トキシック性に基づいたニュースコメントフィルタリングのための投稿者のフォロー関係の活用

中原　輝樹† 　牛尼　剛聡†

† 九州大学大学院 芸術工学府　〒 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，読者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．読者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．しかし，他者のコメントには読者を不快に感じさせるトキシック（有害）コメントが含まれることがあり，ユーザの体験の質を低下させる．このため，コメント閲覧時における読者の体験の質を向上させるために，トキシックコメントの予測を行うことが重要となる．本研究では，コメントの読者の価値観が多様であることを考慮し，読者が過去に低評価したニュースコメントのフィードバックに基づき，読者ごとにトキシックコメントを予測する．提案手法では，読者が不快と感じるコメントの投稿者の傾向を考慮した予測を行うために，Twitter のフォロー関係に基づいてコメントの投稿者の特徴を取得する．評価実験の結果，フォロー関係に基づいて取得したコメント投稿者の埋め込み表現を予測に活用することにより，トキシックコメント予測の個人化の性能が向上することが示された．

キーワード　パーソナライゼーション，ソーシャルメディア，Twitter, BPR, LightGCN

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多い．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].

インターネット上でニュースを読む際，読者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．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．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いる．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し，ニュースの全体像を捉え，ニュースに対する理解を深めることができる．一方，ニュースコメントなどのオンラインコミュニティでは，攻撃的な内容や誹謗中傷を含む発言などの読者にとって不快と感じるコメントが投稿される問題がある．このような読者が不快と感じるコメントは，有害という意味を表す toxic という単語からトキシックコメント (toxic comment) と呼ばれている [3]．本論文では，「トキシックコメント」を攻撃的な内容や誹謗中傷を含む発言などの読者にとって不快と感じるコメントとして定義する．トキシックコメントの例を図 1 に示す．

コメント閲覧時における読者の体験の質を向上させるために，トキシックコメントを予測する手法が提案されている [4–6]．これらの手法では，コメントのテキストなどを入力し，対象のコメントがトキシックであるかを予測する機械学習モデルを利用している．しかし，コメントの読者の価値観などの違いによってトキシックコメントの判定基準は異なる [7].

著者らはトキシックコメント予測の個人化の重要性を検証するために，コメント読者間のコメントに対する評価のばらつきを調査した．クラウドソーシングサービスであるクラウドワークス [3]でアンケートを実施し，ニュースコメントに対するラベル付けを行った．アンケートでは 2022 年 4 月 1 日から 6 月 30

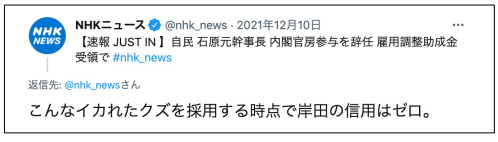


図 1: トキシックコメントの例

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

日の 3ヶ月間に投稿されたニュースコメント 400 件を 50 名のユーザに提示し，不快と感じるかについて {1: 全くそう思わない， 2: あまりそう思わない，3: どちらとも言えない，4: ややそう思う，5: 非常にそう思う } の 5 段階で回答してもらった．これにより，1 件のニュースコメントにつき 50 名のユーザからの回答を取得し，評価値の平均値と標準偏差をコメントごとに算出した．コメントと評価値の平均値・標準偏差の関係の例を表 1 に示す．400 件のコメントの標準偏差をヒストグラムとして表したものを図 2 に示す．

図 2 より，5 段階の評価値の標準偏差は 1.0 付近に集中していることがわかる．つまり，コメント読者による評価値は標準的に 1 程度平均から離れていることが多く，コメント読者間のコメントに対する評価にばらつきが存在していることがわかる．このため，コメントの読者の価値観などを考慮するために，トキシックコメント予測において個人化を行うことが重要であるといえる．

Nakahara ら [8] はこれまでにトキシックコメントの予測において個人化（パーソナライズ）を行い，予測性能の向上を示している．この手法では予測対象のコメントと読者が過去に低評価したコメントの利用に加え，コメントの投稿者の特徴を活用する機械学習モデルを用いた予測を行っている．これにより，読者が不快と感じるコメントの投稿者の傾向を捉えることができ，トキシックコメント予測の性能が向上する．また，コメントの投稿者の特徴を取得する際に過去に投稿したコメントとコメント元のニュースのテキストを利用している．本論文では，コメントの投稿者の特徴はその投稿者が属するコミュニティによって決定されるという仮定をもとに，Twitter のフォロー関係に基づいてコメントの投稿者の特徴を取得し，機械学習モデルによる予測を行う手法を提案する．

本論文の貢献は，以下の通りである．

- ニュースコメントに対する読者のフィードバックから読者の埋め込み表現を生成し，トキシックコメント予測の個人化を行う方法を開発した．
- Twitter のフォロー関係に基づいて取得したコメント投稿者の埋め込み表現を予測に活用することで，トキシックコメント予測の個人化の性能を向上させる方法を開発した．

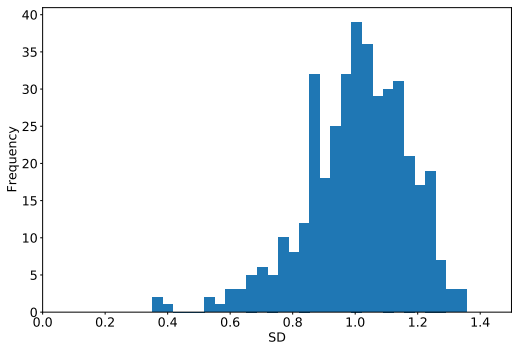


図 2: 評価値の標準偏差の分布

- 実際に投稿されたニュースとコメントからなるデータセットを利用して提案手法が有効であることを示した．

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 章では提案手法について説明する．4 章では評価実験について述べる．5 章で本論文のまとめを述べる．

2 関 連 研 究

2.1 ニュースコメントの理解支援

これまでにニュースコメントの閲覧を支援するためのいくつかの研究が行われている．Ma ら [9] は，代表的なトピックモデルの 1 つである LDA(Latent Dirichlet Allocation) [10] を用いてコメントのトピックを推定することで，ニュースコメントのクラスタリングを行っている．Aker ら [11] は，コメント間の類似度をグラフベースの手法でモデル化することで，ニュースコメントのクラスタリングを行っている．これらの研究では，クラスタリングによってニュースコメントの全体像を読者に理解させることでニュースコメントの閲覧支援を行っている．それに対して本論文での提案手法では，攻撃的な内容や誹謗中傷を含む発言などのトキシックコメントを事前に予測することで読者の体験の質を向上させ，ニュースコメントの閲覧支援を行うことを目的とする．

2.2 トキシックコメント予測

これまでに SNS などのオンラインコミュニティにおけるトキシックコメントを予測する研究や実装が行われている．Georgakopoulos ら [4] は，Wikipedia における編集者の議論ページを対象に，CNN を用いた機械学習モデルによってトキシックコメントの予測を行っている．Hessel ら [6] は，Reddit を対象に LSTM [12] や BERT [13] などのモデルを用いて，論争的な投稿と非論争的な投稿の分類を行っている．Saveski ら [14] は，Twitter における会話の構造と投稿の有害性との関係について分析している．Google Jigsaw はトキシックコメントを検出するシステムを Perspective API として提供している [5]．これらの手法ではコメントの読者の価値観などの違いが考慮されていないため，本論文では予測の個人化を行うことでコメントの読者の価値観などの違いを考慮する手法を提案する．

2.3 トキシックコメント予測の個人化

Sap ら [7] は，評価者のアイデンティティや信条とトキシック性の評価との間に関連性があることを示している．このことは，コメント読者の価値観などの違いを考慮し，トキシックコメントの予測において個人化（パーソナライズ）を行うことの重要性を示している．Kumar ら [15] は，コメントの読者が過去に有害と評価したコメントのフィードバックを利用し，Google Jigsaw が提供する Perspective API のパラメータを読者ごとにチューニングすることによって予測性能が向上することを示している．Nakahara ら [8] は，読者が過去に低評価したコメントだけでなく，低評価したコメントの投稿者の特徴を活用する機械学習モデルを用いた予測を行っている．これにより，読者が不快と感じるコメントの投稿者の傾向を捉えることができ，

表 1: コメントと評価値の平均値・標準偏差の関係の例

ニュース	コメント	平均値	標準偏差
悠仁さま 筑波大学附属高校に入学	ご入学おめでとうございます。	1.14	0.35
弾道ミサイル避難施設に地下鉄の駅 100 か所余りを初指定	馬鹿な女。今の憲法じゃ、避難する前に滅ぼされている！	3.72	0.95
“子ども 2 人自宅放置しパチンコ”乳児死亡で夫婦逮捕	いっそ最初から産むな	2.96	1.34

トキシックコメント予測の性能が向上する．コメントの投稿者の特徴に関して，対象の投稿者が過去に投稿したニュースコメントを入力とした機械学習モデルを利用することにより取得している [16].

2.4 フォロー関係の活用

SNS におけるフォロー関係はユーザの特徴を表す重要な情報のひとつである．これまでにフォロー関係から取得したユーザの特徴を下流タスクで活用する研究が行われている．Li ら [17] は，ユーザのソーシャルネットワークと単語の埋め込み表現を組み合わせることにより，ユーザの属性分類の精度が向上することを示している．本論文では Twitter のフォロー関係からコメント投稿者の特徴を取得し，トキシックコメント予測に活用する．

3 提 案 手 法

3.1 提案手法の概要

本論文では予測対象のニュースコメントと，読者が過去に低評価したニュースコメントのフィードバックに基づいてトキシックコメントを読者ごとに予測することを考える．本論文で提案する手法の概要を図 3 に示す．提案手法ではコメントの投稿者の特徴を考慮することで読者が不快と感じるコメントの投稿者の傾向も捉え，予測性能の向上を目指す．コメントの投稿者の特徴は図 3 の News Comment Encoder および Reader Encoder における Follow Relation Encoder で表現する．ここで，読者 r_t にとってのコメント c_t のトキシック度を予測する機械学習モデルは以下の式 (1) で表される．

$$y = f(x;\theta) \tag{1}$$

ここで x はコメント c_t と読者 r_t のペア (c_t, r_t) のベクトル表現，y はトキシック度を表すスカラー値，θ はパラメータである．提案モデルの学習用データセット D は，入力 x と出力 y のペア (x, y) の集合として、以下の式 (2) で定義される．

$$D = \{(x,y) \mid x = em_c(c_t) \oplus em_r(r_t), c_t \in C, r_t \in R, \; y \in \{1,2,3,4,5\}\} \tag{2}$$

C と R はそれぞれコメントの集合と読者の集合を表す．$em_c(c_t)$ と $em_r(r_t)$ は，それぞれコメント c_t と読者 r_t の埋め込みベクトルを表す．$\oplus$ はベクトルの連結を行う演算子であり，x はコメント c_t の埋め込みベクトルと読者 r_t の埋め込みベクトルを連結したものである．

3.2 News Comment Encoder

News Comment Encoder は，予測対象とするコメントの埋め込みベクトルを生成する機能である．ここでは，単に予測対象とするコメントのテキストを利用するだけでなく，コメントを投稿したユーザの情報も利用する．具体的には，予測対象のコメント c_t の埋め込みベクトル $em_c(c_t)$ を 2 つのベクトルを連結したベクトルとして生成する．

1 つ目のベクトルは予測対象のコメントとコメント元のニュースのテキストを BERT に入力することで得られるベクトルである．形式的には以下の式 (3) で表される関数として定義する．

$$enc^{cmt}(c_t) = \mathrm{BERT}(c_t, news(c_t)) \tag{3}$$

ここで $news(c)$ はコメント c が参照するニュースツイートである．BERT によるコメントとコメント元のニュースのテキストのベクトル化については 3.4 節で述べる．

2 つ目のベクトルは予測対象のコメントを投稿したユーザを表すベクトルであり，3.5 節で説明する投稿者のフォロー関係を活用する手法によって表現する．これより，提案手法における予測対象のコメント c_t の埋め込みベクトル $em_c(c_t)$ は，以下の式 (4) で定義できる．

$$em_c(c_t) = enc^{cmt}(c_t) \oplus enc^{usr}(commenter(c_t)) \tag{4}$$

ここで $enc^{usr}(u)$ は，コメント投稿者 u をフォロー関係を用いてベクトル化する関数を表す．$commenter(c)$ はコメント c を投稿したユーザを表す．

3.3 Reader Encoder

Reader Encoder は対象とする読者の埋め込みベクトルを得るための機構である．ここでは，読者が過去に低評価したコメントの情報を格納するデータベース（フィードバックデータベース）が利用可能であるとする．フィードバックデータベースを利用することで読者 ID に対してそのユーザが過去に低評価したコメントとコメント元のニュース，およびそのコメントを投稿したユーザの投稿者 ID を取得できる．

コメントの読者 r_t のベクトルを生成するために，Reader Encoder では 2 つのベクトルを利用する．

1 つ目のベクトルは読者が低評価したコメントのベクトルである．読者 ID に基づいてフィードバックデータベースから対象の読者が過去に低評価したコメントとコメント元のニュースのペアのテキストを N 件取得し，BERT に入力することで N 個のベクトルを得る．いま，読者 r_t が低評価したコメントの集合を $toxic(r_t) = \{c_1, \ldots, c_N\}$ と表現する．トキシックコメントのベクトル集合 $TC(r_t)$ を以下の式 (5) で定義する．

$$TC(r_t) = \{enc^{cmt}(c) \mid c \in toxic(r_t)\} \tag{5}$$

予測対象のコメント c_t のベクトル $enc^{cmt}(c_t)$ と，対象の読者

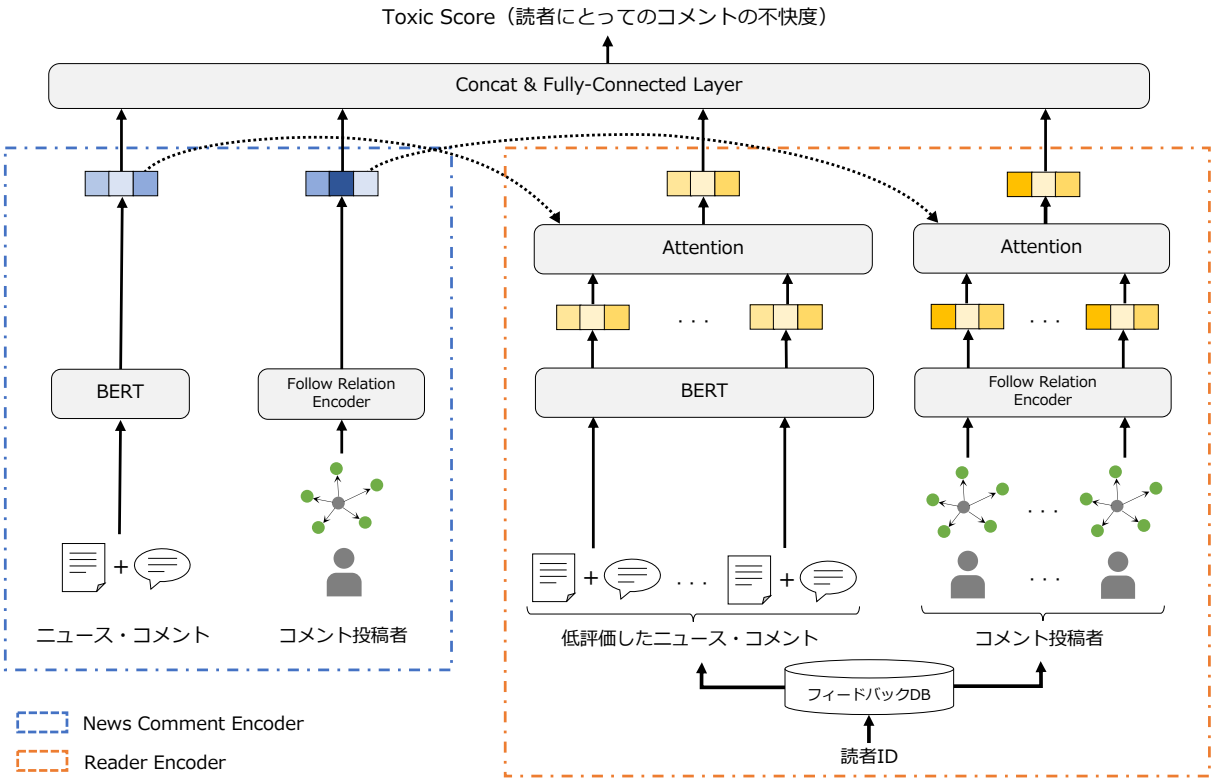


図 3: 提案手法の概要

r_t が過去に低評価したコメントのベクトルの集合 $TC(r_t)$ の Attention を計算することで $TC(r_t)$ を集約し，1 つのベクトルを得る．$enc^{cmt}(c_t)$ と $TC(r_t)$ の Attention を計算することで $TC(r_t)$ を集約する関数は以下の式 (6) で表される．

$$enc^{toxic\_cmt}(c_t, r_t) = \sum_{i=1}^{N} W_i \cdot e(TC(r_t), i) \tag{6}$$

ここで，$e(S, i)$ は集合 S の i 番目の要素を表し，W_i は以下の式 (7) で表される．

$$W_i = \mathrm{Softmax}\left(\frac{qk_i^\top}{\sqrt{d}}\right) v_i \tag{7}$$

q, k_i, v_i はそれぞれ以下の式 (8), (9), (10) で表される．

$$q = W^q \cdot enc^{cmt}(c_t) \tag{8}$$

$$k_i = W^k \cdot e(TC(r_t), i) \tag{9}$$

$$v_i = W^v \cdot e(TC(r_t), i) \tag{10}$$

ここで，W^q, W^k, W^v は学習可能なパラメータである．

2 つ目のベクトルは，読者が過去に低評価したコメントを投稿したユーザのベクトルである．読者 ID に基づいてフィードバックデータベースから対象の読者が過去に低評価したコメントを投稿したユーザを N 件取得し，投稿者のフォロー関係を活用する手法によって N 個のベクトルを得る．ここで，対象の読者 r_t が過去に低評価したコメントを投稿したユーザのベクトルの集合は以下の式 (11) で表される．

$$TU(r_t) = \{enc^{usr}(commenter(c)) \mid c \in toxic(r_t)\} \tag{11}$$

予測対象のコメントを投稿したユーザのベクトル $enc^{usr}(commenter(c_t))$ と $TU(r_t)$ の Attention を計算することで $TU(r_t)$ を集約し，対象の読者が過去に低評価したコメントを投稿したユーザの表現を得る．$enc^{usr}(commenter(c_t))$ と $TU(r_t)$ の Attention を計算することで $TU(r_t)$ を集約する関数は以下の式 (12) で表される．

$$\begin{aligned} enc^{toxic\_usr}(enc^{usr}(commenter(c_t)), TU(r_t)) \\ = \sum_{i=1}^{N} W_i \cdot e(TU(r_t), i) \end{aligned} \tag{12}$$

ここで，W_i は以下の式 (13) で表される．

$$W_i = \mathrm{Softmax}\left(\frac{qk_i^\top}{\sqrt{d}}\right) v_i \tag{13}$$

q, k_i, v_i はそれぞれ以下の式 (14), (15), (16) で表される．

$$q = W^q \cdot enc^{usr}(commenter(c_t)) \tag{14}$$

$$k_i = W^k \cdot e(TU(r_t), i) \tag{15}$$

$$v_i = W^v \cdot e(TU(r_t), i) \tag{16}$$

ここで，W^q, W^k, W^v は学習可能なパラメータである．

3.4 BERT によるニュースとコメントのベクトル化

本節では，BERT によるニュースとコメントのベクトル化について説明する．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で

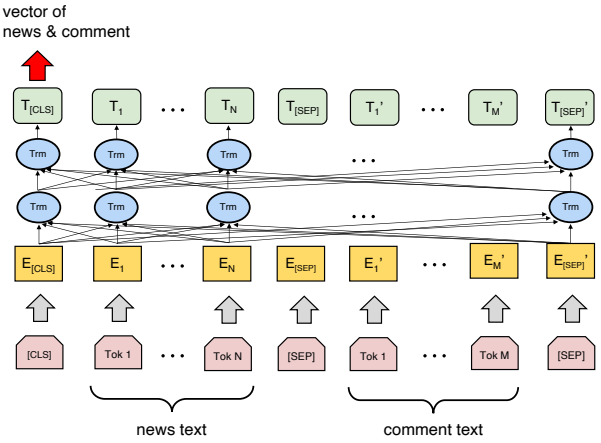


図 4: BERT の構造

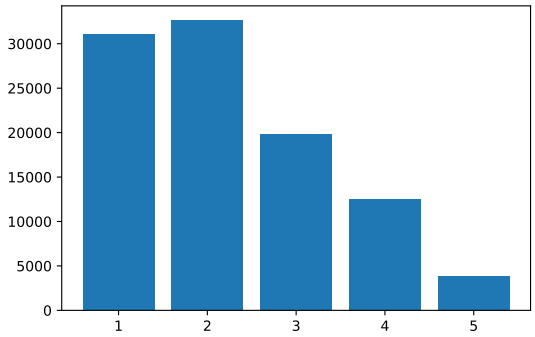


図 5: アンケートで取得したラベルの分布

ある．提案手法では，ニュースとコメントのテキストを以下の形式で結合し，トークンごとに分割したテキストを BERT への入力とする．

[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．BERT では，Transformer [18] の Encoder における Self-Attention 機構により，入力データ内の単語どうしの関係を考慮したベクトルを出力する．提案手法で用いる BERT の構造を図 4 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．提案手法では，最終層の Transformer Encoder の [CLS] トークンの出力ベクトル（768 次元）を BERT の出力として用いる．

3.5 投稿者のフォロー関係の活用

本節では，フォロー関係を用いてコメントの投稿者の特徴を表現する手法について説明する．

3.5.1 Bayesian Personalized Ranking(BPR) を用いたフォロー関係のベクトル化

Bayesian Personalized Ranking(BPR) [19] は，暗黙的な評価値に対する行列分解の手法の一つである．本論文では，コメント投稿者がフォローしているユーザを表す行列を commenter-followee 行列として定義し，commenter-followee 行列 X をコメント投稿者の因子ベクトル P とフォロイーの因子ベクトル Q に $X \approx P \cdot Q^{\top}$ として分解する．コメント投稿者の因子ベクトル P を用いることにより，コメント投稿者の特徴を考慮したトキシックコメント予測を行う．

3.5.2 LightGCN を用いたフォロー関係のベクトル化

LightGCN [20] は GCN(Graph Convolutional Networks) を利用して周辺のユーザやアイテムから直接ベクトル表現を獲得する手法である．本論文ではコメント投稿者とフォロイーからなる二部グラフを定義し，LightGCN を利用して取得したコメント投稿者の特徴をトキシックコメント予測に活用する．

4 評価実験

4.1 使用データ・実験設定

4.1.1 トキシックコメント予測モデル

TwitterAPI を用い，2022 年 4 月 1 日から 2022 年 6 月 30 日にかけて NHK ニュース (@nhk_news)[4]が投稿したニュースのツイート，ニュースに対するユーザの返信，返信ユーザのユーザ ID の 3 つをそれぞれニュース・コメント・投稿者 ID として取得した．

ニュースコメントに対する評価値を取得するため，クラウドソーシングサービスの一つであるクラウドワークスでアンケートを実施した．被験者の総数は 250 名であり，250 名の被験者を 50 名ごとの 5 つのグループに分けた．グループごとに 2022 年 4 月 1 日から 6 月 30 日の 3ヶ月間に投稿されたニュースコメント 400 件を提示し，各ニュースコメントが不快と感じるかについて {1: 全くそう思わない，2: あまりそう思わない，3: どちらとも言えない，4: ややそう思う，5: 非常にそう思う } の 5 段階で回答してもらった．なお，被験者に提示するニュースとコメントは各グループ内で共通である．アンケートで取得したラベルの分布を図 5 に示す．

2022 年 4 月 1 日から 4 月 14 日の 2 週間において 5 段階の評価値が 4 以上である（不快と評価された）ニュースとコメントのペアと，そのコメントを投稿したユーザの投稿者 ID および評価を行ったユーザの読者 ID を格納したデータベースをフィードバックデータベースとして構築した．フィードバックデータベースには，評価値が 4 未満である（不快と評価されなかった）ニュースとコメントは格納しない．提案手法では，読者 ID を入力としてフィードバックデータベースから読者が不快と評価したニュースとコメントのペアと，そのコメントの投稿者 ID を取得する．読者が不快と評価したニュースとコメントのベクトルと，不快と評価したコメントを投稿したユーザのベクトルを利用することにより，予測の個人化を行う．

アンケートで対象としたニュースコメントを投稿した全ユーザのフォロー情報を Twitter API を用いて取得した．取得した

4：https://twitter.com/nhk_news

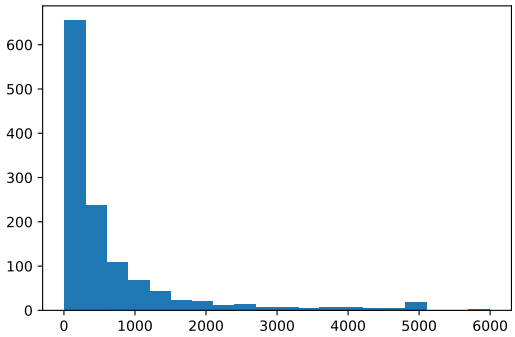


図 6: ユーザのフォロー数の分布

ユーザのフォロー情報から各ユーザのフォロー数を算出し，フォロー数の分布を表したものを図 6 に示す．ユーザのフォロー数の最小値は 0，最大値は 60,459，中央値は 273 であった．

トキシックコメント予測を行うモデルの訓練で利用する教師データについて説明する．教師データの入力データは予測対象のニュースとコメントのペアと読者 ID であり，正解ラベルは読者による 5 段階の評価値である．2022 年 4 月 15 日から 6 月 9 日のニュース・コメントを訓練データ，2022 年 6 月 10 日から 6 月 20 日のニュース・コメントを検証データ，6 月 21 日から 6 月 30 日のニュース・コメントをテストデータとして時系順に分割した．

フィードバックデータベースから 10 件以上のフィードバックを取得できた 150 人のコメント読者を対象にトキシックコメント予測モデルの訓練を行った．ニュースとコメントをベクトル化するための BERT モデルに関しては，東北大学乾研究室による日本語 Wikipedia を用いた事前学習済み BERT モデル [5] を用い，fine-tuning を行った．また，図 3 の 2 箇所の BERT 層に関して，重みは共有せずに訓練を行った．フィードバックを大量に行う行為は読者にとって負担となるため，個人化に用いる読者によるフィードバックのデータ数は少量であることが望ましい．このため，読者 ID の入力時に取得する，過去に低評価したニュースとコメントを 10 件という少量のデータに限定して個人化を行った．

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 の除去
- 記号の除去
- ハッシュタグの除去 (e.g. #nhk_news)

また，コメントのテキストには以下の前処理を行った．

- URL の除去
- 記号の除去
- 絵文字の除去
- メンションの削除 (@ユーザ ID)

4.1.3 使用モデル

提案手法によるトキシックコメント予測の結果を評価・比較するために，本実験では，複数のモデルによる予測結果の比較を行った．各モデルの詳細は以下の通りである．

Follow-BPR

図 3 に示すように，コメント投稿者の特徴を予測に利用する．コメント投稿者の特徴をエンコードするために，BPR によりフォロー関係をベクトル化する．BPR の実装はライブラリ implicit[6] を用い，コメント投稿者のベクトルの次元数は 768 とした．

Follow-LightGCN

図 3 に示すように，コメント投稿者の特徴を予測に利用する．コメント投稿者の特徴をエンコードするために，LightGCN によりフォロー関係をベクトル化する．LightGCN の実装はライブラリ recommenders[7] を用い，コメント投稿者のベクトルの次元数は 768 とした．

Past Comments

上記の 2 つの手法と同様に，コメント投稿者の特徴を予測に利用する．コメント投稿者の特徴をエンコードするために，過去のコメントを利用する [8]．コメント投稿者のベクトルの次元数は 768 とした．

Simple

コメント投稿者の特徴を予測に利用しない．予測対象のニュースコメントと，読者が過去に低評価したニュースコメントのテキストのみを用いて予測を行う．

4.2 コメント投稿者の特徴の可視化

各手法により得られたコメント投稿者の特徴を可視化し，トキシックコメント予測におけるコメント投稿者のフォロー関係の活用の有効性を調べた．アンケートで対象としたニュースコメントを投稿した全ユーザの特徴を各手法により取得し，コメント投稿者の特徴を主成分分析を用いて二次元に圧縮した．次に，各コメントに対する 5 段階の評価値の標準偏差の大きさが上位と下位のコメントを各 50 件取得し，それぞれのコメントの投稿者の特徴を可視化した．可視化したものを図 7 に示す．ここで，評価値の標準偏差が大きいコメントは，読者の価値観などによって受け取り方が大きく異なると捉えることができ，個人化の重要性が高いコメントといえる．反対に，評価値の標準偏差が小さいコメントは個人化の重要性が低いといえる．図 7 より，(a) の BPR により投稿者のフォロー関係をベクトル化した手法では，評価値の標準偏差が大きいコメントと小さいコメントの投稿者の特徴の分布を他の手法と比べて分離できている．このため，BPR によりフォロー関係をベクトル化する手法で得られたコメント投稿者の特徴は, 予測の個人化を行う際に有効であることが示唆された．

4.3 実 験 結 果

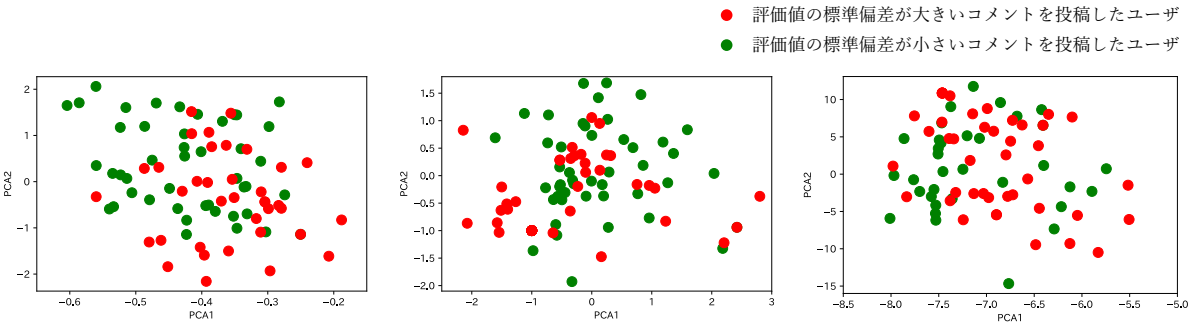
4.3.1 評価値予測の結果

テストデータを対象として評価値の予測を行い，提案手法の有用性を検証した．4.1.3 項で説明した 4 種類の評価値予測モ

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

6：https://github.com/benfred/implicit

7：https://github.com/microsoft/recommenders



(a) Follow-BPR　(b) Follow-LightGCN　(c) Past Comments

図 7: コメント投稿者の特徴の可視化

表 2: 各モデルの MAE（平均絶対誤差）

モデル	MAE
Follow-BPR	**0.756**
Follow-LightGCN	0.763
Past Comments	0.767
Simple	0.780

デルを利用し，被験者が回答した評価値と予測される評価値の誤差を MAE（平均絶対誤差）で評価した．4 種類の評価値予測モデルによる評価値予測の結果を表 2 に示す．表 2 より，Follow-BPR，Follow-LightGCN，Past Comments，Simple の順に小さい予測誤差が得られた．これより，トキシックコメントの予測においてコメント投稿者の特徴を活用することが有効であることがわかった．さらに，提案手法であるユーザが属するコミュニティを表すフォロー関係を用いてコメント投稿者の特徴を取得することが，評価値予測の性能向上に有効であることがわかった．

4.3.2　トキシックコメント分類の結果

読者ごとに予測評価値が上位 k 件のニュースコメントを「トキシックコメント」と判定したときの正解率 (=Precision@k) を求めることで，2 値分類における提案手法の有用性を検証した．正解ラベルに関して，アンケートで 4 または 5 の評価値が付けられたコメントを「トキシックコメント」とし，それ以外の評価値が付けられたコメントは「トキシックコメントでない」とした．SNS などのインターネット上のコンテンツでは，推薦やフィルタリングによって自分の興味があるコンテンツばかりが提供されてしまうフィルターバブルと呼ばれる現象があり，特定の信念が増幅または強化されることが問題となっている [21]．このため，本実験では予測評価値が上位のニュースコメントに対する正解率から予測性能の評価を行うことにより，読者にとって特に有害なコメントのみをフィルタリングすることを目的とした．k=1 から k=10 と変化させたときの Precision@k の値を図 8 に示す．図 8 より，ほとんどの k の値で提案手法である Follow-BPR モデルや Follow-LightGCN モデルから他のモデルと比較して高い Precision の値が得られた．図 7 のコメント投稿者の特徴の可視化では，Follow-BPR モデルが最も投稿者の特徴を分離できていた．しかし，本実験のト

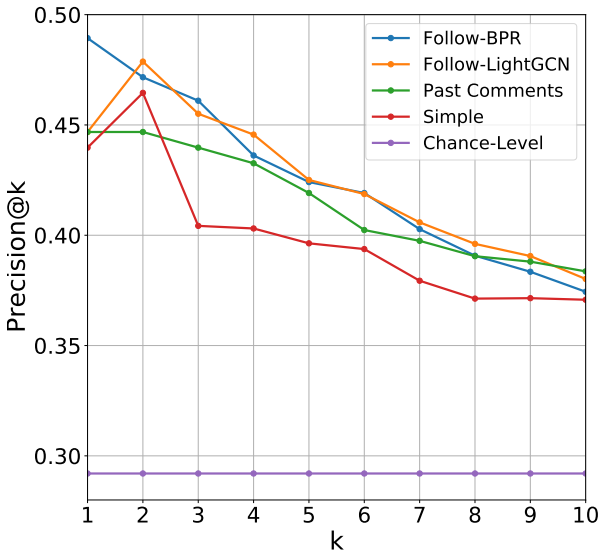


図 8: Precision@k の推移

キシックコメント分類では Follow-LightGCN が Follow-BPR と同様にトキシックコメント分類の性能向上に有効であった．また，k の値が大きくなるにつれて予測評価値が小さいコメントも「トキシックコメント」と判定される．このため，全てのモデルで徐々に Precision@k の値が小さくなり，各モデルの性能の差も徐々に小さくなった．

提案手法で得られた最も高い Precision の値は 0.5 程度であり，2 値分類における性能としては不十分である．しかし，図 5 より今回取得したラベルは不均衡であり，本実験における Precision の値のチャンスレベルは約 0.3(=29,200/100,000) である．このため，ラベルの不均衡を考慮すると，提案手法によって十分な分類結果が得られたといえる．

5　ま　と　め

本論文では，ニュースコメントの読者が過去に低評価したニュースコメントのフィードバックをもとに，トキシックコメント予測の個人化に取り組んだ．提案手法では，予測対象のコ

メントと読者が過去に低評価したコメントの利用に加え，コメントの投稿者の特徴を活用する機械学習モデルを用いた．また，コメントの投稿者の特徴はその投稿者が属するコミュニティによって決定されるという仮定をもとに，Twitter のフォロー関係に基づいてコメントの投稿者の特徴を取得した．評価実験の結果，フォロー関係に基づいて取得したコメント投稿者の埋め込み表現を予測に活用することにより，予測の個人化の性能が向上することが示された．

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.
[2] Natalie Jomini Stroud, Emily Van Duyn, and Cynthia Peacock. News commenters and news comment readers. *Engaging News Project*, pp. 1–21, 2016.
[3] Julian Risch and Ralf Krestel. *Toxic Comment Detection in Online Discussions*, pp. 85–109. Springer Singapore, Singapore, 2020.
[4] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. Convolutional neural networks for toxic comment classification. No. 35 in In Proceedings of the 10th Hellenic Conference on Artificial Intelligence, pp. 1–6, 2018.
[5] Google Jigsaw. Perspective API, 2017.
[6] Jack Hessel and Lillian Lee. Something's brewing! early prediction of controversy-causing posts from discussion features. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1648–1659, 2019.
[7] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, pp. 5884–5906, 2022.
[8] Teruki Nakahara and Taketoshi Ushiama. Personalized prediction of offensive news comments by considering the characteristics of commenters. In Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, 2023. (to appear).
[9] 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.
[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, p. 993–1022, 2003.
[11] Ahmet Aker, Emina Kurtic, A. R. Balamurali, Monica Paramita, Emma Barker, Mark Hepple, and Rob Gaizauskas. A graph-based approach to topic clustering for online comments to news. In Proceedings of the 38th European Conference on IR Research, pp. 15–29, 2016.
[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1977.
[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186, 2019.
[14] Martin Saveski, Brandon Roy, and Deb Roy. The structure of toxic conversations on twitter. In Proceedings of the Web Conference, pp. 1086–1097, 2021.
[15] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In Proceedings of the Seventeenth Symposium on Usable Privacy and Security, pp. 299–318, 2021.
[16] 中原輝樹, 牛尼剛聡. 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. 情報処理学会論文誌データベース (TOD), Vol. 15, No. 3, pp. 99–110, 2022.
[17] Yumeng Li, Liang Yang, Bo Xu, Jian Wang, and Hongfei Lin. Improving user attribute classification with text and social network attention. *Cognitive Computation*, pp. 1–10, 2019.
[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems, pp. 6000–6010, 2017.
[19] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, p. 452–461, 2009.
[20] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 639–648, 2020.
[21] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* penguin UK, 2011.