2.【現在までの研究状況】(図表を含めてもよいので、わかりやすく記述してください。様式の変更・追加は不可(以下同様))
① これまでの研究の背景、問題点、解決方策、研究目的、研究方法、特色と独創的な点について当該分野の重要文献を挙げて記述してください。
② 申請者のこれまでの研究経過及び得られた結果について、問題点を含め①で記載したことと関連づけて説明してください。
なお、これまでの研究結果を論文あるいは学会等で発表している場合には、申請者が担当した部分を明らかにして、それらの内容を記述してください。

●これまでの研究の背景

SNS の発展により、情報を迅速かつ大量に取得し、拡散することで容易に共有できるようになった。一方、悪意により他人を騙すために作られた**フェイクニュース**が拡散されやすくなった。利用者の間で拡散されると、**誤った認識が広がって騙された人々が社会的損害を起こす**という問題があるため、**フェイクニュース拡散の早期抑制が必要とされている**。

●問題点

現在フェイクニュースの拡散抑制のために、有識者が事実関係の確認を行う**ファクトチェック**がとられている。しかしこれは (1) 属人的な作業であること、(2) 拡散されてから調査されることが多いこと、以上の理由より結果を公表するまで時間がかかる。このためフェイクニュースと比べ拡散されず、拡散抑制に繋がらないことが多い。そこで迅速なファクトチェックを行うために、ニューステキストや添付メディア、そしてユーザの反応からディープニューラルネットワーク (DNN) を利用しフェイクニュースを自動で検出する手法が多く開発されている。

しかし、これらの手法において**ユーザの反応は拡散後でしか得られない**ため、早期の検出を想定した場合は評価対象が制限され自動検出の性能が落ちてしまう問題がある。

●解決方策

そこで本研究では DNN の**学習でのみユーザの反応を活用**し、テスト時は**ユーザの反応をコメント生成モデルにより生成・補完**して分類することで、**精度を落とさず早期検出を目指す**ことにした。

●研究目的・研究方法

フェイクニュース早期検出に向け、SNS 上で**ニュースに寄せられたコメントを生成する**ことが、**真偽を分類する精度の向上につながる**ことを示す。本研究はニュースと寄せられたコメントを、ニュースの本文と実際に SNS 上で投稿されたコメント 3 件を 1 ユニットとして扱うことにした。

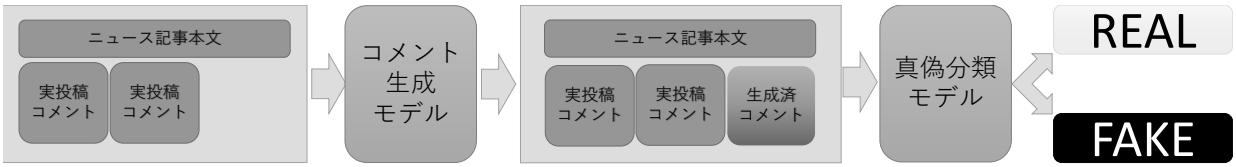

図 1: 分類タスクの流れ。コメント生成モデルで 1 件コメントを生成し真偽分類に活用する。

●特色と独創的な点

- SNS 上の拡散スピードに追いつくことが検出モデルのコンセプトである点
- 真偽を評価する分類タスクに、コメントの生成タスクを組み込んだ点
- 分類性能を大きく失わずに速報性をもつことができる点

●これまでの研究経過及び得られた結果

申請者はデータセットとして FakeNewsNet[3] を使用した。このデータセットは、ファクトチェックで**真偽が評価済である英文ニュース**と、それに **Twitter 上で言及された投稿 (ツイート)** 等をもつ。本研究では最低 3 件以上英文でコメントとしてツイートが寄せられた芸能ニュースを真偽で各 2000 件使用した。拡散の初期段階ではコメントの数は期待できないため、使用するコメントは各 3 件ずつ無作為に選出し、残りは対象から除外した。

生成・分類モデルは、フェイクニュースを自動で作成する Grover モデル [4] を拡張する形で実装した。このモデルはフェイクニュースをドメイン・著者・投稿日・見出し・本文の 5 要素に分け、**ランダムで**

申請者登録名　柳　裕太

（現在までの研究状況の続き）

歯抜けにして予測させる形で生成学習を実現したものである。今回はこれをユニットの 4 要素 (記事本文と 3 件のコメント) での実装を目指し調整を行った。訓練が完了したコメント生成モデルを使い、図 1 の通り**コメントを 1 件欠損させたユニットに生成コメントを付加した上で、Real か Fake か分類させた**。分類モデルは Grover が提供したものを流用した。

その結果、生成コメントを含めた場合の **Fake 記事を見抜いた割合を示す再現率 (Recall) が 0.79** と、欠損のまま分類させたときの **0.75** と、本文単独で分類したときの **0.62** を上回った。つまり、コメントを生成することで**ファクトチェックが必要な疑わしい記事をより多く検出**した [5]。同時に、生成されたコメントで頻出した単語の傾向において真偽で大きな違いはみられなかった。これは、**記事の真偽によってコメント内の単語傾向の差は軽微**であることを意味した。

参考文献

[1] Guardian staff and agencies. Washington gunman motivated by fake news 'pizzagate' conspiracy,12 2016.

[2] John Zarocostas. How to fight an infodemic. *The Lancet*, Vol. 395, No. 10225, p. 676, 2020.

[3] Kai Shu, *et al.* Fakenewsnet: Adata repository with news content, social context and dynamic information for studying fake news on social media. *ArXiv*, Vol. abs/1809.01286, 2018.

[4] Rowan Zellers, *et al.* Defending against neural fake news. *Advances in Neural Information Processing Systems 32*, pp. 9054 − 9065. Curran Associates, Inc., 2019.

[5] Yuta Yanagi, *et al.* Fake news detection with generated comments for news articles. *EasyChair Preprint no. 3190*, EasyChair, 2020.

3. 【これからの研究計画】

(1) 研究の背景

これからの研究計画の背景、問題点、解決すべき点、着想に至った経緯等について参考文献を挙げて記入してください。

● 2. で述べた研究状況を踏まえこれからの研究計画の背景

自然言語処理は、前後の文脈を考慮できる BERT[6] や前節で述べた Grover[4] を始め、より自然な文章が近年は生成できるようになった。前節の研究では実際に Grover を拡張し**コメントを生成することでより多くのフェイクニュースを検出**した。今後は、**より拡散を抑制できるモデルの開発を目指す**。

●問題点・解決すべき点

生成コメントを付加して分類した場合、前節提案モデルが Fake と判断した中で実際に Fake だった割合である精度 (Precision) は 0.59 だった。これは生成コメントを付加せず分類したときの 0.68 と比べ 0.09 ポイント下回る [5]。つまり多くのフェイクニュースを検出することができても、**このモデルは 41%の確率で事実に基づくニュースもフェイクニュースと判定する**。精度と再現率の調和平均である F 値 (F1 score) も、提案モデルは **0.68** と単体で見ても決して高いものではなかった。

同時に、生成されたコメントは文法面でのクオリティが悪いことから、生成コメントから判断の根拠とする説明可能性の提供は難しい。これではいくらフェイクニュースを検出できても、**判断の理由も説明できない狼少年めいたモデル**では利用者の信用を得るのは難しく、拡散の抑制にはならない。

この**分類性能不足**と、**不自然なコメントにより説明可能性を提供できない** 2 点を解決する必要がある。

●着想に至った経緯

実際に早期自動検出モデルを SNS 上で運用する場合を想定した。このモデルの目的は拡散の抑制であるため、**フェイクニュース以上に利用者の信用を得なければ拡散を食い止められない**。そのためには、真偽分類の性能を上げることと、説得力向上のために説明可能性を同時に提供する必要があると着想に至った。

参考文献

[6] Jacob Devlin, *et al.* BERT: Pre-training of deep bidirec-tional transformers for language understanding. In *Proc. of the NAACL-HLT,* pp. 4171 − 4186, Minneapolis, Minnesota, June 2019.

(2) 研究目的・内容（図表を含めてもよいので、わかりやすく記述してください。）

① 研究目的、研究方法、研究内容について記述してください。
② どのような計画で、何を、どこまで明らかにしようとするのか、具体的に記入してください。
③ 所属研究室の研究との関連において、申請者が担当する部分を明らかにしてください。
④ 研究計画の期間中に異なった研究機関（外国の研究機関等を含む。）において研究に従事することを予定している場合はその旨を記載してください。

●研究目的

本研究では、SNS 上でフェイクニュースの拡散を抑制するために、早期自動検出の精度向上と説明可能性の付与を目的とする。具体的には、(1) **生成されたコメントを使用した分類で F 値が 0.8 を上回る手法の確立**と、(2) **生成されたコメントからユーザへ説明可能性を提案する手法の確立**を目指す研究を行う。

●研究方法・研究内容

(1) 生成されたコメントを使用した分類で F 値が 0.8 を上回る手法の確立:

早期検出できてもユーザから信用が得られない狼少年的なモデルではなく、**的確にフェイクニュースだけを発見させることが可能なモデル**を構築する。また、性能向上に向く大規模データセットがない場合は条件に符合するものを自作する。最低でも F 値 0.7 以上を目指す。

(2) 生成されたコメントからユーザへ説明可能性を提案する手法の確立:

SNS 上でフェイクニュースの疑いが強い指摘をする場合を想定し、**生成したコメントを理由付けの題材として活用する**ことを目指す。また、**理由付けの有無によって SNS 利用者からの信用を得られる**ことを主観実験で示す。生成コメントから理由付けが難しいならば、実投稿からの実現を目指す。

●所属研究室との関連

当研究室はエージェント、知的 Web、ソフトウェア工学、データマイニングの 4 つの分野に渡っており、本研究はデータマイニングの一環となる。また、デマや噂の検出を含めても本研究には前例はなく、**申請者が当研究室で初めて本研究に着手**した。

●研究計画の期間中に異なった研究機関 (外国の研究機関等を含む。) において研究に従事することを予定

申請者は**期間中 1 年間北米あるいは欧州の研究施設での活動を予定**している。国内外でフェイクニュース対策研究には温度差が見られ、特に北米と欧州では盛んに研究が行われている (次項目で詳説) ことから、**最前線の研究に従事する**ためにも申請者が現地で研究に従事することが必要である。

(3) 研究の特色・独創的な点

次の項目について記載してください。
① これまでの先行研究等があれば、それらと比較して、本研究の特色、着眼点、独創的な点
② 国内外の関連する研究の中での当該研究の位置づけ、意義
③ 本研究が完成したとき予想されるインパクト及び将来の見通し

●これまでの先行研究等があれば、それらと比較して、本研究の特色、着眼点、独創的な点

ニュースに寄せられそうなコメントを生成する手法は、確率分布に従った潜在変数と正解ラベルを使用して頻出単語を生成する TCNN-URG が提案されている [7]。本研究は**頻出単語を生成するのではなく**、説明可能性に繋ぎやすい**実際に投稿されたようなコメントを生成する**ことを目指している。

また、速報性を維持するためにユーザの反応を補完する弱教師あり学習を活用した手法である MWSS も既に今年提案されている [8]。本研究では**生成する対象をコメントに絞っている**。他のユーザの反応 (リツイート、いいね、反応したユーザ情報) はコメントに比べて説明可能性に繋げにくいためである。

フェイクニュース対策に説明可能性を提供する手法として、記事とコメントから真偽判断の決め手となった部分を評価する dEFEND が提案されている [9]。これは既に投稿されたコメントを対象に含むため、まだコメントが多く寄せられていない状況である**早期検出を目指す場合には向かない**。当研究では、**生成されたコメントから説明可能性を提供する**ことで早期検出を実現する。

●国内外の関連する研究の中での当該研究の位置づけ、意義

風評やデマの自動検出も当該研究に含めると、歴史は決して浅くない。ただし、ここ**数年で社会情勢の変化によって一気に世界的に競争が激化**した。

例えば、Google Scholar 上で 2015 年に投稿された中で “fake news” でヒットする論文は **520 本**に対して、同じ条件で 2019 年に投稿された論文は **15,400 本**と実に **30 倍近くに増加**した。

（研究の特色・独創的な点の続き ）

前項目の通り、特に北米や欧州から頻繁に英語論文が発表されている。先述と同じプラットフォームで2019年で“フェイクニュース”でヒットする論文は**169本**と、1年間で実に日本語の**90倍以上**の英語論文が発表されている。

これは**地域による問題意識の差**の他にも、特に (本研究を含め) 近年機械学習やDNNを活用した研究が多いことも考えられる。これらの手法に必要な記事と真偽データなどを含む**大規模データセットが英語に集中している**のである。フェイクニュース検出の場合、ファクトチェック結果をラベル付けに流用することができるが、北米・欧州に比べて日本国内のファクトチェックは発展途上であるため、日本語データセットが少ない。もしも日本語を研究対象に含める場合、まずはデータセット作りから着手する必要がある。

●本研究が完成したとき予想されるインパクト及び将来の見通し

本研究が完成すると、SNSの利用者へこの情報が事実かフェイクかを判断する新しい判断材料を早い段階からもたらすことができる。また、フェイクニュースがSNS上で**早い段階で説得力がある理由によって指摘**することにより、利用者による**拡散を抑制**することができる。古今東西で虚偽情報を流布しようとする人々は存在するが、**利用者が簡単に騙されないような仕組み作り**を行うことで、**ジャーナリズムと民主主義に対する最大の脅威であるフェイクニュースから人々を守る**ことが可能となる。

参考文献

[7] Feng Qian, *et al.* Neural user response generator: Fake news de-tection with collective user intelligence. In *Proc. of the IJCAI-18*, pp. 3834 − 3840., 2018.

[8] Kai Shu, *et al.* Leveraging multi-source weak social supervision for early detection of fake news. *arXiv*, Vol.abs/2004.01732, 2020.

[9] Kai Shu, *et al.* defend: Explainable fake news detection. In *Proc. of the ACM SIGKDD*, 2019.

(4) 研究計画

申請時点から採用までの準備状況を踏まえ、研究計画について記載してください。

本研究の3年間のスケジュールを以下の表1に示す。

表 1: 本研究の年次計画 (1セルは半期を表す)

項目	1年		2年		3年	
	前期	後期	前期	後期	前期	後期
A. データセットの選定・作成	■	■	■	■	■	
B. コメント生成・真偽分類モデルの実装	■	■	■	■		
C. 真偽分類性能向上		■	■	■		
D. 別言語・ドメインへの対応			■	■	■	
E. 説明可能性の付与				■	■	■
F. 拡散抑止力の評価					■	■

●1年目

A. データセットの選定・作成

以下の各タスクで使用するデータセットを随時選定する。条件はタスクによるが、例えばBならニュースとその真偽、そしてユーザのコメントである。もしも条件を満たすデータセットがない場合は、データセットを自分で集める必要がある。

B. コメント生成・真偽分類モデルの実装

コメントを生成し分類するモデルの実装を引き続き行う。もしも現有モデルの拡張では難しい場合は別の手法からの拡張も検討している。

C. 真偽分類性能向上

本研究では生成コメントを含めた真偽判定において、分類の総合指標であるF値が0.8を上回ることを目指している。データセットの規模拡大やパラメータの調整、分類モデルの変更などで実現を目指す。

（研究計画の続き ）

● 2 年目

D. 別言語・ドメインへの対応

言語やニュースのトピックであるドメインの変動に提案モデルを対応させる。特に日本語対応する場合、形態素解析や事前学習済み日本語単語の分散表現の用意が必要である。また、いずれも同時にデータセットも新たに用意しなければならない。

E. 説明可能性の付与

ユーザに説明可能性を提供するために、生成されたコメントから真偽を判断した材料を取得する。これは分類モデルを拡張することによって実現が可能であると考えている。

● 3 年目

F. 拡散抑止力の評価

実際に SNS 上で提供した時を想定し、分類成績を改善させ説明可能性を付与したモデルが SNS 利用者への意識にどのような影響を与えるか主観評価実験によって評価する。もしも生成されたコメントから説明可能性が得られない場合は、実際に投稿されたコメントや記事から得ることを予定している。

(5) 人権の保護及び法令等の遵守への対応

本欄には、研究計画を遂行するにあたって、相手方の同意・協力を必要とする研究、個人情報の取り扱いの配慮を必要とする研究、生命倫理・安全対策に対する取組を必要とする研究など法令等に基づく手続が必要な研究が含まれている場合に、どのような対策と措置を講じるのか記述してください。**例えば、個人情報を伴うアンケート調査・インタビュー調査、国内外の文化遺産の調査等、提供を受けた試料の使用、侵襲性を伴う研究、ヒト遺伝子解析研究、遺伝子組換え実験、動物実験など**、研究機関内外の情報委員会や倫理委員会等における承認手続が必要となる調査・研究・実験などが対象となりますので手続の状況も具体的に記述してください。

なお、該当しない場合には、その旨記述してください。

コメント取得を予定してしている SNS は Twitter である。Twitter 社は 2020 年 3 月より学術目的で Twitter API の利用を自由化しているほか、取得したツイート ID を含む情報をデータセットとして公開することも学術目的であれば認められている [10]。

また、先行研究が提供したデータセットを使用する場合は、提供者が示すライセンスやポリシーを遵守する。

なお、学習済みモデルの公表は平成 30 年改正著作権法第 30 条 4 号により認められている。

ただし、本研究では主観評価実験として SNS 利用者を対象としたアンケート調査を予定している。この調査により収集したデータは、個人の特定につながる情報を匿名化した上で解析を行い、解析結果の公表に際しては、匿名化を行ったデータを用い、個人情報の漏洩防止に配慮する。

参考文献

[10] Twitter 開発者ポリシーを分かりやすくアップデート, 2020 年 3 月 11 日. (最終閲覧日 2020 年 4 月 19 日)
https://blog.twitter.com/developer/ja_jp/topics/tools/2020/DevPolicyUpdate.html

申請者登録名　柳　裕太

DC

4.【研究遂行能力】研究を遂行する能力について、これまでの研究活動をふまえて述べてください。これまでの研究活動については、網羅的に記載するのではなく、研究課題の実行可能性を説明する上で、その根拠となる文献等の主要なものを適宜引用して述べてください。本項目の作成に当たっては、当該文献等を同定するに十分な情報を記載してください。
具体的には、以下(1)～(6)に留意してください。

(1) 学術雑誌等（紀要・論文集等も含む）に発表した論文、著書（査読の有無を明らかにしてください。**査読のある場合、採録決定済のものに限ります。**）
著者、題名、掲載誌名、発行所、巻号、
pp 開始頁－最終頁、発行年を記載してください。

(2) 学術雑誌等又は商業誌における解説、総説

(3) 国際会議における発表（口頭・ポスターの別、査読の有無を明らかにしてください。）
著者、題名、発表した学会名、論文等の番号、場所、月・年を記載してください。（**発表予定のものは除く。**ただし、発表申し込みが受理されたものは記載してもよい。）

(4) 国内学会・シンポジウム等における発表
(3)と同様に記載してください。

(5) 特許等（申請中、公開中、取得を明らかにしてください。ただし、申請中のもので詳細を記述できない場合は概要のみの記載してください。）

(6) その他（受賞歴等）

申請者はこれまで研究室に配属されてからフェイクニュースの自動検出というトピックに取り組み続けている。研究発表する際には社会情勢の影響もあってか非常に本研究に対する期待感を発表時によく耳にしている。また、自然言語処理技術の急速な発展により、技術的な障壁は年々下がりつつある。

(1) 学術雑誌 (紀要・論文集等も含む) に発表した論文、著書

なし

(2) 学術雑誌等又は商業誌における解説・総説

なし

(3) 国際会議における発表

なし

(4) 国内学会・シンポジウムにおける発表

広学医サイのポスター発表はここに含めていいのか？→高校に問い合わせる
（本研究との直接の関連が皆無）
(以下 1 件　査読なし・口頭発表)

1. ○ 柳裕太、田原康之、大須賀昭彦、清雄一
「画像付きフェイクニュースとジョークニュースの検出・分類に向けた機械学習モデルの検討」、日本ソフトウェア科学会 2018 年度 MACC 研究発表会、大分、2019 年 3 月

(5) 特許等

なし

(6) その他

プレプリント論文
7 月開催の国際学会 INES に投稿中、採録なら (3) に移管予定
一応通知は 5 月 3 日らしいけど多分コロナでぶっ飛ぶからキレそう

1. ○ Yuta Yanagi, Ryouhei Orihara, Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga.
“Fake news detection with generated comments for news articles”. EasyChair Preprint no. 3190, EasyChair, 2020.

申請者登録名　柳　裕太

DC

5.【研究者を志望する動機、目指す研究者像、アピールポイント等】

日本学術振興会特別研究員制度は、我が国の学術研究の将来を担う創造性に富んだ研究者の養成・確保に資することを目的としています。この目的に鑑み、研究者を志望する動機、目指す研究者像、アピールポイント等を記入してください。

●研究職を志望する動機

申請者は嘘が蔓延ることで誰かが謂れのない罪で傷付く社会に大きな問題意識と危機感を抱いている。解決するためには、嘘を発信させないことよりも、嘘を拡散させないユーザの意識醸成が重要と考える。なぜならば、プロパガンダを含めると嘘を流布させようとする人々はどの時代にも存在するためだ。SNSの発展により情報がすぐ拡散されるようになったため、実害が発生するまでの時間が大幅に早まっていることが問題なのである。自然言語処理技術の観点から嘘に騙されない社会作りに必要な技術と知見を迅速に提供することができれば、あるいは利用者がフェイクニュースの拡散を少しでも思いとどまらせることができるかもしれない。そのためには、申請者は研究者として実現方法を検討することが必要である。

●目指す研究者像

申請者が本研究を実現するために必要な研究者像がもつ資質として以下の3つを挙げる。

1. **自分が抱える問題意識や目標から今やるべきことまで切り分ける能力**
2. **今やるべきことの理由を把握しやり切る覚悟**
3. **他者の視点に立って形而上の自分の考えを具体化して説明する配慮**

問題意識のみでは、行動を起こせない。解消には何が必要か、長中短期的な目標が解決への道筋を照らす。それに対して今やるべきことが今後何に繋がるのか、明確な理由がやり切るモチベーションとなる。そして自分の考えを他者と共有する場合は、相手のバックグラウンドを予測した内容にすることで、忌憚なく実りある議論にすることができる。申請者は、日々この3点を常に意識し研究活動に臨んでいる。

●自己の長所

申請者は能動的に自ら目標達成へ働きかけ、自分で立てた計画通りに遂行できる所と考える。これは高校時代の研究活動から既にこの長所が強く出た。申請者がかつて所属していた広尾学園高校の医進・サイエンスコースは、毎年3月に研究成果報告会を行う。所属大学の前期試験の合格発表後、申請者は2週間で研究成果を出し発表を行うことにした。自ら目的達成に最も適したプログラミング言語の選定から独学で文法を習得し、実装作業からポスターの作成まで期限まで立てたスケジュール通り遂行した。

●自己評価をする上で、特に重要と思われる事項

(受賞歴・取得資格)

申請者はプレゼンテーション能力が高く、早稲田大学本位田研究室との合宿におけるグループ発表にて2季連続最優秀発表賞を受賞している。また、本研究のベースとなるコンピュータ・サイエンス技術を示すため、2018年に基本情報技術者を取得した。

(留学経験)

申請者は中学時代では豪州にて5週間、高校時代はUC Davisにて2.5週間、学部1年時にASUにて4週間の語学留学を行っており、定期的に海外で英語学習を行った。また、学部3年時にインドネシアのバンドン工科大学にて現地大学研究室に滞在しスマートシティ構想に関する研究活動を40日間に渡り行った。ここでも自分でロードマップを作成し成果発表まで英語で活動した。

(特色ある学外活動)

申請者は大学入学直後に**プログラミングの講義がないことに危機感**を覚え、自ら2つの行動を起こした。

1つ目は**大学主催の小〜高校生向けプログラミング教室の立ち上げへの参画及び講師活動**である。実際に教える言語(Python)の習得を目的とした輪講に積極的に参加し、講師として開講から2年弱にわたり毎週子供たちのプログラミング活動のメンタリングを行った。

もう1つは**エンジニア活動の開始**である。高校時代の経験を話し、自分を長期インターンシップとして採用してくれる企業を探すことにした。合計**20社に連絡をとり**、アメリエフ株式会社にて1年半エンジニア活動を行った。また、その後は株式会社フィックスターズにて2.5週、株式会社justInCase Technologiesにて1年半以上にわたって活動するなど、**精力的に産業界でも自らの技術を磨く**ようにしている。

申請者登録名　栁　裕太

DC