時間的特徴を捕捉するための動的グラフを利用した誤情報検出

神田　凌弥†　　杉山　一成†

† 京都大学 情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †kanda@db.soc.i.kyoto-u.ac.jp, ††kaz.sugiyama@i.kyoto-u.ac.jp

あらまし　近年，ソーシャルメディア上における様々なトピックの誤情報が，実社会にも重大な影響を及ぼしている．このような誤情報を検出する研究においては，例えば，テキストや画像を用いた手法や，情報の拡散の様子を捉えた静的グラフを利用した手法などが提案されている．これに対して，本研究では，ツイッター上における誤情報検出のため，リツイートとリプライを別々に処理し，時間的特徴の強いリツイート側には動的グラフを適用するモデルを提案する．リツイートグラフの処理には，構造的特徴の捕捉のための Graph Attention Network と，時間的特徴の捕捉のための Temporal Convolutional Networks を採用し，リプライグラフの処理には，誤情報検出に有用な特徴を見出す手法を提案する．実験の結果，本提案手法がベースラインよりも優れており，提案手法の有効性を示すことができた．

キーワード　誤情報, ソーシャルメディア, 動的グラフ

1 はじめに

インターネットが広く社会に普及している今日，ソーシャルメディア上には膨大な量の情報が存在し，人々の生活に役立つ情報を提供するものもあれば，逆に，実社会に悪影響を及ぼしてしまうものもある．例えば，2020 年 8 月 4 日，大阪府公館での大阪府知事の記者会見で，うがい薬を使用したうがいにより，唾液中のウイルスの陽性頻度が低下した，という研究成果が発表され，うがい薬を使用することが広く推奨された [1] ．このニュースによる影響により，各地の薬局やスーパーではうがい薬の品切れが起きたり [2]，違法な転売がおこなわれた [3].

このような誤情報やフェイクニュースは，世界中で大きな問題となっており，ソーシャルメディアのプラットフォームには，これらを事前に検知して警告を促す機能が実装されているものもある．例えば Twitter では,「ツイートに誤解を招く情報や真偽が問われている情報が含まれており，実害につながりうると判断された場合，背景情報を提供するため，そのコンテンツにラベル付けを行う場合がある」というポリシーが設定されている [4] ．しかし，この警告ラベルは，ユーザーのエンゲージメントには，それほど大きな影響を与えないことを示している研究も存在する [5] ．

そのため，近年では誤情報やフェイクニュース，噂を早期に検出し，その拡散を防ぐための研究が活発に行われている．それらの多くはテキストやグラフの情報を用いる手法であるが，時間情報が考慮されていない．しかし，時間情報は，検出において重要な要素の一つであると考えられる．例えば，Wang ら [6] は，Twitter における投稿を自分のフォロワーに共有する「リツイート」について，真実のニュースに関するツイートのリツイートは時間とともに着実に増加していくのに対し，虚偽のニュースに関するツイートのリツイートは急激に増加した後，短時間で一定になる傾向がある，という分析を行なった．このように，拡散の時間的な特徴を捕捉することで，誤情報を効果的に検出することができると考えられる．

本研究では，ある主張に関するツイートに関して，それらのリツイートとリプライをそれぞれ動的グラフと静的グラフに分けて特徴を抽出し，主張の真偽の判定に用いる，という手法を提案する．動的グラフは，時間ごとのグラフの形状から，その時間変化に関する特徴を捉えることができる．前述したように，Wang ら [6] の研究においては，Twitter における「リツイート」は，情報の拡散における時間的特徴をよく表すと考えられる．一方，投稿に対する返信である「リプライ」は，その返信内容に関する感情や立場など，コンテンツ的な情報がより重要であると考えられるため，静的グラフによって，判別に効果的な特徴を抽出することができると考えた．

2 関連研究

2.1 テキスト情報による誤情報検出

テキスト情報の処理に広く用いられているモデルの一つに，attention 機構というものがある．これは，Bahdanau ら [7] によって提案されたモデルであり，ある文における意味理解に重要な単語に対してより高い重要度を割り当てることができるものである．Yang ら [8] が提案した Hierarchical Attention Networks (HAN) は，attention 機構を文レベル，文書レベル，と階層的に適用することにより，誤情報を検出する．これにより，複数の文にわたる単語の意味理解であったり，文書の意味理解に重要な文の抽出が可能となり，誤情報検出における有効性を示した．また，Shu ら [9] は，ニュース記事の本文を HAN に適用したものと，その記事に対するコメントを attention 機構に適用し，二つの出力を co-attention 機構に適用した，deep hierarchical attention ネットワークモデル，dEFEND (Explainable FakE News Detection) を提案している．このモデルは，検出だけでなく，なぜその判断に至ったのかを説明する，説明可能性も備えたモデルである．

一方で，最近の誤情報やフェイクニュースには，限りなく真の情報に近いような書き方が意図的にされているので，テキス

トコンテンツのみで検出することは，困難になってきている．そのため，テキスト以外の情報を利用した手法が提案されてきた．

2.2 テキストと他の情報を利用した誤情報検出

Cui ら [10] は，ニュース記事に対するユーザのコメントからユーザの感情を分析し，ニュース記事内の画像，ニュース記事本文，著者やトピックやキーワードなどのニュース記事のプロフィール情報と組み合わせてフェイクニュースを検出する，SAME (Sentiment-Aware Multi-Modal Embedding for Detectiong Fake News) という手法を提案している．このモデルでは，ユーザーの感情情報の疎な性質と画像情報の密な性質による組み合わせの困難性に対処するため，それぞれに異なるネットワークを適用させ，異なるモダリティの表現に一貫性を強制的に持たせられるように，敵対的ネットワークを適用している．

また，Shang ら [11] は，ニュース記事におけるテキストと画像のような，異なるモダリティの関連性を考慮して真偽の判定を行うために，元のニュースから得られた特徴と，テキスト情報をもとに生成された視覚的特徴，画像情報から生成されたテキスト特徴を用いて真偽を判定する DGExpain というモデルを提案した．

これらの手法は，テキストや画像といった情報を巧みに活用できてはいるが，2.1 節の問題と同様，真の情報と酷似した偽の情報が伝播するようになる可能性が高い．特に，画像に関しては近年の AI による生成技術の発展が著しく，人々が見て瞬時に偽であると判断できる可能性は低くなりつつある．そこで，拡散の様子などのソーシャルメディア特有の特徴を活用することが重要であり，この点に着目した研究も行われている [12] [13].

2.3 静的グラフを利用した誤情報検出

Nguyen ら [14] は，ニュース記事，出版社，ユーザーをそれぞれノードとしてそれらを発行，フォロー，引用，スタンスの 4 種類のエッジで結んだ異種グラフを形成し，GraphSage [15] を用いたグラフ表現学習を利用したモデルによって真偽を判定する，FANG (FActual News Graph) という手法を提案した．この研究は，ユーザーが自分の意見を強化するために，自分と同じ意見の主体と相互作用する，エコーチャンバー [16] や，ユーザーの認知の時間変化にも注目している．

また，Yuan ら [17] が提案した SMAN (Structure-aware Multi-head Attention Network) は，出版社，ニュース，ユーザーの関係がある異種グラフにおいて，出版社とユーザーが提供する情報に対する信頼度を弱いラベルとして活用し，フェイクニュースを早期発見するためのモデルである．出版社，ユーザーの信頼度予測のための表現学習には，自然言語処理における文書の意味表現学習に優れた能力を示すマルチヘッドアテンション機構を採用している．

上述した手法においては，時間変化に注目しているものも多いが，いくつか問題点もある．まず，リツイートとリプライの両方が考慮されている場合，それらを同種のノードやエッジでとらえているものが，ほとんどである．リツイートは，少ない手順で衝動的に情報を拡散できるという性質上，時間変化のような時系列的側面が情報として強いのに対し，リプライは投稿元の意見の吟味や自分の意見との比較，そしてその言語化といった複雑な思考プロセスと計画性が必要である．それらは同時に考慮すべきでなく，分けるべきであるとと考えられる．また，静的グラフで誤情報検出を行うアルゴリズムの多くは，既に大規模なネットワークが構築されていることを前提としている場合が多く，リアルタイム性に欠ける．一方，動的グラフは，誤情報検出において，全体的なグラフが構築される前段階の情報を活用でき，リアルタイム性の向上に有効であると考えられる．第 3 章では，これらの点を考慮した手法を提案する．

3 提案手法

本研究において，我々が提案する手法の概略図を図 1 に示す．ある主張に関連するツイートとそのリツイート，リプライに関して，それぞれ動的グラフと静的グラフを形成する．

3.1 問題定義

ある主張 C と，それに関連する一連のツイート $Tw = Tw_1, Tw_2, ..., Tw_n$ に対して，各ツイート Tw_i の

$$\text{リツイート } RT_i = Rt_i^1, Rt_i^2, ...,$$
$$\text{リプライ } RPi = Rp_i^1, Rp_i^2, ...,$$

が与えられるとする．ここで，Rt_i^j は時間情報を保持している．この時，主張が真であるか偽であるかのラベル

$$y = \{0, 1\}$$

を予測する．

3.2 リツイート情報の処理

リツイートの情報に関して，動的グラフの生成には，Song ら [18] の手法をベースにする．

時刻 $t(1 \leqq t \leqq T)$ における，主張，関連ツイート，リツイートからなるグラフを

$$G(t) = V(t), E(t)$$

とする．ここで，$V(t)$，$E(t)$ はそれぞれ時刻 t におけるグラフのノードとエッジの集合を表す．また，$N(t) = |V(t)|$，この時の重みなし隣接行列を $A(t)$ とし，

$$A(t) = a_{ij}(t)_{N(t) \times N(t)}$$

とする．$a_ij(t)$ は，ノード v_i と v_j が隣接している場合は 1 を，そうでない場合は 0 をとる．また，時刻 t におけるノード v_i の特徴表現を x_i^t と表す．

次に，与えられたノード表現を用いて構造的特徴を捉えるために GAT (Graph ATtention Networks) [19] を用いる．これは，エッジの表現を作成・学習するのではなく，エッジを単純に attention の重みで表現することで，計算速度をより速く行なうためのモデルである．l 層目におけるノード v_i の特徴量を

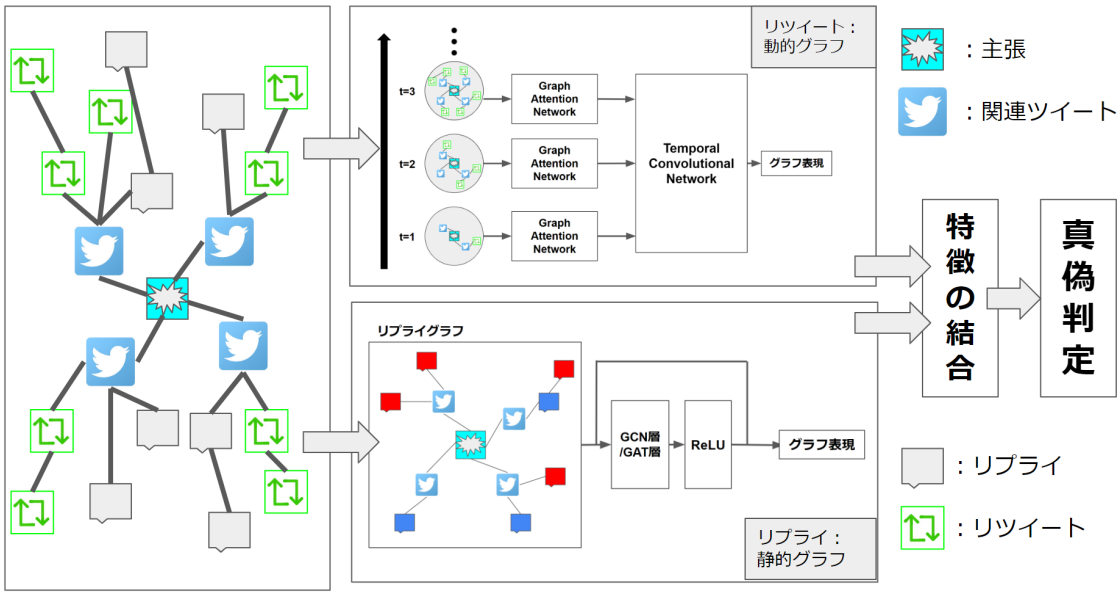

図 1　提案手法のモデル図

h_i^l とする．特徴量は以下の式に基づいて，更新される．

$$h_i^{(l+1)} = \sigma(\sum_{j=\tilde{N}_i} \alpha_{ij}^l h_j^l W^l) \in R^{d^{(l+1)}}$$

ここで，$\tilde{N}_i$ はノード v_i の近傍，α_{ij}^l は l 層目における v_i から v_j への attention 値，h_j^l は $l+1$ 層目から出力された，ノード v_j の特徴表現，$W^l \in R^{d^l \times d^{(l+1)}}$ は学習する重み行列，$\sigma(\cdot)$ は活性化関数を，それぞれ表す．　ここで，l 層目における v_i から v_j への attention 値は，以下のように計算される．

$$\alpha_{ij}^l = \frac{exp(LeakyReLU(\alpha^T[h_i^l W^i \parallel h_j^l W^l]))}{\sum_{k\in\tilde{N}_i} exp(LeakyReLU(\alpha^T[h_i^l W^i \parallel h_j^l W^l]))}$$

ここで，α は重みベクトルの役割を果たす 1 層のニューラルネットワークであり，$\parallel$ は変数の結合 (concatenation) を表す．これにより，$1 \leqq t \leqq T$ におけるこの段階におけるノード v_i の特徴表現は

$$H_i^{(l+1)} = [h_i^{(l+1)}(1), h_i^{(l+1)}(2), ..., h_i^{(l+1)}(T)]$$

と表すことができる．

次に，構造から得られた特徴を入力として，時間変化による特徴を捉えるために，TCN (Temporal Convolutional Networks) [20] を用いる．これは，時系列データに CNN (Convolutional Neural Network) を適用したモデルである．特徴として，TCN の層の数に対して指数関数的に受容野を増加させる dilation 畳み込みと，過去の情報のみによって現在の出力がなされることを保証する causal 畳み込みがある．本モデルは，複数の TCN ブロックが積み重なって構成されている．$r+1$ 番目の TCN ブロックの TCN 層は，以下のように表される．

$$F^{r+1}(U_i^r(t)) = (U_i^r *_{d_{(r+1)}} f)(t)$$

$$= \sum_{z-1}^{r} f_j^T U_i^r(t - d_{(r+1)}j)$$

$$U_i^{r+1} = [..., F^{r+1}(U_i^r(t)), ..., F^{r+1}(U_i^r(T))]$$

ここで，$f \in R^{z\times d^{(l+1)}}$ はサイズ z のフィルター，d_{r+1} は指数関数的に増加する受容野を得るために $(z-1)^r$ に設定される dilation 係数である．フィルター，フィルターのサイズ，TCN ブロックの数を調節することで，このモジュールは柔軟な受容野を獲得することができ，時間情報を十分に探索することができる．　最終的に，こちらからの出力は，時間特徴表現のノード間の平均をとることで得られ，

$$S = \frac{\sum_{i=1}^{N(T)} U_i^{r+1}(T)}{N(T)}$$

となる．

3.3　リプライ情報の処理

リプライのグラフに関する処理は，まずユーザのリプライ同士の意見の立ち位置，すなわち，スタンスの分類を決定するところから始める．

本研究では，Weinzierl ら [21] の手法に基づいたスタンス分類を行なう．この手法は，ユーザーの主張に対する態度は一貫性を持つ，すなわち，ある主張に対して受け入れる意見を持つユーザーは全員同じく同意するはずであり，逆に拒否する考えを持つユーザーは全員同じく否定するはずである，という仮定に基づく．[21] ではこの態度の一貫性のことを AC (Attitude Consistency) と呼ぶ．さらに，スタンスが既知のツイート群で構成される誤情報に関する知識グラフを SMKG (Stance Misinformation Knowledge Graph)，スタンスが不明なツイート群を，TUSM (Tweets with Unknown Stance towards Misinformation) とし，これに含まれるツイート t_1，t_2 のスタンス s_1，s_2 が明らかになるパターンを示した：

- ある主張に対して t_1，t_2 が同じ立場 (賛成か反対か) をとり，かつ t_1 と t_2 の片方がもう片方に賛成している場合
- ある主張に対して，t_1，t_2 が異なる立場をとっており，かつ t_1 と t_2 の片方がもう片方に反対している場合

例えば，ある主張に対し t_1 が賛成，t_2 が反対し，かつ t_1 が t_2 に反対している場合，それぞれのスタンス s_1，s_2 は (Accept, Reject) とわかる．この賛成 (agree)・反対 (disagree) はグラフにおけるエッジとして表現される．[21] ではこれを RTAC (Relation Type that preserves AC) と呼び，$RTAC(s_1, s_2) \in agree, disagree$ である．このときノードはあるスタンス，すなわち主張を受け入れる (accept) か，拒否する (reject) か，を持つツイートを表す．そして，二つのツイート，より具体的にはリプライについて，ノードにスタンスを割り当てた際の，その確かさの信頼度である尤度を考える．このスタンスに関する信頼度の尤度を表すスコアを ACS (Attitude Consistency Score) と呼ぶ．

それぞれスタンス s_1，s_2 をもつ二つのツイート t_1，t_2 について，その AC は保持されるべきである．その関係に関する埋め込みを RE (Relation Embeddings) と表す．この値は，二つのツイート関係 (agree，disagree) によって出力が異なり，それぞれ meagree，medisagree と表される．この埋め込みと，二つのツイートのテキストの表現埋め込み te_1，te_2，そしてスコア関数 f を用いて，二つのツイート間の知識埋め込み表現を以下のように表す．

$$f(te_1, RE(s_1, s_2), te_2)$$

これを用いて，スタンスが既存のツイートから l 連鎖目のツイートの ACS を計算できる．

ツイート t_x に対してスタンス s_x を仮定した際に，$l = 1$ のときの ACS は

$$ACS^1(t_x, s_x) = \sum_{(ty,sy) \in SMKG} \frac{f(te_x, RE(s_x, s_y), te_y)}{|SMKG|}$$

とされ，$l > 1$ の際は，SV={agree, disagree} として，

$$ACS^l(t_x, s_x) = \sum_{t_z \in SMKG, t_z \neq t_x} \sum_{s_z \in agree, disagree} \frac{ACS^{l-1}(t_z, s_z) + f(te_x, RE(s_x, s_z), te_z)}{|TUSM| - 1}$$

と表すことができる．すなわち，$l > 1$ では，そのツイート以前の鎖の ACS の値も考慮して ACS が計算される．これを最大長 L まで考慮する．最終的に，L までの全関係の鎖の可能性を平均化し，仮定したスタンスでの ACS が次式で計算される．

$$ACS^*(t_x, s_x) = \frac{1}{L} \sum_{l=1}^{L} ACS^l(t_x, s_x)$$

その後，尤度の高い方のスタンスが，ツイート t_x のスタンスの予測値として出力される．

$$s_x = softmax_{s_k \in SV} ACS^*(t_x, s_k)$$

このスタンス値を特徴の一つとしてリプライグラフを形成するノード特徴の一部とし，それらをグラフの処理を行うモデルに入力し，リプライグラフ側の出力を決定する．

ここでは，通常の GCN (Graph Convolutional Network) [22] と GAT を比較し，より優れた精度を得た方を適用する．

GCN について，入力する特徴量を X，隣接行列を $\mathbf{A}$，$\mathbf{A}+\mathbf{I}$ の次数行列を $\mathbf{D}$ とし，GCN の k 層目の出力を $\mathbf{H}^k$ とすると，

$$H^0 = \mathbf{X}$$

$$H^{k+1} = f(H^k, A) = f(D^{-\frac{1}{2}}(A + I)D^{-\frac{1}{2}}H^k W_k)$$

ただし，f は活性化関数，W_k は学習されるパラメータである．

何層かの GCN / GAT を経て得られた最終出力を $\mathbf{H}$ とする．

3.4 真偽判定

リツイートのグラフから抽出された情報とリプライのグラフから抽出された情報を結合し，ソフトマックスを計算することで真偽を判定する．

$$\hat{y} = softmax(\sigma(W(S \parallel H) + b))$$

$\parallel$ は特徴の結合を表す．学習の際は，次の目的関数を最小化することで，パラメータを更新させていく．

$$L(\Theta) = -ylog(\hat{y}_0) + (1 - y)log(\hat{y}_1)$$

ただし，$\hat{y} = [\hat{y}_0, \hat{y}_1]$ であり，正解のラベル y が 0 のとき偽，1 のとき真である．また，Θ は学習させるパラメータである．

4 実　　験

4.1 データセット

本研究では，MuMiN (A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset) [23] というデータセットを用いる．このデータセットは，41 種類の言語から収集され，ファクトチェックサイトで真偽が確認された最大で 1 万 3 千件の主張を中心に，それらに関する最大 2,100 万件のツイート，そのツイートへのリプライ，リツイートしたユーザー，ユーザー同士のフォロー関係などの情報を含む．これらは Claim，Tweet，Image，などの 7 種類の異なるノード群から形成される．それらのノードは，disucuss，follows，posted，などの 16 種類のエッジで接続されている．これらのうち，本研究で主に用いるデータは，Claim，Tweet，User，Reply の 4 種類のノードと，discuss，follows，reply_to，posted，retweeted の 5 種類のエッジを用いる．図 2 に，これらの関係を示すとともに，以下において詳細を説明する．

ノ　ー　ド

- **Claim**: 真偽のラベルを持つ主張に関するノード．主張に関するキーワードや情報が発信された日付などを有する．
- **Tweet**: 主張に関連のあるツイートに関するノード．ツイート ID や作成日，リツイート数，リプライ数などを有する．
- **User**: ツイートなどをしたユーザーに関するノード．ユーザー ID やプロフィール文，フォロワー数，フォロー数，ツイート数などを有する．
- **Reply**: あるツイートに対するリプライに関するノード．ツイート ID や作成日，リツイート数，リプライ数などを有する．

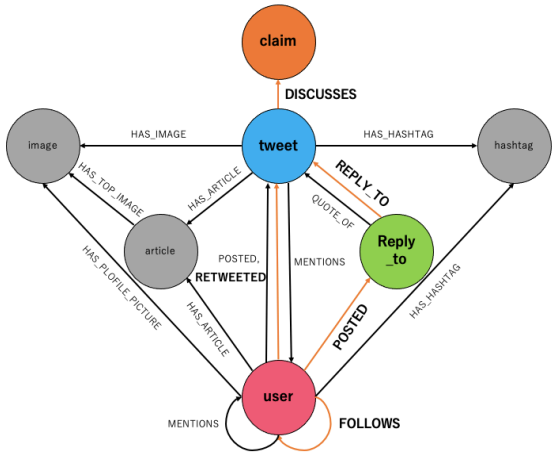

図 2　実験で用いる MuMiN データセットにおけるノードとエッジの関係図

表 1　用いるデータセットの各種の数

	Misinformation	Factual
主張	2,007	158
ツイート	3,690	290
リツイート	105,717	7503
リプライ	174,204	

エッジ

- **discuss**: Claim と Tweet の間のエッジ．そのツイートがその主張に関してのものであるという関係．
- **follows**: User と User の間のエッジ．あるユーザーが別のユーザーをフォローしているという関係．
- **reply_to**: Tweet と Reply の間のエッジ．そのリプライがそのツイートに対してのものであるという関係．
- **posted**: Tweet と User，および Reply と User の間のエッジ．あるツイートやリプライがあるユーザーによって投稿されたものであるという関係．
- **retweeted**: Tweet と User の間のエッジ．あるツイートがあるユーザーによってリツイートされたという関係．

また，リツイートに関して，データセット上のリツイートの情報のみでは，「リツイートのリツイート」，すなわち，あるツイートをあるユーザーがリツイートし，本ツイートを見ていないがリツイートを見たユーザーがリツイートする，という状況を捕捉できないと考えたので，Twitter API を用いて追加でリツイートに関して情報を収集した．しかし，収集した情報のみで「リツイートのリツイート」を考慮するのは困難であるため，次のように仮定する．

ユーザー A のあるツイートに対し，B，C の ユーザーがリツイートしたとき，C のリツイートが B のリツイートよりも後で，B-C 間でフォロー関係が存在し，A-C 間でフォロー関係が存在しないとき，「リツイートのリツイート」である．

この仮定を導入することにより，2 ホップまでのリツイートを捉えることができる．

実験に使用するデータセットに関する各種の数を，表 1 に示す．

4.2　評 価 手 法

評価手法としては，モデルの分別性能を測る F1 を採用する．Misinformation と Factual の比率について，表 1 から Misinformation の割合が高いことがわかり，単純な精度 (accuracy) で評価するのは不適切であると考えた．また，この F1 についても，Misinformation に対しての F1，Factual に対しての F1，その平均の Macro-F1 を採用することで，どの程度偏ったデータセットに対して公平に識別できるかを確認する．データセット MuMiN は，既に学習用，検証用，テスト用に分けられているため，それを採用する．各比率は 8：1：1 である．各手法に対して，3 回実験を行った際の平均を結果として採用する．また，早期発見性を検証するため，ツイートから 8 時間以内のリツイート・リプライを用いて同様に検証する．

4.3　ベースライン

本実験では，以下の手法をベースラインとして比較する．

- **GCN** [22]: グラフ分類器を, [22] に基づいて作成する．
- **GAT** [19]: グラフ分類器を, [19] に基づいて作成する．
- **HeteroGraphSAGE** [23]: この手法は、[23] における分類実験で最高スコアが得られた手法である．MuMiN データセットにおける異種グラフを，Hamilton ら [15] が提案した GraphSAGE に適用し，グラフ分類を行う．
- **PPC** [24]: Liu ら [24] によって提案された，Propagation Pass Classification (PPC) という，Gated Reccurent Unit (GRU) と CNN ベースの手法． この研究に従い，ツイートしたユーザと，そのツイートをリツイートしたユーザの特徴を利用する．また，隠れ層やチャネル数，採用するリツイート数などは，論文に記載の数値で実験を行う．ただし, [24] では，特徴の一つとしてユーザの年齢を用いていたが，MuMiN データセットには記載がないので使用しないものとする．早期検出性の実験において比較する．

4.4　実 験 結 果

本研究では，動的グラフにおけるスナップショットを得るための時間間隔を，時間制限がない場合には，各時間におけるリツイート数を均すことを考え「30 分，1 時間，2 時間，4 時間，8 時間，12 時間，24 時間，1 週間」，早期検出実験の場合では，時間間隔と 8 時間という時間制限を考慮して「1 時間，2 時間，3 時間，4 時間，5 時間，6 時間，7 時間，8 時間」とし，実験を行った．

4.4.1　時間制限なしでの実験結果

まず，時間制限なしの実験で得られた結果を，表 2 に示す．

表 2 から，本論文で提案した手法は，既存の検出手法の精度をおおよそ上回ることが示された．リプライのグラフの扱いに関して，GCN を用いたモデルより，GAT を用いたモデルの方が良い結果を得られたため，以降「提案手法」は，GAT を用いたモデルを指すこととする．特に，Misinformation F1 と Macro F1 という，誤情報検出に欠かせない 2 項目において既存の結果を上回ったことは，この提案手法の有効性を示すことができたと考えられる．

ここで，提案手法の結果が既存のベースラインの結果を上

表 2　時間制限なしでの実験結果

手法	Misinformation F1	Factual F1	Macro F1
Random	0.6466	0.1075	0.3770
GCN [22]	0.8777	0.0452	0.4614
GAT [19]	0.9393	0.1331	0.5361
HeteroGraph SAGE [23]	0.9448	**0.1596**	0.5522
提案手法 (GCN)	0.9426	0.1142	0.5284
提案手法 (GAT)	**0.9538**	0.1585	**0.5561**

回った要因について考察する．提案手法では正しく誤情報と判別できたものの，HeteroGraphSAGE では誤って真の情報だと判断してしまった主張とそのツイート，さらにそのツイートに対するリプライ・リツイートの例を挙げる．

ある誤情報のニュース記事の主張として，イタリアの民主党議員のグループが，イタリアの全国の学校教育のカリキュラムに,「ベッラ・チャオ (日本語訳：さらば恋人よ)」 という，国際的に有名な反ファシスト党の自由とレジスタンスの賛美歌を組み込み,「第二の国歌」のようにしようとしている [25]，というものを挙げる．また，このニュース記事を引用しているツイートは，このニュース記事のタイトルをそのまま載せている内容である．このツイートに対して,「そうなってくれたら嬉しい.」などといった肯定的なリプライが寄せられていた．

リツイートの時間変化に注目してみると，全 10 件のリツイートのうち，7 件はツイートから 2 時間以内に行われているものであり，さらに，うち 4 件はツイートからわずか 1 時間以内に行われているものであった．このような，ツイートが投稿されてから早い段階で急激にリツイートが行われて，時間が経過するにつれて勢いが衰える，というリツイートの時間変化は，誤情報に関するツイートによく見られる傾向である．この時間変化を動的グラフを利用することで特徴としてうまく捕捉でき，提案手法で正しく分類できたものと考えられる．また，リプライとしても，ツイート (で引用されているニュース記事の内容) に肯定的な意見もあり，そのフラグが評価の一因になったと考えられる．一方，HeteroGraphSAGE の手法では，あくまで主張・ツイート・リプライ・ユーザの関係を，ある一点のスナップショットで捉えただけであり，このような時間変化を補足するのは不可能である．

このツイートをしたユーザは，データセットが作成された当時で 20 万人を超えるフォロワーを有しており，影響力もかなり大きいと考えられる．このようなユーザによる誤情報の拡散につながるツイート及び主張の検出ができたことは，本提案手法が有効であることを顕著に示している．

一方で，HeteroGraphSAGE では正しく誤情報と判別できたものの，提案手法では正しく誤情報と判別できなかった事例も存在する．

ある誤情報のニュース記事の主張は，2021 年 4 月に，バイデン現大統領の議会演説がアメリカ本土で放送され，推定 2,690 万人の視聴者が試聴したが，これが 2017 年 2 月にトランプ元大統領が記録した，推定 4,300 万人を大きく下回り，最近の視聴率の歴史の中で最も低いものの一つである [26] ，というものであった．このニュース記事に対して 2 つのツイートが関連づけられており，一つは同年のアカデミー賞の視聴者数と比較して (恐らく) 批判しているもの，もう一つはニュース記事内に用いられた数字を用いてどの程度，トランプ元大統領の時と比較して視聴者が減ったかを強調しつつ，バイデン大統領を批判しているものである．

この主張において提案手法が正しく判別できなかった要因が，それぞれのツイートに対して存在すると考えられる．まず，前者のツイートは，リツイートは誤情報のツイートに見られる，早期に多くリツイートがなされ，後半は緩やかにされる，という傾向が見られた．具体的には，全 5 件のリツイートのうち，4 件は 1 時間以内に行われていた．一方，リプライの方は，絵文字のものであったり，“Who?” というような，スタンスとして意味をなさないようなものがなされており，特徴としての利用が困難であったと考えられる．後者のツイートは，リプライでは，比較に疑問視をするものなどのバイデン大統領を擁護するものや，一方で「(私は) 見なかった」というような批判に同意するものもあり，特徴としては，誤情報のものとして有用だと思われる．しかし，リツイートでは，最初に早い段階で 2 リツイートがあったものの，時間を空けて 2 リツイート，さらに時間をあけて 2 リツイート，というように,「急激にリツイートが行われて，時間が経過するにつれ勢いが衰える」というには少し難しい状況であると考えられるものであった．むしろ，真の情報のリツイートの傾向である「一定のペースでリツイートがなされる」という状況に近いといえる．この 2 つのツイートの，相反するリプライ・リツイートの傾向によって，うまく誤情報と判別できなかったのではないかと考えられる．

一方で，HeteroGraohSAGE の手法でうまく検出できた要因としては，リプライしたユーザ，リツイートしたユーザのそれぞれについて，近傍の特徴から自身の特徴を決定する，という GraphSAGE の特徴の結果だと考える．

GraphSAGE では，各ノードの特徴を学習していくのではなく，近傍のノードから特徴を集約するための関数を学習するので，提案手法に比べ，リツイートしたユーザ・リプライしたユーザ同士の特徴の近さに，より注目できたのではないかと考える．

さらに，隣接ノードのみでなく，近傍のノードから特徴を集約する，という点においても，よりリツイートしたユーザ同士，リプライをしたユーザ同士の関係性や性質の近さを考慮できるものと思われる．

以上のことから，今後，考慮すべき点として，

- リツイートの時間変化のみに依存しすぎないような処理や特徴の利用，
- 構造と隣接のユーザ特徴を利用するだけでなく，離れたユーザからの特徴をうまくできるような仕組み，

の 2 点を挙げることができる．

4.4.2　時間制限ありでの実験結果

表 3 に，8 時間以内のリツイート・リプライについて検出を行った際の各手法の結果を示す．

この結果から，HeteroGraphSAGE と 提案手法は，時間制限なしの場合に比べてさらに良い結果となり，その中でも提案

表 3　時間制限ありでの実験結果

手法	Misinformation F1	Factual F1	Macro F1
GCN [22]	∗(0.9712)	0.000	0.4856
PPC [24]	∗(0.9712)	0.000	0.4856
GAT [19]	0.9288	0.1382	0.5334
HeteroGraph SAGE [23]	**0.9622**	0.1640	0.5631
提案手法	0.9455	**0.2070**	**0.5763**

手法は，Misinformation と Factual の両方を考慮した Macro F1 のスコアで，ベースラインの結果を大きく上回る結果となった．GCN や PPC は，予測が全て誤情報，となってしまった結果，Misinformation F1 が高いにもかかわらず，Factual F1 が 0 となり，平均である Macro F1 が低いような結果となっている．一方で，提案手法の Misinformation F1 に関して，時間制限なしの場合と比べてスコアが減少している上，ベースラインである HeteroGraphSAGE のスコアよりも劣っている．

これらの結果から，

(1)　なぜ，時間制限なしの場合と比べてより良い結果を得られたのか，

(2)　なぜ，時間制限なしの場合と比べて，HeteroGraphSAGE と比較した場合に，Misinformayion F1 スコアと Factual F1 スコアの優劣が逆転したのか，

という点について，考察することが重要であると考えられる．

まず (1) についてだが，これはリツイートやリプライの誤情報と真の情報における違いが，投稿から 8 時間という時間の間において顕著に現れたためであると考える．

これまでに述べた通り，リツイートは投稿されてから早い段階で多く行われ，短時間で一定の値となる，という傾向や，真の情報は一定のペースでリツイートが増えていく，という傾向が，1 日や 1 週間 といった長いスパンに比べて，より明確に現れたと考えられる．また，リプライに関しては，「誤情報に対するリプライ数が真の情報に比べて少ない傾向にある」との報告 [6] もあり，その傾向や構造上の特性を考慮して精度のある検出につながったのではないかと考えられる．

一方で，ツイートは基本的にそのツイートをしたユーザをフォローしているユーザにしか表示されないため，8 時間という短い時間の中では，リプライをしたりリツイートするユーザのほとんどが，ツイートしたユーザのフォロワーであり，そのユーザ同士の特徴が近くなる可能性もあるとも考察できる．そのため，HeteroGraphSAGE においても，時間制限なしの場合に比べて高い精度が得られたと考えられる．

次に (2) について，まず 提案手法 では Misinformation F1 スコアが HeteroGraphSAGE を上回った要因として，時間制限によるリツイートの「停滞」 が判断できなかった場合があったのではないかと考える．ここで，時間制限がある場合で，HeteroGraphSAGE で誤情報を正しく判断でき，提案手法で誤情報と判別できなかった例を示す．

ある誤情報のニュース記事の主張は，アメリカ大統領選挙の際に，バージニア州のフェアファックス群において，票を集計する人の一人が，期日前投票で集計した票数のうち 100,000 票をバイデン氏の方に「誤って」与えてしまったが，すぐに修正された，というものであり，それに対してニュース記事の出版社によるツイッターアカウントからの，ニュース記事の概要とバイデン氏の批判する皮肉の内容の投稿が行われた．表 4 に，そのツイートの 1 時間ごとのリツイート数を示す．

この表によれば，最初の 2 時間で多くリツイートが行われているものの，8 時間が経過するまで，ある程度継続的にリツイートが行われている．しかし，8 時間以降のリツイート数を見てみると，このツイートのリツイートは，投稿から 11 時間後まではある程度のリツイート数を記録し，そこからリツイートがほとんどなくなっている．すなわち，このツイートについては，リツイート数は制限時間である 8 時間を超えても伸びを続け，11 時間までは続いたが，それ以後の伸びが急激に衰えるという，時間スパンが少し長めの誤情報のリツイートの時系列的特徴を持っている．しかし，本研究では，時間制限を 8 時間に設定したため，その衰えを捉えられず，提案手法では，誤って判断してしまったのではないかと考えられる．

このように，誤情報のツイートのリツイートの時間的変化は必ずしも全て同じような時間間隔で行われるとは限らず，さまざまな幅を持って行われることが観察された．したがって，時間制限でスコアが上昇したものの，全体的に精度が向上したわけではないことには注意が必要である．

一方で，提案手法では Factual F1 スコアが大きく上昇している．この要因として考えられるのは，リツイートの時間変化に加えて，リプライにおけるユーザの傾向との真の情報に対する傾向である．リツイートの時間変化は，先に述べた通り，真の情報の場合一定のペースでリツイートが増えていく傾向があるので，それを動的グラフで持って時系列的な情報を捉えて，うまく TCN で畳み込めたのではないかと考えられる．リプライに関しては，フェイクニュースの，ソーシャルメディア上伝播に関する分析を行った研究 [27] において，偽の情報は真の情報に比べ，ツイートの投稿から最初のリプライが届くまでの時間が短い傾向にあり，またリプライ数も有意に異なると報告されている．この傾向は，8 時間という短い時間の中で，リプライに関して真の情報と偽の情報を見分けられる指標の一つとして考えることができる．また，時間制限なしの場合に比べて，提案手法と HeteroGraphSAGE の二つの手法でスコアが上昇したことからも，有効な指標であると考えられる．

5　ま　と　め

本研究では，ソーシャルメディア上における誤情報を検出するための手法として，時間的特徴を捉えるためにリツイートの拡散の様子を動的グラフとしてとらえ，リプライグラフと合わせて誤情報を検出する手法を提案した．実験の結果，提案手法が，特に誤情報の検知や，誤情報と真の情報の検知の両方を加味した結果において優れた性能を発揮し，有効であることを示した．また，誤情報の早期検出に関して，時間制限がない場合に比べて Macro F1 スコアが上昇し，動的グラフの利用が早期検出において，より有効な手法であることを示すことができた．

本研究で提案したモデルでは，対象ツイートのリツイートと

表 4　ニュース記事の概要とバイデン氏の批判する皮肉の内容の投稿に対して行われたリツイートの時間ごとの数

経過時間	1 時間	2 時間	3 時間	4 時間	5 時間	6 時間	7 時間	8 時間	9 時間	10 時間	11 時間	12 時間	13 時間	14 時間
RT 数	14	20	7	7	3	4	6	7	1	10	6	0	2	0

リプライについて、ある程度の間隔のスナップショットを取得した上で真偽を判断した．しかし，スナップショットごとに誤情報である可能性を出力し，時間の経過とともにそのスナップショットを取得して，その可能性を更新する仕組みを動的グラフに適用すれば，時系列情報を活用しながらも，ソーシャルメディアのユーザには，より適合した情報を提供できるものと考えられる．この点に取り組むことは、今後の課題としたい．

文　　献

[1] 朝日新聞デジタル (8 月 4 日): 「うがい薬で唾液中のコロナウイルス減少」吉村知事会見.

[2] 毎日新聞 (8 月 5 日): 都内でも「うがい薬は売り切れ」の張り紙　一夜にして店頭から消えた　「吉村知事会見」を考える.

[3] 毎日新聞 (8 月 5 日): うがい薬転売横行　違法なヨード系はサイト削除、代わりに“透明”が高値.

[4] Twitter 社ヘルプセンター：「twitter 上の警告とその意味」:https://help.twitter.com/ja/rules-and-policies/notices-on-twitter.

[5] Orestis Papakyriakopoulos and Ellen Goodman. The impact of twitter labels on misinformation spread and user engagement: Lessons from trump's election tweets. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*, p. 2541–2551, 2022.

[6] Suhang Wang Dongwon Lee Kai Shu, Deepak Mahudeswaran and Huan Liu. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media, 2020.

[7] Yoshua Bengio Dzmitry Bahdanau, Kyunghyun Cho. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR '15)*, 2015.

[8] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical Attention Networks for Document Classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '16)*, pp. 1480–1489, 2016.

[9] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. dEFEND: Explainable Fake News Detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*, pp. 395—405, 2019.

[10] Limeng Cui, Suhang Wang, and Dongwon Lee. SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '19)*, pp. 41—48, 2019.

[11] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A Duo-Generative Approach to Explainable Multimodal COVID-19 Misinformation Detection. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*, pp. 3623–3631, 2022.

[12] Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. Polarization and Fake News: Early Warning of Potential Misinformation Targets. *ACM Transactions on the Web (TWEB)*, Vol. 13, No. 2, pp. 10:1–10:22, 2019.

[13] Abishai Joy, Anu Shrestha, and Francesca Spezzano. Are you influenced? modeling the diffusion of fake news in social media. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '21)*, pp. 184—-188, 2021.

[14] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*.

[15] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 2017.

[16] Kathleen Hall Jamieson and Joseph N. Cappella. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press, 2008.

[17] Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users based on Weakly Supervised Learning. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING '20)*, pp. 5444–5454, 2020.

[18] Chenguang Song, Yiyang Teng, and Bin Wu. Dynamic Graph Neural Network for Fake News Detection. In *Proceedings of the IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS '21)*, pp. 27–31, 2021.

[19] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR '18)*, 2018.

[20] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, 2018.

[21] Maxwell Weinzierl and Sanda Harabagiu. Identifying the Adoption or Rejection of Misinformation Targeting COVID-19 Vaccines in Twitter Discourse. In *Proceedings of the ACM Web Conference (WWW '22)*, pp. 3196–3205, 2022.

[22] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, 2017.

[23] Dan S. Nielsen and Ryan McConville. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*, pp. 3141–3153, 2022.

[24] Yang Liu and Yi-Fang Wu. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 354–361, 2018.

[25] il giornale (2020 年 9 月 25 日)：　“il pd vuple “bella ciao” a scuola “cantatela con l'inno di mameli” ”.

[26] Deadline (2021 年 4 月 29 日)：　“joe biden ' s address to congress snares 26.9m viewers; way down from trump as abc tops broadcast and cable”.

[27] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical Propagation Networks for Fake News Detection: Investigation and Exploitation. In *Proceedings of the 16th International AAAI Conference on Web and Social Media (ICWSM '20)*, pp. 626–637, 2020.