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 雄一[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。

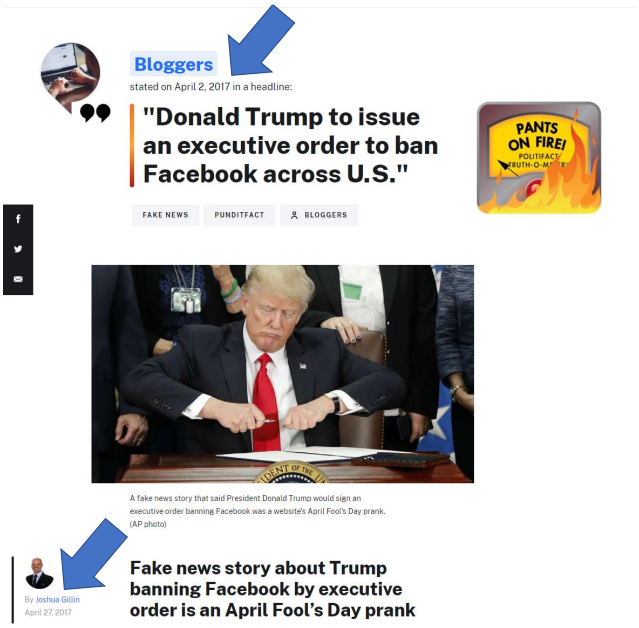

図 1: 北米で行われたファクトチェックの一例。青矢印はフェイクニュース投稿日時とファクトチェック結果投稿日時を示し、両者には 25 日もの間が開いている。

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

[1] 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

© 1959 Information Processing Society of Japan

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。

参考文献

[1] 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. International 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.

© 1959 Information Processing Society of Japan