フェイクニュース検出データセットにおける通時的バイアス

村山 太一　　若宮 翔子　　荒牧 英治
奈良先端科学技術大学院大学
{murayama.taichi.mk1, wakamiya, aramaki}@is.naist.jp

1 はじめに

近年，意図的に広められた誤った記事である「フェイクニュース」が深刻な社会問題の一つとなっている．例えば，2016 年の米国大統領選挙では，Twitter に投稿されたニュースの 25%がフェイクもしくは極端に偏っており，トランプ氏の支持者の活動がフェイクニュースの拡散に影響を与えていた [1]．選挙だけでなく，2011 年の東日本大震災などの自然災害 [2, 3] や，COVID-19 [4] に関連したフェイクニュースが世界各国でソーシャルメディアなどを通じて共有された．これらの問題に対処するため，ソーシャルメディア投稿やニュースコンテンツからのフェイクニュース検出モデルの開発や，検出モデル開発のためのデータセット構築が活発に行われている [5, 6]．

多くのフェイクニュース検出データセットは，現実のフェイクニュースから構成される．現実のフェイクニュースは人々の関心に強く影響される [7] ことから，時期によって流行するトピックが異なる．具体的には，2013 年にはオバマ大統領 [8]，2016 年には米国大統領選挙 [1]，2020 年には COVID-19 [9] など，異なるトピックのフェイクニュースが頻繁に拡散された．そのため，特定時期のイベントや単一のドメインを対象とした，データセットが構築されることが多い．これらのデータセットから学習されるフェイクニュース検出モデルは，学習データセットと同様のニュースに対しては高い検出精度を達成する．その一方で，異なるドメインや将来のフェイクニュースに対してはテキストの語彙情報が大きく異なることから十分に対応できない場合が多い．つまり，時期に偏りがあるデータセットで学習された検出モデルに，特定の人名や組織名を含む文章が入力されると，その文章の真偽に関わらず判定に誤りが生じる可能性がある．これは，同一のドメインデータでも生じる問題であり，頑健なモデル構築の妨げとなる．

本研究では，この問題をデータセットの作成時期に依存することから「**通時的バイアス**」と定義する．本稿では，通時的バイアスがデータセット内の人名などの固有名詞によるものと考え，これらをマスキングすることでバイアスの緩和手法を検討する．はじめに，作成時期の異なるフェイクニュース検出データセットを対象に，ラベルとフレーズ間の相関について分析し，データセット内の人名などの語の偏りを明らかにした．次に，テキストからのフェイクニュース検出タスクを対象に，通時的バイアスの原因と考えられる人名に着目した緩和手法の検討を行った．

2 関連研究

学習データセットにおけるバイアスに関する研究は，攻撃的な言語やヘイトスピーチの検知などを主な目的としており，著者バイアス [10]，アノテータバイアス [11]，ジェンダーバイアス [12, 13]，人種バイアス [14, 15]，政治的バイアス [16] などの様々なバイアスの分析や緩和手法の検討がなされている．Dayanik らの研究 [17] では，データセット内の人名の頻度バイアスに着目している．人名に着目した点は我々と同様であるが，本稿では時間の経過のバイアスについて考慮している点で大きく異なる．å

フェイクニュース検出に近い分野として，含意認識の一種である与えられた証拠から主張の真偽を判定する Fact Verification タスクではバイアスの分析・軽減が取り組まれてる．Schuster ら [18] は頑健な推論モデル構築のためにテキストの語彙の偏りを緩和する手法を提案した．Suntwal ら [19] はテキスト中の語彙を NER タグに置き換えることで，ドメイン外のデータでも頑健な推論モデルの構築を目指した．しかしながら，フェイクニュース検出データセットに対してのバイアス調査や分析は我々の知る限り行われていない．

3 リソース

3.1 データセット

本稿では，フェイクニュース検出に関するドメインや収集時期が異なる以下の 4 つの英語データセットを用いて通時的バイアスの分析・検証を行う．これらのデータセットでは，記事や投稿などのテキストデータの内容に対して，フェイクニュースかどうかを示す 2 値ラベル（Fake / Real）が付与されている．各データセットの詳細を以下に，ラベル数などの情報を付録A.1節に示す．

MultiFC [20] ：複数ドメインのフェイクニュースデータセットであり，38 の事実検証サイトの記事から構成される．全データ 36534 件のうち，本稿では 2015 年以前の truth!，true，mostly true（Real ラベルとする）と false，mostly false（Fake ラベルとする）が付与された 7861 件のデータを用いる．

Horne17 [21] ：2016 年の米国大統領選挙に関連したニュース記事を中心に構成される．Buzzfeed News などのサイトに基づいて，Real，Fake，Satire の 3 値ラベルが付与されており，本稿では Real，Fake のラベルが付与されたデータを用いる．

Celebrity [22] ：有名人を対象としたニュースの事実検証サイトである GossipCop によって検証された記事によって構成される．データセットの記事は 2016，2017 年頃に報道されたものが多く，有名人同士の喧嘩などセンセーショナルな話題が中心である．

Constraint [23] ：Constraint@AAAI2021 Shared Task のサブタスクで用いられたデータセットであり，Twitter などのソーシャルメディアの投稿で構成される．Politifact や Snopes などの事実検証サイトによってラベルが付与されており，2020 年に Twitter で拡散された COVID-19 に関連した投稿が中心である．

3.2 フレーズとラベルの相関

データセットの偏りを検証するために，データセットのフレーズとラベルの相関について調査する．高い出現頻度かつ各ラベルとの相関が高いフレーズ (n-gram) を抽出するために，Local Mutual Information (LMI) [24] を用いる．

$$LMI(w,l) = p(w,l) \cdot \log\left(\frac{p(l|w)}{p(l)}\right), \qquad (1)$$

w はフレーズ，l はラベルとする．条件付き確率 $p(l|w)$ は $\frac{\mathrm{count}(w,l)}{\mathrm{count}(w)}$，$p(l)$ は $\frac{\mathrm{count}(l)}{|D|}$ と算出される．w と l の同時確率である $p(w,l)$ は $\frac{\mathrm{count}(w,l)}{|D|}$ として算出される．$|D|$ はデータセット内の全てのフレーズの出現回数を示す．

MultiFC と Horne17 の bi-gram を対象とした，各ラベルとの相関が高い上位 10 フレーズの LMI を表 1に，Celebrity と Constraint の結果は付録 A.2節に示す．2015 年以前の記事で構成された MultiFC では “barack obama” といった当時の米国大統領が Fake ラベルとの相関が高くなり，2016 年大統領選挙の記事で構成された Horne17 では当時の大統領候補であった “hillary clinton” や “donald trump” が Fake ラベルと高い相関を持つという結果となった．このように，データセットごとに特定の人名とラベルの関係に大きな偏りの存在が明らかになった．どちらか一方のデータセットで学習された検出モデルでは大統領交代などの時代変化に対応できず，もう一方のデータセットでのフェイクニュース検出が上手くいかないことが示唆される．

4 通時的バイアスの緩和の検討

4.1 検討手法

通時的バイアスを緩和し，ドメイン外のデータに対しても頑健なフェイクニュース検出モデルを構築するために，データセットに対する複数のマスキング手法を検討する．本稿で検討するマスキング手法の例を表 2に示す．**Lexicalized** はマスキングされていない通常の入力とする．

Named Entity（NE）Deletion ：Flair [25] の Named Entity Recognition (NER) [26] によって NE とタグ付けされた語彙を削除する．NE に依存しない検出モデルの構築を目的とするマスキング手法の一つである．

Basic NER ：Flair の NER により NE としてタグ付けされた語彙が対応するラベル (PER, LOC など) に置き換える．

WikiD ：Flair の NER により PER とタグ付けされた語彙を，Wikidata [1)] で対応する人物の公的な地位 (P39)，対応がなければ職業 (P106) に置き換える．当時の Wikidata を用いることで，2015 年の記事に登場する “barack obama” と 2020 年の記事に登場する “donald trump” を米国大統領 (Q11696) というラベルに置き換えることができる．これによって，テキストを入力としたフェイクニュース検出モデルが時間経過にも頑健になり，通時的バイアスの緩和が期待

1） https://www.wikidata.org/wiki/Wikidata

表 1　データセットの bi-gram を対象に，各ラベルにおいて LMI が高い上位 10 フレーズと算出された **LMI** と $p(l|w)$ を示す．**LMI** は 10^6 を掛けた値を示す．人名を示すフレーズを太字で記す．Real ラベルは “if you” などの一般的なフレーズとの相関が高い一方で，Fake ラベルでは “donald trump” などの人名と高い相関を示す傾向が見られた．

MultiFC						Horne17					
Real			Fake			Real			Fake		
Bigram	**LMI**	$p(l\|w)$	**Bigram**	**LMI**	$p(l\|w)$	**Bigram**	**LMI**	$p(l\|w)$	**Bigram**	**LMI**	$p(l\|w)$
mitt romney	218	0.69	health care	631	0.64	**trump has**	112	0.82	**donald trump**	605	0.42
if you	217	0.70	**barack obama**	365	0.69	national security	106	0.88	**hillary clinton**	440	0.50
rhode island	190	0.75	**president barack**	337	0.70	would be	104	0.72	i think	292	0.68
new jersey	177	0.67	**scott walker**	258	0.81	people who	92	0.89	united states	258	0.51
john mccain	167	0.73	says president	218	0.78	transition team	88	1.0	have been	230	0.41
no. 1	128	0.86	care law	185	0.80	**mr. trump**	80	0.94	**bill clinton**	208	0.70
voted against	128	0.71	will be	162	0.63	smug style	77	1.0	we are	206	0.56
any other	125	0.61	**hillary clinton**	159	0.67	**george w.**	76	0.90	**hillary clinton’s**	187	0.58
does not	119	0.71	**gov. scott**	148	0.72	republican party	76	0.91	**president obama**	171	0.55
this year	116	0.75	social security	144	0.68	new york	70	0.77	**ted cruz**	149	0.80

表 2　マスキング手法の適用例．ラベル PER に当たる部分を赤字，そのほかの NE を太字で示す．

Lexicalized	18 states including **US UK** and **Australia** request PM **Modi** to head a task force to stop coronavirus
NE Deletion	18 states including and request PM to head a task force to stop coronavirus
Basic NER	18 states including **LOC LOC** and **LOC** request PM **PER** to head a task force to stop coronavirus
WikiD	18 states including **US UK** and **Australia** request PM **Q22337580** to head a task force to stop coronavirus
WikiD+Del	18 states including and request PM **Q22337580** to head a task force to stop coronavirus
WikiD+NER	18 states including **LOC LOC** and **LOC** request PM **Q22337580** to head a task force to stop coronavirus

できる．

WikiD+Del　：Flair の NER により PER とタグ付けされた語彙を WikiD のルールで置き換え，その他の NE とタグ付けされた語彙を削除する．

WikiD+NER　：Flair の NER により PER とタグ付けされた語彙を WikiD のルールで置き換え，その他の NE を対応するラベル (LOC など) に置き換える．

4.2　実験設定

マスキング手法の有効性について検証するために，フェイクニュース検出タスクを取り組む．3 節で取り上げた各データセットの学習データを用いて，各マスキング手法を適用したフェイクニュース検出モデルを作成する．そして，同じドメインや異なるドメインのデータセットに対して，各検出モデルによりどの程度フェイクニュースの検出が可能かを確認する．

モデル　本実験では，Google から公開されている事前学習 $\text{BERT}_{\text{BASE}}$ モデル [27] を用いる．マスキング手法によって追加されたラベル（LOC や Q11696 など）は新たなトークンとして追加し，fine-tuning する．マスキングで用いる Wikidata の詳細は付録A.3節に示す．

データと評価　学習のために，各データセットを全体の 80%を学習データに，20%をテストデータに分割する．検出精度の評価には正解率 (*Accuracy*) を用いる．

4.3　実験結果

各データセットで学習したフェイクニュース検出モデルを，学習したデータセットのテストデータ（ドメイン内）とそれ以外のデータセットのテストデータ（ドメイン外）で評価し，マスキング手法の有用性について検討する．

4.3.1　ドメイン内データ

ドメイン内データに対する各マスキング手法での正解率を表 3に示す．Constraint 以外のデータセットにおいて，マスキング手法を行わない Lexicalized が最も高い精度となった．一方で，Constraint データセットにおいて WikiD が最も高い精度を達成した．また，Lexicalized が最も高い精度となったデータセットにおいても他のマスキング手法による精度と比較して数ポイントの差しか存在しないという結果が得られた．このことから，同一ドメインのデータに対して，各マスキング手法を適用したモデルでフェイクニュース検出を行っても，大きく精度が悪化しないことが示唆される．

4.3.2　ドメイン外データ

ドメイン外データをテストデータとした正解率を表 4に示す．左の列に学習に用いたデータセットを，精度を記載した列はテストに用いた各データセットに対応している．ほとんどのドメイン外のデータに対して，マスキングを行わない Lexicalized

表 3 各マスキング手法におけるドメイン内でのファイクニュース検出精度.

手法	テストデータ			
	MultiFC	Horne17	Celebrity	Constraint
Lexicalized	**0.681**	**0.746**	**0.760**	0.960
NE Deletion	0.656	0.706	0.750	0.959
Basic NER	0.659	0.735	0.750	0.950
WikiD	0.675	0.725	0.730	**0.967**
WikiD+Del	0.660	0.706	0.700	0.959
WikiD+NER	0.660	0.640	0.730	0.957

表 4 各マスキング手法におけるドメイン外でのファイクニュース検出精度.

学習データ	手法	テストデータ			
		MultiFC	Horne17	Celebrity	Constraint
MultiFC	Lexicalized	-	0.706	**0.660**	0.530
	NE Deletion	-	0.706	0.590	0.664
	Basic NER	-	0.725	0.600	0.680
	WikiD	-	**0.746**	0.590	**0.689**
	WikiD+Del	-	0.725	**0.660**	0.669
	WikiD+NER	-	0.632	0.520	0.667
Horne17	Lexicalized	0.504	-	0.670	0.481
	NE Deletion	**0.551**	-	**0.680**	0.553
	Basic NER	0.523	-	0.670	**0.563**
	WikiD	0.525	-	0.620	0.487
	WikiD+Del	0.523	-	0.610	0.515
	WikiD+NER	0.500	-	0.630	0.531
Celebrity	Lexicalized	0.533	0.451	-	0.583
	NE Deletion	0.545	0.529	-	**0.763**
	Basic NER	0.521	0.549	-	0.568
	WikiD	**0.555**	0.549	-	0.724
	WikiD+Del	0.534	0.529	-	0.663
	WikiD+NER	0.525	**0.568**	-	0.598
Constraint	Lexicalized	0.542	0.568	0.580	-
	NE Deletion	0.531	0.588	0.570	-
	Basic NER	0.543	0.568	0.580	-
	WikiD	**0.556**	0.607	0.570	-
	WikiD+Del	0.544	**0.627**	**0.590**	-
	WikiD+NER	0.549	0.607	0.570	-

よりも各マスキング手法が高い精度を達成した.

Lexicalized の精度はドメイン外への適応可能性を示している．例えば，Constraint で学習したモデルは Constraint で 0.96 と高い正解率を達成した（表 3）が，ドメイン外データで学習したモデルでは Constraint のテストデータに適用したところ，約 0.48 – 0.58 と低い正解率となった．このことは，フェイクニュース検出モデルの一般化の難しさを示している．

検討手法の一つである NE Deletion は単純なマスキング手法にも関わらず，12 の実験設定のうち 9 つで Lexicalized よりも高い精度を達成した．一部のデータセットで大幅に精度が向上しており，中でも Horne17 で学習したモデルでは MultiFC と Celebrity の 2 つのテストデータセットで最も高い精度を達成している．Basic NER は Horne17 で学習したモデルで Constraint のテストデータに対して最も高い精度を達成したが，NE Deletion ほどの大きな効果は見られなかった.

Wikidata を活用したマスキング手法である WikiD と WikiD+Del は，Lexicalized と比較して Celebrity をテストデータとした時以外の 10 の実験設定で高い精度を達成した．例えば，MultiFC で学習したモデルで Constraint のテストデータで評価したところ，Lexicalized では 0.530 だったものの，WikiD では 0.689 と大幅な精度向上を達成した．この結果は，Wikidata を活用したマスキング手法が通時的バイアスを緩和し，ドメイン外のデータセットにおいても頑健なモデルを構築できる可能性を明らかにした. 一方で，芸能系のドメインである Celebrity をテストデータとしたとき，Lexicalized と変わらない検出精度であった．これは，学習データが政治系を中心としたものでドメインが大きく異なることから，Wikidata を活用できず大きな効果が見られなかったと考えられる（付録の表7を参照されたい）．また，Wikidata を活用したマスキング手法の WikiD+NER は WikiD+Del と比べ低い精度となっている．つまり，人名以外の Entity，例えば “London” などの地名を “LOC” というタグに置き換えるよりも，削除することで頑健なモデルを構築できることを示す．これは余分な Entity 情報を削除することで，モデルが文体の特徴に着目できることが要因と考えられる.

5 おわりに

本稿は，フェイクニュース検出データセットの作成時期やドメインによって生じる新たなバイアス「通時的バイアス」を定義した．そして，データセット内の人名に大きな偏りがあるのに着目し，人名を Wikidata の情報に置き換えるマスキング手法を提案し，ドメイン外のデータセットに対するフェイクニュース検出タスクに取り組むことによりマスキング手法の有効性について確認した.

今回はマスキングという簡易な手法を検討したが，今後の方向性として，通時的バイアスを緩和するために知識グラフを活用した検出モデルの構築などが考えられる．また，フェイクニュース検出データセットには通時的バイアス以外にも政治的バイアスや人種バイアスが存在すると考えられるため，これらのバイアスを明らかにしていくことも重要な課題の 1 つである.

参考文献

[1]Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, Vol. 10, No. 1, pp. 1–14, 2019.

[2]Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. Rumor diffusion and convergence during the 3.11 earthquake: a twitter case study. *PLoS one*, Vol. 10, No. 4, p. e0121443, 2015.

[3]Takako Hashimoto, David Lawrence Shepard, Tetsuji Kuboyama, Kilho Shin, Ryota Kobayashi, and Takeaki Uno. Analyzing temporal patterns of topic diversity using graph clustering. *The Journal of Supercomputing*, pp. 1–14, 2020.

[4]Gautam Kishore Shahi and Durgesh Nandini. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proc. of ICWSM*, 2020.

[5]Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. Mining disinformation and fake news: Concepts, methods, and recent advancements. *arXiv:2001.00623*, 2020.

[6]Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, Vol. 11, No. 3, pp. 1–37, 2019.

[7]Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. Anatomy of news consumption on facebook. *Proceedings of the National Academy of Sciences*, Vol. 114, No. 12, pp. 3035–3039, 2017.

[8]CNBC News. False rumor of explosion at white house causes stocks to briefly plunge; ap confirms its twitter feed was hacked. https://www.cnbc.com/id/100646197, 2013.

[9]Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of covid-19 misinformation on twitter. *arXiv preprint arXiv:2005.05710*, 2020.

[10]Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. Detection of Abusive Language: the Problem of Biased Datasets. In *Proc. of NAACL*, pp. 602–608, 2019.

[11]Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv:1701.08118*, 2017.

[12]Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *International conference on social informatics*, pp. 405–415. Springer, 2017.

[13]Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proc. of EMNLP*, pp. 2799–2804, 2018.

[14]Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proc. of ACL*, pp. 1668–1678, 2019.

[15]Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proc. of the Third Workshop on Abusive Language Online*, pp. 25–35, 2019.

[16]Maximilian Wich, Jan Bauer, and Georg Groh. Impact of politically biased data on hate speech classification. In *Proc. of the Fourth Workshop on Online Abuse and Harms*, pp. 54–64, 2020.

[17]Erenay Dayanik and Sebastian Padó. Masking actor information leads to fairer political claims detection. In *Proc. of ACL*, July 2020.

[18]Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In *Proc. of EMNLP-IJCNLP*, pp. 3410–3416, 2019.

[19]Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. On the importance of delexicalization for fact verification. In *Proc. of EMNLP-IJCNLP*, pp. 3404–3409, 2019.

[20]Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proc. of EMNLP-IJCNLP*, pp. 4677–4691, 2019.

[21]Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proc. of ICWSM*, Vol. 11, 2017.

[22]Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proc. of COLING*, pp. 3391–3401, 2018.

[23]Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. Fighting an infodemic: Covid-19 fake news dataset. *arXiv:2011.03327*, 2020.

[24]Stefan Evert. The statistics of word cooccurrences: word pairs and collocations. 2005.

[25]Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proc. of COLING*, pp. 1638–1649, 2018.

[26]Alan Akbik, Tanja Bergmann, and Roland Vollgraf. Pooled contextualized embeddings for named entity recognition. In *Proc. of NAACL*, pp. 724–728, 2019.

[27]Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pp. 4171–4186, 2019.

A 付録

A.1 データセットの概要

本稿で扱うデータセットのドメインや各ラベルのデータ数について表 5に示す．

表 5 4 つのデータセットの概要．

データセット	ドメイン	年	Real	Fake
MultiFC	-	–2015	3803	4058
Horne17	政治	2016	128	123
Celebrity	芸能	2016–2017	250	250
Constraint	COVID-19	2020	5600	5100

A.2 データセットの LMI

Celebrity と Constraint の 2 つのデータセットにおいて各ラベルと相関の高い上位 10 フレーズの LMI を表 6に示す．芸能系のニュースで構成された Celebrity では “brad pitt” や “kate middleton” などの著名人が，COVID-19 関係の Constraint では “donald trump” や “bill gates” などの人名が Fake ラベルと高い相関を持つという結果となった．一方で，Real ラベルでは人名や固有名詞が少なく，“i had” や “if you” などの一般的な表現と高い相関を持つという傾向が示された．Fake ラベルと人名が高い相関を持ち，Real ラベルと一般的な表現が高い相関を持つという傾向は 4 つのデータセット全体に見られる傾向である．

A.3 マスキング手法における Wikidata

本実験では，テキストのマスキングのためにデータセットを構成する記事や投稿の時期に対応した Wikidata を用いる．具体的には，MultiFC には 2016 年 1 月 4 日，Horne17 と Celebrity には 2018 年 1 月 11 日，Constraint に対しては 2020 年 12 月 28 日に公開された Wikidata を用いる．

A.4 各データセットにおける Wikidata ラベル

各データセット同士の WikiD による Wikidata ラベルの重複を表 7に示す．これは，左列に示したデータセットの Wikidata ラベルのうち，他のデータセットにある Wikidata ラベルによってどの程度カバーされているかを示したものである．MultiFC と Horne17 データセットは同様の政治トピックであることからお互い高いカバー率である．一方で，

表 6 Celebrity と Constraint データセットの bi-gram を対象に，各ラベルにおいて LMI が高い上位 10 フレーズ．

Celebrity

Real			Fake		
Bigram	**LMI**·10^6	$p(l\|w)$	**Bigram**	**LMI**·10^6	$p(l\|w)$
i think	233	0.90	has been	343	0.55
i doní	164	0.95	do think	214	0.80
they were	102	0.70	an insider	199	0.88
i had	100	0.94	**brad pitt**	163	0.63
so i	100	0.92	insider told	157	0.90
but i	87	0.79	may have	128	0.85
we were	87	0.89	**kate middleton**	124	0.88
what i	75	0.87	they are	122	0.51
i love	70	0.92	**the weeknd**	119	0.62
when i	69	0.92	**kanye west**	113	0.56

Constraint

Real			Fake		
Bigram	**LMI**·10^6	$p(l\|w)$	**Bigram**	**LMI**·10^6	$p(l\|w)$
url url	1378	0.77	a video	591	1.0
rt @user	822	0.93	**donald trump**	569	0.98
total number	650	0.98	has been	569	0.52
more than	635	0.89	**url donaldtrump**	435	1.0
have been	575	0.82	**bill gates**	355	1.0
@user url	449	0.87	video shows	346	0.98
managed isolation	402	1.0	**president trump**	315	1.0
our daily	385	0.99	covid vaccine	293	0.80
states reported	373	1.0	corona virus	275	1.0
update published	367	1.0	social media	275	0.93

Celebrity データセットは芸能というドメインであることから，他のデータセットと比べ低いカバー率である．各データセットにおいて出現回数が上位 3 つの Wikidata ラベルを表 8に示す．

表 7 各データセット同士の Wikidata ラベルのカバー率．

	MultiFC	Horne17	Celebrity	Constraint
MultiFC	-	37.4%	29.4%	30.1%
Horne17	53.6%	-	36.4%	39.9%
Celebrity	38.9%	33.6%	-	33.6%
Constraint	35.3%	34.1%	28.6%	-

表 8 各データセットにおける，出現回数上位 3 の Wikidata ラベル．

Rank	MultiFC	Horne17	Celebrity	Constraint
1	President of the U.S.	President of the U.S.	actor	President of the U.S.
2	U.S. representative	Attorney General of Arkansas	singer	CEO
3	Secretary of State	Secretary of State	television actor	Mayor of London