E-049

深層学習を用いた Twitter 上のツイートに対するファクトチェック支援手法の提案

Fact-Checking Support Methods for tweets on Twitter using Deep Learning

竹井　拓実† 清　雄一† 田原　康之† 大須賀　昭彦†
Takumi Takei Yuichi Sei Yasuyuki Tahara Akihiko Ohsuga

1. はじめに

今日多くの人々が SNS を使い，情報の送受信を簡単に行うことが出来るようになった一方で，SNS の投稿の中には事実と異なった投稿や真偽が疑わしい投稿も数多く存在する．このような情報はデマやうわさと呼ばれ，誤情報伝播や炎上，煽動といった悪影響を及ぼすことがある．これらの情報の真偽については，日本ではファクトチェックイニシアティブ（FIJ）[1]を始めとするファクトチェックサイトによるファクトチェックによって確認される一方で，ファクトチェック作業は人手によるチェック作業が主となっており，情報拡散のスピードに追い付いておらず，中々進んでいないという現状がある．また，自然言語処理モデル BERT は双方向 Transformer ベースの事前学習と転移学習の 2 つの学習を必要とするモデルであり，出力層を追加することで，様々な言語処理タスクに対して当時最高の結果を達成しており[2]，出力の最終層の 1 つ前の層から文章の埋め込み表現を獲得出来ることが知られている．

本研究では FIJ にて，すでにファクトチェックされた Twitter 上のツイートに対して，リプライや引用といった伝播情報を用いるような特徴量抽出や，我々の先行研究[3]では用いていない事前学習済み自然言語処理モデル BERT から得られる埋め込み情報を用いる文章埋め込み表現への変換を用い，「虚偽または誤りと判別されたツイート」であるか，「そうではないツイート」かの 2 値分類を深層学習分類器によって実行した．

2. アプローチ

2.1 本研究の対象データについて

FIJ では真偽判定について，9 つの分類が行われている．このうち，判定結果が保留されている「判定留保」と「検証対象外」を除く 7 つのラベルに対して FIJ では更に「正確・ほぼ正確」・「ミスリード・不正確・根拠不明」・「虚偽・誤り」の 3 つのグループ分けが行われている．FIJ において「正確・ほぼ正確」と判定されているツイートは少量であるため，これら 3 つのグループを分類する際には「正確・ほぼ正確」のツイートが正確に分類されづらくなることが予想される．「正確・ほぼ正確」を除いた 2 つのグループのツイートに対して行う分類は 3 つのグループのツイートが存在する以上，本研究の目的であるファクトチェック支援手法の提案において適切ではないと考えられる．故に本研究では，既存の 3 つのグループ分けから「虚偽・誤り」と判断されたツイートに着目し「虚偽・誤り」と判断されたツイートか，そうでないツイートかを判別する 2 値分類の実験を行うこととした．

† 電気通信大学　大学院情報理工学研究科
University of Electro-Communications
Graduate School of Informatics and Engineering

2.2 全体アプローチ

本手法の全体アプローチに関しては以下（1）～（3）に示す通りである．

（1） FIJ および Twitter 上において検索を行い， ファクトチェック済みツイート及び伝播情報を収集した．

（2） 収集したツイート及び伝播情報に対して，ソースツイートに対する伝播情報に限定し，後述する特徴量の抽出を行い，データセットを作成した．

（3） 作成したデータセットを，Graph Convolutional Network（GCNConv）を用いた深層学習分類器に代入し，正答率（Acc），適合率（Pre），再現率（Rec），F 値の評価指標を得た．

3. データセット

3.1 クリーニング処理

本研究におけるクリーニング処理として以下表 1 にあるような 12 種類の処理を実行した．

表１： クリーニング処理一覧

英数字のみ半角へ変更	URL をカウント, 削除
空白, 改行, !, ?を削除	英字を小文字に変更
@をカウント, 削除	数字を 0 に変更
引用文の削除	文字化け文字列の削除
連続文字（――や ww 等）の削除	絵文字・特殊文字・日付の削除
「」, 『』, 【】で囲まれた文字列を削除	ハッシュタグをカウント, 削除

3.2 特徴量抽出

本研究では Twitter から収集したデータから抽出した，ユーザーのフォロー数，フォロワー数，ユーザーの総ツイート数といったユーザーの特徴やツイートの発信媒体やメディアを含むかどうかといったツイートの特徴を合わせた 22 種類のデータから成る「基本特徴量」とツイート本文を埋め込み表現に直すことによって得られる「テキスト特徴量」の 2 つを抽出し，結合することで特徴量として用いた．テキスト特徴量の抽出には自然言語処理モデル BERT を用い，事前学習済みモデルとして東北大学のモデル[4]と京都大学のモデル[5]を用いて，クリーニング処理を行ったツイート本文を 768 次元の文章埋め込み表現に直した. これらを結合し，本研究では特徴量として用いた．以下，東北大学の事前学習済みモデルを用いてテキスト特徴量を算出したデータセットを Bert，京都大学の事前学習済みモデルを用いてテキスト特徴量を算出したデータセットを Bert-Juman と表記する．

3.3 データセットの内訳

本研究で用いたデータセットに関する詳細な内訳は以下表 2 の通りである．

表 2 ： データセットの内訳

ソースツイート数	100
イベント数	100
正確・ほぼ正確 ソースツイート数	4
誤り・虚偽 ソースツイート数	50
不正確・根拠不明・ミスリード ソースツイート数	46
正確・ほぼ正確　全ツイート数	1,736
誤り・虚偽　全ツイート数	12,989
不正確・根拠不明・ミスリード 全ツイート数	7,761

4. 実験

4.1 実験設定

4 にて作成したデータセットを時系列順に直し，ソースツイートのみ(Source)，全体の 1/6（Little），全体の 1/2（Half），全体（All）を抽出した各データセットを作成した．それらに対して以下図 1, 2 に示すような GCNConv を用いた簡単な分類器を設定し，10 点交差検証にて正答率（Acc），適合率（Pre），再現率（Rec），F 値を算出した．深層学習分類器のロス関数は CrossEntropyLoss，オプティマイザは Adam，エポック数は 50 として実験を行った．図中の $dim_{features}$ は特徴量の総次元数である 790，$dim_{classes}$ は分類クラス数 2 である．

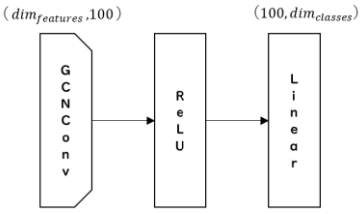

図 1 ： 1 層 GCN モデルの概略図

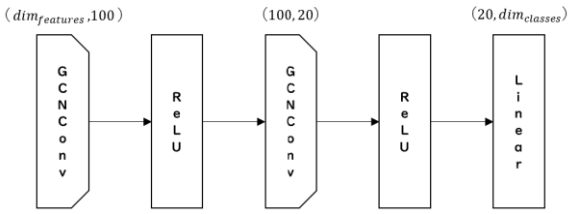

図 2 ： 2 層 GCN モデルの概略図

4.2 実験結果

実験結果は，以下表 3～表 6 のようになった．

表 3 ： データセット Bert の結果（図 1 モデル）

	Acc ↑	Pre ↑	Rec ↑	F 値 ↑
Source	0.470	0.376	0.523	0.410
Little	0.550	0.466	0.556	0.479
Half	0.570	0.517	0.564	0.524
All	0.570	0.521	0.573	0.516

表 4 ： データセット Bert-Juman の結果（図 1 モデル）

	Acc ↑	Pre ↑	Rec ↑	F 値 ↑
Source	0.560	0.572	0.555	0.554
Little	0.580	0.532	0.575	0.540
Half	0.530	0.535	0.535	0.526
All	0.590	0.564	0.564	0.575

表 5 ： データセット Bert の結果（図 2 モデル）

	Acc ↑	Pre ↑	Rec ↑	F 値 ↑
Source	0.530	0.358	0.527	0.404
Little	0.530	0.407	0.527	0.432
Half	0.540	0.412	0.545	0.442
All	0.556	0.563	0.536	0.541

表 6 ： データセット Bert-Juman の結果（図 2 モデル）

	Acc ↑	Pre ↑	Rec ↑	F 値 ↑
Source	0.540	0.531	0.531	0.519
Little	0.540	0.546	0.562	0.546
Half	0.590	0.598	0.609	0.596
All	0.590	0.557	0.560	0.540

5. 考察

実験結果について，多少ばらつきが出たもののおおよその場合においてデータセット内の伝播情報が多い方が良いスコアを出していることが見て取れる．また，Bert-Juman の方が全体的に見て Bert よりも良いスコアを算出した．値のばらつきに関しては交差検証を繰り返し行い，平均値を取ることで解消できると考えられる．また精度が 6 割に達しておらず，手法としては不十分であると思われる一方で，BERT による埋め込み表現を最適な方法で次元圧縮すること，ソースツイート以外の伝播情報を利用した深層学習分類器を構築することに加えて，感情情報の利用や他の事前学習済みモデルを使用した埋め込み表現の獲得が精度向上方法として考えられる．

6. おわりに

今後の展望として，より精度を向上させるためのモデル構築やソースツイート以外の伝播情報の利用，より効果的な次元圧縮手法の思案，実用に向けたファクトチェックシステムの構築が挙げられる．

謝辞

本研究は JSPS 科研費 JP21H03496，JP22K12157 の助成を受けたものです．

参考文献

[1] ファクトチェックイニシアティブ（FIJ） https://fij.info/
[2] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding” 2018. https://arxiv.org/abs/1810.04805
[3] 竹井拓実, 清雄一, 田原康之, 大須賀昭彦, “伝播情報を加味した機械学習による Twitter 上のウワサ判別手法の検討”,電子情報通信学会技術研究報告, 信学技報, 121.298, pp.29-34, 2021
[4] 東北大学 BERT 事前学習済みモデル https://github.com/cl-tohoku/bert- japanese
[5] 柴田知秀, 河原大輔, 黒橋禎夫, “BERT による日本語構文解析の精度向上”,言語処理学会第 25 回年次大会, pp205-208, 2019