E-002

BERT モデルを用いてフェイクニュース検出精度の改善について
Improving Detection Accuracy of Fake News Based on BERT Model

呉 雪瑩 † 仲 思源 † 藤野 巖 ‡
Xueying Wu Siyuan Zhong Iwao Fujino

1. はじめに

近年、インターネットやスマートフォンの普及により、ソーシャルメディアやニュースサイトなどから、情報を即座に発信できる環境が整備されている。一方で、悪意をもって作成された虚偽の情報、いわゆるフェイクニュースが急速に拡散し、社会的混乱や誤情報の拡散を引き起こす事例が増加している。特に政治や医療、災害といった重要分野においては、フェイクニュースが現実の行動に深刻な影響を及ぼす可能性があり、その早期検出が強く求められている。

フェイクニュースの検出においては、従来ニュース本文に加え、SNS 上での拡散パターンやユーザーの反応などを用いるアプローチが広く研究されてきた。これらは確かに有効であるが、ソーシャルメディア上のデータを取得・整備するには時間とコストを要し、速報性が重要なフェイクニュース対策においては必ずしも実用的とはいえない。そこで本研究では、ニュースコンテンツ（記事のタイトルおよび本文）のみを用いて、機械学習によりフェイクニュースを早期に自動検出する手法を検討する。

本研究では、既存のフェイクニュース判定用データセットを用い、BERT モデルをベースとした分類器の構築を行い、2 分類実験と 4 分類実験を実施した。モデルのパフォーマンス向上を目的に、学習率やバッチサイズなどのハイパーパラメーターを調整し、複数回の学習と評価を通じて精度および F1 スコアを向上させた。これにより、従来研究と比較して高い分類性能を達成し、ソーシャルコンテクストなしでも十分な精度でフェイクニュースの自動判定が可能であることを示す。

2. 関連研究

2.1 フェイクニュース検出に関する関連研究

フェイクニュース検出に関する先行研究では、ニュースコンテンツおよびソーシャルコンテクストの両方を特徴量として用いる手法が多く提案されている。Zhou ら［1］は、フェイクニュース検出に関する理論的背景を整理し、複数の観点から検出手法の分類と課題を提示している。また、Raza ら［2］は、Transformer ベースの BERT モデルを用いて、ニュース本文とソーシャルメディアから得られるユーザー情報（投稿文やプロフィール等）を統合した検出モデルを提案し、両者の組み合わせによる精度向上を示した。

これらの研究に共通する特徴は、拡散後のソーシャルメディア上の情報を活用することで検出性能を高めている点にある。しかし、ソーシャルコンテクストの収集には時間を要するため、速報性が求められる初期段階での対応には限界がある。

谷ら［3］は、ニュースの発行直後と拡散後の 2 段階（STEP1・STEP2）に分けた検出手法を提案し、STEP1 ではニュース本文のみを用いて早期判定を行い、STEP 2 でソーシャルコンテクストを追加する構成としている。実験では、5 分後や 10 分後といった初期の段階で高精度な予測が可能であることが示された。

一方、本研究では、ソーシャルコンテクストを一切用いず、ニュースコンテンツ（タイトルと本文）のみを特徴量として用いる。これにより、データ収集の即時性とモデル運用の簡易性を確保しつつ、高い分類性能を実現することを目的とする。また、BERT モデルを用いた深層学習ベースのアプローチにより、文脈理解に優れた特徴抽出を行い、従来手法と比較しても遜色のない、もしくはそれを上回る性能の実現を目指す。

2.2 BERT

BERT (Bidirectional Encoder Representations from Transformers）［4］は、事前学習済みの Transformer ベースモデル［5］であり、双方向の文脈情報を同時に捉えることで、自然言語処理タスクにおいて高い性能を発揮する。BERT はエンコーダ構造を持ち、入力された文中の単語の一部をランダムにマスクし、そのマスクされた部分を予測するという「マスク言語モデル（Masked Language Modeling）」により事前学習を行う。これにより、文全体の意味理解に基づいた深い表現獲得が可能となる。

BERT モデルでは、入力にニュース本文を用い、BERT の出力から得られる[CLS]トークンの最終層出力を用いて分類処理を行う構成を採用している。中間層に全結合層および活性化関数 GELU、Dropout を組み合わせ、過学習の抑制と性能向上を図っている。これにより、BERT はニュースコンテンツのみからでも高い判別性能を発揮できる。

3. データセット

本研究では、FakeNewsNet［6］から公開されている 2 つの主要データソースである Politifact および Gossipcop を対象とし、フェイクニュースと信頼できるニュースの分類タスクに用いるデータセットを構築した。各カテゴリの元データ件数は以下の表 1 の通りである。

表 1　収集したニュース記事の統計情報

データセット	Politifact		Gossipcop	
ラベル	真	偽	真	偽
ニュース（件）	17313	1590	333777	105016

FakeNewsNet から提供されるデータには、ニュースタイトルや URL、ソーシャルメディア関連情報などが含まれるが、本研究ではニュースコンテンツの本文のみを用いることに

† 東海大学大学院情報通信学研究科修士課程
‡ 東海大学情報通信学部情報通信学科

焦点を当てている。したがって、各ニュースに付随する URL を取得し、ウェブ上に実在する元記事の本文情報をスクレイピングによって取得・整形した。

4. 提案手法

本研究では、ニュースコンテンツの本文情報のみを入力とし、事前学習済み BERT モデルをベースとした深層学習分類モデルを構築することにより、フェイクニュース検出精度の改善を試みる。本章では、使用データ、前処理、モデル構成、学習および評価方法について説明する。

4.1 データセットと前処理

本研究では、既存のフェイクニュース検出研究において広く用いられている Politifact および Gossipcop の 2 種類の公開データセットを用いた。各データセットには、信頼できるニュース（Real）および虚偽ニュース（Fake）がラベル付きで収録されている。データの取得後、クラス不均衡の影響を緩和するため、次のような処理を施した。

Politifact_Fake：少数クラスであるため、10 倍に拡張（データ拡張）

Gossipcop_Real：多数クラスであるため、ランダムサンプリングにより 3,300 件に削減

この処理により、学習データにおける各クラスのバランスを調整し、モデルの学習が一部クラスに過剰適応することを防いだ。

4.2 学習設定

学習率は 1.0e-6、エポック数は 6 とした。損失関数にはクロスエントロピー損失を用いている。学習中は GPU 環境（CUDA）を利用し、高速化を図った。また、エポックごとにモデルのパラメーターを保存し、途中からの学習再開（Checkpoint 機能）を可能とすることで、効率的な実験運用が可能な設計とした。

4.3 評価方法

テストデータを用いた評価では、Accuracy（正解率）、Precision（適合率）、Recall（再現率）、F1 スコアを算出し、モデルの性能を定量的に評価した。また、分類レポートを出力することで、Fake / Real 各クラスにおけるバランスや偏りの有無を確認した。

5. 実験と結果

本実験に使った BERT モデルのパラメーターを表 2 に示す。前述のデータセットについて、各種機械学習モデルと BERT モデルを用いて 2 分類と 4 分類の実験結果をそれぞれ表 3 と表 4 に示す。

表 2　BERT モデルのパラメーター

最大シーケンス長	512
学習率	1e-6
エポック数	6
バッチサイズ（学習時）	8
バッチサイズ（推論時）	8

表 3　Real と Fake の 2 分類モデルごとの正解率

モデル	正解率
Naïve Bayes	0.79
XGboost	0.85
SVM	0.86
BERT	0.93

表 4　Politifact_Fake, Politifact_Real, Gossipcop_Fake, Gossipcop_Real の 4 分類モデルの正解率

モデル	正解率
Naïve Bayes	0.82
XGboost	0.86
SVM	0.85
BERT	0.87

谷ら［3］の研究では 2 分類の分析した精度が 89%でしたので、本研究 2 分類の結果では 4%向上し、全体で 93%に達した。データセットを均等に処理して繰り返しパラメーター調整した結果精度が上がった。また、4 分類の実験ではクラス数が増えたため、精度がいくらか下がったが、全体で 87%に達した。

6. まとめと今後の課題

本研究では、ニュース本文の情報のみを用いてフェイクニュースを分類する BERT ベースのモデルを構築し、従来手法と比較して 4%の改善が確認され、高い精度を達成した。一方で、本手法にはいくつかの課題が残されている。

第一に、学習に用いるデータ量が大きく、モデルの訓練には長時間を要した。特に、事前処理および BERT の文脈理解に基づく特徴抽出には計算コストがかかるため、今後は学習の高速化や軽量化に向けた手法の導入を検討する必要がある。

第二に、現段階では本モデルを Python 環境で評価するにとどまっており、社会実装の観点では十分とは言えない。

今後は、本研究で得られたモデルを活用して、ユーザーが任意のニュースをアップロードし、その信憑性を即時判定できるアプリケーションの開発を視野に入れている。これにより、フェイクニュースの早期検出と誤情報拡散の抑制に資する、より実用的なツールの提供が可能になると期待される。

参考文献

[1] Xinyi Zhou, Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 2020.

[2] Shaina Raza, Chen Ding. Fake news detection based on news content and social contexts: a transformer-based approach. *Int. J. Data Sci. Anal.*, 2022.

[3] 谷聡馬, 佐々木裕多, 張建偉. ニュースコンテンツとソーシャルコンテクストを用いたフェイクニュースの早期自動検出. *DEIM Forum 2023*, 1a-8.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of NAACL-HLT*, 2019.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. *Attention is All You Need*. In *Proceedings of NeurIPS*, 2017.

[6] FakeNewsNet : https://github.com/KaiDMML/FakeNewsNet