ニュースソースの違いによるフェイクニュース検出と問題点

岸 祐輝 [1]　中川 翼 [2]　彌冨 仁 [1,2]
[1] 法政大学理工学部応用情報工学科　[2] 法政大学理工学研究科応用情報工学科専攻
{yuki.kishi.4z, tsubasa.nakagawa.5p}@stu.hosei.ac.jp　iyatomi@hosei.ac.jp

概要

フェイクニュースを見破るのは難しい問題であり，機械学習技術を用いた検出が試みられている．しかし学習モデルが未来のニュースに対して適応できず適切な検出ができない点が懸念される．本研究では，機械学習モデルが学習した記事に対し，解析対象とするニュースの配信時期の違いや，ニュースソースが既知であるか否かの違い，タイトルや本文などの解析対象が検出精度にどの程度の影響があるかの検証を 100 万以上を収録する大規模な NELA-GT データセットを用いて解析した．未知のニュースソースに対する検出精度が著しく低下すること，データセットに由来する制限はあるが，著者と本文を組み合わせるのが特に有効だと分かった．

1　はじめに

インターネットや SNS の普及した現代において，誤った情報や誤解を招くフェイクニュースは重要な問題の 1 つであるが，一見して見破るのは難しい．最近では COVID-19 の流行により嘘やデマが大量に発信されてしまい，虚偽の情報を否定するために国によって対応せざるを得なくなった [1]．ニュースが信頼に足りうるか判断するには，様々な情報を入手して多角的に判断する必要があるが，各情報源を精査するには時間がかかるため，自動でフェイクニュースを検知する仕組みが求められている．

フェイクニュースによる混乱を防ぐことを目的とした機械学習技術を用いたフェイクニュース検出の研究が行われている [2, 3]．一般的に，モデルには記事本文やタイトル等ニュース記事に関する情報が入力され，それが本物か偽物かを分類するタスクとして定義される．様々なデータセットが公開され，例えば，主に政治や社会に関する主張や発言を収集したデータセットである LIAR [4] や，2016 年の米国選挙に近い 9 月の平日 7 日間に Facebook より投稿されたニュース記事とそれに対するコメントや反応を収集した BuzzFace [5]，複数のニュースソース（会社などの組織名）から 1 年間かけて収集した NELA-GT [6, 7] データセット等は，この分野の研究で利用されている．

しかし，機械学習を用いたフェイクニュースの自動検出には，未来の出来事に対して検出能が低下する問題が指摘されている [8, 9]．これは時間の経過によるニューストピックや語彙の変化に起因しており，対応が難しい課題となっている．そのため，本質的に頑健なフェイクニュース検出モデルを構築するためには，多岐にわたる記事とそれに対するラベル情報を元に，適切な解析が必要になる．LIAR や BuzzFace 等のデータセットは，記事データ量の少なさ，特定トピックへの限定，収集期間の短さなどが課題である．

本研究では，既知のニュースで学習したモデルが未来のニュースに対して一般化しない可能性について，収集時期の異なる 100 万件を超える大規模なフェイクニュースデータセット NELA-GT-2019 [6] と NELA-GT-2020 [7] を用いて，ニュースの配信時期による影響を検証した．また，ニュースソースが既知であるか否かの違い，タイトルや本文のような要素ごとの違いがどのような影響を与えているかについても検証した．

2　方法

NELA-GT-2019 を既知のニュース，NELA-GT-2020 を未来のニュースとしてニュースの真偽の 2 値分類の分類能の評価を行った．比較実験には，ニュースの配信時期の違い，ニュースソースの違いによって 4 パターンのテストデータを用意し，各パターンに対して記事データの要素を用いた 7 つの組み合わせで検証した．

2.1　データセット

NELA-GT-2019 は 2019 年に 261 のニュースソースから収集された約 118 万件，NELA-GT-2020 は 2020

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　記事データの要素

要素名	内容
id	記事の ID
date	出版日
source	記事を掲載しているソース
title	記事の見出し
content	記事の本文
author	記事の著者
published	掲載された日時
published_utc	出版日の UNIX タイムスタンプ
collection_utc	収集日の UNIX タイムスタンプ
url	記事の URL

表 2　記事データの例

要素名	例
id	21stcenturywire– ... Venezuela
date	2019-01-30
source	21stcenturywire
title	WATCH: Londoners ... Venezuela
content	Journalist Robert ... .\nWatch :
author	21wire
published	2019-01-30 17:59:19+00:00
published_utc	1548889159
collection_utc	1567550189
url	https:// ... -policy-on-venezuela/

年に 519 のニュースソースから収集された約 178 万件の記事データでそれぞれ構成されている．ニュースソースには，0-Reliable，1-Mixed，2-Unreliable の 3 種類のいずれかのラベルが付与され，ラベルが付けられていないソースも存在している．各ソースには複数の記事データが含まれており，各記事データは表 1 に示す 10 種類の要素で構成されている．NELA-GT データセットは，記事そのものにはラベルが付いていないため，実験にはこのデータセットを用いた他の先行研究 [2, 9] と同様，記事が属しているソースのラベルを使用した．よって，同じソースに属している記事はすべて同じラベルを持つことになる．

これらのデータの内，ラベルが Mixed であるものとラベルが付いていないもの，記事の content が保持されていないものを本研究の解析対象から除外した．これらの条件に沿って選定したデータの内，NELA-GT につけられているラベルが本来ソースレベルであることから，NELA-GT-2019 のデータセットをソースを基に 8 : 2 に分割し，8 割を学習データ，2 割をテストデータとした．この 8 割の学習データを今後行う 4 つの実験パターンに共通して使用する．NELA-GT-2020 は学習で使用せず，全てテストデータとして利用した．学習に用いるデータの数を表 3 に，評価に用いるデータの数を表 4 に示す．

表 3　学習に用いるデータの数

		Reliable	Unreliable	合計
NELA-GT-2019	source	66	40	106
	content	304,857	103,113	407,970

表 4　評価に用いるデータの数

		Reliable	Unreliable	合計
NELA-GT-2019	source	17	10	27
	content	140,798	22,886	163,684
NELA-GT-2020	source	96	111	207
	content	491,487	527,575	1,019,062

2.2　学習と評価のパターン

フェイクニュースの検出には BERT [10] を用いた．損失関数には交差エントロピー誤差，最適化手法には Adam [11] を使用し，学習率は 1.0×10^{-5}，バッチサイズは 16, エポック数は 5 に設定した．フェイクニュースの検出能の比較において，以下 4 つのパターンのテストデータに対して評価を行った．

A) NELA-GT-2019 の学習に用いていないデータ（現在の未知のソース）
B) NELA-GT-2020 のデータ全て（未来の記事）
C) NELA-GT-2020 のうち，学習データに含まれていたニュースソースを除外したデータ（未来の未知のソース）
D) NELA-GT-2020 のうち，学習データに含まれていたニュースソースのデータ（未来の既知のソース）

ここで，記事データのうちフェイクニュースの検出に効果的な要素を議論するため，それぞれの条件において，モデルが用いる情報として以下の 7 種の条件（「content」，「title」，「title + content」，「author + content」，「published_utc + content」，「TAP + content」，「FULL」）で評価を行った．TAP は title, author，published_utc を 1 つの文として連結したものであり，FULL は title，author，published_utc，content を 1 つの文として連結したものである．各モデルのフェイクニュース検出能は，accuracy，precision，

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 5 A パターン：NELA-GT-2019 の学習に用いていないデータ（現在の未知のソース）で評価した結果

	Acc	Pre	Rec	F1
content	0.880	0.543	0.907	0.680
title	0.860	0.499	0.732	0.594
title + content	0.864	0.507	0.943	0.660
author + content	0.899	0.590	0.925	0.720
published_utc + content	0.884	0.554	0.871	0.678
TAP + content	0.895	0.581	0.894	0.704
FULL	**0.909**	0.619	0.900	**0.734**

表 6 B パターン：NELA-GT-2020 のデータ全て（未来の記事）で評価した結果

	Acc	Pre	Rec	F1
content	0.607	0.749	0.363	0.489
title	0.578	0.691	0.333	0.449
title + content	**0.632**	0.731	0.456	**0.562**
author + content	0.592	0.775	0.299	0.431
published_utc + content	0.597	0.762	0.324	0.454
TAP + content	0.601	0.792	0.311	0.447
FULL	0.590	0.774	0.294	0.426

recall，F1 スコアで評価した．

3 結果と考察

3.1 時期とソースの違いによるスコア差

実験結果を表 5 から表 8 に示す．A パターン（現在，未知ソース）と B パターン（未来，全ソース）を比較すると未来の記事である B パターンのスコアが 20-30 ポイントほど大きく低下した．1 年違いのニュース記事でも未来のニュースを予測するのは困難になる結果が得られた．C パターン（未来，未知ソース）は 2 値分類であるのにほとんどの組み合わせで $F1 = 0.5$ を切っていることから，学習モデルは未来の未知ソースより配信されたニュースの真偽を判別できていない．一方で，D パターン（未来，既知ソース）は他のパターンを比べて時期によるスコアの減少は少ない．このことから，B パターンや C パターンでスコアが減少する主な原因はニュースソースの違いによるものであり，現状のモデルでは未来の未知のソースに関しては対応できないと言える．ただしデータセットが記事に対してではなくニュースソースに対してラベル情報が付与されている制約も大きいと考えられる．

表 7 C パターン：NELA-GT-2020 のうち，学習データに含まれていたニュースソースを除外したデータ（未来の未知のソース）で評価した結果

	Acc	Pre	Rec	F1
content	0.475	0.788	0.334	0.469
title	0.471	0.802	0.317	0.455
title + content	**0.528**	0.794	0.433	**0.560**
author + content	0.420	0.761	0.241	0.366
published_utc + content	0.451	0.783	0.293	0.426
TAP + content	0.435	0.790	0.255	0.385
FULL	0.410	0.735	0.235	0.356

表 8 D パターン：NELA-GT-2020 のうち，学習データに含まれていたニュースソースのデータ（未来の既知のソース）で評価した結果

	Acc	Pre	Rec	F1
content	0.824	0.644	0.510	0.569
title	0.752	0.451	0.412	0.431
title + content	0.802	0.564	0.574	0.569
author + content	0.874	0.805	0.590	0.681
published_utc + content	0.836	0.704	0.481	0.571
TAP + content	0.874	0.797	0.596	0.682
FULL	**0.886**	0.866	0.588	**0.701**

3.2 利用する情報による予測能

先述の通り，現状のモデルでは未来の未知ソースには対応できず，性能を大きく落としてしまうため，未来の未知ソースが含まれている B パターンと C パターンでは，どの要素がフェイクニュースの検出に有効か比較することはできない．そこで，未来の未知ソースが含まれていない D パターン（未来，既知ソース）の各要素で学習したモデルの結果を比較することで，どの要素がフェイクニュースの検出に有効か検証する．

図 1 に D パターンの各組み合わせの ROC 曲線と AUC を示す．表 8 では content のスコアとあまり変わりのない title + content や published_utc + content も AUC のスコアでは上回っていることから，ニュースのタイトルや出版日はフェイクニュースを検出する上で有効な情報であると言える．特に，すべての情報を組み合わせている TAP + content と FULL と共に，author + content が近い AUC のスコアになっていることから，ニュースの著者情報は最も有力であると言える．著者と記事本文を組み合わせるとモデル性能が向上するのは，信頼できるニュースを配信し続ける著者は基本的に同じ人物であること，信頼で

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

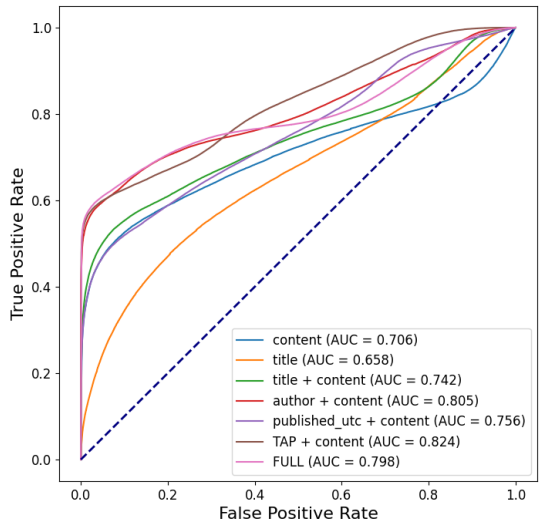

図 1 D パターン（未来の既知ソース）の各組み合わせの ROC 曲線と AUC

きない著者はデマや嘘を配信し続けたり，ニュースの配信が初めてで信頼できるか判別しづらいといった特徴を見分けられていることが理由として考えられる．

4 おわりに

ニュースソースにラベルが付与されている NELA-GT データセットを用いた検証により，既知のニュースソースに対してはフェイクニュースに対して一定の検出能を実現したが，未知のソースに対しては検出ができなかった．また，ニュース本文単体で推定するより他の情報を組み合わせることでフェイクニュースの検出能が増加し，特に著者に関する情報が有効であった．

今回使用した NELA-GT はニュース記事に対してラベルが付けられておらず，ニュース記事の属するソースのラベルで代用して学習を行った．しかし，これはニュース記事そのものの信頼性を保証するものではないため，ニュース記事レベルで信頼性の高いデータを扱えることが望ましい．今後の展望として，ニュース記事レベルの信頼性を補強する手法を探すと同時に，今回の title や content といったニュースを構成する要素とは別に，コメントや賛成票といったニュースに対する人々の反応や評価といった情報であるソーシャルコンテキストを用いて，未来のニュースに対して効果の高い手法を模索する．

参考文献

[1] Usha M. Rodrigues and Jian Xu. Regulation of COVID-19 Fake News Infodemic in China And India. **Media International Australia**, Vol. 177, pp. 125 – 131, 2020.

[2] Özlem Özgöbek, Benjamin Kille, Anja Rosvold From, and Ingvild Unander Netland. Fake News Detection by Weakly Supervised Learning Based on Content Features. In Evi Zouganeli, Anis Yazidi, Gustavo Mello, and Pedro Lind, editors, **Nordic Artificial Intelligence Research and Development**, pp. 52–64, Cham, 2022. Springer International Publishing.

[3] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. FakeBERT: Fake News Detection in Social Media with A BERT-based Deep Learning Approach. **Multimedia tools and applications**, Vol. 80, No. 8, pp. 11765–11788, 2021.

[4] William Yang Wang. “Liar, Liar Pants on Fir”: A New Benchmark Dataset for Fake News Detection. In **Annual Meeting of the Association for Computational Linguistics**, 2017.

[5] Giovanni Santia and Jake Williams. Buzzface: A News Veracity Dataset with Facebook User Commentary And Egos. In **Proceedings of the international AAAI conference on web and social media**, Vol. 12, pp. 531–540, 2018.

[6] Maurício Gruppi, Benjamin D. Horne, and Sibel Adali. NELA-GT-2019: A Large Multi-labelled News Dataset for The Study of Misinformation in News Articles. **CoRR**, Vol. abs/2003.08444, , 2020.

[7] Maurício Gruppi, Benjamin D. Horne, and Sibel Adali. NELA-GT-2020: A Large Multi-labelled News Dataset for The Study of Misinformation in News Articles. **CoRR**, Vol. abs/2102.04567, , 2021.

[8] Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. A Topic-agnostic Approach for Identifying Fake News Pages. In **Companion proceedings of the 2019 World Wide Web conference**, pp. 975–980, 2019.

[9] Shaina Raza and Chen Ding. Fake News Detection Based on News Content And Social Contexts: A Transformer-based Approach. **International Journal of Data Science and Analytics**, Vol. 13, No. 4, pp. 335–362, 2022.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **North American Chapter of the Association for Computational Linguistics**, 2019.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. **CoRR**, Vol. abs/1412.6980, , 2014.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).