Fake Guardian: 日本語フェイクニュース収集システム

久田 祥平*† 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 〒 630–0192 奈良県生駒市高山町 8916-5

E-mail: †{s-hidasa,murayama.taichi.mk1,wakamiya,aramaki}@is.naist.jp

あらまし フェイクニュースは COVID-19 エピデミックや選挙などの社会の様々な分野に大きなダメージを与えている．それに対抗するため，フェイクニュースを発見し，真偽の判定や，拡散状況の調査をおこなうトラッキングシステムが開発されている．英語の豊富な研究資源や活発なファクトチェック機関の恩恵を受けて，これらのシステムは主に，英語やアメリカ社会のイベントを対象としている．一方，日本語のように英語圏以外の言語では，フェイクニュースコーパスやファクトチェック機関が少ないことは，フェイクニュースに対する社会的応用や研究の足かせになっている．そこで我々は，真偽の疑わしい情報に対して注意や疑問を投げかける “Guardian” というユーザに着目することで，低資源下でも効果的に Twitter からフェイクニュースを収集・追跡するシステム **Fake Guardian** を提案する．この Web インターフェイスによって，ユーザがフェイクニュースの現状やそれに対抗する活動としての Guardian の理解を支援する．このシステムは，あらゆる言語に容易に適用可能であり，データセットの構築や責任あるファクトチェックを支援し，様々な国でフェイクニュース研究の促進を期待できる．Fake Guardian を `https://aoi.naist.jp/fakeguardians` で公開している．

キーワード フェイクニュース，デバンキング，Twitter，ソーシャルメディア，ファクトチェック, 自然言語処理

1 はじめに

フェイクニュースは，社会の様々な分野に大きなダメージを与えている．例えば，2016 年の米国大統領選挙では，529 の信頼性の低い情報 [1] が Twitter 上で拡散され，ツイートからリンクされたニュースのうち 25%がフェイクまたは，トランプまたはクリントン支持の極端なバイアスのかかったもので，選挙に影響を与えた可能性があるとされる [2]．最近では，COVID-19 の流行をきっかけとして，健康や政治に関連した疑わしい情報が続々と生産されおり [3]，2020 年の 1 月から 5 月の間で，英語で行われたファクトチェックはおよそ 9 倍に増加している [4]．他にも世界各国でフェイクニュースは社会的な問題を引き起こしており，例えば英国では Brexit に関係するフェイクニュースが数多く広まった [5]．エチオピアでは，暴力事件につながった事例 [6] があり，報道機関の不十分さと民族対立・内戦に伴うフェイクニュースに対する脆弱性が懸念されている [7]．日本においても，東日本大震災に伴うもの [8] や，近年のトイレットペーパーをめぐる混乱が生じている．

この問題に対する社会的な関心の高さから，フェイクニュース対策としてフェイクニュース収集システムの開発がいくつか行われている．例えば，Hoaxy [9] や FakeNewsTracker [10] などがこれらにあたり，ファクトチェック機関のフェイクニュースを収集・可視化し，ジャーナリストや研究者を支援するとともに，一般ユーザが情報の真偽をチェックする手助けをしている．これらのフェイクニュース収集システムの構築の背景には，Politifact[2]や Snopes[3]に代表される，フェイクニュースにすばやく反応し，記事数の豊富なファクトチェック機関や，その結果を活用して作られたフェイクニュースのデータセットなど，豊富な英語の資源を活用していることが挙げられる．一方で，日本には Fact Check Initiative Japan [4]が，複数の報道機関の協力のもとファクトチェック活動を行っている．しかしながら，検証されるニュースの数は日米でも大きく異なっており，米国のファクトチェックサイトの一つである Politifact は 1ヶ月（2021 年 10 月）で 151 件のニュースを検証しているが，ファクトチェック・イニシアティブ・ジャパンは 2 年間（2019 年 9 月から 2021 年 9 月まで）で 279 件のニュースの検証にとどまっている．それぞれの国で拡散されているフェイクニュースの数の差も一つの要因だが，特にこの状況は，日本におけるファクトチェックの需要や取り組みがまだ初期段階であることを示している．このように，日本を始めとする非英語圏では上記のようなリソースが少ないため，フェイクニューストラッカーの開発が難しく、ユーザがソーシャルメディア上で流れる情報を確認するには自主的に調査するしかない．

本研究では，ソーシャルメディア上でフェイクニュースの投稿に対して訂正活動を行う “Guardian” という存在に着目し，リソースが少ない国でも機能するフェイクニュース追跡システム “Fake Guardian” を提案する．Fake Guardian は収集された Guardian のツイートと、Guardian が問題を指摘している

：* equal contribution

2：Politifact: `https://www.politifact.com/`

3：Snope: `https://www.snopes.com/`

4：ファクトチェックナビ：`https://navi.fij.info/`

フェイクニュースを表示する，さらに，そのニュースに関するユーザのフィードバックを得るための投票機能を持つ．また，我々のシステムの Web デモシステムによって，フェイクニュースに対する理解と，ユーザによる有志のフェイクニュース対策活動の周知を促すものである．本システムは Twitter の日本語ツイートに対して機能しているが，将来的には日本語だけでなく様々な言語での適用を目指しフェイクニュース研究のための大規模なデータセットに活用していく予定である．

2 関 連 研 究

2.1 Guardian

我々のシステムで活用するフェイクニュースの投稿に対し訂正や注意喚起活動を行うユーザ “Guardian” は，ファクトチェック URL の自動推薦モデルなどに活用されている [11, 12]．ソーシャルメディアユーザがファクトチェックサイトを引用し，フェイクニュース拡散者に返信する “fact-checking intervention” は，フェイクニュースの拡散を緩和する有用な戦略である [13]．

2.2 フェイクニューストラッキングシステム

ニュースの真偽を確認し，拡散状況を調査することを目的としたツールやシステムがいくつか開発されている．Hoaxy [9] は，事実確認情報とそれらに関連する誤報を収集・追跡するためのフレームワークである．ユーザは興味のあるトピックを検索し，それぞれのトピックの拡散の可視化を確認することができる．FakeNewsTracker [10] は，ソーシャルメディア上のフェイクニュースのデータ収集・検出・可視化を行うシステムである．ファクトチェック機関から検証済みのフェイクニュースの情報源を収集し，その情報源と一致する投稿を検出するシステムである．また，これらのシステムは Hoaxy dataset [14] や FakeNewsNet [15] という形でフェイクニュースのデータセット構築にも利用されている．そのほかにも，ユーザが入力したニュースを Weibo 投稿から追跡しその信憑性を検証するシステムの NewsVerify [16] や，ブラウザの拡張機能で提供されるニュースサイトの信頼性と透明性を 100 点満点で評価し提示する NewsGuard [17] などが存在する．

3 Fake Guardian

本章では，我々のシステム Fake Guardian を，フェイクニュースや Guardian のツイートを収集するためのバックエンドと，収集したツイートをユーザに提示しフィードバックを受けるためのフロントエンドの 2 つの観点から紹介する．図 1 にシステムの全体像を示す．

バックエンドは，フェイクニュースを指摘する可能性が高い Guardian のツイートを収集・処理するために，ツイートクローリング，ノイズ除去，グルーピング，そしてランキングの 4 つのステップで構成されている．フロントエンドは，バックエンドで処理した毎日の Guardian のツイートとフェイクニュースを表示する．さらに，Guardian が指摘したニュースがフェイクかどうかといった情報をユーザが投票できる機能を提供する．

3.1 バックエンド

ツイートクローリング：フェイクの可能性があるニュースを指摘する Guardian のツイートを検索するステップである．観察の結果，Guardian のツイートである可能性が高いパターン，キーワードの組合わせを検索クエリとして，Twitter API を用いてクローリングした．そのパターンは以下の通りである．**“は FAKEWORDS”, “FAKEWORDS です”, “FAKEWORDS である”, “という FAKEWORDS”, “(信じ | 拡散し) ない”** また，FAKEWORDS はこれらの単語のどれかである: **デマ，フェイク，間違い，不正確，誤報，虚偽，事実無根**.

ノイズ除去：このステップでは，言語モデルを用いて，収集したツイートから無関係なツイートやノイズを除去することを目的としている．ノイズ除去のモデルとして，2019 年 9 月から 2020 年 3 月の間に選択したクエリパターンで収集したツイートに対して，Guardian と関係するかどうかをラベル付けした 1200 ツイートを学習データとして，$\text{BERT}_{\text{base-japanese}}$ [18] の事前学習モデルを 10 エポックのファインチューニングを行ったモデルを用いた．

グルーピング：このステップでは，同じフェイクニュースについて言及している Guardian のツイートを集約し 1 つのグループにする．ツイートの内容は日々変化するため，教師ありのグルーピングは困難である．そこで，シンプルかつ堅牢なルールベースのグルーピング手法を実行した．

（1） 同じ URL に関するツイートは，同じグループにまとめる．

（2） 同じツイートに対して返信したツイートは，同じグループにまとめる．

（3） 抽出されたツイートグループ同士の距離を Word Mover’s distance (WMD) [19] を用いて測定し，距離が閾値 0.25 以下のものを同じグループとして設定 [19].

ランキング：ニュースによって，拡散や言及の度合いは異なる．このステップでは，収集したフェイクの可能性が高いニュースの中でも，特に拡散や注目がされているものを取り上げるため，Twitter ユーザから注目されている度合いに応じてランク付けを行う．具体的には，このランキング処理で上位 10 件のニュースをフロントエンドで可視化する．

ここで用いるランク付けの手法として，Glavaš らによって提案された教師なしランキング手法 [20] を採用する．これは，与えられた特徴量の大きさに応じて自動的にランク付けを行うもので，我々のシステムでは，対象となったツイートに付与されたいいね！の数，リツイートの数，そしてリツイートユーザのフォロワー数の割合を計算した Public score の 3 つの特徴を採用している．これら 3 つの特徴のうち，最初の「いいね数」と「リツイート数」が大きければ大きいほど，注目度の高いニュースであると判断している．一方で Public Score は小さいほど注目度が高い，つまり，拡散者以外のフォロワーに拡散されることでより注目されているとして，我々のシステムではランキングの特徴として用いた．

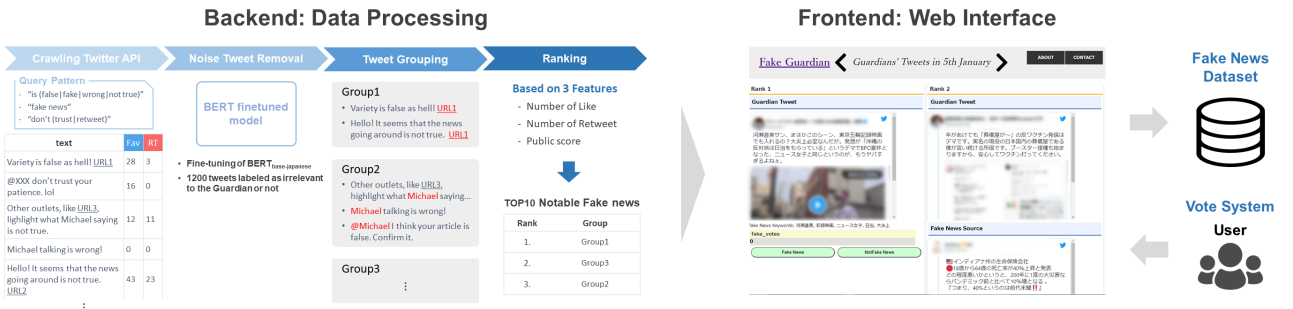

図 1　Fake Gurdian の概要図：Guardian のツイートを収集し注目度の高いニュースを抽出するためのバックエンドと，抽出されたニュースや Guardian のツイートをユーザに示しフィードバックなどを受け取るためのフロントエンドで構成される．

3.2　フロントエンド

フロントエンドでは，バックエンドで処理したデータをシステムを利用するユーザのために，日々のフェイクニュースをランキング形式で確認することができるインターフェイスを設計した．加えて，システムの改良に利用するために，Guardian のツイートがフェイクニュースを指し示すものかどうかを尋ねる投票機能を備えている．

図 2　Fake Guardian が示した 2022 年 1 月 10 日のニュース例

提案システムでは，上位 10 個のイベントを毎日更新して掲載する．図 2 に Fake Guardian が収集し示したツイート例を示す．各ツイートの表示画面は，“Guardian Tweet”，“Fake News Source”，“Keyword”，“Voting System” の 4 つの部分から構成されている．Guardian Tweet は，システムがクローリングして取得した，フェイクニュースを指摘するツイートを示す．News Source には，Guardian Tweet に含まれる URL，引用ツイート，返信ツイートが表示される．Guardian が引用元を明示していない場合や，返信していない場合は，この項目は表示されない．Keyword では，そのフェイクニュースイベントの指摘ツイートを表す特徴的な語を記述し，どのようなニュースかを表す．Voting System は，Gardian のツイートがフェイクニュースを指摘しているものかどうかをユーザに問うもので，ユーザからフィードバックを受け取ることで，より正確なデータの収集を目指すものである．

4　収集システムの有効性

4.1　収集システムの評価

提案システムの有効性を検証するために，ランク付けされたニュースがフェイクニュースかどうかの確認を行った．2021 年 11 月 1 日から 11 月 14 日にランク付けされた Guardian Tweet 122 件を対象に，2 名のアノテーターが以下の観点でアノテーションした．

(a). 収集された指摘ツイートはフェイクであることを指摘するツイートであるか？

(b). 指摘ツイートが指摘している内容は，実際にフェイクやデマであるか？

この結果，2 名の Cohen's Kappa score は 0.73 で一定の合意をとれたことを確認した．なお，2 人のアノテーターが同意しなかったニュースについては第 3 の評価者（著者のうち 1 名）がラベルを付与した．

(a) の結果から，収集したツイートの 77%がニュース記事における虚偽の可能性を指摘していることがわかる．これは，本システムにおける選択パターンやノイズツイートの除去が適切に機能していることを示唆している．また，(b) の結果から，収集したツイートのうち約 52%が本当に嘘であること示している．これらの結果から，本システムはソーシャルメディアユーザが注目するフェイクニュースを大量に収集できることが示唆されたものの，Guardian の指摘の品質に大きく依存している．これらの問題は，ユーザの投票結果を利用することで，改善する可能性がある．

4.2　限　　界

Fake Guardian の現状の限界ついて 2 つの例を示す．まず，図 3 ような，フェイクニュースを指摘している言語的特徴を持

つものの，フェイクニュースと関係のないミーム，ソーシャルメディア上の模倣行為をシステムが収集している．学習データの少なさによる言語的特徴によるノイズ除去の弱点を補う形として，グルーピングで似たような内容のニュースが存在しているか，ランキングで，そのニュースが拡散しているか，3 つの要素により，「注目すべき」フェイクニュースを抽出している．ノイズ除去をすり抜けたミームや宣伝のようなツイートは，広く拡散されリプライが付きやすいことや，似た内容のツイートが出現することが多いことにより，システムの判別が失敗している．

図 3　Fake Guardian が示した 2021 年 12 月 20 日前後の誤認識例：クリスマス前に流行した複数のミームをシステムがフェイクニュースと認識している

次に，Guardian Tweet と News Source の関係が不明瞭であるという点である．Fake Guardian で得られたデータを観察した結果，3 つのパターンが見られた．1 つ目は，Guardian Tweet のみの場合で，これはツイート本文でフェイクニュースの内容を言及とその問題点を指摘する形である．このとき News Source 項目はない．2 つ目は，図 4 のように，Guardian Tweet はリプライや引用リツイートの形で，フェイクニュースと考えられるツイートに対して，問題点を指摘する形である．このとき News Source はフェイクニュースと考えられるツイートを示しており，システムが理想通りに動作している場合と考えられる．3 つ目は，図 5 のように，フェイクニュースと考えられる News source のツイートに対して，問題を指摘している Guardian Tweet という構図をうまく捉えることができていない．Guardian Tweet と News Source は，どちらかのツイートに賛同を示す関係であったり，両方が類似した内容となっていたり，発言の根拠として引用している関係であったりと様々である．このようにフェイクニュースを念頭に置いたシステムであるが，収集したパターンによっては Guardian Tweet が指し示すフェイクニュースが不明瞭である‥そのため，Guardian Tweet と News Source のツイートの関係性を判定する機能を組み込む必要性がある．

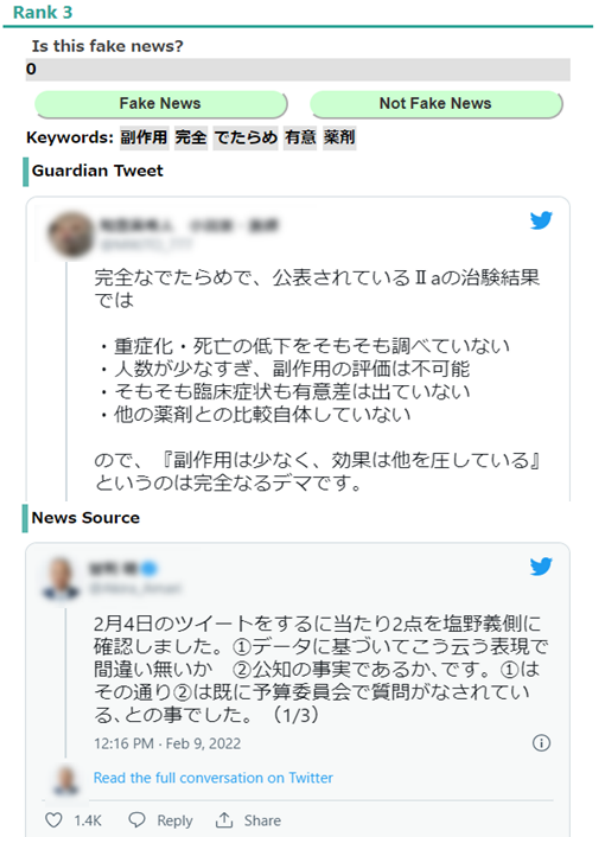

図 4　Guardian Tweet と News Source が指摘と問題点の関係が明瞭な事例

5　おわりに

我々は，Twitter 上でフェイクニュースの投稿に対して訂正活動を行う “Guardian” という存在に着目し，リソースが少ない国でも機能するフェイクニュース追跡システム “Fake Guardian” を提案した．あらゆる言語に容易に適用可能なシステムにより，データセットの構築や責任あるファクトチェックを支援し，様々な国でフェイクニュース研究の手助けをすることを期待する．一方課題も多数存在する．1 つ目は収集システム部分の課題で，Guardian Tweet 自体の収集精度の向上，そのフェイクニュースと Guardian Tweet の関係がの適切どうかの検証が必要である．2 つ目はユーザインターフェースの改良で，Guardian Tweet 通じて，多様な視点や正しい情報を再発信するという目的に適切なデザインを実現する必要がある．3 つ目は，ファクトチェックが行われる前に，有志の Guardian によるのフェイ

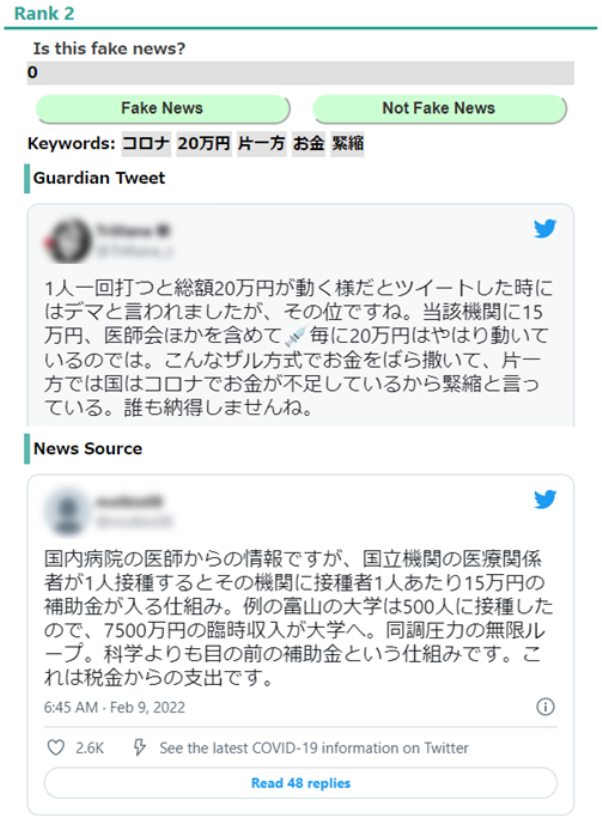

図 5　Guardian Tweet と News Source の関係が不明瞭であり，うまくフェイクニュースを収集できていない考えられる事例

クニュースに対する指摘が，フェイクニュースの拡散や，問題の発生を抑止できるか，一般ユーザに問題の存在を周知できているか，有効な Guardian Tweet は何か検証する必要がある．将来的には，ファクトチェック以外のフェイクニュースへの対策として，トラッキングといったアプローチの有効性の検証も加えていきたい．

文　　献

[1] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. Detection and analysis of 2016 us presidential election related rumors on twitter. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 14–24. Springer, 2017.

[2] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14, 2019.

[3] Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. An exploratory study of covid-19 misinformation on twitter. *Online social networks and media*, 22: 100104, 2021.

[4] J Scott Brennen, Felix Simon, Philip N Howard, and Rasmus Kleis Nielsen. Types, sources, and claims of covid-19 misinformation. *Reuters Institute*, 7(3):1, 2020.

[5] Adam Kucharski. Study epidemiology of fake news. *Nature*, 540(7634):525–525, 2016.

[6] Ashley Lime. A year in fake news in africa. `https://www.bbc.com/news/world-africa-46127868`, 2018. [accessed on 11/09/2021].

[7] European Institude of Peace. Fake news misinformation and hate speech in ehiopia: A vulnerability assesment. `https://www.eip.org/publication/fake-news-misinformation-and-hate-speech-in-ethiopia-a-vulnerability-assessment/`, 2021. [accessed on December 31th, 2021].

[8] 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*, 10(4):e0121443, 2015.

[9] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proc. of the International Conference Companion on World Wide Web*, pages 745–750, 2016.

[10] Kai Shu, Deepak Mahudeswaran, and Huan Liu. Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory*, 25(1):60–71, 2019.

[11] Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 275–284, 2018.

[12] Vim Nguyen and Kyumin Lee. Learning from fact-checkers: Analysis and generation of fact-checking language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–344, 2019.

[13] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*, pages 1395–1405, 2015.

[14] Pik-Mai Hui, Chengcheng Shao, Alessandro Flammini, Filippo Menczer, and Giovanni Luca Ciampaglia. The hoaxy misinformation and fact-checking diffusion network. In *Proc. of International AAAI Conference on Web and Social Media*, pages 528–530, 2018.

[15] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. *arXiv e-prints*, page arXiv:1809.01286, Sep 2018.

[16] Xing Zhou, Juan Cao, Zhiwei Jin, Fei Xie, Yu Su, Dafeng Chu, Xuehui Cao, and Junqiang Zhang. Realtime news certification system on sina weibo. In *Proc. of the International Conference on World Wide Web*, pages 983–988, 2015.

[17] Newsguard. `https://www.newsguardtech.com/`. [accessed on November 9th, 2021].

[18] Hugging face: cl-tohoku/bert-base-japanese. `https://huggingface.co/cl-tohoku/bert-base-japanese`. [accessed on November 9th, 2021].

[19] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *Proc. of International conference on machine learning*, pages 957–966, 2015.

[20] Goran Glavaš and Sanja Štajner. Simplifying lexical simplification: Do we need simplified corpora? In *Proc. of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, 2015.