーザ埋め込み表現の生成

ユーザ埋め込み表現の生成時は，ユーザが過去にコメントを

行ったニュースとそのコメントのペアをユーザ予測モデルに複数個入力し，出力ベクトルの平均をニュースへの反応に基づくユーザ埋め込み表現として取得する．ユーザ u_i が過去に N 件のニュースに対してコメントを投稿した場合，ニュースとそれに対するコメントのペアのテキスト集合を $\{d_{i,1}, d_{i,2}, ..., d_{i,N}\}$ とし，テキスト $d_{i,k}$ に対する提案モデルの出力を $f(d_{i,k})$ とするとき，最終的なユーザの埋め込み表現 e_i を以下の式 (6) で計算する．

$$e_i = \frac{1}{N}\sum_{k=1}^{N} f(d_{i,k}) \tag{6}$$

複数の入力を用いる理由は，1 件のニュース・コメントの情報ではユーザの特徴を捉えることが難しいためである．例えば，「大谷翔平　大リーグ MVP に選出」というニュースに対して，「おめでとうございます」というような反応は一般的な反応であるため，このような反応からはユーザの特徴を捉えることが難しい．そのため，ユーザが過去にコメントを行ったニュースと，そのコメントを複数利用し，モデルの出力値の平均をユーザ埋め込み表現とする．

4　評 価 実 験

4.1　実 験 手 法

4.1.1　データセット

実験に使用するデータの収集は 2021 年 11 月 11 日から 12 月 31 日にかけて TwitterAPI を用いて行い，NHK ニュース (@nhk_news)[3] によるニュースのツイートと，それに対するユーザの返信をそれぞれニュース・コメントとして取得した．ニュース とコメントの例を図 4 に示す．

4.1.2　前　処　理

本研究で用いるニュースのテキストには以下の前処理を行った．

- URL の削除
- 記号の削除
- ハッシュタグ (#nhk_news など) の削除

また，コメントのテキストには以下の前処理を行った．

- URL の削除

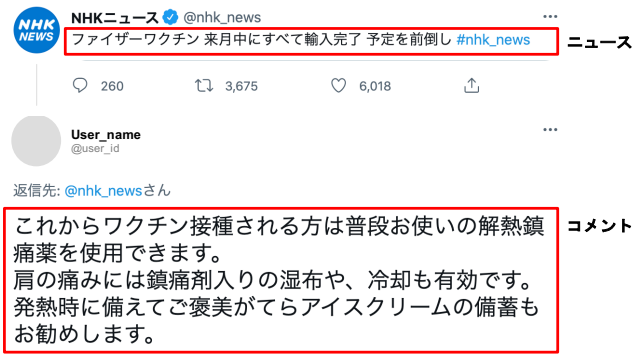

図 4　ニュースとコメントの例

3：https://twitter.com/nhk_news

- 記号の削除
- 絵文字の削除
- メンション (@ユーザ ID) の削除
- 半角カタカナを全角カタカナに置換などの正規化処理

正規化処理には python ライブラリである neologdn[4] を用いた．

4.1.3　ユーザ埋め込み表現を取得するモデルの訓練

提案手法を用いてユーザ埋め込み表現を取得するために，データセットの収集期間内に 50 件以上のコメントを行った 119 人のユーザを対象に提案モデルの訓練を行った．BERT モデルに関しては，東北大学の乾研究室が公開している日本語 Wikipedia を用いた事前学習済み BERT モデル [5] を fine-tuning することでユーザの予測を行った．各ユーザから過去 50 件分のニュースへの反応の投稿を取得し，各ユーザの投稿のうち，40 件を訓練データ，5 件を検証データ，残りの 5 件をテストデータとして利用した．

4.2　実験結果・考察

4.2.1　提案モデルによるユーザ予測

モデルの訓練に用いたユーザを対象に，各ユーザごとに訓練に利用しなかったテストデータ 5 件をユーザ予測モデルに入力し，5 件の出力の平均からユーザの予測を行った．本研究では，入力の形状が異なる 3 種類のモデルによる予測結果の比較を行った．各モデルは以下の通りである．

- 提案モデル

ニュース・コメントのテキストを利用する．また，Segment Embeddings により，各トークンがニュースの文章によるものか，コメントの文章によるものかを表す情報を与え，予測を行う．

- w/o seg モデル

ニュース・コメントのテキストを利用する．また，各トークンがニュースの文章によるものか，コメントの文章によるものかを表す情報を与えずに，予測を行う．

- comment モデル

コメントのテキストのみを利用し，予測を行う．入力はコメントのみの 1 種類であるため，Segment Embeddings の埋め込みは行わない．

ニュースとコメントの対応を表す情報を明示的に与えないモデル (w/o seg モデル) と，ニュースのテキストを利用しないモデル (comment モデル) の 2 種類のモデルを提案モデルと比較

表 1　各モデルの概要

モデル	利用テキスト	Segment Embeddings
提案モデル	ニュース・コメント	○
w/o seg	ニュース・コメント	×
comment	コメント	×

4：https://github.com/ikegami-yukino/neologdn

5：https://github.com/cl-tohoku/bert-japanese

表 2　テストデータに対するユーザ予測結果

	Accuracy	Top-5 acc
提案モデル	**0.546**	**0.773**
w/o seg	0.529	0.731
comment	0.504	0.765

することにより，ユーザの特徴を表現する際に，ニュースとコメントの文章間の対応関係を考慮することが効果的であるかを調べた．テストデータに対するユーザ予測結果を表 2 に示す．

表 2 より，提案モデルではユーザの分類を高い精度で達成でき，ニュースへの反応を基にユーザの特徴を表現できることが分かった．また，提案モデルの精度が他のモデルの精度を上回っており，ユーザの特徴を表現する際に，ニュースとコメントの文章間の対応関係を考慮することが効果的であったと考えられる．

4.2.2　ユーザ埋め込み表現の検証

川口ら [12] の手法を参考に，被験者にアンケートを行い，提案するユーザ予測モデルにより取得したユーザ埋め込み表現の評価を行った．本実験では，データセットから抽出したユーザのペアに対して提案モデルによりユーザ埋め込み表現をそれぞれ取得し，2 人のユーザ埋め込み表現からコサイン類似度を計算することで，ユーザの類似度を予測した．ユーザの類似度を予測する過程を図 5 に示す．

被験者には 2 人のユーザが過去にコメントを行ったニュースと，そのコメントを 3 件ずつ提示し，2 人のユーザの類似度に関して主観評価を行ってもらった．主観評価によってユーザの類似度を予測する過程を図 6 に示す．

提案モデルを用いて予測したユーザの類似度と，被験者の主観評価によるユーザの類似度との相関を調べることで，提案モデルにより取得したユーザ埋め込み表現が，ユーザの実際の

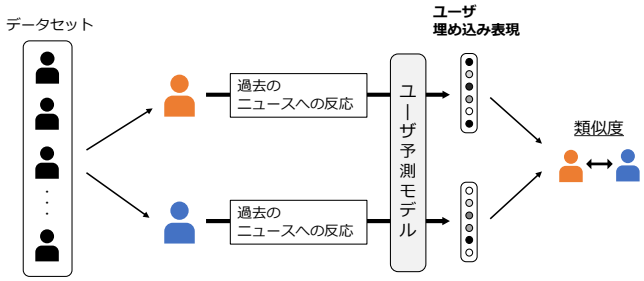

図 5　提案モデルによりユーザの類似度を予測する過程

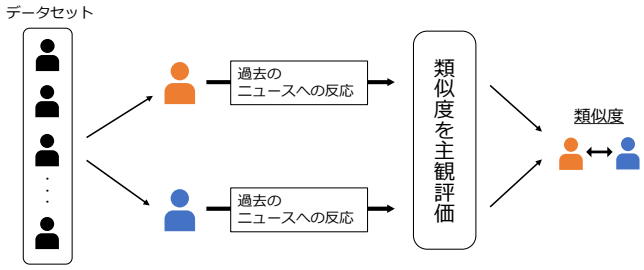

図 6　主観評価よりユーザの類似度を評価する過程

ニュースへの反応を基に表現することができているかについて検証した．本実験では，データセットから 5 件以上の投稿が取得できたユーザを対象に，2 人のユーザのペア 50 組を標本として抽出した．なお，本実験で抽出したユーザは，ユーザ予測モデルの訓練に用いたデータセットに含まれていないユーザである．2 人のユーザのニュースへの反応を図 7 のように提示し，2 人のユーザの類似度を尋ねる以下の質問を行った．

(1) 2 人のユーザの「政治的な立場」はどの程度類似していると感じましたか
(2) 2 人のユーザの「興味関心」はどの程度類似していると感じましたか
(3) 2 人のユーザの「性格」はどの程度類似していると感じましたか

「政治的な立場」「興味関心」「性格」はユーザの属性推定において多く用いられる観点である [13] [14] [15]．3 つの観点からのユーザの類似度を調べることにより，取得したユーザ埋め込み表現が，ユーザの実際のニュースへの反応を基に，どのような観点からユーザの特徴を表現することができるかについて調べた．また，3 つの質問は以下の 5 段階で回答してもらった．

1:「全く類似していない」
2:「あまり類似していない」
3:「どちらとも言えない」
4:「やや類似している」
5:「非常に類似している」

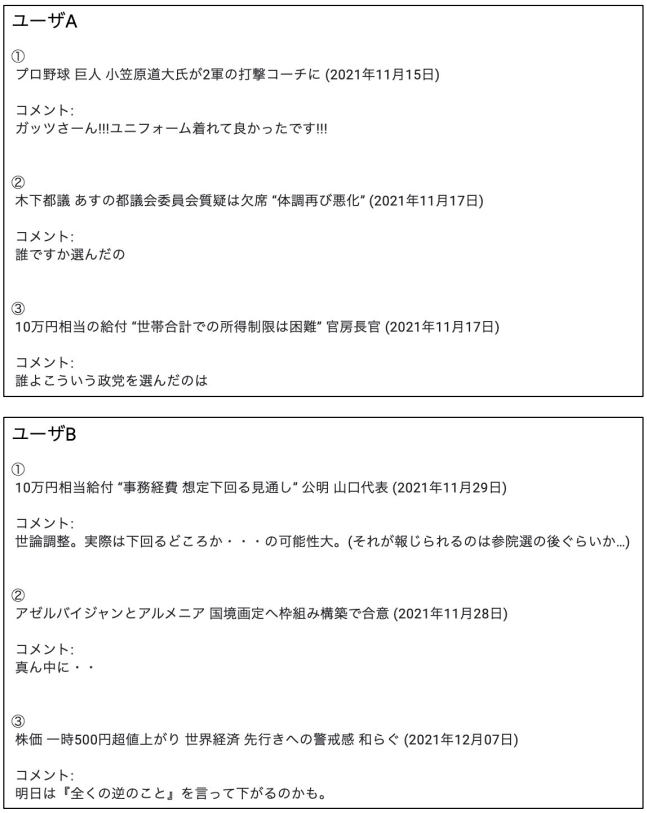

図 7　被験者に提示したニュースとコメントの例

アンケートはクラウドソーシングにより実施し，被験者の総数は 150 名である．1 組のユーザペアにつき 30 人の被験者から回答を取得し，3 つの質問項目ごとに 5 段階の回答を平均した．その結果，1 組のユーザペアにつき 3 種類の観点に関する評価値を得た．また，予測類似度と，3 つの項目の評価値はそれぞれ平均 0，分散 1 となるように標準化を行った．

横軸を提案モデルを用いて予測した類似度の値，縦軸をアンケートの回答から得た評価値として作成した散布図を図 8，9，10 に示す．また，予測類似度と各評価値との相関係数を，ピアソンの相関係数により求めた結果を表 3 に示す．予測類似度は 4.2.1 節の 3 種類のモデルにより算出し，比較を行った．

表 3 に示すように，今回抽出した標本に関して，提案モデル

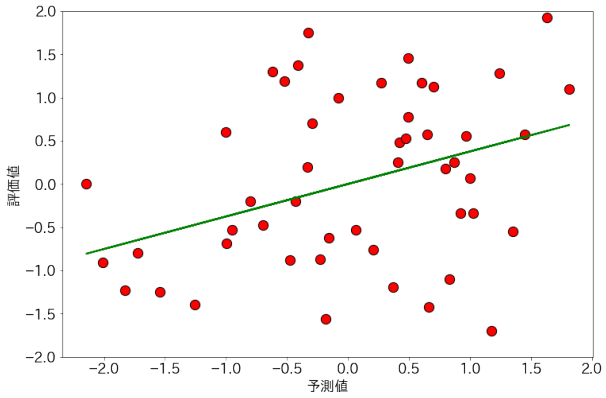

図 8　提案モデルの予測類似度と「政治的な立場」の評価値の散布図

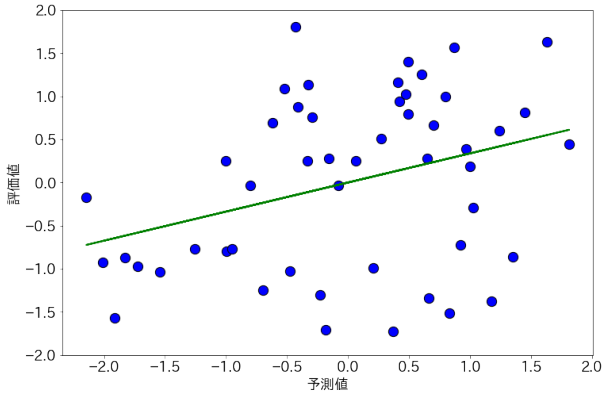

図 9　提案モデルの予測類似度と「興味関心」の評価値の散布図

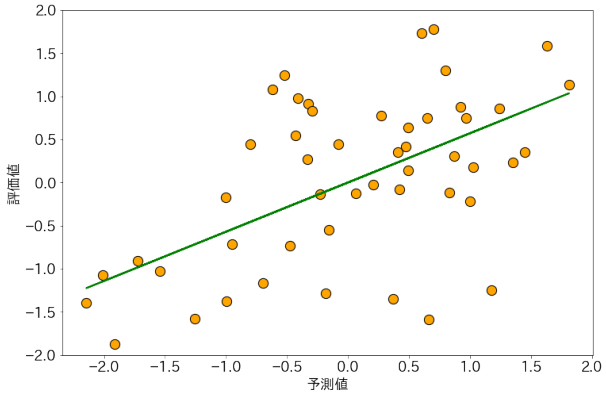

図 10　提案モデルの予測類似度と「性格」の評価値の散布図

表 3　各モデルによる予測類似度と各評価値の相関係数

	相関係数		
	政治的な立場	興味関心	性格
提案モデル	**0.376**	**0.337**	**0.571**
w/o seg	0.293	0.307	0.427
comment	0.132	0.150	0.235

により予測した類似度と，主観評価による「政治的な立場」「興味関心」の類似度には弱い正の相関があり，「性格」の類似度には中程度の正の相関があった．また，提案モデルでは他のモデルよりも各評価値と高い相関を得ることができ，ニュースとコメントの文章間の対応関係を考慮することにより，ユーザの特徴をより適切に表現できることが分かった．

次に，提案モデルによる予測類似度と主観評価による類似度に有意な相関があるかを判断するため，無相関検定を行った．無相関検定とは，標本から得られた結果から母集団の相関係数が 0 かどうかを検定する手法であり，帰無仮説を「母集団における相関係数が 0 である」，対立仮説を「母集団における相関係数は 0 でない」として設定する．p 値が有意水準より小さい場合，母集団における相関係数が 0 であるという帰無仮説が棄却され，母集団において相関があると判断することができる．無相関検定では母集団が正規分布に従っていない場合，相関係数にはスピアマンの順位相関係数を用いる必要がある．データセットにおいて 5 件以上の投稿が取得できたユーザにおける予測類似度をヒストグラムで表したものを図 11 に示す．

図 11 より，予測類似度の値は正規分布に従わないため，無相関検定において相関係数にはスピアマンの順位相関係数を用いた．有意水準 5%とし，無相関検定を行った結果を表 4 に示す．

表 4 に示すように，無相関検定における p 値は「政治的な立場」「興味関心」「性格」の 3 つの項目全ておいて $p<0.05$ となり，有意水準を下回った．このため，提案手法により求めた類似度と，主観評価による 3 つの項目の評価値には有意な相関

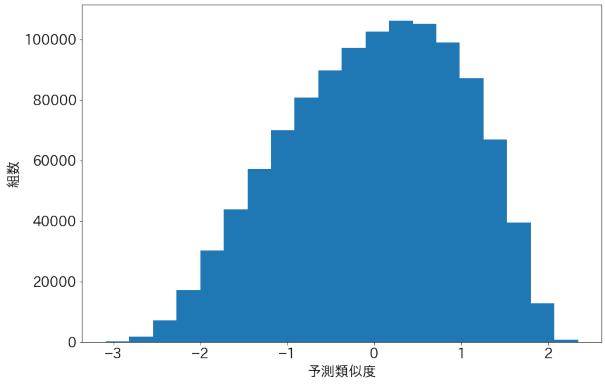

図 11　提案モデルの予測類似度のヒストグラム

表 4　無相関検定の結果

	p 値
政治的な立場	0.031
興味関心	0.042
性格	6.05×10^{-4}

があると判断することができた．この結果から，提案手法により取得したユーザ埋め込み表現は「政治的な立場」「興味関心」「性格」などの複数の観点からユーザの特徴を表すことができていると考えられる．このため，提案手法によるユーザ埋め込み表現を用いることにより，ニュースに対してコメントを行ったユーザの特徴を複数の観点から見つけ出すことができ，コメントを投稿したユーザの特徴を考慮したニュースコメントの閲覧支援を行うことが期待できる．

5　ニュースコメントの閲覧支援

本章では，コメントを投稿したユーザの特徴ごとにコメントの表示を行うことにより，ユーザの特徴を考慮したニュースコメントの閲覧支援手法を提案する．ニュースコメントの閲覧支援手法の概要を図 12 に示す．

5.1　ユーザの特徴に基づくクラスタリング

コメントを投稿したユーザの特徴ごとにコメントの表示を行うために，コメントを投稿したユーザの特徴を基にクラスタリングを行う．クラスタリングには K-means 法 [16] を用いる．コメントを投稿したユーザの特徴は，提案手法により，ユーザの過去のニュースへの反応を利用することで取得する．

5.2　各クラスタからの代表的なコメントの抽出

各クラスタに属する代表的なユーザを選択し，代表的なユーザによるコメントを，ニュースのコメントを読むユーザに提示する．多様な視点からのコメントを提示するため，各クラスタに対してユーザの特徴を基に再度クラスタリングを行い，クラスタの分割を行う．クラスタを分割することにより作成された新たなクラスタから，代表的なユーザによるコメントを選択することでコメントの抽出を行う．代表的なユーザの選択にあたって，各クラスタの重心と，そのクラスタに属する全てのユーザの埋め込み表現との類似度をコサイン類似度を用いて計算する．各クラスタから，重心との類似度が最も高いユーザを代表的なユーザとして選択し，代表的なユーザによるコメントを抽出する．

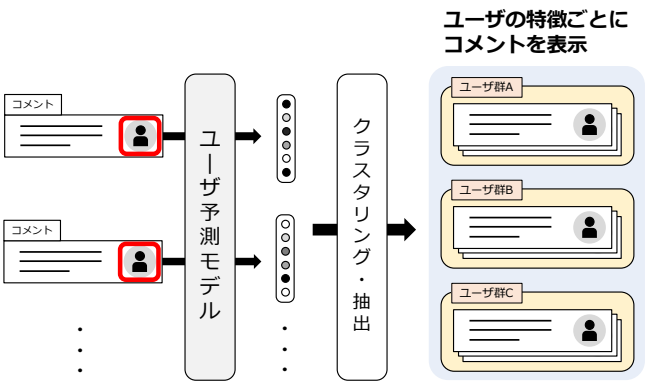

図 12　ニュースコメントの閲覧支援手法の概要

5.3　ニュースコメントの表示結果

「オミクロン株 “ 自覚ない感染者 ” からの拡大に医師が危機感」というニュースツイートへのコメントに対し，ユーザの特徴ごとにコメントの表示を行った結果を表 5 に示す． なお，取得したデータセットから 5 件以上の投稿が取得できなかったユーザによるコメントについては対象外とした．現段階でのコメント表示結果では，各ユーザ群に属するユーザの特徴を示すラベルが無いため，ユーザの特徴を把握することが難しい．このため，今後はクラスタリングと抽出結果に対してラベル付けを行う手法の提案を検討している．

6　ま　と　め

本論文では，ユーザの特徴のモデル化手法と，それに基づいたコメント閲覧支援手法を提案した．提案手法では，ユーザの過去のニュースへの反応を利用したユーザ予測モデルにより，ニュースとコメントの対応関係を考慮したユーザ埋め込み表現を取得可能である．テストデータに対するユーザ予測結果より，提案モデルでは，高い精度でユーザの予測を行うことができており，ユーザのニュースへの反応を基にユーザ埋め込み表現を取得できることを示した．また，提案モデルにより得られたユーザ埋め込み表現の評価を被験者の主観評価により行い，提案するユーザ埋め込み表現が複数の観点からユーザの特徴を表していることを示した．最後に，コメントを投稿したユーザの埋め込み表現を基にコメントのクラスタリング及び抽出を行った．

今後は，クラスタリングと抽出結果に対してラベル付けを行うことにより，解釈可能なコメントの閲覧システムの構築が必要である．また，コメントの閲覧システムによりユーザが正しくニュースへのコメントの全体像を理解することができるかについての評価を行うことを検討している．

謝　　辞

本研究は JSPS 科研費 19H04219 の助成を受けたものです．

文　　献

[1] 新聞通信調査会. 第 14 回メディアに関する全国世論調査, 2021. https://www.chosakai.gr.jp/project/notification/.

[2] Natalie Jomini Stroud, Emily Van Duyn, and Cynthia Peacock. News commenters and news comment readers. *Engaging News Project*, pp. 1–21, 2016.

[3] Ibrahim R Hallac, Semiha Makinist, Betul Ay, and Galip Aydin. user2vec: Social media user representation based on distributed document embeddings. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–5. IEEE, 2019.

[4] Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. *CoRR*, Vol. abs/1405.4053, , 2014.

[5] Xiaodong Wu, Weizhe Lin, Zhilin Wang, and Elena Rastorgueva. Author2vec: A framework for generating user embedding. *arXiv preprint arXiv:2003.11627*, 2020.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional

表 5　ユーザの特徴ごとにコメント表示を行った結果

ユーザ群 A	その通りだ。高齢者や基礎疾患のある人、手術後など免疫力の低い人が感染すると危険だ。 1 型糖尿病とか生まれつき病弱な子供もいるし　感染が拡大したら変異して強毒化する 自覚がないということは症状がないということで何か問題があるのでしょうか。無症状は、本来なら患者とは言いませんよね。
ユーザ群 B	ただの風邪 ワクチン打てば打つほど感染者増えるだろうな 自覚がないんだから仕方ないでしょそれをどうしろと?出歩くなとでも? 自覚ない www 折角ワクチン打って症状を無くしても、まさか悪人にされようとは。
ユーザ群 C	喧しい 大半が自覚なく自然免疫が身に付くならそれでいい なら入国規制しなよ

transformers for language understanding. *CoRR*, Vol. abs/1810.04805, , 2018.

[7] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, Vol. abs/1406.1078, , 2014.

[8] Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 265–274, 2012.

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

[10] Ahmet Aker, Emina Kurtic, AR Balamurali, Monica Paramita, Emma Barker, Mark Hepple, and Rob Gaizauskas. A graph-based approach to topic clustering for online comments to news. In *European Conference on Information Retrieval*, pp. 15–29. Springer, 2016.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, Vol. abs/1706.03762, , 2017.

[12] 川口天佑, 牛尼剛聡. ポピュラリティ推定に基づいた SNS におけるニュースの中立的な理解支援. In *DEIM Forum 2018*, 2018.

[13] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 729–740, 2017.

[14] Zhiheng Xu, Long Ru, Liang Xiang, and Qing Yang. Discovering user interest on twitter with a modified author-topic model. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 422–429. IEEE, 2011.

[15] 山田康輔, 笹野遼平, 武田浩一. 「いいね」「シェア」をした投稿のテキスト情報を利用した SNS ユーザの性格推定. 人工知能学会論文誌, Vol. 35, No. 4, pp. B–K22_1, 2020.

[16] James MacQueen, et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, pp. 281–297. Oakland, CA, USA, 1967.