日本データベース学会 データドリブンスタディーズ
Vol. 5, Article No. 1, 2027 年 3 月

自動ファクトチェック手法のための日本語データセット JAD-AFC

佐々木 佑樹 [1]　北島 信哉 [2]

近年，Web 上で真偽不明な言説が拡散することで社会に混乱が発生しており，言説が真実かどうかを機械的にチェックする自動ファクトチェック（Automated Fact-Checking: AFC）手法の研究が進められている．これまでに，日本固有の文脈に依存した日本語で表記された言説を対象としたデータセットは存在しておらず，AFC 手法の評価が困難であったため，新たに JAD-AFC を提案する．JAD-AFC は，ファクトチェック記事において検証された X 上の投稿をもとに構築しており，クレーム抽出，証拠収集，真実性判定を評価するために必要なデータを含んでいる．JAD-AFC を用いて既存手法による真実性判定の評価を行い，JAD-AFC を評価に利用できること，および JAD-AFC が日本語特有の課題を内包していることを示した．

1　はじめに

近年，デジタル技術の普及に伴い，Web 上での偽情報の拡散が深刻な社会問題となっている．誤情報は意図せず拡散する誤った情報を指す一方で，偽情報は内容が誤っているだけでなく害を及ぼす目的で意図的に拡散される情報と定義されており，偽情報対策の重要性が世界的に高まっている[*1]．また，偽情報は注目を集めやすくするため，テキストだけでなく画像や動画などのメディアデータを含む形式（マルチモーダルとよぶ）で発信されることが多い [1]．偽・誤情報が人々の認識を歪めることで，選挙，公衆衛生，武力紛争，気候変動といった様々な課題において，社会的混乱や対立を引き起こすことが懸念されている．日本では，特に政治，医療健康，災害の分野で偽・誤情報の影響が大きいと指摘されている [2].

このような偽・誤情報に対処するため，発信された言説が真実かどうかを人手によって検証する作業であるファクトチェックが実施されている[*2]．しかし，人手による検証には限界があるため，自動ファクトチェック（Automated Fact-Checking: AFC）の研究が進められており，中でもテキスト，画像，動画などを組み合わせて同時に処理可能なマルチモーダル AFC（Multimodal AFC: MAFC）技術の開発に関心が集まっている [3]．文献 [3] では，AFC 手法は，検証対象となる言説（クレームとよぶ）を抽出するクレーム抽出，証拠情報を収集する証拠収集，収集した証拠情報に基づいてクレームの真実性を判定する真実性判定，判定の妥当性をクレームと証拠情報を用いて作成する正当性説明生成の 4 つのタスクからなるパイプラインとして整理している．

我々の研究チームでも AFC 手法 [4] を提案しているが，AFC 手法を評価するための既存のデータセットは主に英語で表記された言説を対象に構築されており，日本語で表記された言説に対してどの程度有効であるか検証できなかった．以降では，日本語，または英語で表記された言説を対象としたデータセットを，それぞれ日本語データセット，英語データセットとよぶ．SNS（Social Networking Service）への投稿テキストにおける「放射線米ご注意してね！」[*3]のように，日本語は主語の省略や多義的な表現が多く，英語と比べて文脈依存性が高い [5].

そこで本論文では，MAFC 評価用の日本語データセット JAD-AFC（JApanese Dataset for Automated Fact-Checking）を提案する．JAD-AFC は真実性判定タスクの評価を主目的として構築したデータセットであり，日本語で掲載されたファクトチェック記事と，記事内で検証対象となっている X（旧 Twitter）[*4] 上の投稿をもとにデータを作成した．JAD-AFC には真実性判定の評価に用いるラベルに加えて，画像や動画が改ざんされているかどうかや，画像や動画が文脈外で利用されているかどうかを判定するためのメタデータも付与した．作成したデータセットは GitHub で公開[*5]している．

JAD-AFC が MAFC 手法におけるベンチマークとして利用できること，および日本語特有の課題を内包していることを確認するため，既存の英語データセットとの比較評価を実施した．評価の結果，JAD-AFC を対象とした真実性判定の正解率は英語データセットと同等であり，JAD-AFC が MAFC 手法における真実性判定の評価に利用できることを確認した．また，クレームに否定表現を含む場合や，評価対象に動画を含む場合に誤判定が多かったことから，JAD-AFC には日本語における否定表現の理解や動画内の日本語テキスト認識，モダリティ間の関係性認識といった英語データセットにはない課題を内包していることを示した．

以下，2 で関連研究を説明し，3 で提案する日本語データセット JAD-AFC の構築手順を述べる．4 で評価と考察を行い，最後に 5 で結論と今後の課題を述べる．

2　関連研究

本章では，まず AFC のパイプラインを構成する 4 つのタスクと 4 で評価に用いる MAFC 手法を紹介したのち，AFC 手法の評価のための既存データセットについて述べる．

2.1　AFC の主要タスク

AFC 手法は，以下に述べる 4 つのタスクを組み合わせたパイプラインとして整理される [3].

クレーム抽出　SNS 上の投稿やニュース記事から，検証可能な事実を示す主張（クレーム）を抽出する．抽出されるクレームの例は「日本の現在の首都は東京である」といった文である．クレーム抽出タスクの評価では，テキストや画像，動画などを入力とし

[1] 正会員　富士通株式会社 セキュリティサイエンス研究所
sasaki.yuki-01@fujitsu.com
[2] 正会員　富士通株式会社 セキュリティサイエンス研究所
kitajima.shinya@fujitsu.com

[*1] https://www.unic.or.jp/news_press/features_backgrounders/48456/
[*2] https://www.glocom.ac.jp/activities/project/9439
[*3] https://mvau.lt/media/5a7315c0-a94b-4c97-89a4-74d5197949ce
[*4] https://x.com
[*5] https://github.com/FujitsuResearch/japanese-dataset-for-automated-fact-checking

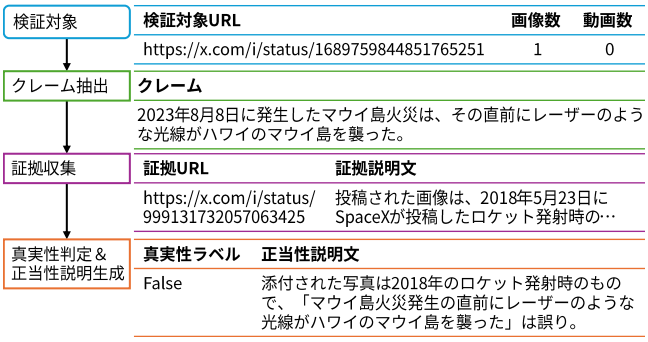

図 1: JAD-AFC の主なデータの例と各タスクの対応関係

て与え，出力されたクレームを正解と比較する．

証拠収集　クレームの真実性を判定するために，信頼できる情報源（オンライン百科事典，ニュースサイトなど）から証拠となる情報を検索し，収集する．証拠収集タスクの評価では，クレームを入力として与え，出力された証拠の文を正解と比較する．

真実性判定　収集された証拠をもとにクレームの真実性を判定し，`Supported`（真実），`Refuted`（虚偽），`NEI`（Not Enough Information: 情報不足で判定不可）など，いずれかの真実性ラベルを出力する．真実性判定タスクの評価ではクレームと証拠の文を入力として与え，出力された真実性ラベルを正解と比較する．

正当性説明生成　証拠をもとにどのように判断して真実性判定を行ったかを説明する文を生成する．正当性説明生成タスクの評価では，クレームと証拠の文，真実性ラベルを入力として与え，出力された正当性説明の文を正解と比較する．

真実性判定以外のタスクでは正解として示される文は一例であるため，単純一致ではなく，意味の類似度や単語の重み付けを考慮した評価指標を利用する必要がある．これらのタスクは相互に依存しているため，AFC 手法の精度向上を行う際にはこれらのタスクを個別に評価するだけでなく，すべてのタスクで一貫性をもって評価できるデータセットが必要になる．例えば，個別のタスクにおいて異なるデータセットを用いて性能改善を行った場合，それぞれのデータセットに特化した性能改善が行われ，パイプライン全体ではかえって性能が低下する可能性がある．

図 1 に，JAD-AFC の主なデータの例と各タスクの対応関係を示す．JAD-AFC では一貫性をもって各タスクを評価できるように，それぞれの事例に対して各タスクの評価に必要な情報をすべて含めている．

2.2　評価に用いる MAFC 手法

本節では，4 で評価に用いる MAFC 手法である DEFAME [6] と 3MFact [7] について説明する．

DEFAME はクレームを表すテキストと画像を入力として，マルチモーダル LLM（Multimodal Large Language Model: MLLM）を用いて証拠収集，真実性判定，正当性説明を行う手法である．図 2 に DEFAME の概要図を示す．DEFAME はどの証拠収集ツールを用いるかを動的に選択する点が特徴である．

3MFact はクレームを表すテキストと動画を入力として，MLLM を用いて証拠収集，真実性判定，正当性説明を行う手法である．図 3 に 3MFact の概要図を示す．3MFact の特徴は，複数のモーダルの情報を総合的に判断して真偽を判定する点で

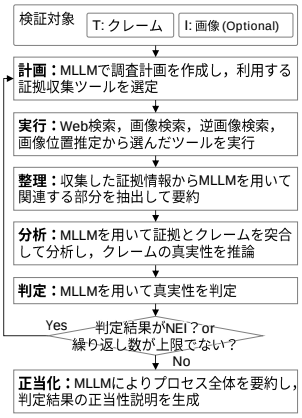

図 2: DEFAME の概要

図 3: 3MFact の概要

ある．TRUE [7] を対象とした真か偽かの 2 値判定において，3MFact は最も高い真実性判定精度を示している．

2.3　既存データセット

表 1 に，JAD-AFC と既存の AFC 手法評価用データセットを比較した結果を示す．表中の**言語**はデータセットが対象としている言説の言語を示し，**モダリティ**はクレームが含む情報の形式を示す．また，各タスクの評価に用いることができるものを ◯，できないものを × で示している．**データ収集終了日時**は，データセットに含まれる情報が収集された最も新しい日時を示している．

表 1 に示すように，日本語データセットでは全タスクを評価できるデータセットは存在しない．JSocialFact [8] は X のコミュニティノートに基づいて作成されたデータセットであり，対象は偽情報であるが，LLM が出力する内容の真実性を評価することを目的としており，各タスクの評価に必要な情報が含まれていない．また，Japanese Fake News Dataset [9] はニュース形式の偽情報であるフェイクニュースの検出を目的としているが，データセットに証拠の情報が含まれていない．日本語言説分解 [10] はテキストを検証可能なクレームに分解するクレーム抽出タスクを対象としており，それ以外のタスクの評価に必要な情報が含まれていない．一方で，JAD-AFC は日本語で表記された言説を対象とした AFC 手法のエンドツーエンド評価を実施することを目的としており，すべてのタスクに必要な情報を含んでいる．

AFC 手法の評価を目的とする英語データセットでは，ほぼすべてのタスクを評価できる包括的なデータセットが複数存在する．AVeriTeC [11] と MOCHEG [12] はテキストベースの AFC パイプライン全体を評価できるデータセットであるが，対象となるクレームはテキストのみである．ClaimReview2024+ [6], AVerImaTeC [13], VERITE [14] は，テキストと画像を含むマルチモーダルなクレームを対象としているが，動画を含むクレームは対象外である．M4FC [15] は日本語を含む多言語データセットであるが，日本語の事例は全体の 0.4%（18 件）と非常に少なく，不十分である．TRUE [7] は動画を含むが，真実性ラベルとして真または偽の 2 値に相当するデータのみを保有し，`NEI` を含まない．一方で，JAD-AFC はテキスト，画像，動画の 3 つのモダリティを網羅し，かつ AFC のパイプライン全体を評価できるという点が他と異なっており，言語や文化が異なる日本語における AFC 手法の評価に適している．

表 1: JAD-AFC と既存データセットの比較（モダリティ: T = テキスト，I = 画像，V = 動画）

データセット名	言語	モダリティ	クレーム抽出	証拠収集	真実性判定	正当性説明生成	データ収集終了日時
JSocialFact [8]	日本語	T	×	×	×	×	2024 年 3 月
Japanese Fake News Dataset [9]	日本語	T	○	×	○	×	2021 年 10 月
日本語言説分解 [10]	日本語	T	○	×	×	×	-
AVeriTeC [11]	英語	T	○	○	○	○	2021 年 12 月
MOCHEG [12]	英語	T	○	○	○	○	2022 年 5 月
ClaimReview2024+ [6]	英語	$T+I$	○	○	○	×	2025 年 1 月
AVerImaTeC [13]	英語	$T+I$	○	○	○	○	2025 年 3 月
VERITE [14]	英語	$T+I$	○	○	○	○	2023 年 1 月
M4FC [15]	多言語	$T+I$	○	○	○	○	2024 年 9 月
TRUE [7]	英語	$T+V$	○	○	○	○	2024 年 11 月
JAD-AFC (Ours)	日本語	$T+I+V$	○	○	○	○	2025 年 10 月

4 における評価では，DEFAME の評価に用いられていたデータセットである AVeriTeC と ClaimReview2024+，TRUE を比較対象に選定した．以降でこれらの詳細について説明する．

2.3.1 AVeriTeC

AVeriTeC [11] は，FEVER 2024[*6]の共有タスクで採用されたテキストのみを判定対象としたデータセットである．50 のファクトチェック機関が検証対象とした 4,568 件のクレームと，証拠 URL，真実性ラベルが含まれている．真実性ラベルとして `Supported`，`Refuted` 以外に，クレーム検証のための証拠が不足していることを示す `NEE`（Not Enough Evidence），複数の証拠間で矛盾があることや証拠の一部を恣意的に抽出していることを示す `Conflicting Evidence/Cherry-picking` のいずれかが付与されている．AVeriTeC では，学習用の `train`，開発用の `dev`，評価用の `test` の 3 つのデータセットが提供されているが，`test` は真実性ラベルが公開されていないため，本論文の評価では DEFAME 同様 `dev` を用いた．

2.3.2 ClaimReview2024+

ClaimReview2024+ [6] は，テキストと画像を判定対象としたマルチモーダルデータセットである．LLM の知識カットオフ以降のデータにおける評価のためのデータセットであり，2023 年 11 月から 2025 年 1 月までに公開されたファクトチェック記事に基づいて作成されている．真実性ラベルとして `Supported`，`Refuted`，`NEI` 以外に，事実を含むが誤解を招く表現であることを示す `Misleading` のいずれかが付与されている．

2.3.3 TRUE

TRUE [7] は，テキストと動画を判定対象としたマルチモーダルデータセットである．著名なファクトチェックサイトである snopes.com で実際に取り上げられた，動画を含むニュース記事から作成されている．真実性ラベルとして，正しいことを示す `True`，誤りであることを示す `False` に加えて，サブラベルとして誤キャプションを示す `Miscaptioned` や真偽が混合していることを示す `Mixture` といった詳細な偽情報の分類が付与されている．本論文における比較評価では真偽の 2 値判定を行うため，`True` と `False` のみを用いた．

[*6] https://fever.ai/2024/task.html

3 提案データセット JAD-AFC

本章では，提案する日本語データセット JAD-AFC の主なデータ項目について説明したのち，データセットを作成した手順について述べ，最後にデータセットの統計情報を示す．

3.1 JAD-AFC のデータ項目

表 2 に JAD-AFC の主なデータ項目とその具体例を示す．以下で，各項目の詳細を説明する．

検証対象 URL ファクトチェック記事において検証対象となった X 上の投稿を一意に特定できる URL であり，検証対象のテキストやメディアデータなどを取得するために用いる．

クレーム 検証対象となった投稿文から抽出されたクレームを表し，クレーム抽出タスクの評価に利用する．

逆クレームラベル 当該のクレームがもとの X 上の投稿と真実性ラベルが反転するように作成されたクレーム（逆クレームとよぶ）かどうかを示し，検証対象の内容とクレームの内容が一致しているかを確認するために用いる．

証拠 URL クレームの真実性を判定するために参照された Web サイトの URL.

真実性ラベル クレームの真実性を表すラベルであり，真実性判定タスクの評価に利用する．

証拠説明文 証拠 URL が指し示す Web サイトから収集した証拠の内容を説明する文で，証拠収集タスクの評価に利用する．

正当性説明文 クレームの真実性判定の理由を示す文であり，正当性説明生成タスクの評価に利用する．

また，X の各投稿には画像と動画が合計で最大で 4 つ添付できるため，画像や動画ごとに以下のラベルを付与している．

改ざんラベル 画像または動画が改ざんされているかどうかを示し，画像・動画の改ざん検出に関する分析に利用する．

OOC ラベル 画像または動画がクレームと異なる文脈（Out-of-Context: OOC）で利用されているかどうかを示し，画像・動画の文脈外利用検出に関する分析に利用する．

AI 生成ラベル 画像または動画が AI により生成された画像か否かを示し，AI 生成画像・動画の検出に関する分析に利用する．

さらに詳細な分析のため, 例えば以下のラベルを付与している．

証拠先行ラベル クレームの検証に必要な証拠情報の URL が検証対象 URL よりも先に公開されているかを表し，検証対象の公

表 2: 主なデータ項目とデータの具体例

項目名	項目の説明	具体例
検証対象 URL	検証対象となった X 上の投稿を一意に特定できる URL	`https://x.com/i/status/168975984...`
クレーム	検証対象となった投稿文から抽出されたクレーム	2023 年 8 月 8 日に発生したマウイ島火災...
逆クレームラベル	もとの X 上の投稿と真実性ラベルが反転するように作成したクレームかどうか	`False`
証拠 URL	クレームの真実性を判定するために参照された Web サイトの URL	`https://x.com/i/status/999131732...`
真実性ラベル	クレームの真実性を表すラベル	`False`
証拠説明文	Web 上の情報源から収集した証拠の内容を説明する文	投稿された画像は、2018 年 5 月 23 日に...
正当性説明文	クレームの真実性判定の理由を示す文	添付された写真は 2018 年のロケット発射...
改ざんラベル	画像または動画が改ざんされているかどうか	`False`
OOC ラベル	画像または動画が文脈外で利用されているかどうか	`True`
AI 生成ラベル	画像または動画が AI により生成されているかどうか	`False`
証拠先行ラベル	証拠情報がクレームより先に公開されていたかどうか	`True`

開直後に真偽判定が可能だったかの分析などに利用できる．

他にも様々な分析を行えるようにするため，JAD-AFC には多数のデータ項目が存在する．詳細は GitHub で公開しているデータセットを参照されたい．

JAD-AFC では，他の X 上の投稿を元としたデータセットと同様に，X の利用規約遵守および個人情報保護のため，X 上の投稿テキストや画像，動画そのものは含んでおらず，検証対象 URL におけるアカウント名の部分は匿名化している．また，X から投稿が削除または非公開とされたことが判明した場合は当該のデータを JAD-AFC から削除し，改版する予定である．

3.2 データセット作成手順

本節では，JAD-AFC を作成した手順について説明する．はじめにどのサイトからデータを収集するかを選定し，次に収集対象とするファクトチェック記事を選定した．その後，選定した記事において検証対象となった X 上の投稿を特定し，投稿から必要な情報を収集し，データセットに必要なデータを作成した．

3.2.1 データ収集元サイトの選定

JAD-AFC において言説の選定対象をファクトチェック記事としたのは，ファクトチェッカによって検証された言説を対象とすることで，自動ファクトチェックに必要なデータを正確に作成するためである．日本語言説のファクトチェック記事を公開している団体のうち，ファクトチェックに関する国際団体である The International Fact-Checking Network (IFCN)[*7]の認証を受けている団体は，InFact，日本ファクトチェックセンター（Japan Fact-check Center: JFC）[*8]，リトマスの 3 団体である．IFCN の認証は，公平性や透明性などの基準の遵守や活動実績が求められる．よって，これらの 3 団体のファクトチェック記事はいずれも信頼性が高いと考えられるが，JFC は JAD-AFC が対象とする SNS 上の多様な言説を主に検証しており，公開記事数が最も多いことから，JAD-AFC の収集元とした．

3.2.2 対象記事の選定

JAD-AFC では，JFC が公開しているファクトチェック記事のうち，医療・健康，国際，政治，災害の 4 カテゴリを対象とした．この理由は，日本において偽情報の影響が大きい分野は政治，医療・健康，災害であり [2]，これらの分野を含めることが重要であると判断したためである．また，これら 4 カテゴリが JFC の公開記事全体の 6 割超を占めており，記事数の少ないカテゴリを除外してもデータセットの網羅性は大きく損なわれないと判断した．

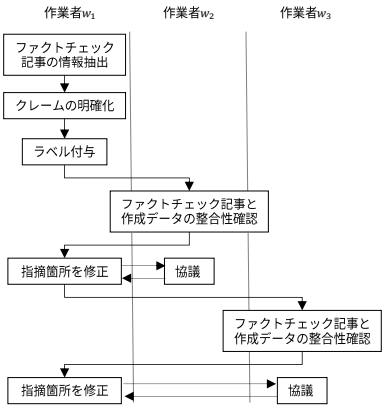

図 4: JAD-AFC におけるデータ作成の体制

また，X の API を利用して機械的に投稿に含まれるテキストや画像，動画，公開日時を取得することを想定しているため，検証対象が X 上の投稿以外の記事は除外した．さらに，データセットの再現性と一貫性を担保するため，検証対象の X 上の投稿が特定できない記事や，検証対象の投稿が X 上から削除または非公開となっている記事，検証対象の投稿以外に依存関係をもつ外部リンクや引用リポストを含む記事も除外した．ここで，検証対象の投稿以外に依存関係をもつものを除外したのは，外部依存のない単一の投稿のみを検証対象とすることでクレーム抽出と証拠収集の複雑化を防ぎ，JAD-AFC を評価に用いた際の考察を行いやすくするためである．

収集対象期間は，データ構築時点までに JFC が公開したファクトチェック記事，すなわち 2022 年 9 月 28 日から 2025 年 10 月 30 日までである．最終的に選定したカテゴリごとのファクトチェック記事の件数は，医療・健康が 61 件，国際が 77 件，政治が 98 件，災害が 22 件だった．

3.2.3 データ作成における品質管理

図 4 に，JAD-AFC におけるデータ作成の体制を示す．図に示す通り，データの品質を確保するため，各ファクトチェック記事

[*7] https://ifcncodeofprinciples.poynter.org/
[*8] https://www.factcheckcenter.jp/

に対して 3 名の作業者の組合せを変えながら 1 次作業者 w_1，2 次作業者 w_2，3 次作業者 w_3 を割り当て，3 人の作業者によって相互確認を行いながら実施した．

w_1 が作成したデータに対し，まず w_2 がファクトチェック記事と作成したデータの整合性を確認し，必要に応じて修正すべき箇所を指摘する．確認の観点は，抽出したクレームが単独で理解できるか，証拠説明文がクレームの真実性判定に適切か，各ラベルの値が妥当か，などである．w_1 は指摘された箇所を修正し，疑義がある場合は w_1 と w_2 が協議を行って修正内容を決定する．その後，w_3 が同様の観点で改めて確認し，必要に応じて修正すべき箇所を指摘する．w_1 は指摘された箇所を修正し，疑義がある場合は w_1 と w_3 が協議を行って修正内容を決定する．

このように，3 名の作業者が相互に確認を行ったうえでデータを作成した．

3.2.4 データ作成

ファクトチェック記事には，検証対象となった X 上の投稿に対する検証プロセスと結果が記載されている．そこで JAD-AFC では，JFC のファクトチェック記事と検証対象となった X 上の投稿をもとに，自動ファクトチェックにおけるクレーム抽出，証拠収集，真実性判定の各タスクに必要なデータを人手で作成した．

以降では，作業者が各データを作成した方法を述べる．

クレーム　ファクトチェック記事に記載されているクレーム単独では理解が困難な場合があるため，作業者が検証対象となった投稿の文脈を補完し，内容を要約することで作成した．1 つの投稿から複数のクレームが作成できる場合は，クレームごとに検証すべき事実が 1 つになるよう，クレームごとに独立したデータとしてデータセットに加えた．ただし，逆クレームラベルが `True` の場合は，3.2.5 で説明する方法でクレームを作成した．

逆クレームラベル　3.2.5 で説明する方法でクレームを作成した場合は `True`，それ以外の場合は `False` とした．

証拠 URL　ファクトチェック記事に証拠として記載されている Web サイトの URL から，クレームの内容に最も関連性の高いものを作業者が抽出した．

証拠説明文　証拠 URL が指し示す Web サイトから，クレームの真実性判定に最も関連性の高いものを作業者が選択し，その Web サイトの内容を要約することで作成した．

正当性説明文　ファクトチェック記事には記載されていないため，ファクトチェック記事で証拠として示されている Web サイトの情報とクレームをもとに作業者が作成した．

真実性ラベル　JFC のファクトチェック記事では検証結果を 5 種類のラベルで判定しているため，記事における判定結果が「正確」または「ほぼ正確」であれば `True`，「証拠不明」であれば `NEI`，「不正確」または「誤り」であれば `False` とした．ただし，記事における判定結果が「正確」または「ほぼ正確」であってもクレームの内容が事実と異なる場合は `False` とするなど，一部例外的な扱いを行った．

改ざんラベル　ファクトチェック記事から，AI によって生成された，または一部が切り取られたり加工されたと判明した場合は改ざんとみなして `True`，そうでない場合は `False` とした．ニュース画面のスクリーンショットのように，発信元で編集された画像や動画からは改ざんされていない場合は，`False` を付与した．

OOC ラベル　作業者によって，画像や動画がクレームと異なる文脈で利用されていると判断された場合は `True`，本来の文脈で利用されていると判断された場合は `False` とした．例えば，災害に関する投稿にもかかわらず，実際の災害とは無関係な画像が利用されている場合は `True` を付与した．

AI 生成ラベル　ファクトチェック記事から AI によって生成されたと判明した場合は `True`，そうでない場合は `False` とした．

証拠先行ラベル　検証対象 URL と証拠 URL が指し示す Web サイトの公開日を作業者が確認し，証拠情報がクレームより先に公開されていたと判明した場合は `True`，そうでない場合は `False` とした．また，いずれかの Web サイトの公開日が不明な場合は `NA`（Not Available）とした．

3.2.5 逆クレーム

ファクトチェック記事はその性質上，検証対象の内容が誤りであることが多いため，真実性ラベルが `False` に偏る傾向がある．そこで真実性ラベルの偏りを是正するため，真実性ラベルが逆となるようなクレーム（逆クレームとよぶ）を作成し，新たなデータとしてデータセットに追加した．具体的には，真実性ラベルが `False` のクレームは `True` に，`True` のクレームは `False` に，`NEI` のクレームは `True` または `False` のラベルが付与されるようなクレームおよび正当性説明文を作業者が作成し，それ以外の項目は元のデータを引き継ぐデータを追加した．

逆クレームの作成方法について，図 1 の「マウイ島火災の直前にレーザーのような光線がマウイ島を襲った」というクレームの例を用いて説明する．真実性ラベルが `False` ということは，そのクレームを否定する証拠が存在している．この例では，証拠は「投稿された画像はロケット発射の画像である」という内容である．作業者はこの証拠をもとに真実性ラベルが `True` となる「投稿された画像は 2018 年 5 月 23 日に SpaceX が投稿したロケット発射時の写真である」という逆クレームを作成する．

この逆クレーム作成手法はラベルの均衡を保つ一方で，証拠に「フェイク画像である」という表現が多くなると，真実性ラベルが `True` のクレームにも「フェイク画像である」という特定の表現が頻出する可能性を内包する．このため，評価対象の手法がこのような特定の表現にもとづいて真実性ラベルの分類を行わないよう留意する必要がある．

3.3 作成したデータの統計情報

表 3 に，JAD-AFC における X 上の投稿から作成したクレームと逆クレームに対する真実性ラベルの内訳をモダリティごとに示す．JAD-AFC では 1 つのデータには必ずクレームが 1 つ含まれるため，クレームの件数とデータの件数は一致しており，JAD-AFC 全体で 612 件のデータが存在する．ファクトチェック記事の件数に対して抽出したクレームの件数が多いのは，1 つのファクトチェック記事に複数の検証対象 URL が含まれる場合や，1 つの検証対象 URL から複数のクレームが作成される場合があるためである．また，逆クレームはすべてのクレームに対して作成しているため，逆クレームの合計件数は X 上の投稿から作成したクレームの合計件数と一致している．

表 3: X 上の投稿から作成したクレームと逆クレームに対する，モダリティごとの真実性ラベルの内訳（モダリティ：T = テキスト，I = 画像，V = 動画，$I \times k$ の k は画像の枚数を表す）

モダリティ	X 上の投稿			逆クレーム			合計
	True	False	NEI	True	False	NEI	
T	0	41	6	47	0	0	94
$T + I \times 1$	1	119	9	128	1	0	258
$T + I \times 2$	0	13	0	13	0	0	26
$T + I \times 3$	0	6	0	6	0	0	12
$T + I \times 4$	0	8	0	8	0	0	16
$T + V$	0	92	6	98	0	0	196
$T + V + I \times 1$	0	2	0	2	0	0	4
$T + V + I \times 2$	0	0	1	1	0	0	2
$T + V + I \times 3$	0	2	0	2	0	0	4
合計	1	283	22	305	1	0	612

4 評価

本章では，JAD-AFC の有効性を検証するための評価について述べる．本論文における評価では，最新の MLLM および MAFC 手法による真実性判定の結果を英語データセットと比較することで，JAD-AFC が AFC 手法の評価に利用できること，および日本語特有の課題を含むことを検証する．以下では，まず評価の内容について説明したあと，評価結果および考察について述べる．

4.1 MAFC 手法を用いた各データセットの評価方法

本節では，MLLM，および Web 検索を用いた MAFC 手法である DEFAME と 3MFact を用いて，それぞれのデータセットの真実性判定の評価を行った際の方法について説明する．

評価指標は正解率（*Accuracy*）である．また，DEFAME および 3MFact の評価結果を詳細に分析するため，データセット全体と正解ラベルごとの正解率および F1 値（*F1-Score*）も算出した．データセット全体の F1 値はマクロ F1 値（*Macro-F1*）であり，正解ラベルごとの F1 値の平均である．ここで，ラベルの取りうる値 l の集合 $\mathcal{L}$ を，DEFAME では $\mathcal{L} = \{\texttt{True}, \texttt{False}, \texttt{NEI}\}$，3MFact では $\mathcal{L} = \{\texttt{True}, \texttt{False}\}$ とし，N を総サンプル数とすると，それぞれの算出方法は以下の式の通りである．

$$Accuracy = \frac{1}{N} \sum_{l \in \mathcal{L}} TP_l \tag{1}$$

$$Precision_l = \frac{TP_l}{TP_l + FP_l} \tag{2}$$

$$Recall_l = \frac{TP_l}{TP_l + FN_l} \tag{3}$$

$$F1\text{-}Score_l = \frac{2 \times Precision_l \times Recall_l}{Precision_l + Recall_l} \tag{4}$$

$$Macro\text{-}F1 = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} F1\text{-}Score_l \tag{5}$$

また，TP，FP，FN の定義はそれぞれ以下の通りである．

- TP_l（True Positive）：正解ラベルが l で，各手法の出力ラベルも l であった件数
- FP_l（False Positive）：正解ラベルが l 以外で，各手法の出力ラベルが l であった件数
- FN_l（False Negative）：正解ラベルが l で，各手法の出力ラベルが l 以外であった件数

4.1.1 MLLM を用いた評価の方法

MLLM はタスクを限定せずに大規模なデータで事前学習されており，追加学習を行わずに知識が必要なタスクに対して高精度な出力を行える．先行研究 [6] においても，MLLM が真実性判定の比較手法として用いられている．真実性判定のためには一般知識や時事情報が求められるため，MLLM は真実性判定のベースライン評価に適している．本論文における評価では，MLLM として Azure OpenAI GPT-4o（2024-11-20）[*9] および Gemini 2.5 Pro (Stable)[*10] を使用した．

付録 A のプロンプト 1 に，MLLM による真実性判定で用いたプロンプトを示す．プロンプト内の `[CLAIM]`，`[INPUT_TEXT]`，`[INPUT_IMAGE]`，`[INPUT_VIDEO]` の部分は，それぞれクレーム，テキスト情報，画像情報，動画情報に置き換えて使用した．入力データに画像や動画が存在する場合は，プロンプトとともに画像ファイルや動画ファイルを MLLM に入力した．また，画像や動画が存在しない場合は，対応する部分を削除して使用した．

4.1.2 DEFAME を用いた JAD-AFC の評価方法

DEFAME は，テキストのみ，またはテキストと画像 1 枚のペアを入力として真実性判定を行う MAFC 手法である．画像の枚数が 2 以上のデータを入力する際は，それぞれの画像とクレームの組合せを別々のデータとして DEFAME に入力し，それらの真実性判定の結果を統合して最終的な真実性判定の結果とした．つまり，DEFAME にクレームといずれかの画像 1 枚を入力した際に真実性判定の結果が `False` であった場合は最終的な真実性判定の結果は `False`，`NEI` であった場合は `NEI`，すべての組合せで `True` であった場合は `True` とした．これは，ファクトチェックでは一部でも偽情報が含まれている場合は偽情報と判定することが妥当であると考えたためである．

また，DEFAME は動画に対応していない．よって，JAD-AFC をもとに，動画を含むデータを除外した 406 件のデータから成るデータセットを JAD-AFC-I とし，DEFAME を用いた評価にはこのデータセットを利用した．評価の際，DEFAME における Web 検索ツールとして Serper[*11]，逆画像検索ツールとして Google Cloud Vision[*12] を使用した．

4.1.3 3MFact を用いた JAD-AFC の評価方法

3MFact はテキストと動画のペアを入力として真実性判定を行う MAFC 手法である．3MFact は真実性判定結果を `True`，`False` の 2 値で出力するため，JAD-AFC の真実性判定ラベルのうち `NEI` のものは `False` とみなして評価に用いた．これは，ファクトチェックにより誤情報拡散を防止する観点から，正しい情報とそれ以外に区別することが重要と考えたためである．よって，JAD-AFC からテキストと動画のみを含むデータを抽出した 196 件のデータから成るデータセットを JAD-AFC-V とし，真実

[*9] https://azure.microsoft.com/products/ai-foundry/models/openai
[*10] https://ai.google.dev/gemini-api/docs/models/gemini-2.5-pro
[*11] https://serper.dev/
[*12] https://cloud.google.com/vision/docs

表 4: DEFAME の出力に対する各データセットの正解ラベル

DEFAME の出力	JAD-AFC-I	AVeriTeC	ClaimReview2024+
True	True	Supported	Supported
False	False	Refuted	Refuted
NEI	NEI	NEE	NEI

表 5: 3MFact の出力に対する各データセットの正解ラベル

3MFact の出力	JAD-AFC-V	TRUE
True	True	True
False	False	False

表 6: JAD-AFC-I と既存データセットの正解率

データセット	GPT-4o	Gemini 2.5 Pro	DEFAME
JAD-AFC-I	0.625	0.804	0.675
AVeriTeC	0.621	0.736	0.656
ClaimReview2024+	0.558	0.784	0.783

表 7: DEFAME による正解ラベルごとの評価結果

正解	AVeriTeC		ClaimReview2024+		JAD-AFC-I	
ラベル	正解率	F1 値	正解率	F1 値	正解率	F1 値
True	0.655	0.741	0.824	0.852	0.591	0.708
False	0.790	0.757	0.761	0.829	0.803	0.737
NEI	0.142	0.225	0.700	0.424	0.200	0.095
全体	0.656	0.574	0.783	0.702	0.675	0.513

性判定ラベルが NEI のものは False に置き換えた上で，3MFact を用いた評価にはこのデータセットを利用した．

評価の際，3MFact における Web 検索ツールとして Serper, Image LLM および Video LLM として MiniCPM-V 2.6 [16], LLM として Azure OpenAI GPT-4o（2024-11-20）を用いた．

4.1.4 既存データセットの評価方法

JAD-AFC と比較するため, 既存データセットとして AVeriTeC, ClaimReview2024+，TRUE を用いた．それぞれのデータセットでは真実性判定ラベルの定義が異なるため，評価の際にはそれぞれの MAFC 手法の出力と同じ意味合いをもつラベルを正解ラベルとし，MAFC 手法の出力と比較した．表 4 および表 5 に，DEFAME および 3MFact を用いて評価を行った際に正解ラベルとした各データセットの真実性判定ラベルを示す．評価の際には，各データセットにおいて表にない真実性判定ラベルをもつデータは除外して評価を行った．

4.2 評価結果

4.2.1 JAD-AFC-I の評価結果

表 6 に，JAD-AFC-I と既存データセットにおける真実性判定手法の正解率を示す．表から，Gemini 2.5 Pro を用いた真実性判定では比較データセットと同等以上の正解率が得られたことがわかる．Gemini 2.5 Pro は高い日本語対応能力を有し，同モデルの知識カットオフは 2025 年 1 月である．一方で，DEFAME でも用いた GPT-4o の知識カットオフは 2023 年 10 月であることから，Gemini 2.5 Pro の方が学習データが新しい．社会的な文脈の認識を必要とする真実性判定では，日本語の認識能力に加えて学習データの収集期間も結果に影響するため，妥当な結果といえる．

評価結果を詳細に分析するため，表 7 に DEFAME による正解ラベルごとの正解率，F1 値を示す．表から，すべてのデータセットで NEI の正解率が著しく低い傾向にあることがわかる．これは，真実性が明確なクレームの判定は高精度に行えるのに対し，複雑な判断を要する場合に判定精度が低下することを示しており，本評価で用いたデータセットにおける共通の課題である．

一方で，F1 値に着目すると，JAD-AFC-I の F1 値は比較データセットと比べて低く，特に正解ラベルが NEI の場合に大きな差があることがわかる．誤判定の傾向を確認すると，JAD-AFC-I において真実性判定ラベルが NEI である 15 件のクレームのうち，10 件が False と誤判定されていることがわかった．これらのクレームは明確に真偽判定できるだけの証拠が存在しないものが多く，実際に収集された証拠を分析したところ，クレームを支持する証拠はデマや陰謀論のような情報が多いことがわかった．そのため，反証する証拠も当該の言説に妥当な証拠がないことを示しているものが多く，結果として反証の方が優勢となり，DEFAME が False と判定しているものが多かった．ただし，JAD-AFC-I においてその傾向が極端に強く見られた要因は明確ではなく，詳細な分析は今後の課題である．

また，JAD-AFC-I では False の正解率が高い一方で，True の正解率は相対的に低い．これに対し，AVeriTeC および ClaimReview2024+ では JAD-AFC-I ほどの正解率の差は見られない．JAD-AFC-I における正解ラベルが True のデータの評価結果を分析したところ，クレームを支持する証拠が収集されているにもかかわらず，正しく判定できていない場合があることがわかった．そのうち約半数では，クレームおよびそれを支持する証拠が「〜ではない」といった否定形で表現されていた．このことから，日本語における真実性判定では否定表現の正確な認識が重要課題であることが示唆された．

表 8: JAD-AFC-V と既存データセットの正解率

データセット	Gemini 2.5 Pro	3MFact
JAD-AFC-V	0.819	0.565
TRUE	0.861	0.790

表 9: 3MFact による正解ラベルごとの評価結果

正解	TRUE		JAD-AFC-V	
ラベル	正解率	F1 値	正解率	F1 値
True	0.629	0.707	0.184	0.308
False	0.899	0.836	0.990	0.682
全体	0.790	0.809	0.565	0.495

4.2.2 JAD-AFC-V の評価結果

表 8 に，JAD-AFC-V と既存データセットにおける真実性判定手法の正解率を示す．表から，Gemini 2.5 Pro を用いた真実性判定では，正解率は 0.8 を超えているものの，TRUE よりも低くなっていることがわかる．これは，JAD-AFC-V の方がデータ収集終了日時が後ろになっており，TRUE よりも新しい事例を多く含んでいるためと考えられる．社会的文脈の理解が求められる真実性判定においては，MLLM の学習データの収集期間が結果に影響するため，本結果は妥当であると考えられる．

一方で，3MFact を用いた評価では，JAD-AFC-V の正解率が TRUE よりも大幅に低くなっている．評価結果を詳細に分析するため，表 9 に 3MFact による正解ラベルごとの正解率，F1 値を示

す．表より，3MFact による評価では JAD-AFC-V における正解ラベルが True の場合の正解率や F1 値が極端に低くなっていることから，判定結果が大きく False に偏っていることがわかる．

誤判定事例を分析した結果，3MFact で用いた MiniCPM-V 2.6 の日本語認識性能が低いため，動画内の情報が正しく取得できていなかった．このことから，JAD-AFC-V を用いることで英語データセットでは顕在化しなかった日本語動画特有の課題に対する評価に有効といえる．また，日本語動画を対象とした AFC の実現には，動画内の日本語字幕や音声情報とテキストを認識するマルチモーダル処理能力が不可欠であることが示唆された．

4.2.3 評価結果のまとめ

以上の結果から，JAD-AFC は日本語環境における MAFC 手法のベンチマーク評価に必要な情報を備えていることが確認できた．特に，最新の MLLM を用いた評価では英語データセットと同等の性能が得られており，JAD-AFC が実用的なデータセットとして機能することが示された．さらに，日本語特有の課題として否定表現の認識や日本語を含む動画の理解といった点が明らかになり，これらの課題に対応することで日本語環境における AFC 手法の性能向上が期待される．

5 おわりに

本論文では，日本語における MAFC 手法の性能評価のためのデータセットである JAD-AFC を提案した．JAD-AFC は，JFC が公開しているファクトチェック記事において検証された X 上の投稿をもとに構築しており，クレーム抽出，証拠収集，真実性判定の性能を評価するために必要なデータを含んでいる．

JAD-AFC の有効性を評価により検証した結果，日本語環境における MAFC 手法のベンチマーク評価に必要な情報を備えており，最新の MLLM を用いた評価では既存の英語データセットと同等の性能を示すことを確認した．また，MAFC 手法を用いた評価により，JAD-AFC は日本語における否定表現や動画内の字幕および音声の認識が難しいといった日本語特有の課題を含んでいることを確認した．

本論文では真実性判定の評価に焦点を当てているが，JAD-AFC はクレーム抽出や証拠収集といった他のタスクの評価にも利用できる．今後，JAD-AFC が真実性判定以外のタスクにおいても有用であることを確認する必要がある．また，JAD-AFC を用いた真実性判定タスクの評価では MLLM を用いることで正解率が約 0.8 に達することがわかったが，さらなる性能向上のために，JAD-AFC を活用した MAFC 手法の提案を行う予定である．

謝辞

この成果は，NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）の委託業務（JPNP22007）の結果得られたものです．

参考文献

[1] N. Dufour, A. Pathak, P. Samangouei, N. Hariri, S. Deshetti, A. Dudfield, C. Guess, P. H. Escayola, B. Tran, M. Babakar, and C. Bregler, “AMMeBa: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild,” arXiv, 2405.11697, May 2024.

[2] 山口 真一，“AI がもたらす with フェイク 2.0 時代の未来と適切な社会的対処”，Nextcom，vol. 2024，no. 59，pp. 4–13，Sep. 2024.

[3] M. Akhtar, M. Schlichtkrull, Z. Guo, O. Cocarascu, E. Simperl, and A. Vlachos, “Multimodal Automated Fact-Checking: A Survey,” in *EMNLP 2023*, pp. 5430–5448, Dec. 2023.

[4] K. Oki, K. Yamashita, and S. Kitajima, “A Hybrid Approach Combining LLMs and Web-Based Information for Automated Fact-Checking,” in *ICWS 2025*, pp. 960 – 962, July 2025.

[5] 萩行 正嗣，河原 大輔，黒橋 禎夫，“外界照応および著者・読者表現を考慮した日本語ゼロ照応解析”，自然言語処理，vol. 21，no. 3，pp. 563–600，June 2014.

[6] T. Braun, M. Rothermel, M. Rohrbach, and A. Rohrbach, “DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts,” in *ICML 2025*, vol. 267, pp. 5383–5417, July 2025.

[7] K. Niu, D. Xu, B. Yang, W. Liu, and Z. Wang, “Pioneering Explainable Video Fact-Checking with a New Dataset and Multi-Role Multimodal Model Approach,” in *AAAI*, vol. 39, pp. 28276–28283, Apr. 2025.

[8] T. Nakazato, M. Onishi, H. Suzuki, and Y. Shibuya, “JSocialFact: A Misinformation Dataset from Social Media for Benchmarking LLM Safety,” in *BigData 2024*, pp. 3017–3025, Dec. 2024.

[9] T. Murayama, S. Hisada, M. Uehara, S. Wakamiya, and E. Aramaki, “Annotation-Scheme Reconstruction for “Fake News” and Japanese Fake News Dataset,” in *LREC 2022*, pp. 7226–7234, June 2022.

[10] 政野 美和，欅 リベカ，欅 惇志，清丸 寛一，中山 功太，堀尾 海斗，源 怜維，橘 秀幸，河原 大輔，“LLM の生成テキストの真偽検証のための日本語言説分解データセットの構築”，研究報告自然言語処理（NL），vol. 2025-NL-265，no. 1，pp. 1–12，Sep. 2025.

[11] M. Schlichtkrull, Y. Chen, C. Whitehouse, Z. Deng, M. Akhtar, R. Aly, Z. Guo, C. Christodoulopoulos, O. Cocarascu, A. Mittal, J. Thorne, and A. Vlachos, “The Automated Verification of Textual Claims (AVeriTeC) Shared Task,” in *FEVER 2024*, pp. 1–26, Nov. 2024.

[12] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, and L. Huang, “End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models,” in *SIGIR 2023*, pp. 2733 – 2743, July 2023.

[13] R. Cao, Z. Ding, Z. Guo, M. Schlichtkrull, and A. Vlachos, “AVerImaTeC: A Dataset for Automatic Verification of Image-Text Claims with Evidence from the Web,” in *NeurIPS 2025*, Dec. 2025.

[14] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, “VERITE: A Robust Benchmark for Multimodal Misinformation Detection Accounting for Unimodal Bias,” *Int. J. Multimed. Inf. Retr.*, vol. 13, Jan. 2024.

[15] J. Geng, J. Tonglet, and I. Gurevych, “M4FC: A Multimodal, Multilingual, Multicultural, Multitask Real-World Fact-Checking Dataset,” arXiv, 2510.23508, Oct. 2025.

[16] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, “MiniCPM-V: A GPT-4V Level MLLM on Your Phone,” arXiv, 2408.01800, Aug. 2024.

付録 A MLLM による真実性判定で用いたプロンプト

```
# Instructions
**Determine the Claim's veracity** by following these steps:
Write one paragraph about which one of the Decision Options
    applies best.
Include the most appropriate decision option at the end and
    enclose it in backticks like `this`.

## Decision Options
- `True`: "The claim is true."
- `False`: "The claim is false."
- `NEI`: "There is insufficient information to determine
    whether the claim is true or false."

# Claim
[CLAIM]

## Source of Claim
[INPUT_TEXT]
[INPUT_IMAGE]
[INPUT_VIDEO]

# Your Judgement
```

プロンプト 1: MLLM による真実性判定プロンプト