CL-006

SNS ユーザの自己紹介文にあらわれる偏りに着目したフェイクニュース検知
Fake News Detection Focusing on Bias of SNS User Profiles

古川 凌也[†‡] 伊藤 大貴[†] 髙田 雄太[†] 熊谷 裕志[†]
Ryoya FURUKAWA Daiki ITO Yuta TAKATA Hiroshi KUMAGAI
神薗 雅紀[†] 白石 善明[‡] 森井 昌克[‡]
Masaki KAMIZONO Yoshiaki SHIRAISHI Masakatu MORII

概要

SNS におけるフェイクニュースの拡散が問題となっている．フェイクニュースを共有するユーザは承認欲求や帰属欲求，自己顕示欲といったヒューマンニーズが強く，自己紹介文に特徴的な単語が現れやすい．本研究では，それら自己紹介文に含まれる単語の偏りに基づき，フェイクニュースを検知する手法を提案する．Twitter 上で同一のニュース URL を投稿する複数ユーザの自己紹介文に含まれる単語から特徴量ベクトルを作成し，機械学習によりその真偽を分類する．日米のリアルおよびフェイクニュースを含む複数のデータセットを用いて実験を行い，平均 90.2%の分類精度を実現した．さらに提案手法がフェイクニュースの検知のみでなく，フェイクニュースの標的とされたユーザ集団の分析等に有効であることを示す．

1. はじめに

近年，ソーシャルネットワーク・サービス（SNS）におけるフェイクニュースの拡散が社会問題となっている．フェイクニュースとして扱われる事象は，デマや陰謀論，プロパガンダなど多岐にわたり，研究者によってさまざまな定義や分類がある[1]．フェイクニュースは，誤報や誤解の広まりとして発生する場合もあるが，主に特定の目的を実現するため，または政治的，個人的もしくは金銭的な利益のために，人々を騙すよう意図的に作成される．例えば，2016 年の米国大統領選挙では，フェイクニュースを掲載する 100 以上のサイトがマケドニアの若者らによって運営されており，彼らはクリック広告によって半年で数万ドルの利益を得たとされる[2].

フェイクニュースへの対策として，メディアや専門家によるファクトチェックが行なわれている．しかしながら，手動によるファクトチェックは専門知識と多大な労力を要する作業であるため，大量に作成されるフェイクニュースに対してスケーラビリティに乏しい．そのため，データマイニングや機械学習などによってフェイクニュースを自動的に検知する手法が研究されており，検知に利用可能な特徴として様々なものが検討されている[1]．これらの特徴は大きく分けてニュース記事のテキストや画像といったコンテンツから得られる特徴とソーシャルネットワークにおけるニュースの伝播パターンといったコンテクストの特徴に分類される．コンテンツの特徴は，フェイクニュースの作成者にとって自由に操作できるものであることから，それのみでは不十分とされる[1]．そのため，コンテクストの特徴と組み合わせることでロバスト性を向上させた手法が提案されているが，モデルが複雑化し，より大量のデータが必要となることが一般的である．したがって，フェイクニュースの検知に効果的でありながら，より軽量な特徴を利用する手法が求められている．

そこで本研究では，フェイクニュースへの対策として，ニュースのテキストやソーシャルネットワークといったデータを使用せず，ニュースを共有した SNS ユーザに暗黙的なつながりがあると見なし，それらユーザの自己紹介文を用いてフェイクニュースを検知する手法を提案する．一般的に SNS ユーザがニュースを共有する動機には，自身の意見や信念の主張，情報共有による他者への貢献，オピニオンリーダや情報のゲートキーパとしての立場の獲得などがあると言われている[3]．特に人々の目を引くようなフェイクニュースを共有するユーザは，これらの動機に関連した承認欲求や帰属欲求，自己顕示欲といった欲求（ヒューマンニーズ）が強い傾向にあると推量される．また，SNS ユーザの性質やふるまいとフェイクニュース共有行動との関連を調査した既存研究では，フェイクニュースを共有しやすいユーザは，人気や社会的なつながりの獲得のためにプロフィールを詳細に記載するといった傾向が示されている[4]．したがって，フェイクニュースを共有しやすいユーザの自己紹介文には，個人の性質や属性，コミュニティ等に関連した特徴的な単語が現れやすいと考えられ，これらの特徴的な単語の偏りがフェイクニュースの検知に利用できる．

評価実験では，日米のリアルおよびフェイクニュースを含む複数のデータセットを用い，提案手法による検知の性能評価を行った．その結果，最高で 95.8%，平均で 90.2%の分類精度が得られ，最新の既存手法と同等またはそれ以上の性能が得られている．また，日本語と英語の両方のデータセットにおいて良好な性能を示しており，提案手法は異なる言語に対しても有効であった．さらにケーススタディでは，自己紹介文に含まれる単語に対するニュースの特徴量ベクトルの分布を可視化し，フェイクニュースの検知に留まらない提案手法の応用例を示している．

本論文の貢献は以下のとおりである．

- フェイクニュースを共有しやすい SNS ユーザの自己紹介文に現れる特徴的な単語の偏りに着目し，軽量かつ高精度にフェイクニュースの検知を行う手法を提案した．
- 実世界のリアルおよびフェイクニュースを含む複数のデータセットを用い，その他の特徴を使用する既存手法と性能を比較する評価実験を行い，提案手法の有効性を示した．

†デロイト トーマツ サイバー合同会社
Deloitte Tohmatsu Cyber LLC
‡神戸大学
Kobe University

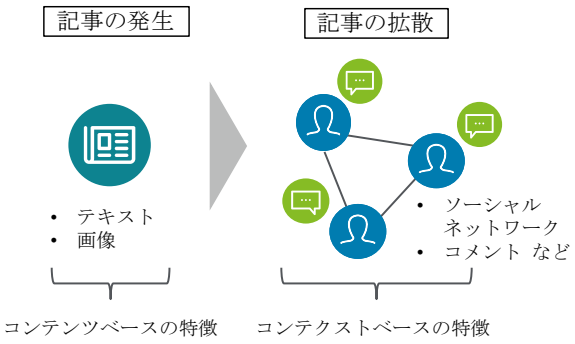


図 1 フェイクニュース検知に利用される特徴

図 2 プロフィール画面の例
（引用: https://twitter.com/twitter）

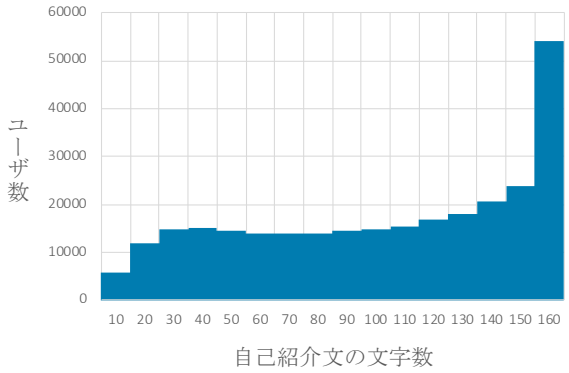


図 3 自己紹介文の文字数の分布

- ケーススタディとして，自己紹介文に含まれる単語に対する提案手法を用いたニュースの特徴量ベクトルの分布を可視化し，フェイクニュースが標的とした可能性のあるユーザ集団の分析といった，検知にとどまらない提案手法の応用の可能性を示した．

2. 関連研究

本章ではフェイクニュースの検知の既存手法について説明する．既存手法は大きく分け，ニュース記事のテキストや画像から特徴を抽出して使用するコンテンツベースの手法とソーシャルグラフやデータ伝播の特徴などを使用するコンテクストベースの手法に分類される．それぞれの手法によって，フェイクニュースを検知可能なタイミングも異なる[1]．フェイクニュースの検知に利用される特徴を図 1 に示す．

2.1 コンテンツベースの手法

テキストの特徴を使用する手法は，使用される単語や文体などの傾向を元にフェイクニュースの検知を行う[5][6]．心理学的な観点から言語表現の特徴を整理した辞書である LIWC[7]などを用いた調査では，フェイクニュースのテキストには，主観的で感情的な記述が多く，砕けた表現や卑語が多用されることが示されている[6]．画像を用いる手法では，画像の統計的な特徴やニューラルネットワークなどを用いて抽出される潜在的な特徴を使用するものがある[8][9][10]．画像の統計的な特徴を用いる手法は，実際のニュースの場合では，ニュースに関連して多様な画像が投稿されるのに対し，フェイクニュースの場合では類似する画像が多く投稿され，分布が偏ることに基づく[8]．ニューラルネットワークを用いた手法では，フェイクニュースではしばしばニュースの内容に無関係な画像が使用されることに着目し，テキストと画像の両方の特徴から両者に生じるギャップを捉えるものなどがある[10]．

コンテンツベースの手法は，記事が発生したタイミングから検知を行うことが可能である．しかしながら，コンテンツの特徴はフェイクニュースの作成者が直接的に操作でき，検知を回避される可能性が指摘されている[1]．

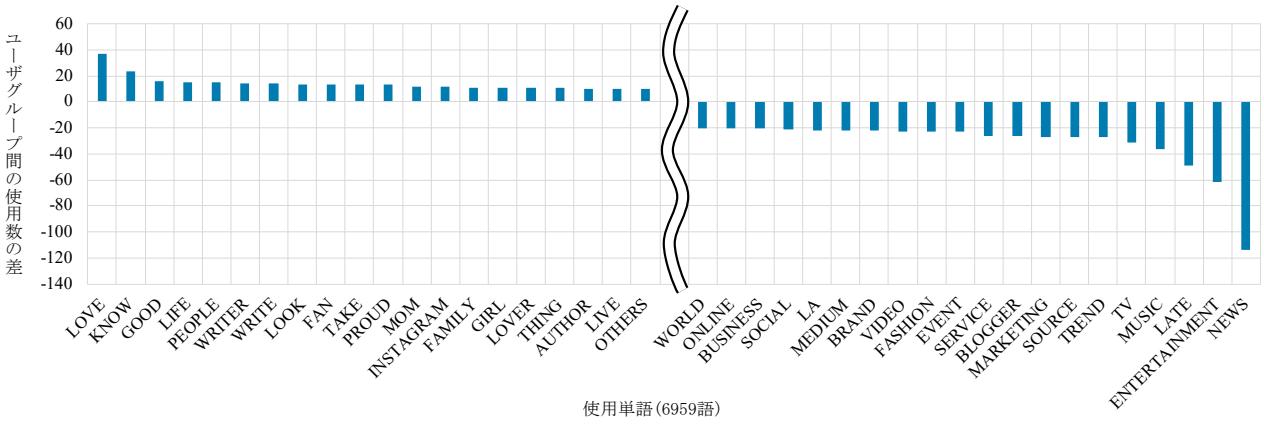
2.2 コンテクストベースの手法

コンテクストベースの手法は，SNS 上でニュースが拡散される際の特徴に着目しており，ソーシャルネットワーク上でのニュースの伝搬パターンやニュースが共有される際に付与されたコメント，ニュースに反応したユーザの特徴などが用いられる．ニュースの伝搬パターンに着目した手法では，フェイクニュースは実際のニュースよりも，より素早く，より多くのユーザに拡散されるとし，伝搬パターンの構造的な特徴を用いて検知を行う[11]．ニュースのコメントに着目した手法では，Co-Attention とよばれるアーキテクチャを用い，ニュースのテキストとコメントを共同で学習することで，説明可能性を持つ検知モデルを構築している[12]．ユーザの特徴を用いる手法では，アカウントの作成時期や位置情報，フォロワ数といった明示的な統計情報や，ツイートの傾向から推測される性別や年齢，政治的なバイアスといった暗黙的な情報を用いる手法が検討されている[13]．その他には，ニュースの発信源の政治的なバイアスや同一のニュースを共有するユーザが形成する暗黙的なネットワークなどの特徴を組み合わせる手法などがある [14][15]．

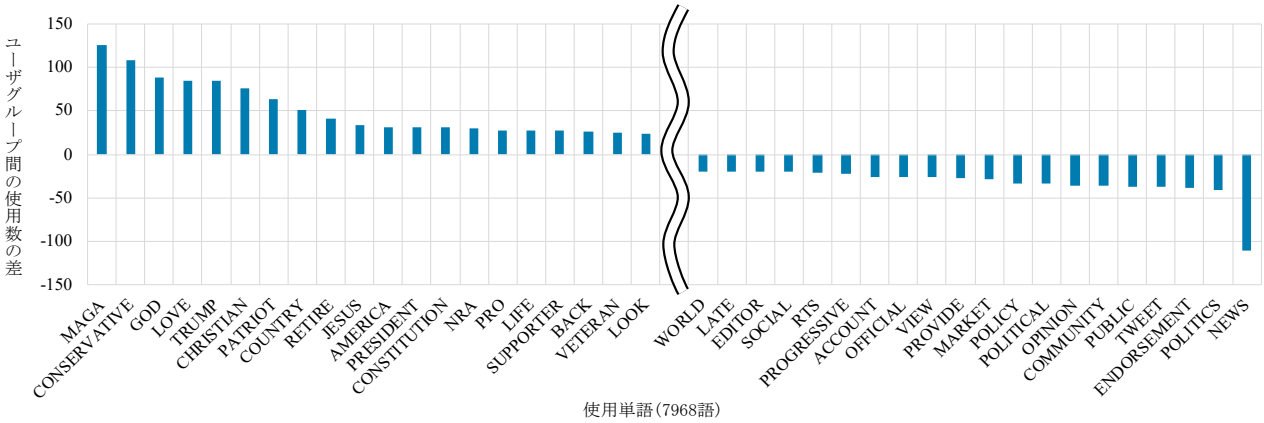
コンテクストベースの手法で利用される特徴は，フェイクニュースの作成者に操作されづらく，よりロバストである言える．しかしながら， 多くの手法ではソーシャルネットワークや関連するユーザのツイート履歴といった特徴を用い，より大量で複雑なデータの収集が必要となることが多い．提案手法はコンテクストベースの手法であり，中でもユーザの特徴と暗黙的なネットワークに着目しているが，使用するデータは自己紹介文のみであり，比較的軽量に収集することが可能である．

3. 自己紹介文に関する事前調査

本章では自己紹介文に関する事前調査を通して，フェイクニュースの検知に利用可能な特徴について検討する．Twitter では，ユーザが自身のプロフィール画面に表示する自己紹介文を最大 160 文字までの範囲で任意に設定することができる．例として 2021 年 6 月 17 日時点の Twitter 社ア



(a) GossipCop(4,000 ユーザを抽出)



(b) PolitiFact(4,000 ユーザを抽出)

図 4 フェイクニュース共有傾向の高いユーザグループと低いユーザグループの間でみられた自己紹介文に使用される単語の傾向の違い（2 種類のデータセットで比較）

カウントのプロフィール画面を図 2 に示す．“What’s happening?!” と記載されている箇所が自己紹介文にあたる．英語圏のリアルおよびフェイクニュースの記事データを含んだ既存のデータセットである FakeNewsNet[16]をもとに，Twitter 上でニュース URL を投稿したすべてのユーザを対象として自己紹介文を収集した．Twitter API[1]を用いてデータの収集を行い，計 337,117 のユーザアカウントのプロフィールデータを取得した．これらユーザのうち，約 83.6%にあたる 281,936 のユーザが自己紹介文を設定していた．設定されていた自己紹介文の文字数の分布を図 3 に示す．多くのユーザが字数制限を最大まで使用した比較的長い自己紹介文を設定していた．

収集した自己紹介文を対象に，ユーザのフェイクニュース共有傾向と当該ユーザの自己紹介文に使用される単語との関連を調査した．手順を以下に述べる．

1. まず，投稿したニュース URL の 50%以上がフェイクニュースの URL であったユーザをフェイクニュース共有傾向の高いグループとし，残りのユーザをフェイクニュース共有傾向の低いグループとして，ユーザを 2 グループに分割した．
2. 次に，特定の記事を共有したユーザに偏らないよう調節しつつ，2 グループの自己紹介文を同数ずつ抽出し，グループごとに単語の使用数を集計した．
3. 各単語についてフェイクニュース共有傾向の高いグループにおける使用数と低いグループにおける使用数との差を取ることで，各グループに特徴的な単語を抽出した．

なお，FakeNewsNet は政治系ニュースのファクトチェックサイトである PolitiFact[2]と芸能ゴシップ系ニュースのファクトチェックサイトである GossipCop[3]の 2 つをデータソースとしており，それぞれ異なる話題のニュースが含まれている．ニュースの話題によって特徴的な単語の傾向が異なる可能性があるため，各ソースからのデータについて個別に結果を算出した．

事前調査の結果を図 4 に示す．グラフの数値は 2 グループでの使用数の差であり，正の値であればフェイクニュース共有傾向の高いグループにおいて使用数が多く，負の値であればフェイクニュース共有傾向の低いグループにおいて使用数が多いことを示している．全単語のうち，それぞれのグループで特に差の大きかった各 20 単語のみ表示している． 図 4(a)と図 4(b)の両者に共通する特徴として，フェ

[1] https://developer.twitter.com/en/products/twitter-api
[2] https://www.politifact.com/
[3] https://www.gossipcop.com/

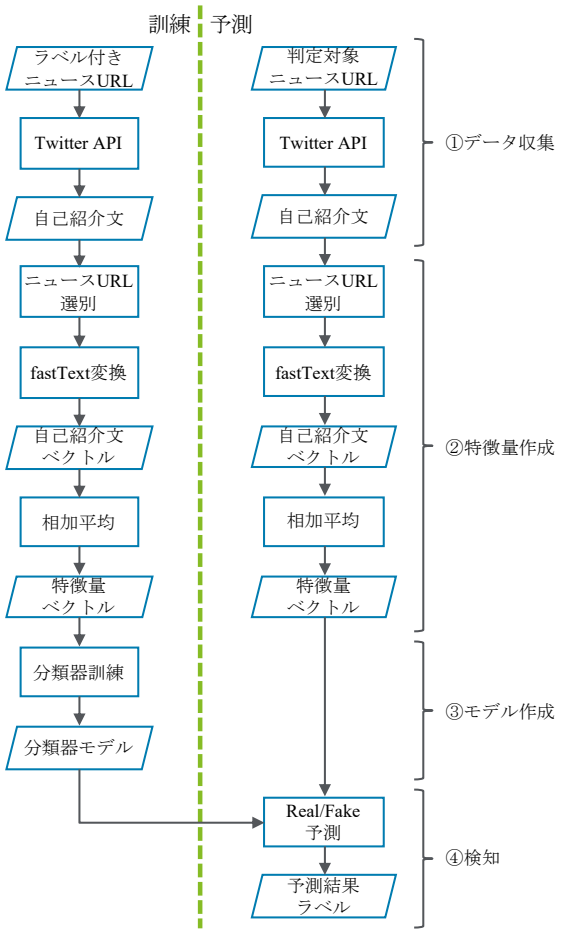


図 5 提案手法の処理の流れ（fastText）

イクニュース共有傾向の高いユーザグループでは LOVE の使用数が大きくなっていることが挙げられる．これは，“I love theater and music.”といった個人アカウントの自己紹介文に含まれるような趣味趣向に関する記載の特徴が表れているものと考えられる．それに対し，フェイクニュース共有傾向の低いユーザグループでは，NEWS や OFFICIAL，MEDIUM といった単語の使用数が大きくなっており，これはニュースメディア等の公式アカウントが多く含まれていることによるものと考えられる．図 4 (b)では，MAGA（Make America Great Again）や CONSERVATIVE，CHRISTIAN といった，米国における保守派の政治思想と関連するキーワードの使用数が大きくなっている．この結果は，フェイクニュース共有傾向と SNS ユーザの特性との関連を調査した先行研究の結果とも一致している[17][18]．以上から，ニュース URL を共有するユーザの自己紹介文にはフェイクニュース共有傾向と関連するキーワードが含まれており，フェイクニュース検知のための特徴として利用することができると考えられる．

4. 提案手法

提案手法は，ツイッター上で同一のニュース URL を投稿した，暗黙的なつながりを持つユーザの自己紹介文に含まれる単語の傾向をもとに記事の特徴量ベクトルを作成し，機械学習によりフェイクニュースの検知を行う．本論文では，単語の出現傾向を捉えた特徴量ベクトルの作成にあたり，事前学習した fastText を用いる手法と TF-IDF を用いる手法の 2 つを検討した．

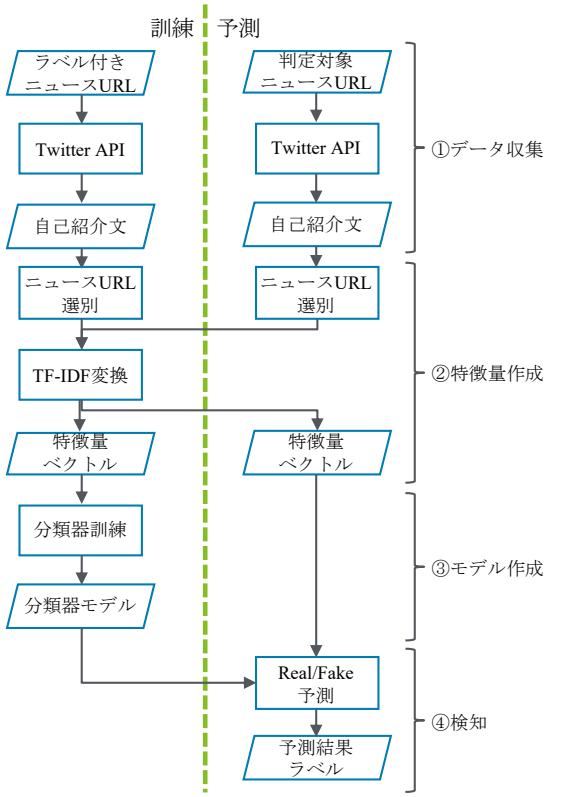


図 6 提案手法の処理の流れ（TF-IDF）

4.1 定義

ニュース URL の集合を$\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$とし，$\mathcal{A}_{train} = \{a_1, a_2, \ldots, a_r\}$をラベル付きのニュース URL，$\mathcal{A}_{test} = \{a_{r+1}, a_{r+2}, \ldots, a_n\}$をラベルが未知のニュース URL とする．これらのニュース URL をツイッター上に投稿したユーザの自己紹介文の集合を$\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$とする．このとき，同一のニュース URL を共有するユーザによる暗黙的なネットワークを$\mathrm{X} \in \{0,1\}^{n \times m}$を用いて表し，$\mathrm{X}_{ij} = 1$のとき，自己紹介文$p_j$に対応するユーザがニュース記事$a_i$の URL を投稿したものと定める．さらに，$\mathcal{P}$の自己紹介文に含まれる$t$件の単語からなる集合を$\mathcal{W} = \{w_1, w_2, \ldots, w_t\}$とし，$\mathrm{P} \in \mathbb{R}^{m \times t}$を自己紹介文に含まれる単語の Bag-of-Words（BoW）とする．

4.2 事前学習した fastText を用いた手法

事前学習した fastText を用いた手法の処理の流れを図 5 に示し，以下に手順を説明する．

① **データ収集** $\mathcal{A}$に含まれるすべてのニュース URL を Twitter API を用いて検索し，$\mathcal{P}$およびXを得る．

② **特徴量作成** 事前学習した fastText を用いて，収集した自己紹介文$\mathcal{P}$をそれぞれ単語の分散表現に変換する．単語w_kに対応するd次元の分散表現を$fastText(w_k) \in \mathbb{R}^{d \times 1}$とし，自己紹介文$p_j$に含まれる単語の分散表現の相加平均を自己紹介文ベクトルU_jとする．すなわち，U_jを次のように定める．

$$U_j = \frac{1}{\sum_{l=1}^{t} \mathrm{P}_{jl}} \sum_{k=1}^{t} fastText(w_k) \cdot \mathrm{P}_{jk}$$

表 1 データセットに含まれる Fake/Real のラベル数

データセット	Fake	Real
PolitiFact	328	277
GossipCop	3,084	14,078
PolitiFact (old-version)	53	95
BuzzFeed (old-version)	56	82
FIJ – 日本語	224	596

次に，$\mathcal{A}' = \left\{a_i | \, a_i \in \mathcal{A}, \, \sum_j^m \mathrm{X}_{ij} \geq \mathrm{N}\right\}$ として，N 件以上の自己紹介文を持つニュース URL を選別する．$a_i \in \mathcal{A}'$を投稿したユーザの自己紹介文ベクトルU_iの相加平均をとり，特徴量ベクトル$V_i^{fastText}$とする．すなわち，$V_i^{fastText}$を次のように定める．

$$V_i^{fastText} = \frac{1}{\sum_{l=1}^{m} \mathrm{X}_{il}} \sum_{j=1}^{m} U_j \cdot \mathrm{X}_{ij}$$

③ **モデル作成** $\mathcal{A}'_{train} = \{a_i | \, a_i \in \mathcal{A}_{train} \wedge \mathcal{A}'\}$ に対応する特徴量ベクトルと Fake/Real ラベルを用いて任意の学習アルゴリズムを訓練し，分類器を構築する．

④ **検知** 構築した分類器を用い，$\mathcal{A}'_{test} = \{a_i | \, a_i \in \mathcal{A}_{test} \wedge \mathcal{A}'\}$に対応するニュースベクトルの Fake/Real ラベルの予測を行い，未知のニュース URL からフェイクニュースを検知する．

fastText を用いた手法は，利用する学習済みモデルに性能が依存するが，単語間の意味的な類似度を加味した特徴ベクトルが算出可能であり，自己紹介文の間の単語の表記ゆれや類義語の使用にも柔軟に対応できる可能性がある．また，新規データを予測する場合にもモデルを再学習する必要がない．

4.3 TF-IDF を用いた手法

TF-IDF を用いた手法の処理の流れを図 6 に示し，以下に手順を説明する．なお，①データ収集、③モデル作成、④検知は fastText を用いた手法と同様の処理であるため説明を割愛し、②特徴量作成のみ説明する．

② **特徴量作成** fastText を用いた手法と同様に$\mathcal{A}'$を選別する．$\mathcal{A}'$に含まれる各ニュース URL について自己紹介文の BoW を合算し，ニュース URL ごとの BoW として $\mathrm{A} \in \mathbb{R}^{|\mathcal{A}'| \times t}$を得る．すなわち，$\mathrm{A}_i$を次のように定める．

$$\mathrm{A}_i = \sum_{j=1}^{m} \mathrm{P}_j \cdot \mathrm{X}_{ij}$$

次に，Aをもとに各単語の TF-IDF を算出し，特徴量ベクトルV_i^{TF-IDF}を求める．すなわち，ベクトルV_i^{TF-IDF}を次のように定める．

$$V_i^{TF-IDF} = \|(v_{i1}, v_{i2}, \ldots, v_{it})\|_2$$

$$v_{ik} = \; tf_{ik} \cdot idf_k$$

表 2 グリッドサーチの探索パラメータ

アルゴリズム	パラメータ	数値
ロジスティック回帰	C	0.01, 0.1, 1, 10, 100, 200, 500,1000
SVM	C	0.01, 0.1, 1, 10 ,100
	gamma	0.01, 0.1, 1, 10 ,100
ランダムフォレスト	n_estimators	10, 100, 200, 500
	min_sample_leaf	1, 2, 3, 4, 5
	min_sample_split	2, 5, 10
	max_depth	4, 5, 6, 7, 8, 9

なお，ニュース URL ごとの単語数の違いによる影響を考慮し，L2 ノルムによって正規化している．$\mathcal{A}'^k = \{a_s | \, a_s \in \mathcal{A}', \mathrm{A}_{sk} > 0\}$を単語$w_k$が出現した記事の集合とし，$tf_{ik}$および$idf_k$を次のように定義する．

$$tf_{ik} = \frac{\mathrm{A}_{ik}}{\sum_{l=1}^{t} \mathrm{A}_{il}}$$

$$idf_k = \log \frac{n+1}{|\mathcal{A}'^k| + 1} + 1$$

TF-IDF ベクトルの次元数は$\mathcal{W}$に含まれる単語数tに応じて大きくなるため，必要に応じて主成分分析(PCA)や特異値分解（SVD）などを適用して次元削減を行う．

TF-IDF を用いた手法では，出現したすべての単語に重み付けをした上で特徴ベクトルに取り込むことができる．しかしながら，新規データを予測する場合には単語ごとに IDF 値の更新が必要となり，再学習を行う必要がある．

5. 評価実験

提案手法の有効性を評価するために，複数のデータセットを用いて実験を行う．

5.1 データセット

実験には英語と日本語の二種類のデータセットを用いた．英語のデータセットには，既存のデータリポジトリである FakeNewsNet[1]を用いた．FakeNewsNet は，ニュースのコンテンツに加え，Twitter ユーザのエンゲージメントといった社会的文脈や時間情報などの多様な特徴を含んでおり，複数の特徴を検知モデルに組み込んだ最新手法のベンチマークとして利用されている [6][10][12][14][16]．また，FakeNewsNet に含まれる記事は，PolitiFact や GossipCop, BuzzFeed といったファクトチェックサイトから収集され，専門家によって Fake/Real のラベル付けがなされている．PolitiFact と BuzzFeed は政治系ニュースを対象とし，GossipCop は芸能系のニュースを対象としていることから，それぞれ異なる話題のニュースについて検証が可能である．なお，FakeNewsNet は 2019 年 5 月に大きく更新がなされており，初期のバージョンは，old-version として公開されている．5.4 節における既存手法との比較のために，本論文では現在のバージョンと old-version の両方を対象とし，各ファクトチェックサイトの記事を含むサブセットごとに実

[1] https://github.com/KaiDMML/FakeNewsNet

表 3 評価結果

		fastText			TF-IDF		
		LR	**RF**	**SVM**	**LR**	**RF**	**SVM**
PolitiFact	Acc.	.851±.005	.853±.030	.868±.017	**.879±.029**	.864±.031	.858±.027
	F1.	.857±.005	.862±.030	.874±.017	**.888±.026**	.870±.033	.868±.026
GossipCop	Acc.	.900±.103	.870±.130	.943±.035	.932±.044	**.958±.014**	.947±.010
	F1.	.809±.144	.769±.151	.867±.070	.846±.078	**.894±.031**	.856±.024
PolitiFact (old-version)	Acc.	.805±.098	.817±.064	.825±.079	.865±.078	**.926±.024**	.899±.061
	F1.	.779±.109	.790±.064	.793±.085	.831±.098	**.893±.038**	.879±.071
BuzzFeed (old-version)	Acc.	.819±.034	**.848±.017**	.826±.066	.833±.030	.840±.071	.848±.028
	F1.	.780±.060	**.825±.018**	.788±.121	.802±.049	.818±.074	.821±.051
FIJ（日本語）	Acc.	.888±.042	.888±.051	.884±.039	.894±.058	.884±.073	**.900±.064**
	F1.	.752±.110	.746±.106	.753±.116	.760±.140	.742±.138	**.771±.144**

験を行った．日本語のデータセットは，認定NPO法人ファクトチェック・イニシアティブ（FIJ）が運営するFactCheck Navi[1]に2020年12月1日時点で「ミスリード」，「不正確」，「根拠不明」，「誤り」，「虚偽」として掲載されていたニュース記事を Fake ラベルのデータとし，NHKや産経新聞，日経新聞，読売新聞，朝日新聞といった複数の国内主要メディアのニュースサイトからRSSで取得したニュース記事をRealラベルのデータとして新たに収集し，手動でラベル付けを行った．各データセットのラベル数を表1に示す． なお本論文では，自己紹介文を設定しているユーザ5名以上によってTwitter上にURLが投稿されたニュース記事を実験対象としている．

5.2 実装

fastTextの事前学習モデルは，公式サイト[2]で配布されている 157 言語のモデルから英語と日本語のものをベースラインとして利用した．これらは Common Crawl[3]と Wikipedia[4]から学習されており，次元数は 300 となっている[19]．TF-IDF ベクトルもこれに揃え，SVD により次元数を 300 に削減した．自己紹介文に含まれる単語の抽出には，英語についてはNLTK[20]を用い，日本語についてはfastTextの事前学習モデルに合わせ MeCab[21]を用いた．分類器となる学習アルゴリズムには，より幅広い検証のためにロジスティック回帰とSVM，ランダムフォレストの3種類を選択し，実装にはscikit-learn[22]を用いた．ラベルの不均衡には，学習時にトレーニングデータをランダムアンダーサンプリングすることで対応した．対象とするニュースURLに最低限含まれる自己紹介文数の値はN=5とした．

5.3 評価方法

各データセットと特徴ベクトル，学習アルゴリズムの組み合わせについて，5×2 の Nested Cross-Validation を行った．具体的には， まずデータセットの20%をテストサブセットとし，80%をトレーニングサブセットとして分割する．次にトレーニングサブセットを用いて 2 分割交差検証によるグリッドサーチを行い，最も精度の良い学習アルゴリズムのハイパーパラメータを選択する．探索したパラメータと数値を表 2 に示す．最後にテストサブセットを用いて，選択したハイパーパラメータでトレーニングサブセットを学習したモデルの性能を評価する．この行程を 5 回繰り返し，得られた値の平均値と標準偏差を報告する．これにより，小規模なデータセットであっても性能評価におけるバイアスを低減することができる[23]．評価指標には，機械学習の分野で一般的に用いられる Accuracy，Precision，Recall，F1 スコアを用いた．

5.4 評価結果

評価結果を表 3 に示す．紙面の節約のため，Accuracy と F1 スコアの値のみ記載している．LR はロジスティック回帰，RF はランダムフォレストの結果である．fastText を用いた場合と TF-IDF を用いた場合を比較すると，全般的にTF-IDFを用いた手法の性能が良好な結果を示している．これはTF-IDFを用いた手法がfastTextを用いた手法に比べ，より多くの出現単語とその重要度を特徴ベクトルに取り込むことができた結果であると考えられる．また，学習済みモデルの語彙に単語が含まれていたとしても，学習元の文書とTwitterの自己紹介文では単語間の意味的関係が異なることにより性能が低下した可能性がある．たとえば，相反する意味を持つ単語であっても，政治や芸能などの同一のドメインで用いられる単語であれば，分散表現における類似度が高くなる場合がある．したがって，新たにTwitterの自己紹介文から語彙を学習したものに事前学習モデルを置き換えるなどによって性能が向上すると考えられる．また，学習アルゴリズムごとの性能を比較すると，データセットと特徴ベクトルの組み合わせごとに，最良の性能を示すアルゴリズムにはばらつきがあった．したがって，分布の異なる学習データを用いる場合には，適切な学習アルゴリズムを選択する必要がある．提案手法は，GossipCop と PolitiFact の両方において良好な性能を示している．これは，PolitiFact が扱う政治的なニュースにおける党派性のように，

[1] https://navi.fij.info/
[2] https://fasttext.cc/
[3] https://commoncrawl.org/
[4] https://www.wikipedia.org/

表 4 既存手法と提案手法の性能比較

		[14]	[6]	[10]	[12]	提案手法 (best)
P	Acc.	-	-	0.874	**0.904**	0.879
	Pre.	-	-	0.889	**0.902**	0.897
	Rec.	-	-	0.903	**0.956**	0.881
	F1.	-	-	0.896	**0.928**	0.888
G	Acc.	-	-	0.838	0.808	**0.958**
	Pre.	-	-	**0.857**	0.729	0.826
	Rec.	-	-	0.937	0.782	**0.977**
	F1.	-	-	**0.895**	0.755	0.894
P (old)	Acc.	0.878	0.892	-	-	**0.926**
	Pre.	0.867	0.877	-	-	**0.928**
	Rec.	0.893	**0.908**	-	-	0.871
	F1.	0.880	0.892	-	-	**0.893**
B (old)	Acc.	0.864	**0.879**	-	-	0.848
	Pre.	0.849	**0.857**	-	-	0.783
	Rec.	0.893	**0.902**	-	-	0.892
	F1.	0.870	**0.879**	-	-	0.825

ニュースを共有するユーザの間に対立構造のような関係がなくとも，提案手法がフェイクニュースを共有するユーザの自己紹介文にあらわれる偏りの特徴を捉えていることを示唆している．したがって，提案手法はニュースの話題を問わず，ドメイン横断的に適用できる．加えて，英語と日本語のようにデータセットの言語が異なっても，提案手法は良好な性能を示している．これは，異なる言語や文化圏においてもフェイクニュースを共有するユーザの自己紹介文の特徴には偏りがあらわれることを示している．したがって，提案手法は多様な言語や文化圏おけるフェイクニュースの検知に有効である．

本論文と同じく FakeNewsNet を用いて評価された最新手法の性能を表 4 に引用する．P は PolitiFact，G は GossipCop，B は BuzzFeed の結果である．データの取得タイミングや実験のセットアップが異なるため，記載の値とは単純に比較はできないが，提案手法の性能を評価するにあたって参考とする．表 5 に比較対象の既存手法が使用している特徴を示す．提案手法はコンテンツの特徴を全く使用していないにも関わらず，ほとんどの結果において最新の手法に匹敵する性能を示している．特に GossipCop の結果では高い Accuracy を示しており，芸能系ニュースを投稿する Twitter ユーザの自己紹介文には特徴が現れやすいことがわかる．既存手法の多くは複数の特徴を活用し，検知モデルに組み込むことで性能を向上させている．提案手法はユーザの特徴とニュース URL でユーザを接続した暗黙的なネットワークの特徴のみで比較的高い性能を発揮している．

6. ケーススタディ

本章では，フェイクニュース検知に留まらない提案手法の応用例を示す．2016 年の米国大統領選挙や同年の Brexit 国民投票の期間中には，他国の内政干渉の手段やプロパガンダとしてフェイクニュースが流布されていたことが示唆されている[24]．そのような偽情報攻撃からユーザを保護するためには，どのようなユーザ集団が標的とされているかを把握することが重要である．

表 5 既存手法が使用する特徴

	特徴	[14]	[6]	[10]	[12]	提案手法
Contents	Text	○	○	○	○	-
	Image	-	-	○	-	-
Context	Comment	-	-	-	○	-
	Explicit Social network	○	-	-	-	-
	Implicit social network	○	-	-	-	○
	User profile	-	-	-	-	○

提案手法の特徴量ベクトルは，自己紹介文に含まれる単語のベクトル表現を元に算出されており，これらは同一の空間上にマッピングされる．したがって，単語のベクトル表現とニュース URL の特徴量ベクトルとの類似度をもとに，あるキーワードを自己紹介文に含むユーザ集団の周辺にどのようなニュースが分布しているかを知ることができる．PolitiFact のデータセットを用い，米国における代表的な政治イデオロギーである「CONSERVATIVE」と「LIBERAL」に対するニュースの分布を調べた．UMAP[25]によって 2 次元空間上で可視化した結果を図 7 に示す．CONSERVATIVE の周辺に，ほとんどがフェイクニュースによって構成されたクラスタが形成されている．それに対し，LIBERAL の周辺にもフェイクニュースは分布しているが，CONSERVATIVE ほど顕著な偏りは見られない．これは，CONSERVATIVE を自己紹介文に含むユーザのみに支持されやすいような，極端に偏ったフェイクニュースが多く存在したことを示唆している．提案手法を用いることで，フェイクニュースが標的とした可能性のあるユーザ集団を分析することができ，フェイクニュースに対する積極的な防御やユーザ保護につながる．

7. 結論

本論文では，ニュースの URL を共有した SNS ユーザの自己紹介文を集約し，出現単語の偏りを特徴として捉えることで，フェイクニュースを検知する手法を提案した．既存手法の多くは，ニュースのコンテンツの特徴と，ニュースが拡散したあとのコンテクストの特徴を組み合わせることで検知性能を向上させているが，モデルの複雑さや学習データの収集が課題となっている．それに対し，提案手法はニュースを共有したユーザに暗黙的なつながりがあると見なし，それらユーザの自己紹介文のみを用い，軽量でありながらも高い性能を発揮している．

評価実験では，複数のデータセットを用い，提案手法とは異なる特徴を用いる既存手法との性能比較を行った．PolitiFact のデータセットを用いた結果では，87.9%の分類精度を示し，ニュースのコンテンツの特徴を用いる既存手法[10]を上回る結果となった．コンテンツの特徴とコンテクストの特徴を組み合わせる手法[12]に及ばないものの，

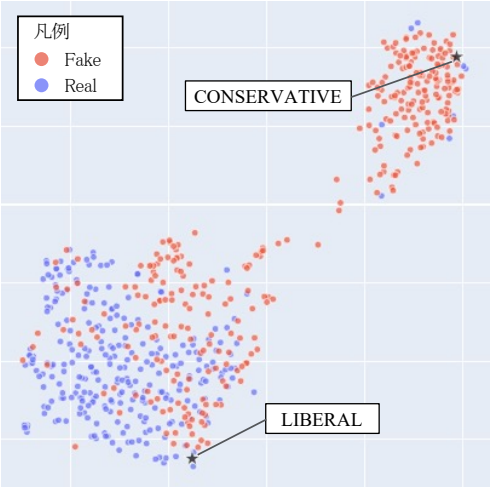


図 7 PolitiFact データセットにおける
ニュースの特徴量ベクトルの分布

提案手法はコンテクストの特徴のみで相応の性能が出ていることからコンテンツの特徴を用いる手法と組み合わせることでさらなる性能向上が期待できる．また，GossipCop のデータセットを用いた結果においては 95.8%の分類精度を示し，他の手法を上回る結果となった．

ケーススタディでは，フェイクニュースの検知に留まらない提案手法の応用例を示した．提案手法のように，ユーザに着目した手法を用いてフェイクニュースを分析することで，フェイクニュースの標的となりやすいユーザの保護といった，より積極的なフェイクニュース対策につながる．

今後の展望としては，一つに，様々な種類のフェイクニュースに適した検知モデルを開発することがあげられる．政治や芸能以外にも，科学や医療，災害など，フェイクニュースには様々な類型やドメインがあり，効果的な検知モデルはそれぞれ異なる可能性がある．したがって単に利用可能な特徴を組み合わせるのではなく，モデルの複雑性やデータ量などのトレードオフを考慮しながら，適切な検知モデルを選択することが課題となる．また，SNS などのプラットフォームにおける対策としては，フェイクニュースを検知するのみでは十分でなく，ユーザを意図的な偽情報攻撃から，いかに保護するかが重要となる．そこで，もう一つの研究の方向性として，フェイクニュースが標的とするユーザ集団の特定や偽情報攻撃の兆候を検知する手法の開発などがあげられる．具体的には，ケーススタディにおける提案手法の応用を発展させ，フェイクニュースの分布の偏りといった構造の特徴を利用することが考えられる．

参考文献

[1] X. Zhou and R. Zafarani, “A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities,” ACM Comput. Surv., vol. 53, no. 5, pp. 1–40, Sep. 2020.
[2] H. Allcott and M. Gentzkow, “Social Media and Fake News in the 2016 Election,” J. Econ. Perspect., vol. 31, no. 2, pp. 211–236, May 2017.
[3] M. J. Metzger, A. J. Flanagin, P. Mena, S. Jiang, and C. Wilson, “From dark to light: The many shades of sharing misinformation online,” Media Commun., vol. 9, no. 1, pp. 134–143, Feb. 2021.
[4] S. Talwar, A. Dhir, P. Kaur, N. Zafar, and M. Alrasheedy, “Why do people share fake news? Associations between the dark side of social media use and fake news sharing behavior,” Journal of Retailing and Consumer Services, vol. 51, pp. 72–82, Nov. 2019.
[5] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” presented at the Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017.
[6] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, “Fake News Early Detection: A Theory-driven Model,” Digital Threats: Research and Practice, vol. 1, no. 2, pp. 1–25, Jun. 2020.
[7] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: LIWC and computerized text analysis methods,” J. Lang. Soc. Psychol., vol. 29, no. 1, pp. 24–54, Mar. 2010.
[8] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, “Novel visual and statistical image features for microblogs news verification,” IEEE Trans. Multimedia, vol. 19, no. 3, pp. 598–608, Mar. 2017.
[9] Y. Wang et al., “EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection,” in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom, Jul. 2018, pp. 849–857.
[10] X. Zhou, J. Wu, and R. Zafarani, “SAFE: Similarity-Aware Multi-modal Fake News Detection,” in Advances in Knowledge Discovery and Data Mining, 2020, pp. 354–367.
[11] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.
[12] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, “dEFEND: Explainable Fake News Detection,” in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, Jul. 2019, pp. 395–405.
[13] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, “The role of user profiles for fake news detection,” in Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, Aug. 2019, pp. 436–439.
[14] K. Shu, S. Wang, and H. Liu, “Beyond News Contents: The Role of Social Context for Fake News Detection,” in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, Melbourne VIC, Australia, Jan. 2019, pp. 312–320.
[15] N. Ruchansky, S. Seo, and Y. Liu, “CSI: A Hybrid Deep Model for Fake News Detection,” in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, New York, NY, USA: Association for Computing Machinery, 2017, pp. 797–806.
[16] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media,” Big Data, vol. 8, no. 3, pp. 171–188, Jun. 2020.
[17] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, “Fake news on Twitter during the 2016 U.S. presidential election,” Science, vol. 363, no. 6425, pp. 374–378, Jan. 2019.
[18] A. Guess, J. Nagler, and J. Tucker, “Less than you think: Prevalence and predictors of fake news dissemination on Facebook,” Sci Adv, vol. 5, no. 1, p. eaau4586, Jan. 2019.
[19] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, “Learning Word Vectors for 157 Languages,” arXiv [cs.CL], Feb. 19, 2018.
[20] S. Bird, “NLTK: the natural language toolkit,” in Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, 2006, pp. 69–72.
[21] KUDO and T, “MeCab: Yet Another Part-of-speech and Morphological Analyzer,” http://mecab.sourceforge.jp, 2006, Accessed: Jun. 01, 2021.
[22] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, no. 85, pp. 2825–2830, 2011.
[23] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” BMC Bioinformatics, vol. 7, p. 91, Feb. 2006.
[24] W. L. Bennett and S. Livingston, “The disinformation order: Disruptive communication and the decline of democratic institutions,” Eur. J. Commun., vol. 33, no. 2, pp. 122–139, Apr. 2018.
[25] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” arXiv [stat.ML], Feb. 09, 2018.