# 記事コメント生成によるフェイクニュースの早期検出

柳 裕太 $^{1,a)}$  折原 良平 $^{1,b)}$  清 雄 $^{-1,c)}$  田原 康之 $^{1,d)}$  大須賀 昭彦 $^{1,e)}$ 

概要: SNS 上でフェイクニュースが拡散されて事実と異なる風評が広がりやすくなった。誤った風評に騙された人々が社会的損害を与えるためこの問題は深刻である。ファクトチェックがフェイクニュース対策として行われているが、属人的な作業である上に時間がかかるため、フェイクニュースと比べ拡散されにくい課題がある。自動でフェイクニュースを検出することが広く研究されており、記事に加えてリツイートやリプライといったソーシャルコンテキストが検出性能を改善することが確認されている。しかしながら、ソーシャルコンテキストは SNS ユーザの拡散によって生まれる情報であるため、同じく検出に時間がかかる。我々はフェイクニュースの早期検出に向けて、ソーシャルコンテキスト情報として記事へのコメントを生成することで検出を補助するフェイクニュース自動検出モデルを提案する。コメント生成モデルと真偽分類モデルは記事とコメントを併せ持つデータセットから学習される。検証時は実在コメント件数を制限した状況から新たにコメントを生成した上で真偽分類を補助させる。実際に生成コメントを付加して分類した場合と、付加せず分類した場合を比較した結果、生成コメントを付加した方がより多くのフェイクニュースを検出した。これは、我々の提案したモデルが早期検出に向くことを示唆している。

# 1. 序論

現代において、ニュースといった情報の入手と拡散が簡単にできるソーシャルメディアは生活の重要な一部となった。その中には信憑性に乏しい情報が含まれており、特に悪意によって読者を騙して誤った風説を作るために作られた情報であるフェイクニュースがある。

フェイクニュースの実例として、特に今年は新型コロナウイルス感染症 (COVID-19) にまつわる誤った風説がソーシャルメディア上で広く流布された。WHO 局長はこの問題を"インフォデミック"と呼び、フェイクニュースはウイルスそのものよりも早く簡単に拡散されると警戒を呼びかけている。[1] また、フェイクニュースによってオンライン上で誤った風説が広がった結果、オフライン上へ大きな影響を与えたこともある。ワシントン DC で発生したピザ屋で銃乱射事件が発生した際、被疑者はインターネット上でのフェイクニュースに端を発する児童ポルノ疑惑が犯行の動機であることが報道されている[2]。以上より、フェイクニュースの拡散によって読者が事実に基づく正しいニュースへのアクセスが難しくなるため、民主主義の根幹を揺る



"Donald Trump to issue an executive order to ban Facebook across U.S."



FAKE NEWS PUNDITFACT 名 BLOGGERS



A fake news story that said President Donald Trump would sign an executive order banning Facebook was a website's April Foot's Day prank. (AP photo)

Fake news story about Trump banning Facebook by executive order is an April Fool's Day prank

図 1: 北米で行われたファクトチェックの一例。青矢印はフェイクニュース投稿日時とファクトチェック結果投稿日時を示し、両者には 25 日もの間が開いている。

がしてしまう。現在、フェイクニュース検出に向けて有識者が事実関係を確認して結果を公表するファクトチェックが行われている。

図1はファクトチェックの一例である[3]。この実例のように、ファクトチェックは属人的な作業であることに加え

電気通信大学

UEC, Chofu, Tokyo 182–8585, Japan

a) yanagi.yuta@ohsuga.lab.uec.ac.jp

b) orihara@acm.org

c) seiuny@uec.ac.jp

d) ohsuga@uec.ac.jp

e) tahara@uec.ac.jp

IPSJ SIG Technical Report

て結果公表まで時間がかかるため、フェイクニュースその ものに比べて拡散されにくい。このため、機械学習によっ てフェイクニュースを自動で検出する研究が行われている。

自動検出にあたって困難な点は、フェイクニュースは人々を騙すために巧妙なつくりをしていることが挙げられる。このため、単純なルールベース手法による検出は難しい。検出性能の向上において、記事そのものがもつ情報に加えてソーシャルメディア上での反響を示すソーシャルコンテキスト(リツイート・いいね・リプライなど)を考慮することが有効であることが先行研究で示されている[4]。しかしながら、ソーシャルコンテキストはユーザの拡散によって生まれるため、この場合も早期の検出には向かない。これに対して、ニュースに対してソーシャルメディア上で寄せられるコメントで発生しやすい単語を、条件付き変分オートエンコーダ(CVAE)で生成する手法も提案されている[5]。この手法は、記事から確率分布とラベルを元に隠し変数を介して生成を行っている。

本研究では、記事と実際に記事に寄せられたコメントから信憑性の学習を行い、記事と限られた数のコメントから別のコメントを予測させた上で真偽を判断するモデルを提案する。このモデルは、フェイクニュースそのものを生成するモデル [6] を拡張する形で実装することでコメントの生成を実現する。学習では記事と実際に記事に寄せられたコメントを3件、更に真偽ラベルを入力するが、テスト時は記事に加えて実際に寄せられたコメントは2件に制限し、真偽ラベルは入力しない。

我々は提案モデルの検出性能を実際に投稿された情報を もつデータセットによって検証した。

### 2. 関連研究

フェイクニュースの検出 (真偽分類) は、対象をスパム [7] や風評 [8]、そして虚偽広告 [9] を含めると新しいトピックではない。これまでの研究 [10], [11], [12] に倣い、意図的に作成され、明確に誤りであると確認できるニュースをフェイクニュースと定義する。

#### 2.1 フェイクニュース検出

ニュース記事がもつ情報からフェイクニュースを検出する手法は多く提案されている。文字情報からは、フェイクニュースが独自の書かれ方をする上に感情的な表現を多用することから、文章のスタイル [13] や感情的表現の頻度 [14] を考慮する手法がある。また、ディープニューラルネットワーク (DNN) によって検出性能が改善された報告 [15], [16], [17] も多い。

ソーシャルコンテキストを考慮した手法も多く提案されており、扱うコンテキストの種類によってユーザベース [18], [19], [20] ・投稿ベース [21], [22], [23] ・ネットワークベース [24], [25] の 3 種類に分けられる。

ソーシャルコンテキストの共通した問題点として、ソーシャルコンテキストはユーザの拡散によって生まれる情報であるため早期検出に向かない点が挙げられる。早期検出の実現へ、TCNN-URGという2層の畳み込みニューラルネットワークとCVAEによるユーザレスポンス生成器を組み合わせたモデルも提案されている[5]。ニュース記事を畳み込みニューラルネットワークで特徴化してから隠れ変数を算出し、寄せられたコメントとして尤もらしい単語群を生成することで検出性能が改善されることが報告されている。しかしながら、TCNN-URGはあくまで尤もらしい単語を生成することに限られ、実際のコメントそのものは生成していない。

#### 2.2 フェイクニュース生成

自然言語生成タスクの1つとして、架空のニュース記事を作成するGroverモデルがある[6]。このモデルはニュース記事データセットから記事をドメイン・著者・投稿日時・見出し・本文の5要素に分け、無作為に歯抜けにさせた記事の残り部分から歯抜け部分を予測させることで訓練している。興味深い点として、Groverモデルで生成した記事の方が実在の記事よりも読者が信じる傾向であることが報告されていたことがある。本研究ではこのモデルを拡張することで、より自然なコメントを生成することを目指した。

## 3. 提案手法

先行研究により、自然言語文章生成モデルは言語モデル問題の1つとされており、式1のように文章 $x=\mathbf{w}_1^T=(\mathbf{w}_1,\mathbf{w}_2,...,\mathbf{w}_T)$ はある単語 $\mathbf{w}_t$ が生成される前の単語群 $\mathbf{w}_1^{t-1}$ による条件付き確率の総積であると定義されている。

$$p(x) = p(\mathbf{w}_1^T) = \prod_{t=1}^{T} p(\mathbf{w}_t | \mathbf{w}_1^{t-1})$$
 (1)

提案モデルによる文章生成の流れは図2の通りである。第2.2の通り、ベースとなったGroverモデルは記事を5要素に分けて学習が行われており、生成及び分類学習において、各要素の始点と終点には開始及び終了トークンが付加されている。本研究ではこれらの要素を記事本文とそれに寄せられた3件のコメントに置換することで実装する。ベースとなったモデルに倣い、提案モデルは以下の同時分布として定義する。

 $p(\text{article}, \text{comment\_1}, \text{comment\_2}, \text{comment\_3})$  (2)

コメント生成学習時は、ベースとなった Grover モデル と同様に記事とコメントのセットを 2 つの集団に分け、無 作為に歯抜けにする。コメントの場合は 10%、記事本文の 場合は 35%の確率で歯抜けにしてから一方の集団から学習

IPSJ SIG Technical Report

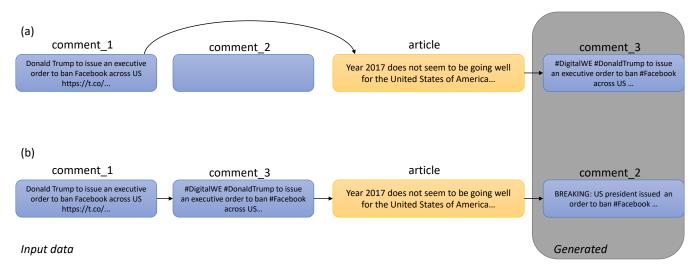


図 2: 提案モデルのコメント生成例。(a) は記事と 1 件の実際に寄せられたコメントからコメントを生成している。(b) は (a) で生成したコメントを含めた状況で更にコメントを生成している。



図 3: 実験の流れ

を行い、もう一方での生成におけるクロスエントロピー誤差を最小化するように訓練される [6]。提案モデルの目的は記事ではなく SNS 上で記事に寄せられたユーザの反応を生成することである。

記事とコメントのセットの末尾にはセットの終端を意味するトークンである [CLS] を追加し、またこのトークンが真偽を分類する際に使われる。これは Grover モデルがベースとしている GPT-2 がとる手法 [26] と同一である。図は実際の記事とコメントのセットを真偽分類するまでの流れを示している。まず、記事に寄せられたコメント群から実験に使用するために無作為に 3 件選出し、コメント生成の学習を行う。真偽分類する際には、3 件の実際に投稿されたコメントから 1 件削除してから生成されたコメントを追加してから真偽の分類を行う。また、同時に生成コメントを追加しなかった状況で分類を行った際の結果との比較も行った。

## 4. 実験

#### 4.1 単語生成の傾向

まずはじめに、正しいニュースとフェイクニュースの 記事とコメントのセットから生成されたコメントの傾向

を調べた。いずれのニュース記事とコメントのセットは FakeNewsNet データセット [27] から取得した。このデータ セットは北米で英文記事を対象にファクトチェックを行う 団体である PolitiFact(政治ニュース中心) と GossipCop(芸 能ニュース中心) の判断結果を元に正しいニュースとフェ イクニュースのラベルが付けられている。我々はまず実験 手法に合わせるために記事に対して最低3件以上コメント が寄せられているセットに対して無作為に3件選出した。 また、PolitiFact から真偽で各 200 セットを用意して学習 を行った。実際に生成されたコメントに対して、すべてア ルファベットを小文字にした上で単語ごとの出現回数と出 現確率を算出した。また、算出するにあたって記号(クォー テーションやピリオド、コンマなど)や URL の削除を行っ たほか、"a" や "is" といったストップワードは NLTK[28] が提供するメソッドを使用して除外した。なお、コメント の収集元が Twitter であることから、Twitter 独自の用法 をもつ記号 (ハッシュタグ#やメンション@)、またコロン は例外として除外しなかった。以上の処理を行った結果、 以下の特徴が得られた。

 真偽問わず最も頻度が高い単語は "via" であり、真偽 全体の単語のうち約 1.5%を占めた。

表 1: 分類成績

 入力データ	精度	再現率	F値
記事本文のみ	0.647	0.615	0.631
+ 実在コメント 2 件	0.682	0.750	0.714
+ 生成コメント1件	0.590	0.790	0.675

• それに続いて "trump" と "obama" が続いたが、いず れも割合は 1%を下回った。

また、真偽における傾向差として以下の違いが見られた。

- "via" は真偽単独で見てもそれぞれで最も高い頻度で 生成されていた。
- フェイクにおける "via" の生成頻度は正しい場合に比べて約 2 倍であり、その差は約 0.9 ポイントとフェイク頻出上位 10 単語中最多だった。
- "breaking:"という単語が"via"に次いで2番目に頻度 の差が高い単語であり、その差は0.7ポイントだった。

# 4.2 検出における生成コメントの影響

生成コメントの有無によって真偽分類の結果に影響が出るか調べた。ベースラインとして2つの入力データを用意した。1つは生成されたコメントを入力せず、記事と実際に投稿された2件のコメントから分類させた場合、もう1つは実際に投稿されたコメントも入力せず、記事のみから分類させた場合であった。この実験では、PolitiFactでは十分な量の学習を行うにはセット数が少なかったため、GossipCopから真偽で各2000セットを用意して行った。実験結果は表1の通りである。提案モデルは再現率において全体ベストとなったものの、精度においては生成モデルを使わない方が優秀であることが読み取れる。また、生成されたコメントは共通して文法面にさらなる改善の必要性が残された。

#### 参考文献

- John Zarocostas. How to fight an infodemic. *The Lancet*, Vol. 395, No. 10225, p. 676, 2020.
- [2] Guardian staff and agencies. Washington gunman motivated by fake news 'pizzagate' conspiracy, 12 2016.
- [3] Joshua Gillin. Politifact fake news story about trump banning facebook by executive order is an april fool's day prank, Apr 2017.
- [4] Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. Rumor detection with hierarchical social attention network. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, pp. 943–951, New York, NY, USA, 2018. ACM.
- [5] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. Neural user response generator: Fake news detection with collective user intelligence. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp. 3834–3840. Inter-

- national Joint Conferences on Artificial Intelligence Organization, 7 2018.
- [6] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, Advances in Neural Information Processing Systems 32, pp. 9054–9065. Curran Associates, Inc., 2019.
- [7] Hua Shen, Fenglong Ma, Xianchao Zhang, Linlin Zong, Xinyue Liu, and Wenxin Liang. Discovering social spammers from multiple views. *Neurocomputing*, Vol. 225, pp. 49–57, 2017.
- [8] Z. Jin, J. Cao, Y. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In 2014 IEEE International Conference on Data Mining, pp. 230–239, 12 2014.
- [9] Hen-Hsen Huang, Yu-Wei Wen, and Hsin-Hsi Chen. Detection of false online advertisements with dcnn. In Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, pp. 795–796, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.
- [10] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. SIGKDD Explor. Newsl., Vol. 19, No. 1, pp. 22–36, September 2017.
- [11] Natali Ruchansky, Sungyong Seo, and Yan Liu. Csi: A hybrid deep model for fake news detection. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, pp. 797–806, New York, NY, USA, 2017. ACM.
- [12] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 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*, KDD '18, pp. 849–857, New York, NY, USA, 2018. ACM.
- [13] Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. CoRR, Vol. abs/1702.05638, , 2017.
- [14] Chuan Guo, Juan Cao, Xueyao Zhang, Kai Shu, and Miao Yu. Exploiting emotions for fake news detection on social media. CoRR, Vol. abs/1903.01728, , 2019.
- [15] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics.
- [16] Hamid Karimi and Jiliang Tang. Learning hierarchical discourse-level structure for fake news detection. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3432–3442, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [17] Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. Multi-source multi-class fake news detection. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1546–1557, Santa Fe, New Mexico, USA, August 2018. Association for Com-

#### 情報処理学会研究報告

IPSJ SIG Technical Report

- putational Linguistics.
- [18] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In Proceedings of the 20th International Conference on World Wide Web, WWW '11, pp. 675–684, New York, NY, USA, 2011. ACM.
- [19] K. Shu, S. Wang, and H. Liu. Understanding user profiles on social media for fake news detection. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 430–435, 4 2018.
- [20] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profile for fake news detection. CoRR, Vol. abs/1904.13355, , 2019.
- [21] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In AAAI, 2019.
- [22] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. ArXiv, Vol. abs/1704.07506, , 2017.
- [23] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. News verification by exploiting conflicting social viewpoints in microblogs. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, pp. 2972–2978. AAAI Press, 2016.
- [24] Liang Wu and Huan Liu. Tracing fake-news footprints: Characterizing social media messages by how they propagate. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, pp. 637–645, New York, NY, USA, 2018. ACM.
- [25] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. Fake news detection on social media using geometric deep learning. CoRR, Vol. abs/1902.06673, , 2019.
- [26] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.
- [27] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. ArXiv, Vol. abs/1809.01286, , 2018.
- [28] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics.