

言語的特徴に依存しない汎用的なフェイクニュース検知手法の実現に向けて

小野原 覚^{1,a)} 三村 守¹

概要：2016 年の米国大統領選挙を契機にフェイクニュースの拡散が社会問題となり、フェイクニュースを検知することは喫緊の課題となった。フェイクニュースを検知するため、言語的特徴を用いた機械学習モデルによる検知手法が提案されている。しかし、これらの手法は特定のデータセットのみで精度が評価されており、言語的特徴に依存しているため汎用性に課題がある。そこで本研究では、言語的特徴に加え、主張の根拠、情報源の信頼性、扇情的な表現、個人へのレッテル付け等、10 種類の特徴の有無を大規模言語モデルに評価させ、それらの特徴を用いた言語的特徴に依存しない検知モデルの実現を試みた。未知のデータセットを用いて言語的特徴だけを使用したモデルとの検知精度を比較したところ、提案モデルの汎用性が向上していることを確認した。一方で、フェイクニュース検知に貢献した特徴を分析したところ、データセット毎にばらつきがあり、検知精度の向上には検証の余地があることが判明した。

キーワード：フェイクニュース検知、大規模言語モデル、特徴量、汎用性

Towards the Realization of a Generic Fake News Detection Method Independent of Linguistic Features

SATORU ONOHARA^{1,a)} MIMURA MAMORU¹

Abstract: The spread of fake news became a social issue following the 2016 U.S. presidential election, making the detection of fake news an urgent task. To address this, detection methods using machine learning models based on linguistic features have been proposed. However, these methods have been evaluated only on specific datasets, and their reliance on linguistic features poses challenges for generalizability. In this study, we aimed to develop a detection model that does not rely solely on linguistic features. In addition to linguistic features, we evaluated the presence of ten other characteristics using a large language model—such as the basis of claims, source credibility, sensational expressions, and labeling of individuals. We then incorporated these features into the detection model. When comparing the accuracy of this proposed model with one that uses only linguistic features on an unseen dataset, we confirmed that the generalizability of our model had improved. On the other hand, an analysis of the features that contributed to fake news detection revealed variability across datasets, indicating that further investigation is needed to improve detection accuracy.

Keywords: fake news detection, large language model, features, generality

1. はじめに

2016 年の米国大統領選挙において、フェイクニュースがソーシャルメディア上で大規模に拡散し、主要メディア

による記事を上回る注目を集め、大きな社会問題となったことを契機に「フェイクニュース」という言葉が注目され始めた。フェイクニュースの定義は、様々であるが、嘘やデマ、陰謀論やプロパガンダ、誤情報といったインターネット上を拡散して現実世界に負の影響をもたらす現象として一括りにされている [1]。フェイクニュースの拡散は現代社会における深刻な問題であり、その検知は喫緊の課題

¹ 防衛大学校研究科
National Defence Academy of Japan
a) em64006@nda.ac.jp

である。そこで、フェイクニュースの社会への影響を抑えるべく、様々な手法でフェイクニュースを検知する研究がされてきた。特に機械学習モデルを用いたフェイクニュース検知は主要な研究分野であり、自然言語処理技術や深層学習を用いたモデルが提案されている[2][3]。

先行研究には、精度の高いモデルが多く提案されているが[4][5]、それらの研究の大半は、特定のデータセットのみで精度が評価されている。また、政治、健康、スポーツ等を複合させたデータセットを作成し、検知モデルの検証を行っているものの[2]、検知モデルの汎用性を検証した提案は少ない。フェイクニュース検知の手法として、言語的特徴を利用した研究がされているが、フェイクニュースに共通する言語的な特徴は少なく、検知モデルの汎用性には改善の余地があることが報告されている[5][6]。

そこで本研究では、言語的特徴に加え、主張の根拠、情報源の信頼性、扇情的な表現、個人へのレッテル付け等、10種類の特徴の有無を大規模言語モデル(LLM)に評価させ、それらの特徴を用いた言語的特徴だけに依存しない検知モデルの実現を試みる。作成したモデルが未知のデータセットに対し、どの程度フェイクニュースを検知できるかを確認し、提案モデルの汎用性について評価する。また、フェイクニュースを検知するために、どのような特徴が貢献したかを検証する。本研究の研究課題(RQ)の設定は次のとおりである。

- (RQ1) フェイクニュースから抽出した言語的特徴だけに依存しない検知モデルに汎用性はあるか？
- (RQ2) フェイクニュースを検知するために有効な特徴はなにか？
- (RQ3) フェイクニュース検知に有効な特徴は異なるデータセットで共通しているか？

本研究では、これらの課題を解決するため、異なる4つのデータセットを用意して提案モデルの検知精度を評価した。本研究の主な貢献は次のとおりである。

- (1) 提案モデルは従来モデルと比較して、4つのうち3つのデータセットで高い検知精度を示し汎用性が高かった。一方、特定のデータセットをほとんど検知できなかつたことから改善の余地がある。
- (2) 10種類の特徴のうち、情報源の信頼性、主張の裏付け、扇情的な表現の3つの特徴は、いずれのデータセットでも高い貢献度を示した。
- (3) いずれのモデルでも検知に有効な特徴があった。一方、データセット毎に有効な特徴はばらつきがあり、検知精度の向上には検証の余地がある。

2. 関連研究

ソーシャルメディアの普及により、フェイクニュースに限らず、真偽不明の情報が簡単に拡散されている。真偽不明の情報は膨大であり生成される速度は増大していること

から、その真偽を個人が検証することは非常に困難である。そのため機械学習技術を用いて、フェイクニュースを検知する手法が数多く提案されている。

Ahmed らは、テキストベースのデータセットに対し、TFとTF-IDFを使って特徴を抽出し、線形ベースの分類器を用いたフェイクニュースの検知モデルを提案した[4]。Sastrawan らは Word2Vec, GloVe, fastText 等の単語埋め込みを用いて特徴を抽出し、CNN-RNN ベースの検知モデルを提案した[7]。これらの研究では ISOT FAKENEWS と呼ばれるフェイクニュース分類の代表的なデータセットを使用し Accuracy が 0.9 を超えている。Garg らは TF-IDF 等3つの手法を用いて単語や特殊文字の数、可読性等の特徴を抽出し、機械学習モデルでフェイクニュースの検知を行う手法を提案した[8]。これらはいずれの研究も、それぞれ高い検知精度を示したもの、各データセットは独立に検証されており、汎用性についての言及がなかった。Hakak らはニュース記事から単語数や文長等の統計的特徴と固有表現の26種類の特徴を抽出し、ランダムフォレスト等の機械学習モデルに入力する手法を提案している[5]。この手法は特定のデータセットでは高精度な検知精度を示したが、別のデータセットでは検知精度が低下しており、汎用性がないという課題を残している。

近年では、フェイクニュース検知に LLM を使った手法も提案されている。Xu らは BERT モデルを活用してニュースの書式(長さ、大文字の頻度)や内容から特徴を抽出して検知精度を向上させた[9][10][11]。しかし、いずれも複数のデータセットで検知精度を向上させているが、訓練データとテストデータに同じデータセットを使用しており、別のデータセットに対する検証は実施していない。石丸らは、3つの特徴の異なるデータセットを BERT を用いて分類し、各データセットの訓練モデルが別のデータセットの分類に対しても有効であるかを検証した[6]。しかし、フェイクニュースに共通する言語的特徴は少なく、フェイクニュース検知モデルは、訓練データの特徴に依存し、未知のデータに対する汎用性については課題があると結論付けた。

そこで、本研究では特徴の異なるデータセットから言語的特徴だけに依存しない検知モデルの実現を試みるとともに、未知のデータセットの検知に対しても有効であるかを検証し、提案するモデルの汎用性の有無を確認する。

3. 関連技術

本研究では、LLM の1つである Llama3 を、ニュース記事の特徴量を抽出するツールとして活用する。このアプローチの技術的な基盤となるのが Transformer モデルと、その核となる自己注意機構である。このセクションでは、本研究で使用する技術の詳細について述べる。

3.1 Transformer モデル

Transformer は、Google の研究者によって開発されたニューラルネットワーク技術であり [12]、特に自然言語処理の分野において広く利用されている。これは従来の再帰型ニューラルネットワーク (RNN) や畳み込みニューラルネットワーク (CNN) が抱えていた、長距離の文脈依存関係を捉えることの困難さや、計算の並列化が難しいという課題を解決した。Transformer は図 1 のようなエンコーダ層とデコーダ層を持ち、自己注意機構を複数回、並列に実行するマルチヘッド注意機構を備えている。これにより文中の各単語が、他の単語との程度関連しているかを学習するとともに、複数の異なる視点から単語間の関係性を同時に分析することで、より豊かな文脈理解を可能にしている。

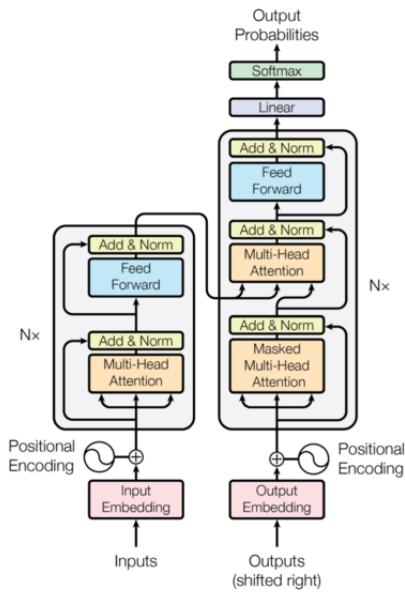


図 1 transformer モデルの構造

3.2 Llama

Llama は meta 社が開発したオープンソースの LLM であり、transformer モデルを基盤としている。本研究において使用した Llama3-8B-instruct モデルは、80 億のパラメータで学習した基本モデルに加えて、様々な指示や質問に対して適切に応答できるよう追加学習されている。本モデルは、より大規模なモデルと比較して少ないメモリと計算能力で動作する利点を持ち、かつ対話形式のユースケースで優れたパフォーマンスを発揮するとともに、テキスト生成に特化した利点を持つ。この特徴を使って、単に単語や文法といった表層的な言語的特徴を捉えるだけでなく、ニュース記事の論理的な一貫性、文法的欠陥、論理的矛盾、感情的な誇張、情報源の不明確さといった、より抽象的で複雑な概念を評価できる能力が期待できる。本研究では

Llama3-8B-instruct を特徴抽出器として用いることで、従来のモデルが捉えることができなかったフェイクニュースの特徴を定量的に評価し、汎用的なモデル構築を試みる。

4. 提案手法

4.1 概要

本研究におけるフェイクニュース検知手法の概要を図 2 に示す。提案手法の検知精度検証のため、40,000 以上のニュース記事で事前学習済みの BERT ベースのフェイクニュース検知モデルとの精度を比較する。

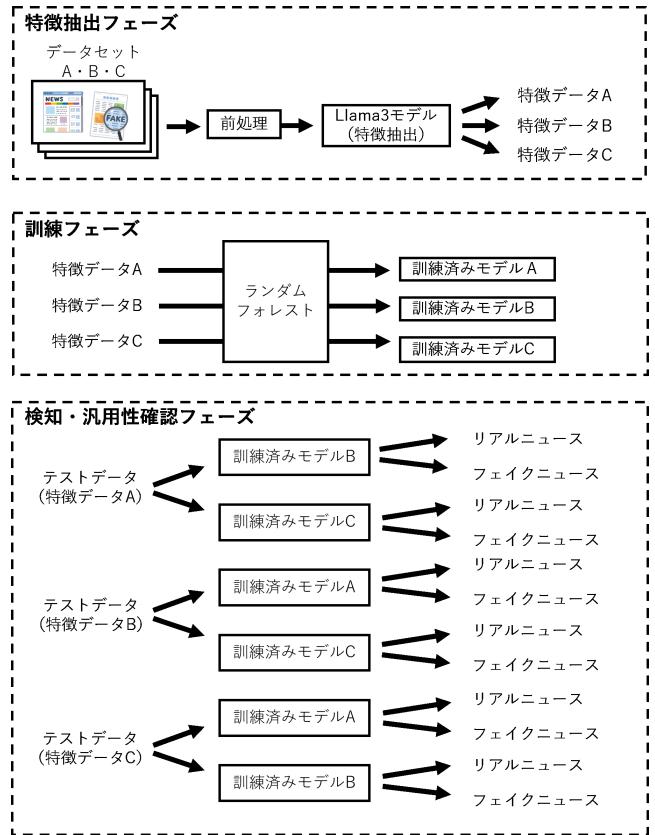


図 2 提案手法の手順

本研究には異なる 4 つのデータセットを使用する。各データセットに対して前処理を行い、Llama3 に各データセットを入力し、特微量を抽出する。抽出した 1 つのデータセットの特微量を用いて、ランダムフォレストを訓練する。訓練が終了したモデルに対して、別のデータセットを用いてフェイクニュース検知をそれぞれ行い、未知のフェイクニュースの検知精度を評価して、提案手法の汎用性を確認する。

以下、各手順の詳細を説明する。

4.2 前処理

前処理では、Llama3 が各ニュースの特微量を抽出するために、データクレンジングを行った。データクレンジングは、各データセットの完全性を確保するため、データ

セット内の空行及び置換文字を削除し、REALあるいは、FAKEのラベル、記事のタイトル及びテキスト以外の項目を削除した。

4.3 特徴抽出フェーズ

従来の機械学習モデルでは、ユニークな単語の数、単語の出現頻度等を使ってフェイクニュース検知を行っていたが、近年では LLM の登場により、高度な自然言語理解能力と推論能力を活用して、深い意味を理解して検知することが可能になった。本研究では、ニュース記事からフェイクニュース特徴の特徴を抽出し、多角的な視点でフェイクニュースを検知する。

まず、表 1 に示すプロンプトを Llama3 に入力し、表 2 に示す 10 種類のフェイクニュースの特徴を選定した。次に、Llama3 にニュース記事のタイトルと本文を与えて、10 種類のフェイクニュースの特徴について、0~100 のスコアを付与するよう表 1 のとおり指示して、定量的に評価した。Llama3 が特徴を定量的に評価できなかった場合は、各特徴量のスコアをデータセット全体の中央値で補完し、モデルの学習に影響を及ぼさないようにした。

表 1 入力プロンプト

目的	入力プロンプト
特徴選定	You are an expert in journalism research, specializing in identifying and analyzing fake news. What other types of information or characteristics are particularly useful in identifying fake news? Please provide specific examples and elaborate.
特徴量の定量的評価	You are an AI assistant specialized in detecting fake news. Your task is to analyze the provided news article's title and text, and evaluate the degree of each of the 10 features defined below. For each feature, assign a **score ranging from 0 (indicating the characteristic is completely absent or strongly indicative of real news) to 100 (indicating the characteristic is extremely prominent and strongly indicative of fake news)**. A higher score means the article exhibits that specific fake news characteristic more strongly.

4.4 訓練フェーズ

各データセットにおいて、Llama3 モデルによって抽出された特徴データを分類器であるランダムフォレストモデルに入力して、訓練済みモデルを作成する。具体的には、Llama3 モデルで抽出した 10 次元の各スコアを特徴ベクトルに変換し、各ニュースの真のラベル（リアルニュース／フェイクニュース）とともにランダムフォレストモデルに入力し、それぞれのデータセットに対する訓練済みモデルを作成する。

表 2 特徴抽出

No.	Features of fakenews
1	Linguistic Flaws (Grammar, Spelling, Awkward Phrasing)
2	Logical Inconsistency (Internal Contradictions, Illogical Flow)
3	Low Source Credibility (Unreliable or Vague Sources)
4	Lack of Supporting Evidence (Insufficient Justification)
5	Emotional/Sensational Language (Inflammatory Tone)
6	Lack of Specificity / Vagueness (Ambiguous Details)
7	Overuse of Rhetorical Questions (Leading Questions)
8	Extreme Language / Overuse of Emphatic Symbols (Exaggerated Phrasing)
9	Contextual Distortion/Misuse (Manipulated Context)
10	Labeling / Ad Hominem Attack (Personal Attacks)

4.5 検知、汎用性確認フェーズ

1 つのテストデータに対して、訓練フェーズで作成した訓練済みモデルを用いてフェイクニュース検知を行う。これを全てのテストデータに対して行い、検知精度をデータセット毎に評価する。また、フェイクニュース検知のために事前学習されたモデルである BERT ベースの BERT モデル [10] [13] と RoBERTa モデル [11] [14] の検知精度を提案モデルの検知精度を比較することで、提案モデルの汎用性を確認する。

4.6 各特徴の重要度

本研究では 10 種類のフェイクニュースの特徴を抽出したが、各特徴がどの程度フェイクニュース検知に貢献したかを検証する。検証には置換重要度 (Permutation Importance : PI) を用いた。この手法は学習済みのモデルにおいて、特定の特徴量をランダムにシャッフルした際にモデルの予測性能（本研究では F1 スコア）がどれだけ低下するかを測定することで、各特徴の貢献度を評価することができるため、フェイクニュース検知に必要な汎用的な特徴を評価することが可能である。具体的には、元の検知精度を基準に、各特徴を個別にシャッフルした際の性能低下量を算出して、特徴重要度として評価する。

5. 検証実験

5.1 データセット

本研究では、ジャンルの異なる 4 つのデータセットを使用した。データセットの内訳を表 3 に示す。

ISOT FAKENEWS は、Ahmed らが提案した英語のフェイクニュースに関するデータセットであり、本物のニュースは Reuters.com から、フェイクニュースは Politifact と Wikipedia によってフラグが立てられた信頼性の低いウェブサイトから収集された [4]。記事のタイトル、テキスト、記事の公開日等の情報があり、世界的なニュースおよび政

表 3 データセット内訳

データセット名	リアルニュース	フェイクニュース
ISOT FAKENEWS	21417	23481
COVID-19	5600	5100
McIntire	3171	3164
MisInfoText	19203	18556

治に関するニュースが多数を占めている。

COVID-19 は、Patwa らが提案した新型コロナウイルス感染症に関するフェイクニュースのデータセットで、Facebook, Tewtter および Instagram 等の 10700 件のソーシャルメディアへの投稿および記事から収集されている [15]。本物のニュースは、WHO(World Health Organization) 等の公式アカウント、CDC(Centers for Disease Control and Prevention) 等の医療機関から発信された情報が収集され、フェイクニュースは、新型コロナウイルス感染症に関するソーシャルメディア上の記事の中から、Politifact, Snopes, Boomlive などの有名な事実確認サイトでフェイクニュースと判定された情報が収集されている。

McIntire は、フェイクニュース検知のためのベンチマークデータセットとして知られ、政治や経済を中心とした記事から構築されている [16]。本物のニュースは New York Times, WSJ, Bloomberg, NPR, Guardian 等、フェイクニュースは Kaggle で公開された記事から収集されている。

MisInfoText は、Torabi Asr らがラベルの信頼性がフェイクニュース検知に与える影響を分析し構築したデータセットである [17]。本研究では、Politifact から収集したデータを使用し、2 値分類とするため true, mostly true, half-true のラベルを REAL とし、false, mostly false, pants on fire! のラベルを FAKE とした。

5.2 評価指標

本研究で使用する評価指標について説明する。評価指標を算出するために必要な値を表 4 に示す。評価指標には、Accuracy, Precision, Recall, F-measure の 4 種類を用いた。各評価指標の定義については、式 (1)~(4) に示すとおりである。

表 4 評価指標の算出

		真の結果	
		フェイク	リアル
予測結果	フェイク	True Positive(TP)	False Positive(FP)
	リアル	False Negative(FN)	True Negative(TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F\text{-measure} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

5.3 実験環境

この実験で使用した実験環境及び使用した主なライブラリを表 5 に表す。

表 5 実験環境

CPU	IntelCore i9-14900KF 3.20GHz
GPU	NVIDIA GeForce RTX 4090
Memory	128GB
OS	Windows11 Home
Cuda	11.8
使用言語	Pyhon3.8.9
Transformers	4.46.3
Scikit-learn	1.3.2

5.4 実験

5.4.1 検出精度比較

各データセットにおいて、フェイクニュース検知のために 40,000 以上のニュース記事で事前学習されたモデルである BERT [10] [13], ROBERTa [11] [14] と、提案手法のモデルにおけるフェイクニュース検知精度の比較を行う。データセット別の検知精度比較結果を図 3, 図 4, 図 5 及び図 6 に示す。各グラフの縦軸は分類モデル、横軸は各評価指標の値である。本研究では、検知器の精度を検証することから、Accuracy と F1 値に注目する。

ISOT FAKENEWS をテストデータとした場合、提案モデルは COVID-19 で訓練した場合の Accuracy と F1 は 0.91, McIntire で訓練した場合の Accuracy と F1 は 0.87 と事前学習モデルより精度は劣ったものの、9 割近い検知ができていることを確認した。一方で、MisInfoText で訓練した場合の Accuracy は 0.81, F1 は 0.79 と検知精度が突出して劣ることを確認した。

COVID-19 をテストデータとした場合、提案モデルの Accuracy は約 0.7 と事前学習モデルより高い検知精度であった。F1 値は MisInfoText で訓練した場合を除き、事前学習モデルより提案モデルの汎用性が高いことを確認した。

McIntire をテストデータとした場合、提案モデルの Accuracy はいずれも 0.7 を超え、事前学習モデルより汎用性が高いことを確認した。しかし F1 値に注目すると MisInfoText で訓練した場合は、事前学習モデルに劣ることを確認した。

MisInfoText をテストデータとした場合、提案手法によるモデルの Accuracy は約 0.5, F1 値は約 0.3 であり、ほとんど検知できないことを確認した。

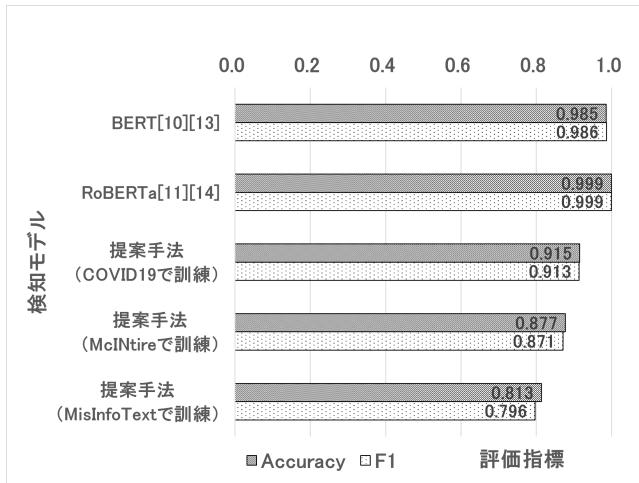


図 3 ISOT FAKENEWS の検出精度比較

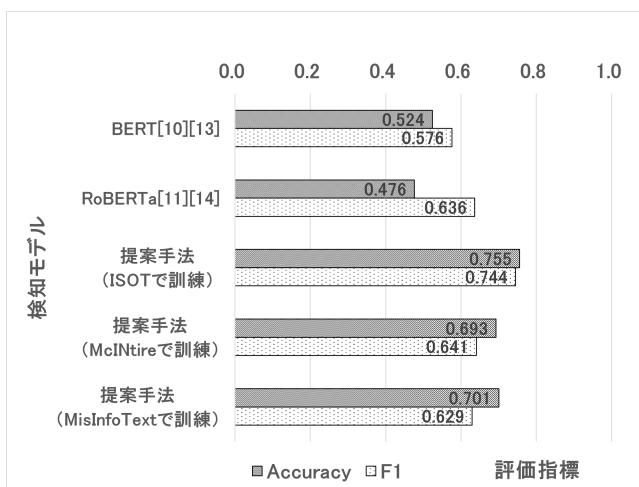


図 4 COVID-19 の検出精度比較

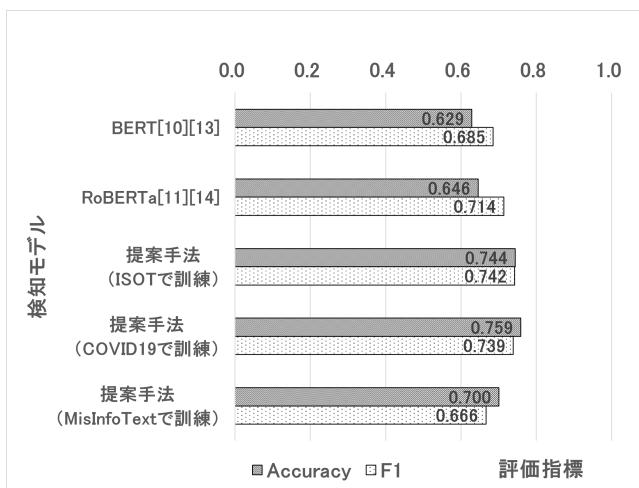


図 5 McIntire の検出精度比較

5.4.2 フェイクニュース検知に重要な特徴

Llama3 に評価させた 10 種類のフェイクニュースの特徴が、フェイクニュース検知にどれだけ貢献したのかを評価した。評価結果を表 6, 表 7, 表 8 及び表 9 に示す。

ISOT FAKENEWS と McIntire で訓練したモデルは、

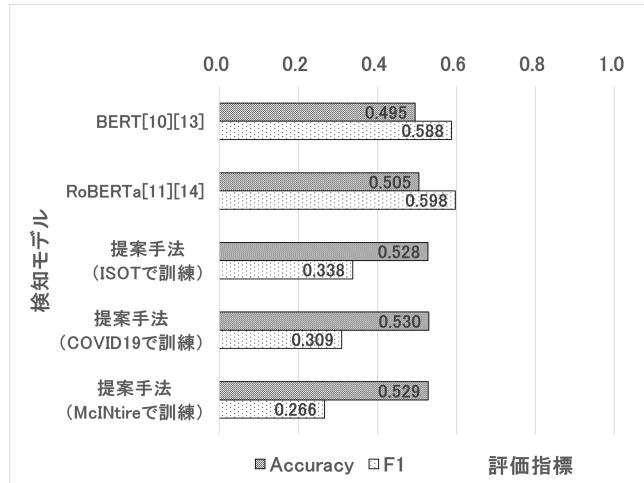


図 6 MisInfoText の検出精度比較

MisInfoText を検知した際を除き, Linguistic Flaws, Logical Inconsistency, Low Source Credibility, Lack of Supporting Evidence の 4 つの特徴の貢献が高かった。MisInfoText を検知した際は、いずれの特徴も貢献度が低いことを確認した。

表 6 ISOT FAKENEWS で学習した場合の特徴貢献度

特徴	COVID19	McINTire	MisInfoText
Linguistic Flaws	0.019	0.033	-0.044
Logical Inconsistency	0.005	0.013	-0.013
Low Source Credibility	0.050	0.019	-0.064
Lack of Supporting Evidence	0.082	0.037	-0.100
Emotional/Sensational Language	0.062	0.041	-0.043
Lack of Specificity / Vagueness	0.005	-0.004	-0.005
Overuse of Rhetorical Questions	0.002	0.018	-0.012
Extreme Language / Overuse of Emphatic Symbols	0.000	-0.004	0.004
Contextual Distortion/Misuse	0.000	0.009	-0.005
Labeling / Ad Hominem Attack	0.000	0.001	0.000

COVID-19 で訓練したモデルは、MisInfoText を検知した際を除き, Low Source Credibility, Lack of Supporting Evidence, Emotional/Sensational Language, Contextual Distortion/Misuse の 4 つの特徴の貢献が高かった。MisInfoText で訓練したモデルは、Low Source Credibility, Lack of Supporting Evidence, Emotional/Sensational Language, Contextual Distortion/Misuse の 4 つの特徴の貢献が高かったものの、その他の特徴も貢献していることを確認した。

表 7 COVID-19 で学習した場合の特徴貢献度

特徴	ISOT	McIntire	MisInfoText
Linguistic Flaws	0.000	0.002	-0.001
Logical Inconsistency	0.028	0.013	-0.021
Low Source Credibility	0.084	0.021	-0.081
Lack of Supporting Evidence	0.126	0.042	-0.111
Emotional/Sensational Language	0.246	0.108	-0.044
Lack of Specificity / Vagueness	0.069	0.022	-0.091
Overuse of Rhetorical Questions	0.012	0.004	-0.006
Extreme Language / Overuse of Emphatic Symbols	0.001	-0.001	-0.002
Contextual Distortion/Misuse	0.118	0.036	-0.064
Labeling / Ad Hominem Attack	-0.003	0.001	0.000

表 9 MisInfoText で学習した場合の特徴貢献度

特徴	ISOT	COVID19	McIntire
Linguistic Flaws	-0.004	-0.001	-0.003
Logical Inconsistency	0.087	0.004	0.021
Low Source Credibility	0.083	0.047	0.025
Lack of Supporting Evidence	0.130	0.038	0.044
Emotional/Sensational Language	0.089	0.054	0.018
Lack of Specificity / Vagueness	-0.042	-0.026	-0.018
Overuse of Rhetorical Questions	0.040	0.001	0.007
Extreme Language / Overuse of Emphatic Symbols	0.067	-0.005	0.020
Contextual Distortion/Misuse	0.132	0.024	0.047
Labeling / Ad Hominem Attack	0.080	0.002	0.024

表 8 McIntire で学習した場合の特徴貢献度

特徴	ISOT	COVID19	MisInfoText
Linguistic Flaws	0.069	0.004	-0.039
Logical Inconsistency	0.005	-0.003	-0.010
Low Source Credibility	0.109	0.052	-0.083
Lack of Supporting Evidence	0.104	0.011	-0.147
Emotional/Sensational Language	0.120	0.026	-0.082
Lack of Specificity / Vagueness	-0.003	-0.061	-0.044
Overuse of Rhetorical Questions	-0.005	0.000	-0.002
Extreme Language / Overuse of Emphatic Symbols	-0.005	-0.010	0.002
Contextual Distortion/Misuse	-0.006	0.000	-0.001
Labeling / Ad Hominem Attack	-0.014	0.000	-0.004

6. 考察

6.1 提案手法の汎用性

実験の結果、提案手法のフェイクニュース検知モデルは、事前学習モデルと比較して ISOT FAKENEWS のデータセットでは検知精度が劣ったものの、他のデータセットでは高い検知精度を示し、汎用性が高いことを示した。ISOT FAKENEWS でも検知精度が劣ったが、9割近い検知精度を保持することができた。一方、MisInfoText データセットでは、ほとんど検知ができていなかった。以上の

ことから、提案した手法は未知のデータセットにも対応可能な汎用性がある一方、特定のデータセットへの汎用性については改善の余地がある。

6.2 Llama3 が評価できなかった記事数

Llama が 3 フェイクニュースの特徴を評価できなかった記事数を表 10 に示す。本研究では Llama3 を特徴抽出器として用いたが、Llama3 は入力する記事のトークン数が最大 8192 トークンまでであり、これを超える長文の場合に特徴を評価できなかったほか、記事の内容に情報が不足している場合等に、Llama3 が特徴の有無を評価できていない可能性がある。Llama3 の後継に Llama3.1 や Llama4 があり、これらは最大処理可能トークンが飛躍的に増加するとともに、高度なタスクを実行できることから、これらの使用により検知精度が向上できる可能性がある。

表 10 Llama3 が評価できなかった記事数

データセット名	総記事数	評価できなかった記事数
ISOT FAKENEWS	44898	2
COVID-19	10700	0
McIntire	6335	20
MisInfoText	37759	1656

6.3 特徴重要度の共通性

各データセットにおける特徴の重要度を確認した結果、10 種の特徴のうち Low Source Credibility, Lack of Supporting Evidence, Emotional/Sensational Language の 3 つの特徴は、いずれのモデルでも高い貢献度を示しており、

一定の共通性があることがわかった。一方、検知に全く貢献していない特徴やデータセット毎に重要な特徴が異なっていたことから、検知精度の向上には特徴を見直す必要性がある。

6.4 研究倫理

本研究で使用した4つのデータセットはすべて公開されている。また、提案モデルの実装に使用したライブラリやLlama3も無償公開されている。したがって、本研究と同様の環境構築は容易であり、本研究の再現性は高いと考える。

6.5 研究の限界

本研究では4つのデータセットを用いて実験を行った。これらはファクトチェック機関によってラベル付けされたものを含むが、フェイクニュースの定義が多様であること、データセットの一部はリアルかフェイクだけではなく、half-trueのような曖昧なラベルも存在し、ラベルの信頼性については検証の余地が残されている。これらの課題に対応するため、今後は複数の異なるデータセットで検証を実施する必要がある。

7. おわりに

本研究では、言語的特徴に加え、主張の根拠、情報源の新体制、扇情的な表現、個人へのレッテル付け等10種類の特徴の有無を大規模言語モデルであるLlama3に評価させ、言語的特徴だけに依存しない検知モデルの実現を試みた。未知のデータセットを用いて言語的特徴だけを使用したモデルとの検知精度を比較したところ、提案モデルの汎用性が一部向上していることを確認した。一方で、フェイクニュースの検知に貢献した特徴を分析したところ、一定の共通性はあるものの、データセット毎にばらつきがあり、検知精度の向上には検証の余地を残した。

今後の課題として、Llama3をさらに高性能にしたLlama4モデルや別のLLMの使用、プロンプトエンジニアリングの改善等による精度の高い特徴抽出、より効果的な特徴の採用等により、更なる汎用性の高いモデルの構築を試みる。

昨今のフェイクニュースでは、テキストベースだけでなく、ディープフェイク画像を活用し人間の目では見分けるものが困難な場合も増加している。大規模言語モデルは自然言語理解能力だけでなく、画像から特徴を抽出することが可能なため、画像とテキストの両方から情報を統合することで、より包括的な分析を可能にする可能性も考えられる。

参考文献

- [1] 総務省. 総務省—令和元年版情報通信白書—フェイクニュースを巡る動向. <https://www.soumu.go.jp/johotsusintoeki/whitepaper/ja/r01/html/nd114400.html>.
- [2] Junaed Younus Khan, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:100032, 2021.
- [3] Nishant Rai, Deepika Kumar, Naman Kaushik, Chandan Raj, and Ahad Ali. Fake news classification using transformer based enhanced lstm and bert. *International Journal of Cognitive Computing in Engineering*, 3:98–105, 2022.
- [4] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques, 2017.
- [5] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Kumar Reddy Madikunta, and Wazir Zada Khan. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 117:47–58, 2021.
- [6] Mamoru Mimura and Takayuki Ishimaru. Analyzing common lexical features of fake news using multi-head attention weights. *Internet of Things*, 28:101409, 2024.
- [7] I. Kadek Sastrawan, I.P.A. Bayupati, and Dewa Made Sri Arsa. Detection of fake news using deep learning cnn-rnn based methods, 2021.
- [8] Sonal Garg and Dilip Kumar Sharma. Linguistic features based framework for automatic fake news detection. *Computers & Industrial Engineering*, 172:108432, 2022.
- [9] Xiaochuan Xu, Zeqiu Xu, Peiyang Yu, and Jian Wang. A Hybrid Attention Framework for Fake News Detection with Large Language Models. *arXiv preprint arXiv:2501.11967*, jan 2025.
- [10] Álvaro Ibrain Rodríguez and Lara Lloret Iglesias. Fake news detection using deep learning. *arXiv preprint arXiv:1910.03496*, 2019.
- [11] T. Poovozhi, Syed Afsar Ahamed, Thota Sankeerth, Suram Suresh Reddy, and Tempalli Ganesh. Fake news detection using bert & roberta. *International Journal For Research in Applied Science and Engineering Technology (IJRASET)*, 13:2875–2880, April 2025.
- [12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [13] Aymene. Fake-news-detection-bert-based-uncased. Hugging Face, n.d.
- [14] jy46604790. Fake-news-bert-detect. Hugging Face, n.d.
- [15] Parth Patwa, Mohit Bhardwaj, Vineeth Guptha, Gitanjali Kumari, Shivam Sharma, Srinivas PYKL, Amitava Das, Asif Ekbal, Shad Akhtar, and Tanmoy Chakraborty. Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts, 2021.
- [16] George McIntire. Fake and real news dataset. *GeorgeMcIntire's Github*, 2018.
- [17] Fatemeh Torabi Asr and Maite Taboada. Big data and quality data for fake news and misinformation detection. *Big Data & Society*, 6(1):1–14, 2019.