修士論文

テキストの単位の違いを考慮したメディアバイアスの検出

LI Donghao

主指導教員 白井 清昭

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和８年３月

Abstract

In the contemporary digital era, the landscape of information consumption has undergone a radical transformation, enabling individuals to access an unprecedented volume of different information with remarkably low temporal costs every day. However, this accessibility is inextricably linked to the pervasive risk of media bias, a phenomenon where information is disseminated in a non-objective manner specifically designed to steer the recipient's perceptions, interpretations, and judgments toward a particular direction. Within this context, political bias poses a particularly severe risk, as it has the capacity to obstruct healthy democratic discourse, distort the fundamental processes of public opinion formation, and exert profound, long-term deleterious effects on electoral behavior and civic engagement. To address these concerns, research on automated media bias detection methods utilizing Natural Language Processing (NLP) technology has flourished. Nevertheless, a significant obstacle to the practical application of these technologies remains: the "inconsistency of granularity of text." This phenomenon occurs during the machine learning phase of bias detection models when the granularity of text in the training data significantly differs from that in the test data. In such cases, the discrepancy in volume of information resulting from the inconsistent granularity leads to a marked degradation in the model's discriminatory performance. This study aims to resolve the inconsistency of granularity of text and enhance the overall performance of automated media bias detection. In this research, the detection of media bias is defined as a binary classification problem, where the system must determine whether a given input text constitutes a "Biased" instance containing bias or a "Neutral" instance characterized by objective reporting.

To achieve effective granularity alignment, this research proposes two distinct modules: the "Text Expansion Module" and the "Text Reduction Module." The Text Expansion Module is specifically engineered for scenarios where the primary unit of the granularity of text is a "sentence." The module identifies the original source article to which the target sentence belongs and expands the textual context by concatenating the sentences that immediately precede and follow the target sentence. To ensure that the expansion process only incorporates contextually relevant information and avoids the introduction of irrelevant information, the module calculates the semantic similarity between the original sentence and its neighboring sentences. Textual expansion is only executed if the similarity score exceeds a pre-defined threshold, which is rigorously optimized using validation data. The Text Reduction Module is applied when the granularity of text is at the "document" level. In this process, the system leverages a Large Language Model (LLM) provided with sophisticated instructions to identify and remove sentences deemed irrelevant to the assessment of media bias or its underlying characteristics.

The module generates a condensed summary or "reduced" version that distills the text down to its most critical informational elements. By employing precise prompt engineering, the module transforms lengthy, redundant documents into concise documents that retain high-density information directly related to bias judgment.

In situations where there is an inconsistency of granularity of text between the training and test datasets, these two modules are utilized to achieve precise granularity alignment through two distinct strategies: the "Single-Approach" and the "Double-Approach." The Single-Approach adjusts the length of the test data only. Specifically, when the training data is document-level and the test data is sentence-level, the text expansion module is employed to expand the test data. Conversely, if the training data is sentence-level and the test data is document-level, the text reduction module is used to condense the test data. The Double-Approach, on the other hand, involves the simultaneous adjustment of both the training and test data. By applying the text reduction module to document-level texts and the text expansion module to sentence-level texts across both datasets, the approach seeks to bring both textual sets toward a standardized, median length.

After the adjustment of the text length through these proposed modules, the refined text is fed into an LLM to be converted into an "intermediate representation" that describes the bias with respect to multiple dimensions. These dimensions include tone and language, sources and citations, coverage and balance, agenda and framing, and examples and analogies. Rather than attempting to predict a binary label directly from raw text, this methodology utilizes the intermediate representation to ensure a high degree of explainability regarding the final judgment. The core innovation of this research lies in the dynamic adjustment of the input text's granularity prior to the generation of the intermediate representation. Furthermore, another improvement in this study is the adoption of a configuration where the intermediate representation is input into BERT, acting as a classifier, rather than relying on the simple majority voting method utilized in the baseline model, where samples similar to the test data are retrieved from the training set and the final label is determined by a majority vote.

To evaluate the effectiveness of the proposed methods, four public datasets are used: BABE and BASIL as the sentence-level datasets, while FlipBias and MFC as the document-level datasets. The performance of the models is evaluated using four evaluation metrics: Precision, Recall, and the F1-score, as well as the Macro F1-score.

The experimental results demonstrated the effectiveness of the proposed Text Expansion and Text Reduction modules in resolving the inconsistency of granularity of text. Specifically, in cases where the training data was at the document-level

and the test data was at the sentence-level, it was confirmed that the Double-Approach that adjusted the granularity in both the test and training data significantly improved bias detection performance. Additionally, for cases where the training data was at sentence-level and the test data was at document-level, the results revealed that the Single-Approach that utilized text reduction on the test data worked well for bias detection. These findings underscore the critical importance of granularity alignment of texts in the training and test data.

Furthermore, an ablation study was conducted to verify the individual contributions of each component within the proposed method. Regarding the refinement of the judgment model, it was confirmed that replacing the simple majority voting method used in the baseline with BERT as a classifier significantly contributed to the improvement of detection performance. In addition, the effectiveness of the Text Expansion and Text Reduction modules themselves was validated through the ablation study. Quantitatively, it was shown that excluding these modules prevented the resolution of the inconsistency of granularity of text, subsequently leading to a decline in media bias detection performance.

Finally, a detailed case study and error analysis were performed to clarify the limitations and advantages of the proposed method. While the Text Expansion Module succeeded in supplementing context and background information, it was found to trigger a "dilution effect" in some biased samples, where the additional information tends to lower the relative density of the biased text. Conversely, while the Text Reduction Module effectively stripped away unnecessary information from long documents, it suggested a "neutrality vulnerability" wherein the loss of neutral context increased the likelihood of neutral instances being misclassified as biased. These insights provide valuable guidance for model design in future research.

The major contribution of this thesis is to introduce a new perspective, granularity alignment of text, to the field of media bias detection and to demonstrate its effectiveness from both quantitative and qualitative analysis.

概要

デジタル時代において，人々は日々膨大な情報に接し，様々な情報を簡単に入手できる．しかし，受け手の認識や判断を特定の方向へ誘導するために，客観的ではない情報を発信するメディアバイアスの存在が問題となっている。特に政治的バイアスは，民主主義的な対話を妨げ，世論形成や選挙行動に長期的な影響を及ぼすリスクを孕んでいる．これに対し，自然言語処理技術を用いたメディアバイアスの自動検出手法の研究が盛んに行われているが，実用化に向けた大きな課題として「粒度不整合問題」が存在する．これは，バイアス検出モデルを機械学習する際，訓練データとテストデータとでテキストの粒度が異なる場合，情報量の差がモデルの判定性能を低下させる現象である．本研究は，この粒度不整合問題を解消し，メディアバイアスの自動検出の性能を向上させることを目的とする．本研究におけるメディアバイアスの検出とは，与えられたテキストがバイアスを含む Biased 事例かバイアスを含まない Neutral 事例かを判定する二値分類問題と定義する．

まず，粒度整合を実現するため，本研究では「テキスト拡張モジュール」と「テキスト縮退モジュール」という二つのモジュールを提案する．一つ目のテキスト拡張モジュールは，テキスト粒度が「文」である場合に適用される．対象文が属する元の記事を特定し，その前後に出現する文を連結することでテキストを拡張する．ただし，適切な文脈を表す文のみを拡張するため，元の文と前後の文の類似度を計算し，それが閾値以上のときにテキスト拡張を行う．閾値は事前に検証データを用いて最適化する．二つ目のテキスト縮退モジュールは，テキスト粒度が「文書」である場合に適用される．ここでは大規模言語モデル（LLM）に対して「メディアバイアスの有無や特徴の判断に無関係と考えられる文を除去し，重要な情報のみを保持した要約を生成する」よう指示し，テキストを縮退する．これにより，バイアス判定に深く関連する情報を含み，かつテキスト長が短い文書に変換する．

訓練データとテストデータでテキスト粒度が異なるとき，2 つのモジュールを用いて両者の粒度を整合させる．この際,「シングル・アプローチ」と「ダブル・アプローチ」の 2 つの手法を提案する．シングル・アプローチでは，テストデータのテキストの長さを調整する．すなわち，訓練データの粒度が文書でテストデータの粒度が文のとき，テキスト拡張モジュールを用いてテストデータのテキストを拡張する．また，訓練データの粒度が文でテストデータの粒度が文書のとき，テキスト縮退モジュールを用いてテストデータのテキストを縮退する．一方，ダブル・アプローチでは，訓練データとテストデータの両方のテキストの長さを調整する．テキスト粒度が文書のときはテキスト縮退モジュールを，テキスト粒度が文のときはテキスト拡張モジュールを用いることで，双方のテキスト長を中間的な長さに近づける．

次に，これらのモジュールによって調整されたテキストは，LLM へと入力され，文体と語調，情報源の選択，報道のバランス，フレーミング，および例示の用い方

といった複数の側面からバイアスの有無を説明した「中間表現」へと変換される．テキストからバイアスの有無を直接予測するのではなく，この中間表現を介することで，判定結果に対する高度な説明可能性を持つ．本研究の核心は，入力テキストの粒度をあらかじめ動的に調整してから中間表現を生成させる点にある．さらに，ベースライン手法ではテストデータに類似したサンプルを訓練データから検索し，類似サンプルに付与されたラベルの多数決によってバイアスの有無を判定する Voting 方式を採用していたが，本研究では中間表現を入力としバイアスの有無を判定する分類モデルを BERT によって学習する手法を採用する．

評価実験では，文単位のデータセットとして BABE および BASIL を，文書単位のデータセットとして FlipBias および MFC を使用し，提案手法の有効性を検証した．評価指標には，Precision，Recall，F1，および Macro F1 の 4 つを採用した．

実験の結果，粒度不整合を解消するために提案したテキスト拡張モジュール・テキスト縮退モジュールの有効性を確認した．特に訓練データが文書単位でテストデータが文単位であるケースにおいて，双方にモジュールを適用する「ダブル・アプローチ」によってバイアス検出の性能が大きく向上したことを確認した．また，訓練データが文単位でテストデータが文書単位であるケースについては，テストデータのテキスト縮退による「シングル・アプローチ」がバイアス検出に特に有効であった．総じて，訓練データとテストデータのテキストの粒度を合わせることの重要性が示された．

さらに，アブレーション研究を通じて提案手法の構成要素の有効性を検証した．まず，判定モデルの改良に関して，従来のベースライン手法が採用していた単純な多数決（Voting）方式を BERT による分類へと変更したことが，バイアス検出の性能向上に大きく寄与していることを確認した．加えて，テキスト拡張およびテキスト縮退という各モジュール自体の有効性もアブレーション研究によって確認された．これらのモジュールを排除した場合，テキスト粒度の不整合が解消されず，検出性能の低下を招くことが定量的に示された．

最後に，詳細なケーススタディおよびエラー分析を行い，提案手法の利点と欠点を明らかにした．テキスト拡張モジュールは文脈と背景情報を補完できる一方で，Biased 事例において，バイアスを示唆する文の量が相対的に低下する「希釈効果」を確認した．また，テキスト縮退モジュールは，長い文書からバイアス判定に不要な情報を削除する一方，中立的な文脈が失われることで生じる「中立事例の脆弱性」により，Neutral 事例が Biased へと誤判定される可能性が高まることを確認した．これらの知見は，今後の研究におけるバイアス検出モデル設計の指針となる有益な情報を提供するものである．

本論文の主たる貢献は，テキスト粒度のアライメントという新たな視点をメディアバイアス検出に導入し，その有効性を定量的・定性的な両面から明らかにした点にある．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

デジタルメディアプラットフォームの急速な普及により，人類社会はこれまでにないほど容易にニュースや情報へアクセスできる時代を迎えている．高度に発達したデジタル環境においては，情報の伝播が地理的・時間的制約を超えて加速し，人々は極めて低い時間的コストで膨大な情報に接することが可能となった．一方で，Rodrigo-Ginés らが指摘するように，デジタル時代の特徴の一つは情報が迅速かつ受動的に消費される点にある [1] ．すなわち，多くのユーザは個々の情報を十分に精査することなく，大量のニュースや意見に晒されており，このような環境は，情報の流通を効率化する一方で，偏向したコンテンツが拡散しやすい土壌を生み出している．

こうした状況下において，特定の視点や一方が有利となるよう情報が選択的・体系的に提示される「メディアバイアス」は，デジタル空間における重要な社会的課題となっている．例えば，特定の政党や政治家に対して有利な印象を与える書き方をした記事はメディアバイアスの典型例である．メディアバイアスは，ソーシャルメディアを介して急速に拡散し，個人の認識形成や意思決定に深刻な影響を及ぼす．先行研究では，メディアバイアスが選挙結果に影響を与える可能性 [1] や，公衆衛生政策に対する国民の受容態度を左右することが示されている [2]．さらに，メディアが反対意見や多様な視点を取り入れなくなることでバイアスが長期的に固定化されることは，民主主義的なプロセスの健全な機能や社会の持続的な発展を脅かす直接的な要因となり得る．

メディアバイアスが世論形成において決定的な役割を果たしていることを踏まえると，メディアコンテンツに内在するバイアスを検出することは喫緊の課題である．しかし，Rodrigo-Ginés らによる近年の包括的なサーベイ研究 [1] によれば，既存のメディアバイアス検出用データセットは，規模や情報源の限定性，対象イベントの偏り，英語圏および米国政治領域への集中といった問題を抱えている．さらに，文書単位と文単位が混在したアノテーション設計や，統一的な評価基準の欠如により，異なる研究間での直接的な性能比較が困難であるという課題も報告されている．

特に問題になるのは，バイアス検出モデルを機械学習する際，訓練データとテストデータとでテキストの粒度が異なる場合である．テキストの粒度とは，ここでは「文」または「文書」を指す．バイアスを検出したい対象テキストは文 (もし

くは文書) であるが，バイアス検出モデルを学習するためのデータセットとして文書 (もしくは文) 単位のものしか用意できない場合がありうる．訓練データとテストデータとでテキストの粒度が異なるとき，バイアス検出の性能が低下する．

1.2 目的

本研究の主な目的は，メディアバイアス検出において，訓練データとテストデータのテキスト粒度の不一致によって生じる粒度不整合問題を解消し，特に政治的バイアスに対するバイアスの自動検出の性能を向上させることである．

具体的には，本研究は以下の三つの技術的アプローチを通じて，この目的の達成を目指す．

第一に，大規模言語モデル (Large Language Model; LLM) による中間表現の生成と活用である．生のテキストをそのまま分類器に入力するのではなく，LLM を用いて，バイアスを示唆する指標である中間表現を生成する．これにより，テキストの表層的特徴に依存することなく，フレーミングやトーンといった高次の政治的バイアス要因を明示的に表現可能な特徴量を抽出する．

第二に，テキスト粒度の双方向同期（拡張と縮退）である．訓練データとテストデータ間に存在する粒度の不一致を解消するため，入力テキストのテキスト粒度を調整する二種類のモジュールを導入する．テキストが文である場合には，意味的類似度に基づいて周辺文脈を付与する文脈拡張を行い，文単位では捉えにくい潜在的な意図や背景情報を補完する．一方，テキストが長い文書である場合には，重要文抽出による情報の縮退を適用し，バイアス判定に直接寄与しない情報を除去する．

第三に，BERT[3] による高度な判定モデルの構築である．ベースライン手法である IndiVec[4] が採用していた多数決（Voting）による単純な手法を見直し，本研究では BERT による分類器を導入する．粒度が同期された中間表現を BERT に入力することで，複雑な文脈下におけるバイアスの有無をより精緻に分類する．

提案手法の有効性をテキスト粒度の異なる複数のデータセットを用いて検証する．さらに，本研究では数値的な性能評価にとどまらず，詳細なケーススタディおよびエラー分析を実施する．統計的傾向の分析と具体例に基づく定性的検討を通じて，提案手法が有効に機能する条件および限界を明らかにし，今後のメディアバイアス研究に資する知見を提供する．

1.3 本論文の構成

本論文の構成は以下の通りである．第 2 章では，本研究に関連する既存研究について紹介する．第 3 章では，本研究で提案する粒度整合に基づくメディアバイアス検出手法について詳述する．第 4 章では，提案手法の有効性を検証するため

の評価実験を行い，その実験結果および考察について述べる．最後に，第 5 章では，本研究の成果を総括し，提案手法の意義と限界について述べるとともに，今後の課題および展望について述べる．

第2章　関連研究

本章では，本論文に関連する先行研究について述べる．2.1 節では，メディアバイアスの種類と政治的バイアス定義に関する研究を概観する．2.2 節では，バイアス検出に関する既存研究を紹介する．2.3 節では，これらの研究を踏まえた上で，本研究の特徴について論じる．

2.1　メディアバイアスと政治的バイアス

メディアバイアスの定義については，研究分野によって異なり，かつ時に相互に矛盾している [5]．Rodrigo-Ginés らは，既存の膨大な関連研究を分析した結果，メディアバイアスを「ジャーナリズムの形式をとりつつも，情報の選択や提示において客観性から逸脱し，受け手の認識や判断を特定の方向へ誘導する傾向を持つ現象」と定義している [1]．同研究によれば，メディアバイアスには以下の三つの本質的要素が共通して認められる [1]．第一は，受け手の意見や態度に影響を与えようとする意図性である．第二は，特定のメディアまたはジャーナリストにおいて反復的・一貫的に観察される持続性である．第三は，記憶に残るストーリーを構築する過程において，意図的または無意識的にバイアスが付加され得る構成的性質である．

広義のメディアバイアスには，広告主への配慮によって報道内容が選択・歪曲される広告バイアス（Advertising bias）[6] や，科学的根拠に乏しい主張を助長する反科学バイアス（Anti-science bias）[7] など，多様な形態が含まれる．その中でも，政治的バイアス（Political bias）は，特定の政党，政治家，政策，あるいは政治的イデオロギーに対して，報道が一貫して肯定的または否定的な傾向を示す現象として位置付けられる．政治的バイアスは，メディアバイアスの一般的特徴を共有しつつ，その適用対象を政治領域に限定したものであり，民主主義社会における世論形成や選挙行動に長期的な影響を及ぼすことが指摘されている [8]．

政治的バイアスはニューステキスト上で複数の言語的・構造的手法を通じて具現化される．例えば，特定の政治家や政策に対して肯定的または否定的な形容詞や動詞を選択的に用いる表現上の偏りは，ステートメント・バイアス（Statement bias）[9] として捉えられ，主に語調や語彙選択の観点から分析可能である．また，一方の政党の不祥事のみを強調して報じる量的偏りや，対立する議論において片方の主張のみを提示する質的偏りは，カバレッジ・バイアス（Coverage bias）[10]

およびコンテント・バイアス（Content bias）として整理される．さらに，どの政治課題をニュースとして取り上げるかを選別し，特定の責任主体を曖昧化する手法はゲートキーピング・バイアス（Gatekeeping bias）[9] に該当する．加えて，事実関係を歪曲した例示や誇張表現は，ディストーション・バイアス（Distortion bias）[11] と呼ばれ，政治的プロパガンダにおいて典型的に観察される．

メディアバイアスの検出タスクに関するデータセットについては，Rodrigo-Ginés らが指摘するように，既存のデータセットの多くは，こうした多様なバイアス形態の中でも，とりわけ政治的バイアスを分析対象として設計されている [1]．

2.2 バイアス検出手法

2.2.1 深層学習アプローチ

深層学習の発展に伴い，メディアバイアス検出の手法は，人手による特徴量エンジニアリングを必要とする従来の機械学習から，テキストから特徴表現を自動学習するニューラルネットワークモデルへと大きく移行した．初期には RNN(Recurrent Neutral Network) や LSTM(Long Short-Term Memory) が用いられたが，近年では BERT に代表される事前学習済み言語モデルが主流となっている．大量のテキストから事前学習された言語モデルはテキストをその意味を表す抽象表現に変換する能力が高いため，バイアス検出の性能を劇的に向上させた．

Chen らは，従来のニューラルネットワークモデルが，未知のイベントに関する記事に対するバイアス検出の性能が低いという課題を指摘し，単語レベルの表面的な情報だけに頼るのではなく，記事全体における「バイアス記述の分布」に着目した手法を提案した [12]． 具体的には，BERT を用いて文単位のバイアスを抽出し，その出現頻度，位置，および順序情報をガウス混合モデルでモデル化した．このアプローチにより，特定の単語に依存せず，記事全体で構造的に表現されるバイアスを捉えることに成功し，BERT が文脈理解だけでなく，このような構造的バイアスの検出にも有効であることを示した．

Maab らは，BERT を用いた文レベルの政治的バイアス検出において，ターゲットを意識した文脈拡張の手法を提案した [13]．文脈拡張とは対象文の前後に出現する文を文脈として獲得しバイアス検出モデルの入力に追加する手法だが，従来の手法が固定長文脈 (前後 N 文) といった単純な手法で文脈を定義することでノイズとなる情報が文脈として取得されることがあるという問題に着目し，バイアスの対象を考慮して慎重に文脈を探索・拡張するアルゴリズムを開発した．BASIL データセットを用いた実験の結果，この手法と BERT を組み合わせることで，文脈依存性の高いバイアス表現に対しても State-of-the-Art（SOTA）の精度を達成しており，文脈の適切な取得がバイアス検出の性能向上に貢献することを実証した．

2.2.2 大規模言語モデルを用いたアプローチ

近年，Transformer ベースのアーキテクチャの進化に伴い，GPT シリーズや Llama，Gemini といった大規模言語モデル（LLM）を用いたバイアス検出および緩和の研究が急速に進展している．これらのモデルは，従来の機械学習手法では捉えきれなかった微妙なニュアンスや高度な言語的特徴を学習する能力を有している．

具体例として，Wen と Younes は，メディアバイアス識別ベンチマーク（MBIB）[14] を用い，ゼロショットの GPT-3.5 と，BART や ConvBERT などのファインチューニングされたモデルの能力を比較した [15]．その結果，GPT-3.5 は文書レベルの文脈バイアスの検出において，ファインチューニングされたモデルと同等の性能を発揮することを示した．しかし，フェイクニュース，人種，ジェンダー，認知バイアスといったより微細なバイアスの検出には依然として困難が伴うことも明らかにした．また，文書または長いテキストに対するバイアス検出には課題が残されているとしている．

一方で，LLM の判定プロセスが人間の認知とどのように整合するかという点については，課題も指摘されている．Banik らは，イスラエル・パレスチナ紛争に関するニュース記事を対象に，人間によるアノテーションと LLM を含む複数の言語モデル（BERT，RoBERTa，GPT-5，FLAN-T5 など）の判定結果を比較分析した [16]．彼らの実験によると，ファインチューニングされた RoBERTa モデルが人間のラベルと最も高い整合性を示した一方で，GPT などの生成モデルは Zero-shot 設定でも高い性能を発揮した．しかし，質的な分析からは，人間とモデルの推論プロセスに体系的な乖離があることを明らかにした．具体的には，人間が文脈的なフレーミングや話者の帰属に基づいてバイアスを判断するのに対し，モデルは「テロリスト」や「占領」といった感情的な単語に過度に依存する傾向が見られた．

こうした LLM の「ブラックボックス性」や「適応性」の課題に対し，Lin らは，IndiVec と呼ばれる新しいフレームワークを提案した [4]．IndiVec では，LLM を用いてテキストからいくつかの観点からバイアスの有無を説明するテキスト (中間表現) を抽出し，それらを大規模なベクトルデータベースに格納する．推論時には，入力テキストから生成された中間表現とデータベース内の中間表現を照合し，検索された中間表現に付与されたラベルの多数決（Voting）によってバイアスの有無を決定する．この手法は，特定のデータセットに依存するファインチューニングを行わないため，異なるドメインやデータセットに対しても高い汎化性能を持つ．また，中間表現を参照することで，どの指標に基づいてバイアスのを判定したかが明示されるため，高い説明可能性を有している点に特徴がある．

2.3 本研究の特徴

本研究は，既存手法が抱える「テキスト粒度の不整合問題」の解決を目指し，以下の三つの独自のアプローチを探求する．

第一に，訓練データとテストデータにおける「テキストの単位（粒度）の不一致」を解消するアプローチである．従来のメディアバイアス検出モデルの多くは，訓練データとテストデータのテキストの粒度が一致していることを前提としている．しかし，実世界におけるバイアス検出の対象となるテキストは，短い文から長い文書まで多岐にわたり，訓練データとテストデータとでテキスト粒度が異なることが起こりうる．本研究では，この粒度不整合が判定精度を低下させる要因のひとつであるとみなし，訓練時と推論時におけるテキストの粒度を揃えることを目指す．

第二に，前述の課題を解決するため，テキスト拡張およびテキスト縮退モジュールを統合した双方向のテキスト粒度アライメント手法を提案する．Maab らが提案したターゲット指向の文脈拡張手法 [13] に着想を得つつ，本研究では文間の意味的類似度に基づいて記事内の関連情報を柔軟に補完する文脈拡張手法を提案する．これに加え，長い文書に対しては LLM を用いてバイアス判定に重要な情報を保持したまま文書を縮退させる手法を提案する．両者を組み合わせたアプローチにより，文単位の短いテキストから文書単位の長い記事まで，粒度の差が大きいデータセットに対しても，高い精度でバイアスを検出することを目指す．

第三に，判定対象のテキストを直接分類器に入力するのではなく，LLM によって生成された中間表現を BERT による分類器の入力とする．Wen と Younes の研究が示すように，生のテキストのみを用いたゼロショットの LLM 検出には限界があり，特にフェイクニュースや認知バイアスといった微妙なバイアスの特定には困難が伴う [15]．本研究では，IndiVec のフレームワークを応用し，テキストを多角的に分析・構造化した中間表現を BERT の入力に用いる．これにより，BERT の入力長制限という技術的課題を克服しつつ，LLM によって生成されたバイアスの有無の分析結果を BERT に与えることでバイアス検出の性能向上を狙う．

第3章　提案手法

本章では，本研究のテキストの粒度の違いを考慮したメディアバイアス検出手法の詳細について述べる．ここでのメディアバイアス検出とは，与えられたテキストに対し，それがバイアスを含むか否かを判定する二値分類問題と定義する．また，バイアスありのラベルを Biased，バイアスなしのラベルを Netural と記す．

本章の構成は以下の通りである．3.1 節では，本研究のベースとなる手法 IndiVec について紹介する．3.2 節では，本研究の核である粒度整合のためのモジュールについて詳述する．3.3 節では，異なる粒度設定に対する上述のモジュールの応用の詳細について説明する．3.4 節では，BERT に基づくバイアス判定モデルについて述べる．

3.1　本研究のベースとなる手法：IndiVec

本研究では，Lin らによって提案された IndiVec[4] を提案手法の基盤となるバイアス検出フレームワークとして採用する．IndiVec は，大規模言語モデル（LLM）を活用し，テキスト内に潜在する微細なメディアバイアス情報を Indicator と呼ばれるメディアバイアス指標として明示化する手法である．

従来のメディアバイアス検出手法の多くは，特定のデータセットを訓練データとしてモデルを学習する教師あり学習に基づいており，ドメイン外のデータセットに対する適応性と汎用性が低いという課題が指摘されている．これに対し IndiVec は，分類器を直接学習するのではなく，バイアス有無を表す Indicator から構成されるデータベースを構築し，これに基づいてバイアスを判定するという点に特徴がある．このような構造により，IndiVec は従来の検出手法より高い適応性および汎用性を達成すると同時に，予測結果の根拠となる指標を明示的に提示できる点で，より高い解釈可能性を実現している．

本節では，IndiVec の基本概念と構成要素について概説する．

3.1.1 IndiVec の概要

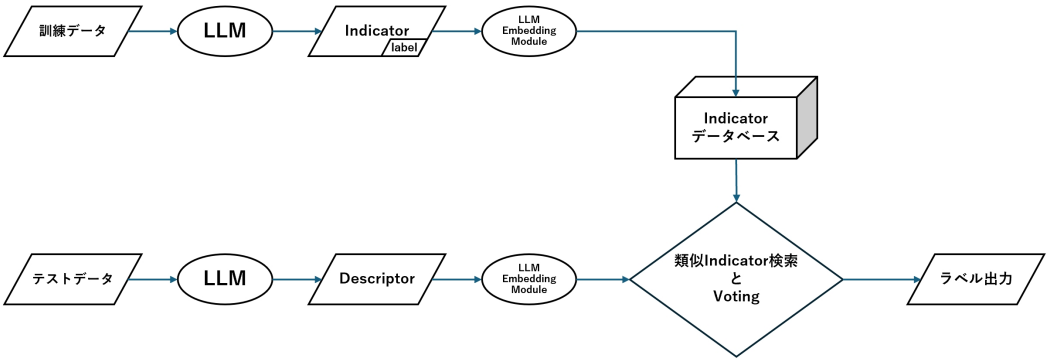

図 3.1: IndiVec の処理フロー

IndiVec の全体的な処理フローを図 3.1 に示す．本フレームワークは，大きく Indicator データベース構築段階とバイアス判定段階の二つの段階から構成される．

まず，Indicator データベース構築段階では，プロンプトを設計し，訓練データにおける文書とバイアスのラベルを LLM に与え，当該文書がバイアスを含むか否か，また含む場合はどのような観点でバイアスを表出しているか，を説明する文章（Indicator）を生成する．生成された Indicator を埋め込み表現に変換し，データベースを構築する．

次に，バイアス判定段階では，テストデータにおける文書に対して同様のプロンプトを LLM に与え，バイアスの有無や特徴を説明する文章（Descriptor）を生成する．ただし，このとき正解ラベルの情報はプロンプトとして与えない．Descriptor も同様に埋め込み表現に変換し，データベース内の Indicator との類似度をコサイン類似度で算出する．各テスト文の Descriptor に対して最も似ている Indicator を 5 つ取得し，それらに付与されたラベルの多数決（Voting）により，テスト文のラベルを決定する．

3.1.2 Indicator と descriptor の生成

IndiVec において，メディアバイアス検出を直接的な分類問題として扱うのではなく，テキスト中にメディアバイアスが表出しうる複数の側面を言語的に記述した中間表現を介してバイアス判定を行う．この中間表現は，訓練データから生成される場合には Indicator，テストデータから生成される場合には Descriptor と呼ばれる．

LLM には，メディアバイアスが表出しうる複数の分析観点を考慮するよう設計されたプロンプトが与えられる．具体的には，文体と語調，情報源の選択，報道のバランス，フレーミング，および例示の用い方といった側面から，バイアスの有無及びその特徴を説明するよう指示される．以下に，この五つの分析観点をまとめる．

バイアスに関する五つの分析観点

1. **Tone and Language**：語調や感情的表現の有無
2. **Sources and Citations**：引用される情報源の選択の偏り
3. **Coverage and Balance**：論点の網羅性および視点のバランス
4. **Agenda and Framing**：特定の解釈を誘導する枠組みの有無
5. **Examples and Analogies**：例示や比喩による偏り

1. Indicator の生成　Indicator は訓練データを用いた few-shot 設定の下で，LLM により生成される．プロンプトには，対象テキストに加え，上記の五つの分析観点に基づいて文書のバイアスの有無を説明する Indicator の例，および対象テキストの正解ラベルが含まれている．これにより，生成される Indicator には上記の五つの分析観点が明示的に反映され，文書のバイアスの有無とその理由を説明することが可能となる．その結果，Indicator は，バイアスの性質を直接的に表現する詳細な説明として機能する．

2. Descriptor の生成　Descriptor の生成においても，Indicator の生成時にほぼ同様のプロンプトを用い，few-shot 設定の下で，LLM により生成される．ただし，プロンプトには，文書の正解ラベルは与えられない点が Indicator の生成時とは異なる．これにより，生成される Descriptor は，Indicator と同一の構造を有する中間表現となる．

以下に，Indicator と Descriptor の生成時に使用するプロンプトを示す．

Prompt Format for Indicator and Descriptor Generation

Example1:
Input:[Biased][content]
Output:[Biased]This article exhibits bias in several ways:
1. Tone and Language:[content]
2. Sources and Citations:[content]
3. Coverage and Balance:[content]
4. Agenda and Framing:[content]
5. Examples and Analogies:[content]

Example2:(With the same format with but different label[Neutral])

Task: Determine why the article is Biased or Neutral.
Give your output in the same format as the examples above:
- **Start with** `Output: [Label]` *where Label is either* `[Biased]` *or* `[Neutral]`.
- **Follow with** a multi-point explanation, analyzing the article based on:
  1. Tone and Language
  2. Sources and Citations
  3. Coverage and Balance
  4. Agenda and Framing
  5. Examples and Analogies
- **Finish with** a summary sentence beginning with `"Overall, ..."`

***Input:** [ ]text_input*

3.1.3 Votingによるメディアバイアス予測

IndiVec のバイアスの有無の判定は，生成された Indicator および Descriptor を埋め込み表現に変換し，類似度計算と Voting を組み合わせることで実現される．

ここで，IndiVec における類似度計算および Voting に用いられる記号を定義する．まず，あらかじめ構築された Indicator 集合を $\mathcal{I} = \{i_1, i_2, \ldots, i_{|\mathcal{I}|}\}$ とする．各 Indicator の $i_j \in \mathcal{I}$ は，埋め込み抽出手法（例：OpenAI Embeddings）を用いて N 次元のベクトル表現 $v_j \in \mathbb{R}^N$ に変換され，Indicator ベクトルデータベース $\mathcal{V}_\mathcal{I} = \{v_1, v_2, \ldots, v_{|\mathcal{I}|}\}$ を構成する．

次に，入力文書 c に対して生成される Descriptor の集合を $D^c = \{d_1^c, d_2^c, \ldots, d_{|D^c|}^c\}$ と定義する．各 Descriptor の $d_j^c \in D^c$ も，Indicator と同一の埋め込み抽出手法を

用いてベクトル表現 $v_j^c \in \mathbb{R}^N$ に変換される．

Descriptor ベクトル v_j^c と Indicator ベクトル v_k（$k \in \{1, 2, \ldots, |\mathcal{I}|\}$）との間の類似度は，次式で定義されるコサイン類似度により計算される．

$$\text{Distance}(v_j^c, v_k) = \frac{v_j^c \cdot v_k}{\|v_j^c\| \|v_k\|} \tag{3.1}$$

この類似度スコアに基づき，Descriptor と最も距離が近い，すなわち意味的に最も近い上位の 5 個の Indicator を検索し，取得する．最終的なメディアバイアスラベルは，これら上位 5 個の Indicator に付与されているラベルを用いた多数決により決定される．この類似度計算・検索と Voting に基づくアプローチにより，IndiVec は未知のデータセットに対しても，類似したバイアス傾向を有する Indicator を参照することで，精度の高い判定を実現している．

本研究では，この IndiVec の特徴となる中間表現生成プロセス（Indicator/ Descriptor 生成）を継承しつつ，次節以降で述べる「粒度整合モジュール」,「粒度アライメント処理」,「BERT 分類器」を導入することで，バイアス検出の性能向上を目指す．

3.2 粒度整合のためのモジュール

IndiVec は，ドメイン外のデータセットに対しても高い適応性を示し，従来の教師あり学習と比較して優れた汎用性を達成しているメディアバイアス検出フレームワークである．一方で，同じドメインのテストデータに対しては，他のバイアス検出手法（特に教師あり学習の手法）と比べて性能が劣ることが指摘されている [4]．

この性能差の一因として，IndiVec が Indicator と Descriptor という中間表現を介してバイアス予測を行う際，それらを生成する元テキストのテキスト粒度が一致していない点が考えられる．すなわち，異なる粒度を有するテキストに対して同一のプロンプトを適用する場合，LLM による理解や生成結果にばらつきが生じ，Indicator と Descriptor の長さの一致性，また意味的対応関係が不十分となる可能性がある．そのため，本研究では両者の違いを考慮したバイアス検出手法を提出する．

上述の IndiVec の特徴となる中間表現生成プロセス（Indicator/Descriptor 生成）を継承したうえで，事前に訓練データとテストデータの文長の差を小さくするアプローチ，すなわちテストデータおよび訓練データの前処理として，テキスト粒度の整合を目的としたモジュールを導入する．具体的には，元のテキストが文のときに情報を補完するテキスト拡張モジュールと，元のテキストが文書のときに冗長な情報を削除するテキスト縮退モジュールを設計し，訓練データとテストデータ間に存在する文書長および意味表現の差異を緩和することを目指す．

以下，テキスト拡張モジュールおよびテキスト縮退モジュールの詳細について説明する．

3.2.1 テキスト拡張モジュール

本研究では，ある文の前後に出現する文は，当該文と一定の意味的関連性を有しており，元の文に対して補助的な情報を提供しうると仮定する．このような補助情報（文脈情報）は，LLM がバイアスの有無や特徴を判断する際の重要な根拠となると考えられる．

テキスト拡張モジュールは，文単位のデータセットを対象とし，個々の文の背景に存在する文脈情報を補完することを目的とする．ここでは，テキスト粒度が文であるデータセットについて，それぞれの文は文書から抽出されていることを仮定する．テキスト拡張モジュールは，データセット内の文について，それが出現する元の文書における前後の文を連結し，長いテキストを生成する処理を行う．これにより，拡張された文集合は文書との文長の差が縮小されると同時に，Indicator・Descriptor 生成時に生じうる情報不足の問題を緩和し，より十分な分析観点を有する中間表現の生成が期待される．

本モジュールの具体的な処理フローを図 3.2 に示す．

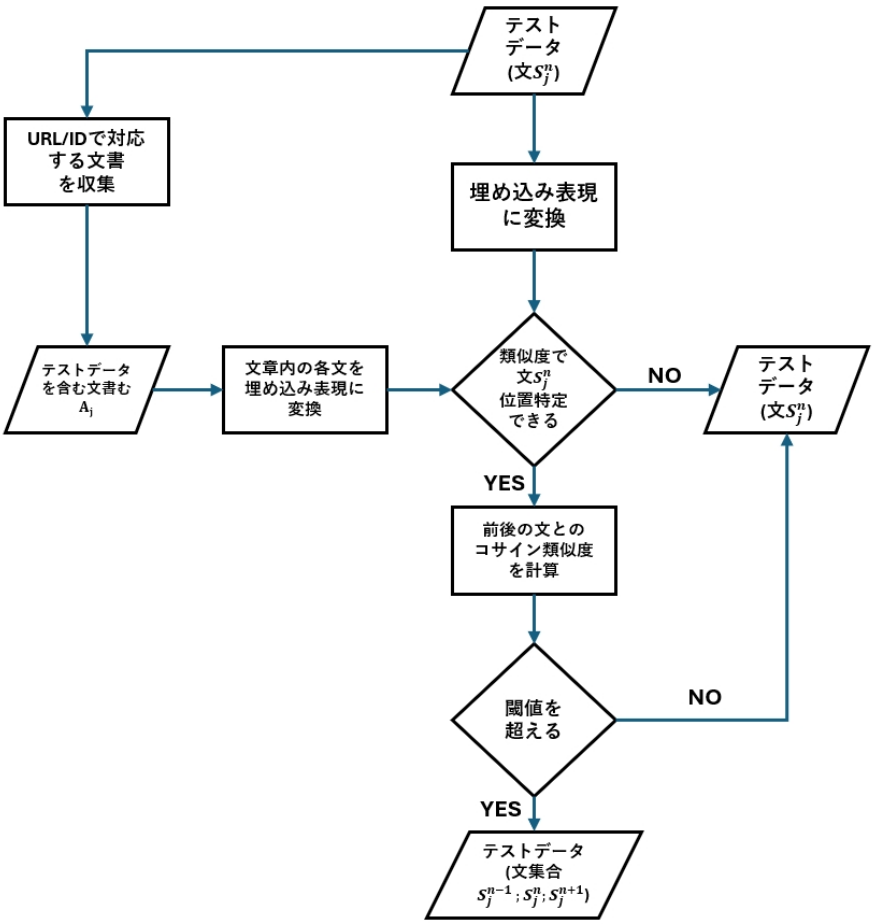

図 3.2: テキスト拡張モジュールの処理フロー

本プロセスは，以下の 3 つのステップで構成される．

1．元文書の収集と埋め込み変換　まず，テストデータとして与えられた文 S_j^n に対し，データセットに付随するメタデータ（URL や記事 ID）を用いて，その文が含まれる元の文書（記事）A_j を特定し，取得する．ここで，j は入力文のインデックスを，n は元文書内における当該文の位置を表す．

続いて，位置特定を行う準備として，入力文 S_j^n および文書 A_j 内に含まれる全ての文を，埋め込みモデルを用いて埋め込み表現に変換する．

なお，元文書を特定できない場合には，テキスト拡張は行わず，元の文 S_j^n をそのまま出力とする．

2．文書内における位置特定　次に，文書 A_j 内における S_j^n の正確な出現位置（インデックス n）を特定する．

テキストの表記ゆれやノイズを考慮し，単純な文字列一致ではなく，ベクトル間

のコサイン類似度計算を用いる．文書内の各文と入力文との類似度を計算し，類似度が最大値の場合，その文を対応する位置 n として特定する．

3. 閾値の設定とテキストの結合　位置 n が特定された場合，その前後の文（S_j^{n-1} および S_j^{n+1}）を拡張候補とする．ここで，無関係な文脈が混入することを防ぐため，意味的整合性の検証を行う．具体的には，元の文 S_j^n と候補文とのコサイン類似度を計算し，その値が事前に設定した閾値 τ を超える場合にのみテキストの結合を行う．

拡張後のテキスト $S_{j\_expanded}$ は以下のように定義される．

$$S_{j\_expanded} = \begin{cases} [S_j^{n-1}; S_j^n; S_j^{n+1}] & \text{if } \mathrm{Sim}(S_j^n, S_j^{n\pm1}) > \tau \\ S_j^n & \text{otherwise} \end{cases} \tag{3.2}$$

ここで，$[;]$ はテキストの連結操作を表す．最終的に，得られた $S_{j\_expanded}$ が中間表現の生成プロセスへの入力となる．

3.2.2　テキスト縮退モジュール

大規模言語モデルは，一般に長文入力に対しても優れた処理能力を有するものの，比較的小規模なモデルにおいては，入力中にバイアス判定タスクに関連しない情報が多く含まれる場合，重要な判断根拠への注意が分散し，生成される Indicator や Descriptor の品質が低下する可能性があると考えられる．そのため，データセットにおけるテキストが文書のとき，その中から重要な情報のみを抽出して文長を短くする処理を行う．本モジュールの処理フローを図 3.3 に示す．

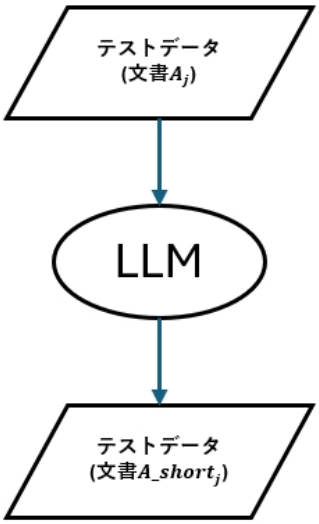

図 3.3: テキスト縮退モジュールの処理フロー

テキスト縮退モジュールは，一般的な要約とは異なり，新たな表現を生成することを目的とせず，入力文書からメディアバイアス判定に寄与する中間表現生成に向けて，より簡潔で，バイアス判定に関連する情報の密度が高いテキストを抽出することを目的としている．同モジュールは，前節のテキスト拡張モジュールとは対照的に，情報量を削減する方向から訓練データとテストデータ間の粒度差の緩和を図るモジュールである．

本研究では，この縮退処理を LLM による zero-shot 設定の生成タスクとして定式化する．具体的には，入力として与えられる文書 A_j に対し，事前に設計したプロンプトを用いて，LLM に対し「メディアバイアスの有無や特徴の判断に無関係と考えられる文を除去し，重要な情報のみを保持した要約を生成する」よう指示する．プロンプトの設計意図は以下の通りである．

- **役割設定:** LLM に対し「メディアバイアス検出の専門家」という役割を付与し，専門的な視点での分析を促す．
- **情報の選別:** ニュース記事を簡略化する際，メディアバイアスの分類に潜在的に関連する文のみを保持するよう指示する．
- **関連性と一貫性:** 「Biased（バイアスあり）」か「Neutral（中立）」かの判定に影響を与えない文を削除対象として明示する一方で，段落としての論理的一貫性を保つよう制約を課す．

具体的には，以下に示すプロンプト P_{sum} を用いる．

テキスト縮退のためのプロンプト (P_{sum})

```
You are a media bias detection expert.
Your task is to simplify the following news paragraph by keeping
only the sentences that are potentially relevant to media bias
classification.
Remove any sentence that does not affect whether the paragraph
is [Biased] or [Neutral], while keeping the paragraph logically
coherent.

News paragraph:
"{text_input}"

Simplified paragraph:
```

このプロンプトは，zero-shot 形式で構成されており，LLM が有する一般的な言語理解能力に基づいて，バイアス判定に寄与する情報を選択的に抽出することを目的としている．

入力文書を A_j，LLM による生成プロセスを LLM$(\cdot)$ とすると，生成される縮退テキスト A_short_j は次式で表される．

$$A\_short_j = \mathrm{LLM}(A_j, P_{sum}) \tag{3.3}$$

その結果として得られる縮退文書 A_short_j は，元の文書 A_j と比較して文長が短縮される一方，バイアスに関連する主要な内容を保持しており，後続の中間表現の生成において，より安定した生成を可能にすると考えられる．

3.3 テキスト粒度アライメント

本節では，3.2 節で述べた「テキスト拡張モジュール」及び「テキスト縮退モジュール」を活用し，訓練データとテストデータのテキスト粒度の差を小さくする手法について説明する．以下，この処理を「テキスト粒度アライメント」と呼ぶ．

1 章で述べた通り，実世界のメディアバイアス検出においては，訓練データとテストデータの間でテキストの「単位（粒度）」が一致しない状況が想定される．このような不整合は，IndiVec における中間表現（Indicator および Descriptor）間の意味的対応関係を弱め，バイアスの有無の予測性能を低下させる一因となると考えられる．

そこで本節では，3.2 節で導入したテキスト拡張およびテキスト縮退モジュールを組み合わせることで，この粒度の不一致を緩和するための具体的な処理プロセスについて述べる．本研究では，データセットの構成に応じて，訓練データのテキスト粒度が文書，テストデータのテキスト粒度が文であるケース I と，訓練データのテキスト粒度が文，テストデータのテキスト粒度が文書であるケース II の 2 つのケースを想定する．そして，それぞれのケースにおいてテストデータのみを

処理する「シングル・アプローチ（Single）」と，訓練データおよびテストデータの双方を処理する「ダブル・アプローチ（Double）」を提案する．

3.3.1 ケースI：訓練データが「文書」，テストデータが「文」の場合

本ケースでは，より長く詳細な背景情報を持つ「文書」から生成された Indicator に対し，比較的短く情報が限定的な「文」から生成された Descriptor を整合させることを目的とし，粒度の不一致がバイアス判定に与える影響を緩和する．具体的には，テストデータのみに対して文書拡張処理を行うシングル・アプローチ (アプローチ I-1) と，テストデータと訓練データの両方について処理を行うダブル・アプローチ (アプローチ I-2) の 2 つのアプローチを設計する．

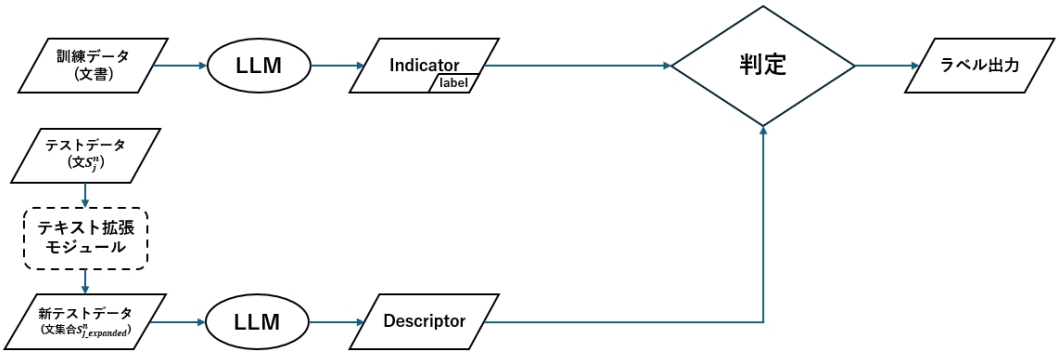

図 3.4: アプローチ I-1：テストデータの拡張による粒度整合

アプローチ I-1：テストデータの拡張　図 3.4 に示すように，本アプローチではテストデータのみに対し前処理を行う．

訓練時には文書単位の訓練データから Indicator を生成する．一方，テスト時には，文単位のテストデータである各文 S_j^n に対し，テキスト拡張モジュールを適用し，周辺文脈を補完した「拡張文（文集合 $S_{j\_expanded}^n$）」を生成する．この拡張プロセスにより，テストデータ側の情報不足を補完し，Descriptor 生成時に，より多様かつ十分な分析観点を抽出することが可能となる．

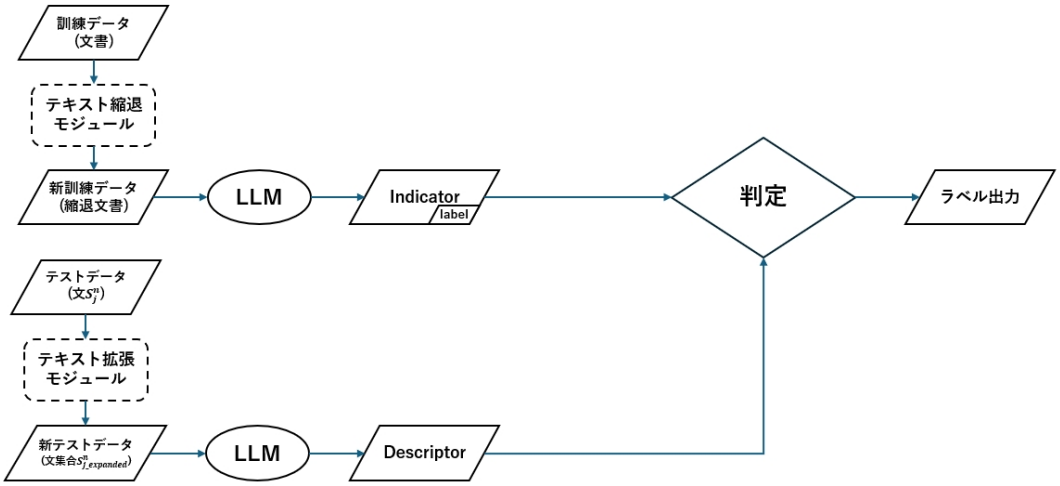

図 3.5: アプローチ I-2：訓練データの縮退とテストデータの拡張の併用

アプローチ I-2：訓練データの縮退とテストデータの拡張　図 3.5 に示すように，本アプローチでは訓練データおよびテストデータの双方に対し前処理を適用する．

具体的には，アプローチ I-1 の処理に加え，訓練データである「文書」に対してテキスト縮退モジュールを適用する．これにより，訓練文書から判定に寄与しない冗長な情報・ノイズを除去した「縮退文書」を生成し，そこから Indicator を生成する．この処理により，Indicator はより判断に関連した情報を含むことが期待される．

これにより，訓練データ側では情報量を削減し，テストデータ側では情報量を補完することで，双方のテキスト粒度を中間的な長さに近づけ，テキスト粒度の不整合を解消する．

3.3.2　ケース II：訓練データが「文」，テストデータが「文書」の場合

本ケースでは，比較的短く局所的な情報を含む「文」から生成された Indicator と，より長く多様な情報を含む「文書」から生成された Descriptor を整合させることを目的とし，粒度の不一致がバイアス判定に与える影響を緩和する．

本研究では，本ケースに対してもケース I と同様に，二つのアプローチを設計する．テストデータのみに対して文書縮退処理を行うシングル・アプローチ (アプローチ II-1) と，テストデータと訓練データの両方について処理を行うダブル・アプローチ (アプローチ II-2) である．

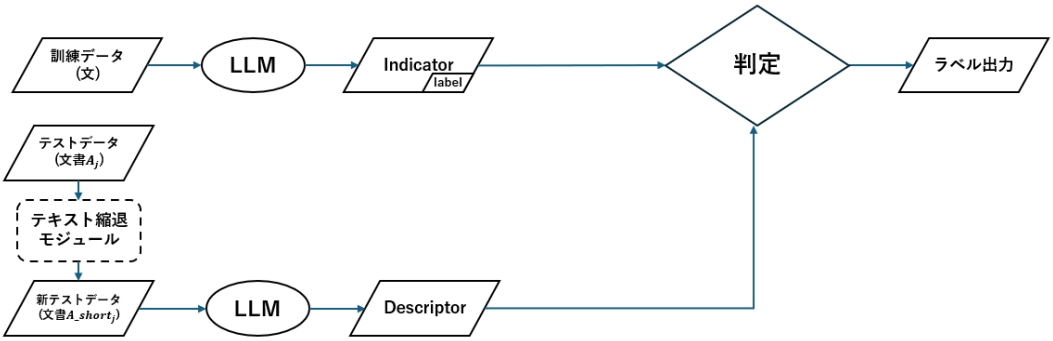

図 3.6: アプローチ II-1：テストデータの縮退による粒度整合

アプローチ II-1：テストデータの縮退　図 3.6 に示すように，本アプローチでは，テストデータのみに対し前処理を行う．

訓練時には，文単位の訓練データから Indicator を生成する．一方，テスト時には，文書単位のテストデータである各文書 A_j に対し，テキスト縮退モジュールを適用し，バイアス判定に寄与しない不要な文を削除した「縮退文書（A_short_j）」を生成する．この縮退プロセスにより，テストデータ側の情報密度を高め，Descriptor 生成時に，より正確な分析観点を抽出することが可能となる．

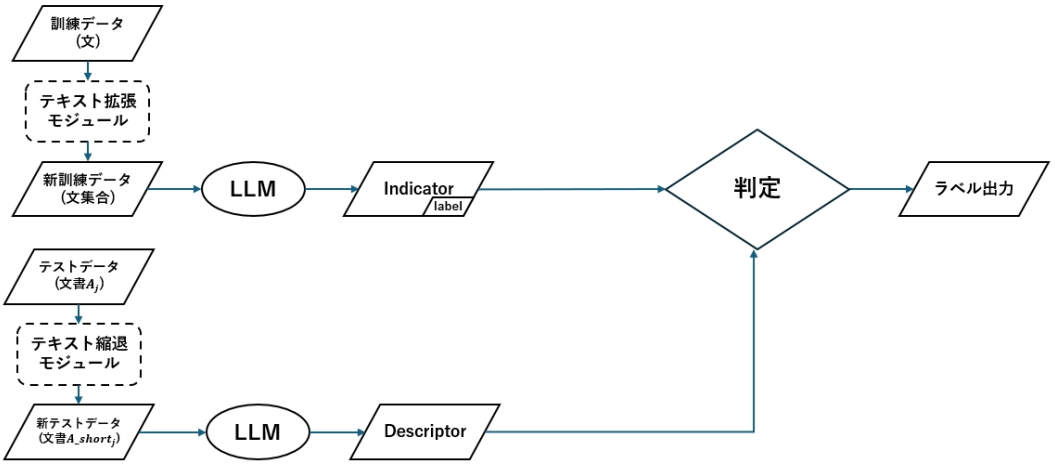

図 3.7: アプローチ II-2：訓練データの拡張とテストデータの縮退の併用

アプローチ II-2：訓練データの拡張とテストデータの縮退　図 3.7 に示すように，本アプローチでは訓練データおよびテストデータの双方に対し前処理を適用する．

具体的には，アプローチ II-1 の処理に加え，訓練データである「文」に対してテキスト拡張モジュールを適用する．これにより，断片的な文に周辺文脈を付与した「文書」を構築し，そこからより情報密度の高い Indicator を生成する．これにより，訓練データ側においても，文単位テキストが持つ局所的な情報の不足を緩和し，より文書レベルに近い情報構造を持つ Indicator を構築することが可能となる．

これにより，訓練データ側では情報量を補完し，テストデータ側では情報量を削減することで，双方のテキスト粒度を中間的な長さに近づけ，テキスト粒度の不整合を解消する．

3.4 BERT を用いたバイアス判定モデル

IndiVec では，生成された Indicator および Descriptor 間のコサイン類似度に基づく多数決（Voting）により，バイアスの有無を判定していた．一方で，この手法は中間表現間の相対的な近さに依存するため，より複雑な意味的関係を十分にとらえられない可能性がある．

これに対し本研究では，長文理解および文脈理解に優れた BERT モデルを分類器として導入する．これにより，Indicator と Descriptor の間の意味的関係をより適切に捉え，メディアバイアス検出の性能が向上するかを検証する．

本節では，BERT モデルの構造を概説し，続いて本研究における入力表現及び学習目的について説明する．

3.4.1 BERT の概要

本研究では，メディアバイアス検出の最終分類器として，Devlin らにより提案された BERT[3] を用いる．

BERT は双方向の自己注意機構を備えており，入力された全トークンの相互関係を全レイヤーで並列的に処理する．バイアス判定タスクにおいては，入力シーケンスの先頭に特殊トークンである `[CLS]` を配置する．この `[CLS]` トークンに対応する最終層の出力ベクトル $C \in \mathbb{R}^H$（H は隠れ層の次元数）は，入力テキスト全体の意味を表すベクトル表現として扱われる．

最終的な分類ラベル y は，このベクトル C を標準的な全結合層および softmax 関数に入力することで得られる．

$$P(y|C) = \text{softmax}(W \cdot C + b) \tag{3.4}$$

ここで，W および b は学習パラメータである．

3.4.2 入力表現と学習時の損失関数

BERT に対する入力は，バイアスの有無を判定する対象テキストそのものではなく，Indicator または Descriptor とする．BERT をファインチューニングする際には，訓練データから生成された Indicator を入力とし，正解のバイアスのラベル (Biased または Neutral) を出力として与える．BERT を推論に用いる際には，テストデータから生成した Descriptor を入力として与える．通常の BERT によるバイアス検出とは異なり，元のテキストの代わりに，LLM によって生成したバイアスの有無を 5 つの観点から説明した文書，すなわち Indicator または Descriptor を入力とする点に提案手法の特徴がある．また，IndiVec とは異なり，3.3 節で述べた手法（4 つのアプローチのいずれか）で粒度を調整したテキストから Indicator または Descriptor を生成する．

実世界のデータセットにおいてはラベル分布が不均衡であることが多いため，ファインチューニング時には，損失としてクロスエントロピー損失関数（Cross-Entropy Loss）を用い，BERT の全パラメータをメディアバイアス検出タスクに最適化する．

第4章　評価

4.1　実験設定

4.1.1　データセット

本研究では，異なるテキスト粒度およびバイアス定義を有する複数の公開データセットを用いて提案手法の有効性を検証する．使用したデータセットの概要を表 4.1 に示す．

BABE [17] および BASIL [18] は，文単位でメディアバイアスがアノテーションされたデータセットである．BABE は各文に対して *Biased* または *Non-biased* の二値ラベルが付与されている．一方，BASIL は文単位で *Lexical Bias* および *Informational Bias* といった異なる種類のバイアスを区別している．本実験ではこれらのラベルを *Biased* と *Neutral* の 2 つに変換して使用する．BABE と BASIL のテキスト長の平均値はそれぞれ 32.6 と 24.1 トークンで比較的短いテキストで構成されている．

これに対し，FlipBias [19] および MFC [20] は，文書単位でラベルが付与されたデータセットである．FlipBias では，同一ニュースイベントに対して政治的立場の異なるメディアによる記事が収集されており，各文書には *Left*, *Center*, *Right* の三値ラベルが付与されている．MFC は，より大規模な文書単位データセットであり，各記事に対して *Pro*, *Neutral*, *Anti* のスタンスラベルが付与されている．これら三値以上のラベルを持つデータセットについては，（*Center/Neutral*）を *Neutral*，それ以外を *Biased* とする二値分類タスクのデータセットとして再構成して実験を行う．FlipBias と MFC のテキスト長の平均値はそれぞれ 909 と 260 トークンであ

表 4.1: 実験で使用するデータセット

Dataset	Bias Level	Bias Label	#Inst.	Avg Len	% Biased
BABE	Sentence	Biased, Non-biased	3,674	32.6	49.3%
BASIL	Sentence	Lexical Bias, Informational Bias , Non-biased	7,984	24.1	19.6%
FlipBias	Article	Left, Center, Right	6,447	909	76.5%
MFC	Article	Pro, Neutral, Anti	37,623	260	84.5%

り，文単位のデータセットと比べてテキストが長く，広範な文脈理解を必要とする点が特徴である．

これらのデータセットは，文単位と文書単位という異なる粒度設定を含んでおり，訓練データとテストデータの単位が一致しない状況を自然に構成できる．したがって，本研究が対象とする粒度不整合の問題設定に対する提案手法の有効性を評価できる．

また，表 4.1 に示すように，BABE における *Biased* ラベルの割合は 49.3% とほぼ均衡しているが，その他の 3 つのデータセットはいずれもラベル分布に偏りが見られる不均衡データセットである．そのため，3.4.2 項で述べた通り，学習時にはクロスエントロピー損失関数を用いることで，ラベル分布の不均衡を考慮した最適化を行っている．

4.1.2 ベースライン手法

本研究では，提案手法の有効性を検証するため，以下の 4 つのベースライン手法と比較実験を行う．

IndiVec [4]
: Lin らにより提案された IndiVec フレームワークをそのまま用いた手法である．訓練データから生成された Indicator を埋め込み表現に変換し，テストデータから生成された Descriptor とのコサイン類似度に基づいて上位 k 個の Indicator を検索し，多数決（Voting）により最終的なバイアスラベルを決定する．本手法は，中間表現を用いたバイアス検出の代表的手法として位置付けられる．

IndiVec (CL)
: IndiVec の中間表現（Indicator/Descriptor）を利用しつつ，Voting による判定を提案手法と同様の BERT 分類器に置き換えた手法である．ただし，本研究で提案するテキスト拡張やテキスト縮退といった前処理モジュールは適用しない．これにより，分類器自体の影響と，提案する粒度整合モジュールの影響を切り分けて評価するを目的としている．CL は Classification の略称である．

FT_BERT
: 中間表現を介さず，ニューステキストの原文を直接 BERT に入力してバイアスを検出する手法である．この手法との比較を通じて，人間のような多角的な分析観点に基づいた中間表現を抽出することの意義を検証する．なお，FT は Fine-tuning を表す．

Zero-shot LLM
: 大規模言語モデルを用いた zero-shot 設定によるバイアス判定手法である．

テストデータをそのまま LLM に入力し，事前に設計したプロンプトを用いて，各テキストが *Biased* か *Neutral* かを判定する．本手法では，訓練データやタスク固有のファインチューニングを一切行わず，LLM が有する一般的な言語理解能力に基づく判定性能を評価する．

4.1.3 評価指標

本研究では，提案手法の有効性を定量的かつ多角的に検証するため，メディアバイアス検出タスクにおいて広く用いられる以下の 4 つの指標を評価に採用する．

- **Precision (精度):** モデルが「Biased」と予測したデータのうち，実際に「Biased」であった割合．
- **Recall (再現率):** 実際に「Biased」であるデータのうち，モデルが正しく「Biased」と予測できた割合．
- **F1-score:** Precision と Recall の調和平均であり，両者のトレードオフを考慮した総合的な指標．

各指標の算出式は以下の通りである．

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.3}$$

ここで，TP，FP，FN はそれぞれ真陽性，偽陽性，偽陰性を表す．

- **Macro-F1** さらに，本研究においては，Macro-F1 を評価指標として導入する．4.1.1 項で述べた通り，メディアバイアス検出に用いられるデータセットは，ラベル分布（Biased / Neutral）が不均衡である場合が多い．単純な正解率（Accuracy）では，多数派のクラス（例：Neutral 事例）を常に予測するモデルが高く評価される．Macro-F1 は，各クラス（本タスクでは Biased および Neutral）の F_1 スコアを個別に算出し，それらの算術平均をとる指標である．

$$\text{Macro-}F_1 = \frac{1}{C} \sum_{i=1}^{C} F_{1,i} \tag{4.4}$$

ここで，C はクラス数，$F_{1,i}$ は各クラスの F_1 スコアを表す（本研究においては $C = 2$）．これにより，サンプル数の少ないクラスの判定能力を平等に評価することが可能となる．

4.1.4 実装詳細

提案手法および各比較手法の具体的な実装条件を以下に述べる．

表 4.2: モデル学習および推論におけるハイパーパラメータ設定

対象モデル	パラメータ項目	設定値
BERT	Base Model	`bert-base-uncased`
	Learning Rate	2×10^{-5}
	Batch Size	8
	Epochs	3
	Loss Function	Weighted Cross-Entropy
Llama-3.1-8B-Instruct	Max New Tokens	300
	Temperature	1.0
	Top-p / Top-k	0.9 / 50
	Number of Returns	1
Embedding & DB	Embedding Model	`all-MiniLM-L6-v2`
	Vector Database	`Chroma DB`

BERT 分類器の設定　本研究の提案手法および比較手法で用いる BERT モデルとして，Hugging Face の `bert-base-uncased`[1]を採用した．モデルのファインチューニングにおけるハイパーパラメータは，表 4.2 に示す通り，学習率を 2×10^{-5}，バッチサイズを 8，エポック数を 3 と設定した．荷重減衰（Weight Decay）は 0.01 とし，最適化アルゴリズムには AdamW を用いた．また，4.1.1 項で述べたラベル分布の不均衡に対処するため，損失関数には重み付きクロスエントロピー損失（WCE）を導入し，少数クラスの分類精度を考慮した学習を行っている．

なお，MFC を除く各データセットは学習データ数が比較的小規模であるため，過学習のリスクよりも十分な学習を優先するため，早期終了（Early Stopping）戦略は適用せず，すべての実験設定において固定エポック数で学習を行った．

LLM の設定　中間表現（Indicator/Descriptor）の抽出，およびテキストの拡張・縮退モジュールで用いる LLM として，オープンソースの大規模言語モデルである `Llama-3.1-8B-Instruct` を使用した．テキスト生成時のパラメータとして，`max_new_tokens` を 300，`temperature` を 1.0，`top_p` を 0.9，`top_k` を 50 に設定した．

本実験では生成される中間表現が事前に定義した出力形式を満たすことを重視した．LLM による生成では，稀に出力形式の不備が発生する場合がある．これを

[1] https://huggingface.co/bert-base-uncased

回避し，有効な中間表現を得るため，各入力に対して最大 5 回までのリトライ処理を実装し，所定の形式を満たした最初の出力を採用した．

Indicator/Descriptor の基盤設定　Indicator と Descriptor を埋め込みに変換する際に用いるテキストの埋め込みモデルとして，計算効率と精度のバランスに優れた `all-MiniLM-L6-v2`[2] を採用した．また，生成された Indicator を格納・検索するためのベクトルデータベースとして `Chroma Vector Database` を構築した．これにより，Descriptor と Indicator 間の高速かつ高精度な意味的類似度検索を実現している．

計算環境　すべてのモデルの学習および推論は，NVIDIA 製 Quadro RTX 5000 を搭載した計算環境で実施し，Python を用いて実装した．

4.2　予備実験：閾値 τ の設定

表 4.3: BABE データセットにおける類似度閾値 τ と拡張サンプル数の関係

類似度閾値 τ	総サンプル数	一つ前の文を拡張	一つ後の文を拡張	拡張されたサンプル数
0.2	5,374	2,429	3,072	3,496
0.3		1,855	2,438	2,929
0.4		1,230	1,634	2,239
0.5		651	907	1,266
0.6		261	410	611
0.7		68	109	171

表 4.4: BASIL データセットにおける類似度閾値 τ と拡張サンプル数の関係

類似度閾値 τ	総サンプル数	一つ前の文を拡張	一つ後の文を拡張	拡張されたサンプル数
0.2	7,984	6,275	6,275	7,504
0.3		4,531	4,531	6,214
0.4		2,720	2,720	4,248
0.5		1,314	1,314	2,259
0.6		444	444	829
0.7		104	104	204

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

3.2.1 項で述べたように，提案手法におけるテキスト拡張モジュールでは，文単位のデータセットを対象とし，対象文の前後に出現する文を元の文に追加する際の判断基準として，類似度閾値 τ を導入した．具体的には，前後文と対象文との意味的類似度が τ を超えた場合にのみ，テキストの結合を行う．

本研究では，適切な閾値を設定するための事前分析として，異なる τ の値に対して，前後文が連結可能と判定されたサンプル数を測定した．BABE および BASIL データセットにおける結果を表 4.3 と表 4.4 にそれぞれ示す．τ の値が大きくなるにつれて，拡張可能と判定されるサンプル数が単調に減少することが確認できる．これは，より厳しい意味的類似度条件を課すことで，入力文と強く関連する文のみが拡張対象として選択されるためである．

特筆すべき点として，BASIL データセットにおいては「一つ前の文を拡張」したサンプル数と「一つ後の文を拡張」したサンプル数が常に同数となっている．これは BASIL がニュース記事を構成する連続した文をサンプルとして持つためである．具体的には，文 S_i が文 S_{i+1} の「前の関連文」として判定される条件は，文 S_{i+1} が文 S_i の「後の関連文」として判定される条件と同一である．一方，文が独立してサンプリングされている BABE では前の文を拡張するケースと後の文を拡張するケースの数は同じではない．

$\tau = 0.6$ および $\tau = 0.7$ の設定では，拡張後のサンプル数が大幅に減少し，テキスト拡張処理の有効性を実験的に検証することが困難であった．そのため，これらの高い閾値設定に対しては，以下に述べるバイアス判定実験は実施していない．

次に，実際のバイアス判定性能に基づいて最適な閾値を決定するため，検証データを用いた実験を行った．この実験では，文書単位のデータセットである FlipBias と MFC を訓練データとし，文単位のデータセットである BABE と BASIL をテストデータとして，テストデータの拡張処理を行うシングル・アプローチによってバイアスを検出する．BABE および BASIL をテストデータとした場合の検証結果をそれぞれ表 4.5 および表 4.6 に示す．$\tau = 0$ の設定では常に拡張し，$\tau = \infty$ の設定では常に拡張しないことを表す．

表 4.5: BABE 検証データを用いた閾値 τ の最適化

Training set	Metric	$\tau = 0$	0.2	0.3	0.4	0.5	∞
FlipBias	Precision	54.0	55.0	**58.2**	58.0	54.5	54.6
	Recall	70.4	75.4	**77.4**	75.5	72.2	69.1
	F1	61.1	63.6	**66.4**	65.6	62.1	61.0
	Macro F1	55.2	56.3	**60.0**	58.7	55.3	56.2
MFC	Precision	56.2	56.7	58.2	**59.6**	57.3	56.7
	Recall	66.1	**71.1**	70.9	71.0	68.5	65.1
	F1	60.8	63.1	64.0	**64.8**	62.4	60.7
	Macro F1	57.9	58.5	59.7	**60.5**	58.7	58.6

表 4.6: BASIL 検証データを用いた閾値 τ の最適化

Training set	Metric	$\tau = 0$	0.2	0.3	0.4	0.5	∞
FlipBias	Precision	20.5	22.2	23.3	**23.5**	22.5	22.1
	Recall	60.1	65.2	**66.9**	64.3	64.7	63.7
	F1	30.6	33.2	**34.5**	34.4	33.4	32.8
	Macro F1	43.0	43.2	45.4	**47.1**	45.8	44.8
MFC	Precision	20.5	23.1	23.0	**23.7**	23.3	23.1
	Recall	53.4	**62.4**	58.6	54.8	58.2	56.1
	F1	29.6	**33.7**	**33.7**	33.1	33.3	32.7
	Macro F1	45.1	46.3	48.3	**49.5**	48.8	48.6

これらの検証データにおける評価，特に Macro F1 の指標に基づき，本研究では訓練データが FlipBias，テストデータが BABE の場合は $\tau = 0.3$ を採用し，それ以外のすべての粒度設定においては $\tau = 0.4$ として，本実験を行う．

4.3 訓練データが「文書」，テストデータが「文」の場合の評価

4.3.1 実験結果

ケース I（訓練データが文書単位，テストデータが文単位）における実験結果を報告する．表 4.7 および表 4.8 は，それぞれテストデータとして BABE および BASIL を用いた場合の評価結果を示している．Proposed(Single) と Proposed(Double) は，それぞれ，シングル・アプローチ (テストデータのみテキスト拡張処理を行う) とダブル・アプローチ (テストデータはテキスト拡張処理，訓練データはテキスト縮退処理を行う) による提案手法を表す．

表 4.7: 実験結果（ケース I）：訓練データが文書，テストデータが文の場合 (テストデータ: BABE)

Training set	Metric	BABE				
		IndiVec	FT_BERT	IndiVec (CL)	Proposed (Single)	Proposed (Double)
FlipBias	Precision	56.3	48.9	55.9	57.7	**61.7**
	Recall	63.2	**100**	67.2	78.3	75.3
	F1	59.5	65.6	61.0	66.4	**67.8**
	Macro F1	58.0	32.8	57.9	59.4	**64.0**
MFC	Precision	**57.3**	47.6	56.6	54.5	53.4
	Recall	66.4	**88.6**	61.1	69.3	69.3
	F1	61.5	**61.9**	58.8	61.0	60.3
	Macro F1	**59.4**	36.7	58.1	55.0	53.5

まず，テストデータとして BABE を用いた結果について述べる．訓練データが FlipBias の場合，Proposed（Double）は F1 が 67.8，Macro F1 が 64.0 と，すべての手法の中で最も高い性能を示した．Proposed（Single）も F1=66.4 と高い値を示しているが，ダブル・アプローチよりやや劣る．

一方，FT_BERT は Recall が 100 と極めて高い値を示すものの，Precision が 48.9 にとどまり，その結果として Macro F1 は 32.8 と大きく低下している．これは，予測が一方のクラスに偏っている可能性を示唆している．

訓練データを MFC に変更した場合でも，同様の傾向が確認できる．IndiVec は Macro F1=59.4 と比較的安定した性能を示しているのに対し，FT_BERT は Recall が 88.6 と高い一方で，Macro F1 は 36.7 にとどまっている．提案手法は，Precision と Recall のバランスを維持しつつ，シングルおよびダブルの両手法で IndiVec フレームワークおよび他手法と同程度の F1 を達成している．ただし，Macro F1 は IndiVec と比べてやや劣る．

表 4.8: 実験結果（ケース I）：訓練データが文書，テストデータが文の場合 (テストデータ: BASIL)

Training set	Metric	BASIL				
		IndiVec	FT_BERT	IndiVec (CL)	Proposed (Single)	Proposed (Double)
FlipBias	Precision	23.7	20.3	22.6	22.9	**25.1**
	Recall	51.0	**100**	57.6	63.8	61.8
	F1	32.4	33.8	32.4	33.7	**35.7**
	Macro F1	47.2	16.9	47.2	46.6	**50.7**
MFC	Precision	22.9	21.2	24.0	**25.2**	24.7
	Recall	51.0	**87.0**	51.6	61.2	61.8
	F1	31.6	34.1	32.7	**35.7**	35.3
	Macro F1	49.2	31.7	50.5	**51.0**	50.1

次に，テストデータとして BASIL を用いた結果について述べる．訓練データが FlipBias の場合，Proposed（Double）は F1=35.7，Macro F1=50.7 を記録し，IndiVec および他手法を上回る性能を示した．Proposed（Single）も F1=33.7，Macro F1=46.6 と，既存手法と同等以上の結果を達成している．

一方で，FT_BERT は BABE の場合と同様に Recall が 100 と非常に高い値を示すが，Precision は 20.3 と低く，Macro F1 は 16.9 にとどまっている．この傾向は，訓練データを MFC に変更した場合でも一貫して観測される．

MFC を訓練データとした場合には，Proposed（Single）が F1=35.7，Macro F1=51.0 と最も高い値を示しており，Proposed(Double) はこれに近い性能を示している．IndiVec（CL）も IndiVec と比較して一貫した改善が確認できる．

表 4.9: 実験結果（ケース I）：ゼロショット LLM との比較結果

Method			Test set	
			BABE	BASIL
Zero-shot LLM			**73.2**	**42.7**
Training set	FlipBias	Proposed (Single)	66.4	33.7
		Proposed (Double)	67.8	35.7
	MFC	Proposed (Single)	61.0	35.7
		Proposed (Double)	60.3	35.3

表 4.9 は，提案手法とゼロショット設定の LLM との性能比較を示している．ゼロショット LLM では，学習データを一切用いず，テスト文をそのまま入力し，バイアスの有無を直接判定させている．

BABE データセットにおいては，ゼロショット LLM が F1 スコア 73.2 を達成しており，FlipBias および MFC を訓練データとする提案手法を上回る結果となった．また，BASIL データセットにおいても，ゼロショット LLM が F1 スコア 42.7 を記録し，提案手法と比較して高い性能を達成している．

この結果は，BABE および BASIL がいずれも文単位で構成されたデータセットであり，比較的短いテキスト中に明示的なバイアス表現を含む傾向が強いことに起因すると考えられる．Llama は大規模な事前学習を通じて，言語的バイアスや評価的表現に関する知識を獲得しており，ゼロショット設定においても，プロンプトとの整合性に基づいた短文レベルの判断を高精度に行うことが可能である．

一方で，本研究の提案手法は，単にバイアスの有無を分類するだけでなく，判断の根拠となる Descriptor を明示的に生成できる点に特徴がある．このような中間表現に基づく設計は，モデルの判断根拠を明らかにし，メディアバイアス検出における説明可能性を高める役割を果たす．

4.3.2 アブレーション研究

本項では，提案手法を構成する各要素の寄与を明らかにするため，アブレーション解析を行う．具体的には，判定モデルを Voting から BERT 分類器へ変更したことによる効果と，本研究の核心であるデータ処理アプローチ（テキスト拡張および縮退モジュール）の有効性について検証する．

4.3.2.1 BERT 分類器の有効性検証

まず，判定モデルを従来の IndiVec が採用している多数決（Voting）から，学習可能な BERT 分類器へ置き換えたことによる有効性を検証する．表 4.7 および表 4.8 において，オリジナルの IndiVec と IndiVec (CL) を比較する．

分析の結果，半数の設定において F1 スコアの向上が確認された．例えば，FlipBias を訓練データとして BABE を評価した場合，F1 スコアは 59.5 から 61.0 へ

と改善した．Recall に着目すると，訓練データが MFC かつテストデータが BABE であるケースを除き，すべての設定において IndiVec (CL) の Recall が IndiVec を上回ることが確認された．一方で，Precision に着目すると，訓練データが MFC かつテストデータが BASIL であるケースを除き，Precision が IndiVec より僅かに劣る傾向も観察された．これらの結果から，BERT 分類器が Descriptor に内在するバイアス表現を柔軟に捉える一方で，Voting に比べて境界付近のサンプルをより積極的にバイアスありと判定する傾向を持つためであると考えられる．

4.3.2.2　データ処理アプローチの有効性検証

本研究の主要な提案であるテキスト拡張および縮退モジュールの有効性を検証するため，粒度整合を行わず Voting を BERT 分類器へ置き換えた IndiVec(CL)，テストデータのみを拡張した Proposed (Single)，および訓練データ・テストデータ双方を処理した Proposed (Double) を比較する．

まず，IndiVec (CL) と Proposed (Single) を比較すると，すべての設定において Proposed (Single) の Recall および F1 スコアが大きく向上した．これは，文単位のテストデータに対し，意味的に関連の強い周辺文を付与することで，バイアス判定に必要なコンテキストが補完され，モデルがより正確にバイアスの存在を捉えられるようになったためと考えられる．

次に，訓練データにテキスト縮退モジュールを適用した Proposed (Double) の結果に注目する．表 4.7 および 4.8 において，FlipBias を訓練データとした場合，Proposed(Double) は Proposed (Single) より Precision，F1，Macro F1 を一貫して向上させた．一方で，MFC 訓練時にはこのような顕著な変化は確認できなかった．

この差は各データセットの平均文長に起因すると考えられる．4.1.1 項で述べた通り，FlipBias の平均長は 909 トークンであり，MFC の 260 トークンと比較して大幅に長い．そのため，FlipBias のような長い文書に対しては，テキスト縮退モジュールによるノイズ除去が BERT 分類器の学習効率を大きく向上させる一方で，もともと比較的短い MFC では情報の圧縮による影響が限定的であったと推測される．

特に，FlipBias を訓練データ，BABE をテストデータとした場合において，粒度整合のない IndiVec (CL) の Macro F1 57.9 に対し，Proposed (Double) が 64.0（+6.1 ポイント）と大幅な性能向上を達成した点は重要である．これは，訓練側の情報の「縮退」とテスト側の情報の「拡張」を同時に行うことで，双方のテキストの粒度を中間的な値で整合させる双方向的なアプローチが，モデルの汎化性能を向上させることを裏付けている．

4.3.3 議論と考察

前項までの実験結果およびアブレーション研究に基づき，本項では提案手法がメディアバイアス検出における粒度不整合問題をいかに解決したかについて深く議論する．各データ処理モジュールの寄与を総括すると，以下の三点に要約される．

1. **拡張モジュールの寄与（情報の補完）**：文単位の判定において不足していたコンテキストを意味的類似度に基づく周辺文の付与によって解消した．これにより，文単体では識別困難な潜在的バイアスの検出（Recall の向上）を可能にした．

2. **縮退モジュールの寄与（ノイズの除去）**：長い文書からバイアス判定に無関係な情報を排除することで，訓練データから生成された中間表現（Indicator）が，文から生成された中間表現（Descriptor）と意味的に高い整合性を持つ状態へと誘導された．これにより，文書全体のノイズが分類モデルに悪影響を与えることを防ぎ，判定性能の向上に寄与した．

3. **双方向整合の相乗効果**：訓練データとテストデータという異なるソースからの入力を同じ「情報の粒度」を持つ共通の表現空間へと誘導した．特に粒度差の大きい FlipBias において顕著な改善が見られた事実は，単一の処理（Single）よりも双方向の処理（Double）が，モデルの汎化性能を最大化させる有効な手段であることを示している．

次に，上記の考察をより詳細に解明する．具体的には，誤分類の統計的傾向に基づく定量的な分析に加え，具体的な事例を用いたケーススタディ，およびモデルが判定に失敗したエラー分析を通じた定性的な考察を行う．これらの多角的な検証を通じて，提案手法の優位性と現在の限界，および今後の改善に向けた示唆を明らかにする．

以下では，ケース I において高い有効性を示した Proposed(Double) について，訓練データに FlipBias，テストデータに BABE を用いたときの実験結果を対象に詳細な分析を行う．

4.3.3.1 誤分類の統計的傾向

まず，誤分類の傾向を統計的に分析する．具体的には，正解ラベルと各手法の予測結果の組み合わせに基づき，以下の四つのケースに分類した．各ケースについてサンプルをランダムに 10 例ずつ抽出し（計 40 例），人手による詳細な確認を行った．

- N✓：正解が Neutral．IndiVec は誤分類したが，Proposed(Double) は正解．
- B✓：正解が Biased．IndiVec は誤分類したが，Proposed(Double) は正解．

- N×：正解が Neutral．Proposed(Double) が誤分類したが，IndiVec は正解．
- B×：正解が Biased．Proposed(Double) が誤分類したが，IndiVec は正解．

第一に，拡張文が元の文のバイアス判定にどの程度寄与しているかを評価するため，拡張部分の有用性を 1〜5 の 5 段階で主観的に評価した．以下，このスコアを「function スコア」と呼ぶ．具体的な評価基準は以下の通りである．

5 段階の評価基準

1：拡張文が元の文とほぼ無関係，または判定に寄与しない
2：周辺情報を補足するが，バイアス判断への影響は限定的
3：背景説明や事実関係の補足として一定の寄与がある
4：バイアスの根拠となる情報（引用，対立構造等）を明確に補強
5：拡張文が判定結果を左右する決定的な情報を含む

結果を表 4.10 に示す.「拡張あり」と「拡張なし」はテキスト拡張処理によって前後の文の付与がされた事例とそうでない事例の数,「拡張失敗」は類似度マッチングに基づく原文特定プロセスにおいて記事内での位置にずれが生じ，拡張された文が実際の前後の文ではない事例の数,「function」は拡張ありの事例に対する function スコアの平均を表す.

表 4.10: 拡張文に対する分析結果

	拡張あり	function	拡張なし	拡張失敗
N✓	6	1.83	4	
B✓	5	3.4	4	1
N×	5	2.6	4	1
B×	4	3.75	5	1

「拡張失敗」の事例が 3 件確認されたが，全体として前後文を追加してもバイアスの有無が変化したと人手で判断された例は 1 例のみであり，拡張アプローチが原文のバイアス構造を損なうことなく文脈を補完できていることが示された．

抽出事例および対応する Descriptor を詳細に確認した結果，拡張された文脈の内容は，Descriptor の各分析項目を明確に反映していることが確認された．具体的には，以下の 3 つの傾向が観察された．

1. **Biased 事例における「希釈効果」**： Biased と正しく判定された例（B✓）において，拡張文の平均点数は 3.4 と高く，内容の背景や詳細な証拠がバイアス予測を強化していることが確認された．一方で，誤判定（B×）の平均点数も 3.75 と高い．人手で判断する際に，最終的な予測の正誤に関わらず，拡張文が元の出来事の背景，証拠，詳細説明を補足する傾向が確認された．

   正しく分類された Biased の事例 $B\checkmark$ の例と対応する中間表現（Descriptor）を確認すると，特に，補完された背景情報が主に判定基準である「*Sources*

and Citations」や「*Coverage and Balance*」を強化し，判定を支える重要な根拠となった．

一方で $B\times$ では，拡張文に出典の明記や対立意見が含まれることで，LLM が「情報のバランスが確保されている」と過剰に解釈し，「*Agenda and Framing*」に反映される傾向が確認された．その結果，原文に含まれる本来のバイアス表現を相対的に軽視してしまう「希釈効果」が生じ，結果として Neutral と誤判定される傾向が確認された．

2. **Neutral 事例の「脆弱性」と伝染効果**： 正しく分類された Neutral の事例 $N\checkmark$ では，追加された文脈が原文に対して例示的な情報や，より具体的な事実，専門家・団体・人物の言及を付加しており，これらの内容が「*Sources and Citations* 」に反映されることが確認された．拡張文脈の中に反対意見や対立的表現が含まれる場合には，「*Coverage and Balance*」における記述が Neutral ラベルに沿う方向で強化される傾向が観察された．

   しかし，誤分類した $N\times$ では，追加された前後の文脈自体に強い扇動的表現が含まれる場合，LLM が「*Agenda and Framing*」において「記事全体に特定の意図がある」と判断してしまう「伝染効果」が見られた．これらの結果は，Neutral な文が周辺文脈の影響に対して相対的に脆弱であることを示唆している．

3. **生成の不確実性**： 同一の入力文であっても，実行ごとに Descriptor の内容や予測結果が異なる例が観察された．これは LLM のサンプリングプロセスにおける確率的揺らぎに起因するものと考えられる．本研究では多様な解釈を許容するために temperature を 1.0 に設定しているが，一貫性の観点からは今後はて temperature 最適化が課題として挙げられる．

総じて，拡張されたテキストの内容は，生成された Descriptor に一貫して反映されており，特に 「*Sources and Citations*」 および「*Coverage and Balance*」 に関する内容はバイアス判定の精度向上に寄与する一方，「*Agenda and Framing*」 に関する情報は誤判定を引き起こす要因となり得ることが示された．具体的な事例については，以降のケーススタディおよびエラー分析において詳述する．

表 4.11 は，Proposed(Double) と IndiVec の全体的なラベル予測の割合を示している．

表 4.11: 全体的な予測傾向の比較

	Biased 予測率	Neutral 予測率
Proposed(Double)	60.7%	39.3%
IndiVec	56.1%	43.9%

テキスト拡張の適用後も予測割合に劇的な変動は見られなかったものの，Biased と判定される割合が約 4.6 ポイント上昇したことが確認された．この結果は，Neu-

tral とラベル付けされた文が，周辺文脈の追加によってバイアスの兆候を含む情報に影響を受けやすく，先に述べたように，追加情報に対して相対的に「脆弱」である可能性を示唆している．

4.3.3.2　ケーススタディ

表 4.12 は，テキスト拡張モジュールの適用によって判定結果が Biased から Neutral へと適正に修正された成功例である．本事例の中間表現（Descriptor）の全文は付録 A.1 に示す．

表 4.12: テキスト拡張による判定精度の向上事例

対象文 (Original Text, Gold: Neutral)
As the Black Lives Matter movement grows, companies like Nike, Target, and Google have made Juneteenth a paid holiday.
最終的判定: Biased (誤判定)

拡張後のテキスト (Merged Text, Threshold: 0.3)
Hundreds gathering to celebrate Juneteenth a few years ago Source / Fox News. As the Black Lives Matter movement grows, companies like Nike, Target, and Google have made Juneteenth a paid holiday. Michelle Gordon, a senior lecturer at Emory University, believes if Juneteenth became a holiday it could be one of the most American of all holidays.
最終的判定: Neutral (正解)

この具体例について考察する．閾値を 0.3 に設定した文脈拡張の結果，対象文は元の文書の冒頭の一文であると判断され，前方には見出しおよび報道機関の出典情報が，後方には原文と関連する専門家の評価文が追加された．

原文のみから生成された中間表現では，「*Sources and Citations*」の観点において情報源が明示されておらず，記述の客観性を判断することが困難であると評価された．また，Black Lives Matter 運動との関連付けが，一方的に肯定的なトーンとして解釈され，結果として Biased と誤判定された．

一方，拡張後テキストから生成された中間表現では，「*Sources and Citations*」に「Fox News」という報道機関名および「Emory University」に所属する専門家が明示的に追加されたことで，当該記述が「事実に基づく報道」であると再定義された．その結果，記事全体の文脈が中立的に評価され，最終的に Neutral と正しく判定された．

4.3.3.3 エラー分析

表 4.13 は，テキスト拡張モジュールの適用によって判定結果が Neutral から Biased へと誤判定された失敗例である．本事例の詳細な中間表現（Descriptor）の全文は付録 A.2 に示す．

表 4.13: テキスト拡張による判定精度の低下事例

対象文 (Original Text, Gold: Neutral)
A Texas law requires contractors who work for or do business with the state to certify that they do not boycott Israel or Israeli-occupied territories.
最終的判定: Neutral (正解)

拡張後のテキスト (Merged Text, Threshold: 0.3)
In response to the BDS movement, 26 states have passed laws seeking to deter businesses and individuals from participating in it. For example, a Texas law requires contractors who work for or do business with the state to certify that they do not boycott Israel or Israeli-occupied territories. The American Civil Liberties Union has filed lawsuits challenging the Texas law and similar laws in three other states, saying they violate the right to free speech.
最終的判定: Biased(誤判定)

この例の対象文は，テキサス州における特定の法制度について事実のみを簡潔に記述したものであり，感情的表現や評価的立場を含まないことから，Descriptor においても 「*Tone and Language*」,「*Agenda and Framing*」 の観点で中立的であると評価され，正しく Neutral と判定された．

しかし，文脈拡張モジュールによって前後文が付加された結果，BDS（Boycott, Divestment and Sanctions）運動に対する州法の動きや，American Civil Liberties Union（ACLU）による訴訟といった，政治的・対立的文脈を含む情報が新たに混入した．この拡張後テキストから生成された中間表現では,「*Coverage and Balance*」および 「*Agenda and Framing*」 の観点において，特定の立場を暗に支持しているとの解釈がなされ，結果として Biased と誤判定された．

4.4 訓練データが「文」，テストデータが「文書」の場合の評価

4.4.1 実験結果

ケース II（訓練データが文単位，テストデータが文書単位）における実験結果を報告する．表 4.14 および表 4.15 は，それぞれテストデータとして FlipBias お

よび MFC を用いた場合の評価結果を示している．

表 4.14: 実験結果（ケース II）：訓練データが文，テストデータが文書の場合 (テストデータ: FlipBias).

Training set	Metric	FlipBias				
		IndiVec	FT_BERT	IndiVec (CL)	Proposed (Single)	Proposed (Double)
BABE	Precision	80.2	**84.1**	79.0	77.3	77.4
	Recall	54.6	39.8	63.2	**63.5**	61.3
	F1	65.0	54.0	**70.2**	69.7	68.4
	Macro F1	48.5	47.4	49.2	**50.2**	49.9
BASIL	Precision	81.3	**82.5**	79.1	77.4	77.5
	Recall	41.2	45.0	**62.2**	61.9	60.9
	F1	54.7	58.3	**69.6**	68.8	68.2
	Macro F1	44.5	49.0	49.2	**50.0**	49.9

まず，テストデータとして FlipBias を用いた結果について述べる．訓練データが BABE の場合，IndiVec（CL）は F1=70.2 を記録し，IndiVec および FT_BERT を大きく上回る性能を示している．また，Proposed（Single）および Proposed（Double）も，それぞれ F1=69.7，68.4 と IndiVec（CL）に近い性能を達成している．Macro F1 においてはそれぞれ 50.2，49.9 という数値が得られ，IndiVec（CL）と比べてわずかな改善が確認できる．

一方，FT_BERT は Precision が 84.1 と高い値を示すものの，Recall が 39.8 にとどまり，F1 は 54.0 と低下している．この傾向は，高精度だが予測が保守的である (Neutral と判定しやすい) ことを示唆している．

訓練データが BASIL の場合でも同様の傾向が確認できる．IndiVec（CL）は F1=69.6 を達成しており，Proposed（Single）および Proposed（Double）も，それぞれ F1=68.8，68.2 と僅差で続いている．Macro F1 においては，Proposed（Single）が 50.0 と最も高い値を示している．この結果より，テキスト粒度の整合性を向上させることで，文単位のデータセットで学習した情報を文書単位のデータセットの判定により効果的に活用できていることが分かる．

表 4.15: 実験結果（ケース II）：訓練データが文，テストデータが文書の場合 (テストデータ: MFC).

Training set	Metric	MFC				
		IndiVec	FT_BERT	IndiVec (CL)	Proposed (Single)	Proposed (Double)
BABE	Precision	84.9	**85.3**	85.1	84.6	84.8
	Recall	49.0	33.7	59.4	**63.5**	60.2
	F1	62.1	48.3	70.0	**72.5**	70.4
	Macro F1	43.3	37.1	46.9	**47.4**	46.7
BASIL	Precision	82.7	**87.9**	85.0	85.0	84.8
	Recall	28.2	18.1	58.9	**61.6**	58.6
	F1	42.0	30.1	69.6	**71.4**	69.3
	Macro F1	33.2	28.7	46.6	**47.4**	46.3

次に，テストデータとして MFC を用いた結果について述べる．訓練データが BABE の場合，Proposed（Single）は F1=72.5，Macro F1=47.4 と最も高い性能を示しており，IndiVec（CL）および他の手法を上回る結果となった．特に Recall が 63.5 と高く，文書におけるバイアスの検出漏れを軽減できている．

同様に，訓練データを BASIL とした場合においても，Proposed（Single）は F1=71.4，Macro F1=47.4 を記録し，すべての手法の中で最も高い性能を示した．FT_BERT は Precision が高い一方で Recall が著しく低く，結果として F1 および Macro F1 の双方で低い値となった．

表 4.16: 実験結果（ケース II）：ゼロショット LLM との比較

Method			Test set	
			FlipBias	MFC
Zero-shot LLM			61.0	64.9
Training set	BABE	Proposed (Single)	**69.7**	**72.5**
		Proposed (Double)	68.4	70.4
	BASIL	Proposed (Single)	**68.8**	**71.4**
		Proposed (Double)	68.2	69.3

表 4.16 は，提案手法とゼロショット設定の LLM との比較結果を示している．ケース II においては，FlipBias および MFC の両テストデータにおいて，提案手法がゼロショット LLM を上回る F1 スコアを達成している．特に，訓練データが BABE の場合，Proposed（Single）は FlipBias に対して F1=69.7，MFC に対して F1=72.5 を記録しており，ゼロショット LLM（FlipBias: 61.0，MFC: 64.9）を大きく上回る結果となった．同様の傾向は，訓練データが BASIL の場合にも確認できる．

この結果は，長い文書を対象としたバイアス判定において，ゼロショット LLM による直接判定よりも，中間表現への変換およびテキスト粒度の整合処理を行う提案手法の方がより適したアプローチであることを示唆している．LLM は広範な事前学習知識を有する一方で，文書内に含まれるバイアスの判定に寄与しない情

報（ノイズ）の影響を受けやすい．これに対し，提案手法はテキスト縮退モジュールを通じて「判定に最も寄与する文」に焦点を絞ることで，文書全体から本質的なバイアス表現を抽出できたと考えられる．

4.4.2 アブレーション研究

本項では，4.3.2 項と同様に，提案手法を構成する各要素の寄与を明らかにするため，アブレーション解析を行う．すなわち，判定モデルを多数決（Voting）から BERT 分類器へ変更したことによる効果と，本研究の核心であるデータ処理アプローチ（テキスト拡張および縮退モジュール）の有効性について検証する．

4.4.2.1 BERT 分類器の有効性検証

まず，判定モデルを従来の IndiVec が採用している多数決（Voting）から，BERT 分類器へ置き換えた効果を確認する．表 4.14 および 表 4.15 において，オリジナルの IndiVec と IndiVec (CL) を比較する．すべての場合において，特に Recall と F1 スコアにおいて劇的な向上が確認された．これに加え，F1 スコアや Macro F1 の向上も確認された．例えば，BASIL を訓練データ，MFC をテストデータとした場合，F1 スコアは 42.0 から 69.6 へと，Macro F1 は 33.2 から 46.6 へと向上した．

この要因として，文単位の訓練データから生成される Indicator の性質が挙げられる．文単位の短いテキストから抽出された Indicator は情報量が少ないと考えられる．従来の IndiVec で用いられている Voting は，個々の Indicator と Descriptor の意味的な距離を評価するため，情報の不足やノイズの影響を受けやすい．

これに対し，BERT 分類器は，事前学習によって獲得された高度な言語理解能力に基づき，5 つの分析観点（側面）の間に存在する相関関係や，文脈に応じた重み付けを動的に学習することが可能である．つまり，Indicator の中から「どの分析観点の組み合わせがバイアス判定に重要か」を柔軟に識別できるようになったことが，バイアス判定の性能向上に繋がったと考えられる．すなわち，情報の粒度が細かい（文単位の）学習データを用いる場合，単純な類似度計算よりも，ニューラル分類器による文脈的な学習が有効であることを示唆する．

4.4.2.2 データ処理アプローチの有効性検証

次に，テキスト拡張および縮退モジュールの寄与を検証するため，IndiVec (CL) と Proposed (Single/Double) を比較する．ケース II において最も顕著な結果は，IndiVec (CL) と Proposed (Single) の性能差である．Proposed (Single) は，テストデータである文書に対して「テキスト縮退モジュール」を適用し，バイアス判定に重要な文のみを抽出する手法である．表 4.15 において，すべての場合においては最高の Macro F1 スコアを達成している．Macro F1 が向上したことは，バイ

アスを含む陽性（Biased）サンプルの検出力が向上しただけでなく，バイアスを含まない陰性（Neutral）サンプルについても，文書内の不要なノイズに惑わされることなく正しく識別できるようになったことを示唆している．

一方で，Proposed (Single) と Proposed (Double) を比較すると，Macro F1 は同程度もしくは Proposed(Double) が僅かに下回る傾向が見られた．Proposed (Double) では，訓練データである「文」に対して拡張モジュールを適用し，テキストの粒度を大きくする処理を行っている．しかし，本実験の結果からは，訓練側の文を無理に拡張するよりも，テスト側の文書を縮退させてテキストの粒度を「文（訓練データセットにおけるテキスト粒度）」に合わせる方が，有効的な粒度不整合解消の手段であると言える．

4.4.3 議論と考察

前項までの実験結果およびアブレーション研究に基づき，本項では提案手法がメディアバイアス検出における粒度不整合問題をいかに解決したかについて深く議論する．各データ処理モジュールの寄与を総括すると，以下の二点に要約される．

1. **縮退モジュールの寄与**： 長い文書からバイアス判定に不要な情報を削除することで，文単位の中間表現で学習した分類器が本来の識別能力を発揮できる環境を整えたことが，判定性能の向上に寄与したと言える．

2. **粒度整合の効果の非対称性**： 訓練データが文，テストデータが文書の場合においては，訓練データの拡張よりもテストデータの縮退が判定の信頼性を高める上でより効果的な戦略であることが証明された．

以下の分析では，4.3.3 項と同様に，誤分類と統計的分析，ケーススタディ，エラー分析を行う．実験で高い性能を示した Proposed (Single) について，訓練データに BABE，テストデータに FlipBias を用いたときの実験結果を対象に詳細な分析を行う．

4.4.3.1 誤分類の統計的傾向

ここでは，4.3.3.1 と同様に，提案手法における縮退モジュールの影響を定性的に分析するため，生成された要約文，対応する Descriptor および予測結果に着目し，統計的かつ定性的な分析を行う．正解ラベルと予測結果の組み合わせに基づき，N✓，B✓，N×，B× の四つのケースに分類し，各 10 事例（計 40 事例）を抽出して分析を進める．

まずはテキスト縮退処理によって得られた要約文に着目する．要約文が原文の情報をどの程度保持しているかを,「意味内容の正確性（情報の欠落）」および「政

治的傾向の保持」の 2 観点から 5 段階で主観評価した．具体的な評価基準を以下に示す．

5 段階の評価基準

1：核心的情報の欠落，事実誤認，または政治的立場の逆転が認められる
2：情報の断絶や引用・対立意見の消失により，原文の論理や傾向が著しく変質している
3：主要な論点は把握可能だが，一部の重要な文脈の欠落により解釈の多義性が生じている
4：軽微な詳細の省略はあるが，原文の論旨および政治的傾向を概ね正確に反映している
5：原文の事実関係，論理構造，およびバイアスの有無を完全に保持している

2 つの観点による評価点の平均を表 4.17 に示す．

表 4.17: 要約文に対する質的分析結果

	意味内容の正確性	政治的傾向の保持
N✓	5	4.8
B✓	4.7	5
N×	4.4	4.8
B×	4	3.4

分析の結果，正解例（N✓, B✓）では要約文が原文の政治的傾向（バイアスの有無）を極めて高い水準で保持していることが確認された．一方で誤判定例（B×）では，政治的傾向の保持スコアが 3.4 と低く，要約の過程でバイアス判定に不可欠な表現が欠落していることが示唆された．詳細を以下に述べる．

1. **Biased 事例における「中立化」現象**：正解ラベルが Biased であるにもかかわらず誤判定された事例（B×）では，要約プロセスによる情報の欠落が顕著であった．具体的には，1. 偏見的な見解，2. 強い政治的背景を持つ引用，3. 批判的なコメント，といったバイアスの根拠となる表現が削除されたことで，文章が「中立化」してしまったことが判明した．これにより，政治的傾向の保持スコアが大幅に低下し，誤判定を招いたと言える．

2. **Neutral 事例における成功と失敗の要因**：正解ラベルが Neutral で予測も正解した事例（N✓）では，要約によって「政治的傾向とは無関係な出来事」や「冗長な引用」が効果的に除去されている．これにより，ノイズが排除されたことが正しい予測に寄与したと考えられる．

   一方で誤判定された事例（N×）では，要約文自体は原文の傾向を保持しているものの，具体的な情報源や引用元が省略されたり，事実の一部のみが抽出されたりしたことで，LLM が Descriptor 生成時に「情報の偏り（不十分な根拠）」と誤認し，判定が Biased に変化した可能性が高い．

次に，要約による情報圧縮の度合いを評価するため，以下の式で定義される縮退率（CR）を算出した．

$$CR = 1 - \frac{\text{要約文の文字数}}{\text{原文の文字数}}$$

N✓，B✓，N×，B× のそれぞれのカテゴリについて，該当するサンプルの平均 CR を表 4.18 に示す．

表 4.18: カテゴリごとの平均圧縮率（CR）

	平均圧縮率 (CR)
N✓	0.6125
B✓	0.4789
N×	0.3810
B×	0.4310

表 4.18 より，特に正解が Neutral 事例において，正解例の CR（0.61）が不正解例（0.38）を大きく上回っていることが分かる．これは，圧縮率が高まるほど政治的傾向とは無関係な「ノイズ（不要な背景情報や引用）」が効果的に除去され，LLM が正確な Descriptor を生成しやすくなったためと考えられる．

原文の平均長を確認したところ，正解例が 4456 文字であったのに対し，不正解例は 4908 文字と，不正解例の方が文長が大きい傾向が見られた．原文の長さが長くなるほど，要約の過程でバイアスを示唆する重要な情報が失われるリスクが高まる傾向も確認された．

さらに，縮退モジュールを適用した Proposed (Single) と，原文をそのまま入力した IndiVec が生成した Descriptor の品質を比較し，要約処理が Descriptor の生成に与えた影響を考察する．Descriptor の品質は，元の記事の内容に対する正確性と網羅性の観点から 5 段階で主観評価した．評価基準を以下の通り定義する．

Descriptor の品質評価基準

1：内容が無関係，または矛盾している
2：大きな誤解や誤りを含んでいる
3：部分的に正確だが，重要情報の欠落や一般論への偏りが見られる
4：概ね正確で主要内容を網羅しているが，細部に欠けがある
5：正確かつ包括的で，記事の核となる概念と詳細を完全に捉えている

評価結果を 4.19 に示す．

表 4.19: 中間表現（Descriptor）の品質評価比較

	Proposed(Single)	IndiVec
N✓	4.3	2.3
B✓	3.2	2.3
N×	1.6	4.6
B×	2.7	3.6

Proposed(Single) は IndiVec と比べて正しい判定をするときは Descriptor の品質が高く，判定を誤ったときは低い傾向が見られる．また，同一の Descriptor 内において 5 つの評価項目が常に同一の傾向を示すわけではなく，LLM がハルシネーションを生じさせているわけではないことが観察された．各事例における具体的な考察は以下の通りである．

1. **Biased 事例における「頑健性」**：正解ラベルが Biased の場合，縮退モジュールを適用した Proposed (Single) と，原文に基づく IndiVec とで，Descriptor の品質スコアにそれほど大きな差は見られなかった．最終的な予測が誤っていた事例であっても，生成された Descriptor には部分的に正確な判断が含まれているケースが多く観察された．

   特に「*Agenda and Framing*」については，記事の背後にある意図や構造的なフレーミングを一貫して捉えることができており，これは要約文と原文の政治的傾向が概ね一致していることに起因すると考えられる．また，LLM は否定的・攻撃的，あるいは感情的な単語の検出を得意としており，これらを「*Tone and Language*」の判断根拠として効果的に利用していることが確認された．

   一方で，要約の過程で一部の論点や引用情報が省略されるため，「*Sources and Citations*」に関する正確性は，原文に基づく場合と比較して低下する傾向にある．しかし，本分析においては，この情報の欠落が「Biased」という最終判定を下す能力に致命的な影響を及ぼすケースは限定的であった．

2. **Neutral 事例における「脆弱性」と誤誘導**：正解ラベルが Neutral の事例では，正しく分類された例（N✓）と誤分類された例（N×）の間で，品質スコアに顕著な差異が確認された．特に誤分類例（N×）の品質スコアが 1.6 と極めて低い値を示したことは注目に値する．

   この要因は，要約の過程で関連する引用や根拠情報，および対立的な観点の記述が削除されたことにある．その結果，LLM は「関連する引用や反対意見が十分に提示されていない」と判断する傾向が観察された．この特徴は「*Coverage and Balance*」の評価にダイレクトに反映され，中立的な記事を「偏った観点しか提示していない」と見なしてしまうことがあった．

このように，反対意見の欠落という要約由来の情報不足が，LLM を「Biased」判定へと誤誘導する主要因となっている可能性が高い．事実，正しく Neutral と判定された事例（$N\checkmark$）においても，「*Coverage and Balance*」の評価は必ずしも高くなく，要約による情報の欠如が Neutral 判定の安定性を損なう「脆弱性」とみなせることが明らかとなった．

表 4.20 は Proposed(Single) と IndiVec の全体的な予測ラベルの分布を示している．

表 4.20: 全体的な予測傾向の比較結果

	Biased 予測率	Neutral 予測率
Proposed(Single)	63.4%	36.6%
IndiVec	53.5%	46.5%

Proposed (Single) では IndiVec と比較して，Biased と判定される割合が約 10 ポイント増加し，一方で Neutral と判定される割合が同程度減少していることが確認された．この結果は，先に分析した Neutral 事例における「脆弱性」と深く関連していると考えられる．Neutral な文章においては，要約の過程で引用情報や反対意見（多角的な視点）が省略されやすく，その結果，生成された Descriptor が「*Coverage and Balance*」において不十分であると LLM に判断され，判定が Biased 側へシフトする傾向が強まったと言える．

一方で，本来 Biased な文章においては，こうした要約による情報欠落の影響は相対的に小さい．バイアス判定の根拠となる感情的な単語や特定の政治的フレーミングといった「強いシグナル」は，縮退モジュールを経ても文章の核として保持されやすいため，判定結果に大きな変動は生じにくいと推測される．

4.4.3.2 ケーススタディ

ここでは，長い文書に対してテキスト縮退モジュールを適用することで，判定結果が Neutral から Biased へと適正に修正された成功例を考察する．この例では，トランプ陣営に対する監視権限の乱用を告発する共和党作成のメモ公開をめぐり，公開を推進する下院共和党・トランプ政権と，内容の不正確さや機密保持の観点から反対する FBI・司法省との対立に関する記事が判定対象となっている．元文書，要約，およびそれらに対応する中間表現の全文は，長いため，見やすさのために付録 B.1 に示す．

テキスト縮退の結果，得られた要約は原文中の文のみから構成され，文字数は原文の約 1/4（6209 から 1563）にまで削減された．この過程において，以下の情報が主に削除されている．第一に，背景的な分析要素として，「ソ連の KGB」との類比表現が削除された．第二に，具体的な証拠の詳細として，Fusion GPS，ヒラリー・クリントン陣営による資金提供，および捜査官 Strzok と Page の短信ス

キャンダルに関する記述が除去された．第三に，出典情報として，括弧内に付記されていた引用元（例：*via The Hill*，*via The Washington Post*）が削除された．

原文から生成された中間表現では，「*Sources and Citations*」 の観点において，FBI の立場やホワイトハウスの対応方針が言及され，*The Washington Post* や *The Hill* といった複数の報道機関が引用されている．この形式的な「A の主張に対して B が反論する」という構造により，LLM は記事全体の客観性を過大評価する傾向を示した．その結果，「*Agenda and Framing*」 の観点からは，右派的，あるいは親トランプ的なバイアスを示唆する根本的な語りの枠組みが相対的に捉えにくくなっており，結果として記事に内在する強い政治的バイアスが看過され，最終的に Neutral] と誤判定された．

一方，重要文抽出による縮退の結果，ソ連の KGB との類比や特定の捜査官のスキャンダルといった「扇動的な背景情報」および「客観性を装うための補助的な引用」が削除された．このプロセスにより，記事の論理構造は「FBI への攻撃とトランプ政権の正当化」に単純化された．その結果，「war」「last-ditch effort」「damning」といったバイアスを示唆する単語が短いテキストで強調される形で残された．中間表現においても，当該記事の基本的な論述の構造がトランプ政権寄りであることが明確に把握され，特に「*Agenda and Framing*」 の観点に強く反映された結果，最終的に Biased と正しく判定された．

4.4.3.3　エラー分析

ここでは，長い文書に対してテキスト縮退モジュールを適用した結果，判定が Neutral から Biased へと誤判定された失敗例を考察する．. この例では，2012 年の米国大統領選挙におけるオバマ候補とロムニー候補による第 2 回大統領討論会を前にした事前の戦略分析および情勢報道に関する記事が判定対象となっている．元の文書，要約，およびそれらに対応する中間表現の全文は長いため，見やすさのために付録 B.2 に示す．

テキスト縮退の結果，得られた要約は原文中の文のみから構成され，文字数は原文の約 1/7（8457 から 1221）にまで大幅に削減された．原文では，民主党側（Gibbs，Axelrod，Cardona）と共和党側（Gillespie，Portman，Navarro）の双方の発言が複数引用されており，両者の発言数のバランスが保たれていた．また，オバマ候補の第 1 回討論会に対する「評価が芳しくなかった」という言及も，当時広く共有されていた事実に基づくものであり，特定候補への攻撃的表現ではなかった．一方，縮退後の要約では引用構造に偏りが生じ，民主党関係者の発言が全体の約 75% を占める構成となった．この変化は，原文において確保されていた引用の量的・質的な均衡が部分的に損なわれる結果となった．

この構造的変化は生成された Descriptor にも明確に反映された．まず，「*Tone and Language*」 の観点では，「languid」という表現が「lazy（怠惰）」と同義に解釈され，否定的・攻撃的な単語として誤って評価された．しかし，2012 年当時の政治

的文脈において，これらの単語は第 1 回討論会での消極的なパフォーマンスを要約するための事実記述的表現であり，縮退による文脈情報の欠落が，単語の誤解釈を引き起こしたと考えられる．次に,「*Sources and Citations*」 の観点では，オバマ陣営の補佐官（Gibbs，Axelrod）の引用が「オバマに有利な視点を提示するための意図的操作」として解釈された．一方で，縮退後テキスト内に残存していたロムニー陣営の反論といった重要な対抗的要素は過小評価され，引用した発言数の偏りのみを根拠として偏向ありと断定された．さらに,「*Coverage and Balance*」の観点では，縮退過程において，ロムニー候補の戦略的優位性や弱点を分析する段落が削除された結果，残されたテキストが「オバマ候補の改善」に過度に焦点を当てたフレーミングであると誤認された．これは，長文の一部を切り出す操作が，意図せず記事全体の均衡構造を歪めてしまうリスクを示している．

以上より，本来は Neutral であった文章が，大幅なテキスト縮退によって民主党と共和党の意見数のバランスが失われ，その結果として生じた「不自然な欠落」が中間表現に反映され，著者の政治的意図として誤って解釈されたと考えられる．このような誤認識により，最終的に Biased と誤判定された．

第5章　おわりに

5.1　本論文のまとめ

本論文では，ニュースメディアにおけるバイアス検出において，訓練データとテストデータの情報単位（粒度）の不一致によってバイアス検出モデルの性能が低下する問題（粒度不整合問題）に着目し，これを解決する手法を提案した．

提案手法では，既存の IndiVec をベースにしつつ，テキストの単位が「文」である場合には拡張モジュールを，単位が「文書」である場合には縮退モジュールを適用した．これにより，訓練データとテストデータのテキスト粒度を近づけたうえで，LLM を用いて中間表現を生成した．テストデータのみテキスト粒度を調整する「シングル・アプローチ」と，訓練・テストデータの両方に対してテキスト粒度を調整する「ダブル・アプローチ」を提案した．また，従来の多数決方式（Voting）を BERT 分類器に置き換えた．文単位のデータセットと文書単位のデータセットを用いた評価実験により，提案手法の有効性を確認した．

本研究の主要な成果および結論は，以下の通りである．

第一に，粒度不整合を解消するために提案したテキスト拡張モジュール・テキスト縮退モジュールの有効性を実証した．評価実験の結果，特に訓練データが文書単位でテストデータが文単位であるケースにおいて，双方にモジュールを適用する「ダブル・アプローチ」によってバイアス検出の性能が大きく向上したことを確認した．これにより，訓練データとテストデータテキストの粒度を合わせることの重要性が示された．また，訓練データが文単位でテストデータが文書単位であるケースについては，テストデータのテキスト縮退による「シングル・アプローチ」がバイアス検出に特に有効に機能することを明らかにした．

第二に，オリジナルの IndiVec の Voting による判定モジュールを BERT 分類器に置き換える改良を行い，その有効性を検証した．実験の結果，LLM が生成した中間表現を BERT で分類するアーキテクチャは，テストデータと類似した中間表現による単純な多数決よりもバイアスの微細な特徴やニュアンスを捉える上で優れており，バイアス検出の性能向上に寄与することを確認した．

第三に，詳細なケーススタディとエラー分析による提案手法の定性的分析を行った．この分析を通じて，テキスト拡張がバイアスを中和してしまう「希釈効果」や，テキスト縮退によって中立的な文脈が失われることで生じる「中立事例の脆弱性」など，テキスト粒度調整に伴う重要な副作用を確認した．これらの知見は，バイアス検出における説明可能性を向上させるだけでなく，今後の研究におけるモデ

ル設計の指針となる有益な情報を提供するものである．

以上の通り，本研究はテキスト粒度のアライメントという新たな視点をメディアバイアス検出に導入し，その有効性を定量的・定性的な両面から明らかにした．本手法は，自動バイアス検出システムの信頼性を高めるだけでなく，将来的にニュース読者が情報の偏向性を客観的に理解するための強力な支援技術となることが期待される．

5.2 今後の課題

本論文の今後の課題を述べる．実用化およびさらなる精度向上に向けて，主に四つの課題が残されていると考えている．

第一に，中間表現の生成に用いる LLM の規模の拡大が挙げられる．本研究では計算リソースの制約により，LLM として `Llama-3.1-8B-Instruct` を使用したが，この規模のモデルではプロンプトの指示を完全には遂行できず，中間表現の生成においてフォーマットの逸脱や余計な内容の混入が散見された．今後は，70B 以上の大規模パラメタを持つモデルや GPT-4 などのより高度な推論能力を持つモデルを用い，モデルの規模がバイアス検出の性能や中間表現の質に与える影響を精査する必要がある．

第二に，より多様なデータセットを用いた提案手法の有効性の検証である．本研究では先行研究との比較のため BABE，BASIL，FlipBias，MFC の 4 つのデータセットを対象としたが，提案手法の汎用性を確立するためには，NewsWCL50[21] のような異なるドメインや政治的背景を持つデータセットに対しても同様の有効性が得られるかを確認することが不可欠である．

第三に，ルールベースと意味的類似度を組み合わせたハイブリッドなテキスト拡張モジュールの探求である．本研究のテキスト拡張モジュールではコサイン類似度に基づいた判定を採用したが，今後は固有表現抽出（NER）を活用し，特定のエンティティ（地名，団体名，人名等）に着目したルールベースの手法の導入が考えられる．具体的には，特定のメディアバイアス辞書に基づいてエンティティを検索し，前後文から同一のエンティティを抽出してコンテキストを動的に拡大する，あるいはテキストに追加する文脈長を調整するといったアプローチにより，ノイズ混入を抑制しつつ精緻な拡張手法を探求したい．

第四に，Neutral 事例の脆弱性に対する対策とプロンプトエンジニアリングの改善である．本研究のエラー分析を通じて，特にテキスト縮退モジュールの適用時に Neutral 事例が誤判定されやすい傾向が明らかになった．今後は，Neutral 事例が持つ言語的・構造的特性を詳細に分析し，判定に不可欠な「均衡情報」を保持するためのより適切なプロンプト設計や，中立性の判定基準を再定義した推論プロセスの開発が求められる．

参考文献

[1] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, Vol. 237, p. 121641, 2024.

[2] Jeanette B Ruiz and Robert A Bell. Understanding vaccination resistance: vaccine search term selection bias and the valence of retrieved information. *Vaccine*, Vol. 32, No. 44, pp. 5776–5780, 2014.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[4] Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. Indivec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1038–1050, 2024.

[5] Felix Hamborg, Karsten Donnay, and Bela Gipp. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, Vol. 20, No. 4, pp. 391–415, 2019.

[6] Jakob-Moritz Eberl, Markus Wagner, and Hajo G Boomgaarden. Party advertising in newspapers: A source of media bias? *Journalism Studies*, Vol. 19, No. 6, pp. 782–802, 2018.

[7] M. Gardner. *The Night Is Large*. St. Martin's Griffin, 1997.

[8] Matthew Gentzkow, Jesse M Shapiro, and Daniel F Stone. Media bias in the marketplace: Theory. In *Handbook of media economics*, Vol. 1, pp. 623–645, 2015.

[9] Tim Groeling. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, Vol. 16, No. 1, pp. 129–151, 2013.

[10] Jakob-Moritz Eberl, Hajo G Boomgaarden, and Markus Wagner. One bias fits all? three types of media bias and their effects on party preferences. *Communication Research*, Vol. 44, No. 8, pp. 1125–1148, 2017.

[11] Robert M Entman. Framing bias: Media in the distribution of power. *Journal of communication*, Vol. 57, No. 1, pp. 163–173, 2007.

[12] Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. Detecting media bias in news articles using Gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4290–4300, 2020.

[13] Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. Target-aware contextual political bias detection in news. *arXiv preprint arXiv:2310.01138*, 2023.

[14] Martin Wessel, Tomas Horych, Terry Ruas, Akiko Aizawa, Bela Gipp, and Timo Spinde. Introducing mbib-the first media bias identification benchmark task and dataset collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2765–2774, 2023.

[15] Zehao Wen and Rabih Younes. Chatgpt vs media bias: A comparative study of gpt-3.5 and fine-tuned language models. *arXiv preprint arXiv:2403.20158*, 2024.

[16] Chen Wei Kuo, Kevin Chu, Nouar AlDahoul, Hazem Ibrahim, Talal Rahwan, and Yasir Zaki. Neutralizing the narrative: Ai-powered debiasing of online news articles. *arXiv preprint arXiv:2504.03520*, 2025.

[17] Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. Neural media bias detection using distant supervision with babe-bias annotations by experts. In *Findings of the association for computational linguistics: EMNLP 2021*, pp. 1166–1177, 2021.

[18] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. In plain sight: Media bias through the lens of factual reporting. *arXiv preprint arXiv:1909.02670*, 2019.

[19] Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. Learning to flip the bias of news headlines. In *Proceedings of the 11th International conference on natural language generation*, pp. 79–88, 2018.

[20] Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pp. 504–514, 2019.

[21] Felix Hamborg, Anastasia Zhukova, and Bela Gipp. Automated identification of media bias by word choice and labeling in news articles. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 196–205, 2019.

付 録A　バイアス判定の事例（ケースI）

A.1　ケーススタディ：テキスト拡張による判定改善事例

対象文（Original Text）

正解ラベル：Neutral
As the Black Lives Matter movement grows, companies like Nike, Target, and Google have made Juneteenth a paid holiday.

Descriptor（Original）

Prediction: Biased

1. *Tone and Language*: The article lacks a balanced tone and uses celebratory language, indicating a positive stance on the companies' decision.
2. *Sources and Citations*: No sources or citations are provided, making it difficult to assess the objectivity of the information.
3. *Coverage and Balance*: The article focuses solely on companies embracing Juneteenth as a paid holiday, neglecting any potential criticisms or opposing viewpoints.
4. *Agenda and Framing*: The article frames the companies' decision as a positive step, without exploring alternative perspectives or considering the broader implications.
5. *Examples and Analogies*: No specific examples or analogies are used, but the tone suggests a pro-Juneeteenth narrative.

Overall, this article exhibits bias by promoting a positive view without presenting a balanced analysis.

拡張後テキスト（Merged Text）

Hundreds gathering to celebrate Juneteenth a few years ago Source / Fox News. As the Black Lives Matter movement grows, companies like Nike, Target, and Google have made Juneteenth a paid holiday. Michelle Gordon, a senior lecturer at Emory University, believes if Juneteenth became a holiday it could be one of the most American of all holidays.

Descriptor（After Expansion）

Prediction: Neutral

1. *Tone and Language*: The article presents a factual account of companies making Juneteenth a paid holiday without expressing a personal opinion or emotion.

2. *Sources and Citations*: The source, Emory University, provides an expert opinion without promoting a specific agenda or ideology.

3. *Coverage and Balance*: The article mentions Nike, Target, and Google as examples, and quotes Michelle Gordon, a neutral source. It doesn't favor one company over another.

4. *Agenda and Framing*: The article simply presents information about companies making Juneteenth a paid holiday without trying to persuade the reader.

5. *Examples and Analogies*: The article provides factual examples of companies taking action, but does not use metaphors or emotional comparisons.

Overall, the article remains neutral by presenting facts and quotes from a neutral source without expressing bias.

A.2　エラー分析：テキスト拡張に起因する誤判定事例

対象文（Original Text）

正解ラベル：Neutral
A Texas law requires contractors who work for or do business with the state to certify that they do not boycott Israel or Israeli-occupied territories.

Descriptor（Original）

Prediction: Neutral

1. *Tone and Language*: The tone is informative and objective. It states the law and the requirements without emotion or bias.
2. *Sources and Citations*: The article doesn't rely on emotional or ideological sources; instead, it focuses on the law and its implications.
3. *Coverage and Balance*: The article doesn't take a stance on the boycott or the law; it merely presents the information as a factual matter.
4. *Agenda and Framing*: The framing is straightforward, presenting the law as a neutral fact without advocating for or against it.
5. *Examples and Analogies*: The article doesn't use analogies or emotionally charged comparisons; it sticks to the facts of the law.

Overall, this article is a neutral, informative piece that presents the law and its requirements without bias.

拡張後テキスト（Merged Text）

In response to the BDS movement, 26 states have passed laws seeking to deter businesses and individuals from participating in it. For example, a Texas law requires contractors who work for or do business with the state to certify that they do not boycott Israel or Israeli-occupied territories. The American Civil Liberties Union has filed lawsuits challenging the Texas law and similar laws in three other states, saying they violate the right to free speech.

Descriptor (After Expansion)

Prediction: Biased

1. *Tone and Language*: The article presents the BDS movement as an opposition to Israel without adding any context that might justify the boycott, and states that 26 states have passed laws against the BDS movement, which may not be neutral as it could be interpreted that the BDS movement is doing something bad.

2. *Sources and Citations*: The article only cites the American Civil Liberties Union (ACLU) without presenting any opposing views or opinions.

3. *Coverage and Balance*: The article does not provide equal coverage of the BDS movement and the laws against it. The focus is primarily on the laws and the ACLU's response, implying that the laws are unjust.

4. *Agenda and Framing*: The framing of the article suggests that the BDS movement is under attack and that the laws are an attempt to silence free speech, without providing any context about the boycott or the reasons behind it.

5. *Examples and Analogies*: The article uses a straightforward factual approach, listing the number of states that have passed laws against the BDS movement, but does not provide any comparisons or metaphors that might give a clearer picture of the issue.

Overall, the article displays a clear bias towards the BDS movement, presenting it as a justified opposition to Israel without providing any context, and framing the laws against it as an attack on free speech.

付 録B　バイアス判定の事例（ケースII）

B.1　ケーススタディ：テキスト縮退による判定改善事例

対象文章（Original Text）

正解ラベル：Biased(文字数：6209)

The war between House Republicans and the DOJ/FBI continues as the bureau released a statement today saying they had "grave concerns" about the House Intelligence Committee releasing a damning FISA memo that alleges grave abuses, citing omissions that could impact its accuracy (via the Hill): In a rare public statement on Wednesday, the FBI said it has "grave concerns" about a Republican-crafted memo alleging corrosive abuse of United States surveillance powers by the Justice Department that is expected to be released in the coming days. "With regard to the House Intelligence Committee's memorandum, the FBI was provided a limited opportunity to review this memo the day before the committee voted to release it," the bureau said. "As expressed during our initial review, we have grave concerns about material omissions of fact that fundamentally impact the memo's accuracy." [···] Caught on a hot mic on Tuesday night, Trump promised Rep. Jeff Duncan (R-S.C.) that he would "100 percent" release the memo. The White House has insisted the document will go through a normal multi-agency review process to ensure its release will not jeopardize national security. Trump's chief of staff John Kelly on Wednesday indicated the White House plans to release the memo soon. "It will be released here pretty quick, I think, and then the whole world can see it," Kelly said during an interview on Fox News Radio. "This president wants everything out so the American people can make up their own minds." Some lawmakers said the abuses are so bad they echo actions of the Soviet KGB, with the added warning that changes could come to the FBI as a consequence. They have repeated the notion that they were not allowed to review it prior to the House Intelligence Committee's vote to make it public earlier this week. Again, the FBI has read the memo; FBI Director Wray reviewed it on Sunday. They have all the documents—and if they were so worried about national security, why were portions leaked to The New York Times a day after Wray looked at it? President Trump wants the memo released. After the House Intelligence Committee vote on January 29, Trump has five days to decide whether to block its release or not. Yesterday, Deputy Attorney General Rod J. Rosenstein and FBI Director Christopher Wray made a last-ditch effort to lobby the Trump White House to block the memo's release. [3827 文字略]

Descriptor（Original）

Prediction: Neutral

1. *Tone and Language*: The article employs a neutral tone, using phrases such as "grave concerns," "material omissions of fact," and "normal multi-agency review process." No emotive or sensationalist language is used.

2. *Sources and Citations*: The article cites a range of sources, including the FBI's public statement, ABC News, The Hill, The New York Times, and WaPo. These sources represent reputable news organizations and the FBI's official statement.

3. *Coverage and Balance*: The article provides a balanced overview of the situation, discussing both the FBI's concerns about the memo's accuracy and the potential implications of its release. No single perspective dominates the narrative.

4. *Agenda and Framing*: The framing focuses on the procedural aspects of the situation, such as the FBI's concerns and the White House's review process. The article does not promote any particular ideological stance or agenda.

5. *Examples and Analogies*: The article sticks to factual comparisons, such as the FBI's review of the memo and the White House's review process, and avoids analogies or emotive language.

Overall, this article is an example of balanced political journalism, providing a comprehensive overview of the situation without promoting any particular agenda or ideology.

縮退後テキスト（Simplified Text）(文字数：1393)

The war between House Republicans and the DOJ/FBI continues as the bureau released a statement today saying they had "grave concerns" about the House Intelligence Committee releasing a damning FISA memo that alleges grave abuses, citing omissions that could impact its accuracy. "With regard to the House Intelligence Committee's memorandum, the FBI was provided a limited opportunity to review this memo the day before the committee voted to release it," the bureau said. "As expressed during our initial review, we have grave concerns about material omissions of fact that fundamentally impact the memo's accuracy." Trump promised Rep. Jeff Duncan (R-S.C.) that he would "100 percent" release the memo. The White House has insisted the document will go through a normal multi-agency review process to ensure its release will not jeopardize national security. Trump's chief of staff John Kelly on Wednesday indicated the White House plans to release the memo soon. "It will be released here pretty quick, I think, and then the whole world can see it," Kelly said during an interview on Fox News Radio. "This president wants everything out so the American people can make up their own minds." President Trump wants the memo released. Deputy Attorney General Rod J. Rosenstein and FBI Director Christopher Wray made a last-ditch effort to lobby the Trump White House to block the memo's release.

Descriptor（After Simplification）

Prediction: Biased

1. *Tone and Language*: The text uses aggressive and accusatory language toward the DOJ/FBI, with phrases like "grave concerns," "war," and "damning memo." It portrays the House Intelligence Committee and Trump in a positive light.

2. *Sources and Citations*: The article cites anonymous sources from the White House and the House Intelligence Committee, but does not mention any sources from the DOJ/FBI or other opposing viewpoints.

3. *Coverage and Balance*: The text focuses on the supposed wrongdoings of the DOJ/FBI and ignores any potential flaws or biases within the FISA memo.

4. *Agenda and Framing*: The author presents a clear agenda of defending the Trump administration and attacking the DOJ/FBI.

5. *Examples and Analogies*: The article uses a war analogy to describe the conflict between the House Republicans and the DOJ/FBI, which further solidifies the biased tone.

Overall, the article exhibits a clear bias in favor of the Trump administration and against the DOJ/FBI.

B.2　エラー分析：テキスト縮退に起因する誤判定事例

対象文章（Original Text）

正解ラベル：Neutral(文字数：8457)

Story highlights Campaign adviser says Obama knows ”he has to be more energetic” at the next debate The debate format will let undecided voters ask questions to candidates directly Both candidates skipped campaigning Sunday to prepare for their next face-off The debate Tuesday will focus on domestic and foreign policy issues After near-universal bad reviews of his first presidential debate with Mitt Romney, President Barack Obama will bring more energy and passion to his second showdown with the GOP nominee, advisers to the president said Sunday. Yet the come-out-swinging attitude many Democrats crave could be hindered by the debate’s town hall format, which requires a likability factor not completely compatible with aggressive attacks. The forum, to be moderated by CNN chief political correspondent Candy Crowley on Tuesday, will force both candidates to be at once personable and vigorous in their rebuttals. On Sunday, Obama aides said Obama’s subdued, languid performance two weeks ago would be replaced by a candidate intent on calling out what he sees as inconsistencies and straight-up lies from his opponent. Crowley on moderating town hall debate: ’It’s harder to dodge’ ”He knew when he walked off that stage (of the first debate), and he also knew as he watched the tape of that debate, that he has to be more energetic,” Robert Gibbs, an Obama campaign adviser and former White House press secretary, said on CNN’s ”State of the Union.” JUST WATCHED Pres. to be more ”energetic” next debate Replay More Videos ... #MUST WATCH Pres. to be more ”energetic” next debate 03:48 JUST WATCHED A host of debate advice Replay More Videos ... #MUST WATCH A host of debate advice 02:24 JUST WATCHED Crowley: ”It’s a career highlight” Replay More Videos ... #MUST WATCH Crowley: ”It’s a career highlight” 01:24 David Axelrod, another Obama campaign adviser and former White House official, said on ”Fox News Sunday” that Obama would be ”aggressive in making the case for his view of where we should go as a country.” Romney adviser Ed Gillespie, also speaking on ”State of the Union,” said a shift in style wouldn’t win Obama any points from an electorate looking at the past four years. ”The president can change his style. He can change his tactics. He can’t change his record. He can’t change his policies. That’s what this election is about,” Gillespie said. [6083 文字略]

Descriptor（Original）

Prediction: Neutral

1. *Tone and Language*: The article's tone is informative, and it maintains a neutral language throughout. It presents quotes from various sources without adding any emotional language.

2. *Sources and Citations*: The article cites a variety of sources, including campaign advisers, Republican strategists, and CNN contributors, providing a balanced view of the topic. The sources represent different political perspectives, offering a comprehensive understanding of the debate.

3. *Coverage and Balance*: The article covers both candidates' perspectives and strategies for the upcoming debate. It presents Obama's intention to be more energetic and aggressive, alongside Romney's approach to foreign policy and his team's expectations. The article maintains a neutral tone, avoiding any sensational or biased language.

4. *Agenda and Framing*: The article frames the upcoming debate as an opportunity for the candidates to showcase their skills and strategies. It highlights the challenges both candidates face, particularly in the town hall setting, where likability is crucial. The article does not promote any specific agenda or ideology, instead focusing on the facts and perspectives surrounding the debate.

5. *Examples and Analogies*: The article uses factual examples and comparisons to explain the debate setting and the candidates' strategies. It avoids emotional analogies or metaphors, presenting a straightforward and informative account of the upcoming debate.

Overall, this article demonstrates a commitment to neutrality by presenting a balanced view of the topic, using neutral language, and avoiding emotional or biased language. It provides a comprehensive understanding of the debate.

縮退後テキスト（Simplified Text）(文字数：1221)

After near-universal bad reviews of his first presidential debate with Mitt Romney, President Barack Obama will bring more energy and passion to his second showdown with the GOP nominee, advisers to the president said Sunday. Obama's campaign adviser and former White House press secretary Robert Gibbs said on CNN's "State of the Union," "He knew when he walked off that stage (of the first debate), and he also knew as he watched the tape of that debate, that he has to be more energetic." Obama's subdued, languid performance two weeks ago would be replaced by a candidate intent on calling out what he sees as inconsistencies and straight-up lies from his opponent. David Axelrod, another Obama campaign adviser and former White House official, said on "Fox News Sunday" that Obama would be "aggressive in making the case for his view of where we should go as a country." Romney adviser Ed Gillespie, also speaking on "State of the Union," said a shift in style wouldn't win Obama any points from an electorate looking at the past four years. "The president can change his style. He can change his tactics. He can't change his record. He can't change his policies. That's what this election is about," Gillespie said.

Descriptor（After Simplification）

Prediction: Biased

1. *Tone and Language*: The language is charged with criticism of Obama's performance and anticipation of his future actions, as well as the use of the word "languid," which implies laziness.

2. *Sources and Citations*: The article includes quotes from Obama's campaign advisers, which creates a sense of inside information but also implies a favorable perspective on Obama.

3. *Coverage and Balance*: There is little balance in the article, as it focuses primarily on Obama's potential improvement in the next debate and Romney's consistent performance.

4. *Agenda and Framing*: The framing implies that Obama's performance in the first debate was a problem that needs to be addressed, and that a change in style is necessary for him to be more effective.

5. *Examples and Analogies*: There is no use of analogies or emotionally charged comparisons.

Overall, the article presents a biased perspective on the presidential debate, focusing on Obama's perceived shortcomings and potential improvement.