東南アジアにおける研究の実践から見た大規模言語モデルの利活用について

八木暢昭[1][2]

《要 旨》

本稿は、東南アジアの政治研究における非構造化テキストの構造化という課題に対し、大規模言語モデル（LLM）の実践的利活用を検討したものである。従来の研究は人手のラベリングに依存し、大規模データへの対応と一貫性に課題が生じていた。本稿では、⑴OpenAI API で汎用モデルを用いる方法と、⑵BERT をファインチューニングして用いる方法を、同一タスク（マレーシア政治のニュース／コメントを 12 個のトピック、4 つの極性で分類するという事例）で比較しながら提示する。前者ではプロンプトエンジニアリングを施しながら、温度パラメータ（temperature）を 0 にし、Structured Outputs による JSON 形式の出力設定で多様性と出力形式を制御することで、少数の例示で高精度かつ一定の再現性を得る利点があるが、完全な再現性を得ることには限界がある。Facebook と Reddit の計 43,546 件から抽出したテスト 50 件に対し、GPT-4o-mini で分類した結果の重み付き F1 スコアは 0.8752 であった。後者の BERT では Human-in-the-Loop を核に、k 平均法による多様性サンプリングと予測確率に基づく不確実性サンプリングを反復して決定境界付近を重点的にアノテーションしていき、データを 405 件、605 件、1,205 件、そして 1,705 件へと拡充した。加えて肯定的／中立的なデータが少ないというデータの偏りを補正するため、OpenAI API で 200 件の人工的なデータを作成してデータ拡張を行った。その結果、重み付き F1 は 0.6898、0.7134、0.8212、0. 8476 と改善し、最終的にハイパーパラメータのチューニングをしたモデルでは 0.8606 に到達して GPT-4o-mini に近い性能となった。トピック別では政治的リーダーシップや行政パフォーマンスなど、単一文に賛否が並存しやすい領域では性能が伸び悩み、実務上ではラベル設計の再考が有効であることが示唆された。再現性の許容度が高い場面では API の活用が有効であり、長期の安定運用やバージョン固定、データ拡張による性能の向上を目指す場面では BERT のファインチューニングが有利である。加えて、人間による出力の監視とデータの統計的な品質管理、大規模言語モデルを適用する領域に関する知見とデータサイエンスの協働が不可欠である。本報告は LLM を用いた社会科学における定量的な分析に対する実務的なロードマップを提示した。

はじめに

東南アジアにおいてもソーシャルメディアは急速に普及しており、インドネシアでは選挙ツールとしても重要な役割を果たしている（岡本・八木・久納、2024）。ソーシャルメディア上のやり取りは画像や動画などもあるが、中でもテキストは広く発信されているデータである。そのため、テキストデータは人々の考えを知る上でも重要な役割を果たす。

マレーシアの政治研究においてもテキストデータを対象とした分析はいくつか見られる。例えばカスマニらは#PRU13 や#PRU14 というハッシュタグを含むツイートを対象にコンテンツの分析を行い、総選挙に関して人々がどのような意見を発信し

[1] 京都大学大学院 アジア・アフリカ地域研究研究科 博士課程学生 yagi.nobuaki.74z@st.kyoto-u.ac.jp
[2] 本論文はアジア政経学会 2025 年度秋季大会にて提出した報告論文である。

ているかを分析している（Kasmani, 2020; Kasmani, Sabran and Ramle 2014）。このコンテンツ分析では手作業でラベル付け(coding)がされている。シーインナサミーとマナフも類似したラベル付けを行い、ソーシャルメディア上で発信される憎悪を含んだ政治的発信について分析している（Chinnasamy and Manaf, 2018）。手作業でのラベル付けがマレーシアにおける政治研究で広く見られるが、こうした分析にはいくつかの課題がある。一つ目は手作業であるが故に分析の拡張性に乏しいというものである。上述した研究では研究者やコーダーが手作業でラベル付けを行うがその際にはあらかじめ定められたルールに則ってラベル付け担当者間で齟齬なく分類がされる必要がある。コーディング担当者が一人で、かつデータが小規模であれば一貫性を保つことは容易かもしれない。しかしデータが大規模になると数日にわたってラベル付けを行うか（同一人物でも日を跨ぐとアノテーションの基準がずれる可能性がある）、複数人で分担する必要がある。このような状況ではラベル付けの一貫性が損なわれる蓋然性が高くなる。

しかし、テキストデータはそのままでは定量的な分析に適さない。テキストデータのようにそのままでは機械学習のような処理では扱いづらいデータを非構造化データと呼び、オズデミールによると、これを「構造化された形式に変換すること」が重要である（Ozdemir、2023: 17）。ヴァジャラらはこうしたテキストデータを扱うための方法論の一つに「人間の言語を分析し、モデル化し、理解する方法を扱うコンピュータサイエンスの一分野」である自然言語処理を挙げている（Vajjala ほか、2022: 4）。こうした方法論を用いることでテキストデータの分類や情報の抽出、要約が可能であり、実際に社会科学に適用するための概説もある。例えばグリマーらによる『Text as Data』が挙げられる（Grimmer, Roberts and Stewart, 2022）。

こうした方法論の発展がある一方で、ミュラー・ハンセンらが指摘するように、特定の学問領域のみならずテキストマイニングなどの学際的知識が必要であるといったことからテキストデータの機械的な処理を社会科学に適用する動きは緩慢である（Müller-Hansen et al., 2020）。すなわち、テキストデータの重要性に比して、技術的な困難を含め様々な障壁からこうしたデータの研究現場における利活用が追いついていない状況であると考えられる。

テキストマイニングなどの伝統的な方法論がある一方で、ヴァスワニらによって近年ではトークン間の依存関係を高速かつ柔軟に考慮できる仕組みであるアテンション（Attention）メカニズムを中核とするトランスフォーマー（Transformer）が提唱された（Vaswani et al., 2017）。これを基礎にしてデヴリンらやチャンらによって、テキストの分類を得意とするバート（BERT）や要約を得意とするペガサス（PEGASUS）などの各種大規模言語モデルが開発されており（Devlin et al., 2018; Zhang et al., 2019）、構造化されていないテキストデータを機械的に扱うことが技術的に容易となっている。また、大規模言語モデルに関する知見も体系的に整備されており、そうした技術を扱うための障壁も下がってきている。例えばタンストールらは Transformer を使った大規模言語モデルの実装に関するテキストを書いている（Tunstall, Von Werra and Wolf、2022）。こうした各種のタスクを得意とする大規模言語モデルが開発されてきた一方で、ブラウンらやアキアムらによって生成モデルとしての GPT（Brown et al., 2020; Achiam et al., 2023）が提案・開発されてきた。「モデルの大規模化にともなって、従来はファインチューニングを行わないと解けないと思われていた多くのタスクが、モデルにプロンプトと呼ばれるテキストを入力して後続するテキストを予測するという単純な方法で解けること」（山田ほか、2023: 60）がわかってきた。GPT はこうしたプロンプト、特にいくつかの例示を与えるような文脈内学習を用いることで様々なタスクを解くことができる。本報告では特定のタスクの処理を得意とするモデルの構築とともに、プロンプトを与えることで様々なタスクを処理できる汎用的なモデルの利活用について Human-in-the-Loop 機械学習（Monarch、2023）や生成系 AI を用いたデータ拡張（Zhao, Chen and Yoon, 2023）といったデータセットの準備において重要な技術も含め、マレーシアの政治的トピックを抽出するというタスクを例にとって議論することで、大規模言語モデルを用いた社会科学領域での研究における技術的障壁を下げることを目的としている。

機械学習の議論ではモデルに関するものが多いが、中でも重要な構成要素はデータであるといっても良い。オズデミールによると、「問題を深く表現し、よく構造化されたデータがなければ、正確で構成なモデルを保証することは事実上不可能」（Ozdemir 2023: 4）である。確かにハイパーパラメータのチューニングや適切なモデルの選定といった要素は重要であるが、一方でガイガーらが指摘するように、「無意味なデータを用いると無意味な出力が生み出される」（Garbage in, garbage out）といった考えが広く浸

透していることからも、「教師あり機械学習では、高品質な分類器を得るためには高品質な訓練データが必要である」(Geiger et al., 2019: 1)。機械学習のモデリングや実装についてはすでに数多くの論文や書籍で議論されているため、特にBERT のファインチューニングについてはいかにして良質なデータを効率的に用意するか、という問題に焦点を当てたい。

Ⅰ OpenAI API の利活用

GPT は特にGPT-3 からGPT-4 以降、Few-shot 学習といった文脈内学習を用いることで要約や分類といった様々なタスクに対応することができる汎用的なモデルとなってきた。文脈内学習ではモデルのパラメータを更新しない（山田ほか、2023）ため、BERT に見られるようなファインチューニングとは異なり、容易に特定のタスクにモデルを適用することが可能となる。

GUI ベースで広く用いられているOpenAI のサービスはChatGPT であるが本報告ではGPT を簡易にプログラムから呼び出せるOpenAI API（OpenAI 2020）を用いる。これは定量的な分析を行うためのデータの構造化処理において、大規模なデータを繰り返し処理する際に再現性の確保や自動化による効率性といった利点を当該API が有しているためである。

OpenAI API ではパラメータとして出力の多様性を制御するための温度（temperature）を受け付けている。温度パラメータには0 から2 までの値を設定でき、0 であれば最もランダム性（randomness）が低くなり、2 であれば最もランダム性が高くなる。温度パラメータの設定を2 に近づけることで、出力が多様になる。データの構造化を行う際には温度パラメータを0 に設定することで多様性を限りなく低減させることで、出力をある程度制御することができる。ただし、温度パラメータによる制御は完全ではない。これは機械学習システムの特性であるが、ルールベースのシステムとは異なり出力に不確実性が伴うことには注意が必要である。フエンが指摘するように、機械学習は人間がパターンを定義する従来のソフトウェアとは異なり、データからパターンを学習するシステムである（Huyen、2023）。そのため、明確に定義されたパターンに基づく確実な出力が得られる保証はなく、出力された結果は確実なものではない。もし不確実性が許容できない領域に関する分析を行うのであればルールベースの手法を採用することも考慮に入れる必要があるだろう。それでも、ある程度の不確実性を許容できる場合であれば効率性の観点でOpenAI API のような汎用性の高いモデルを簡易に扱えるシステムを用いることは有用である。大規模な並列処理が可能であり、情報処理が速く、長期間情報を正確に記憶できる人工知能（鈴木、2023: 10）を社会科学における営為に適用することは人間の持つ弱点を当該技術によって補完し、更なる知識の探索を可能にする。

ある程度の不確実性を許容できる場合、機械学習を活用することで単純な分類問題などを高速に、そして一定のパフォーマンスを維持して処理することができる。これがOpenAI API といった機械学習アプリケーションを用いる利点の一つである。大量のデータを繰り返し処理する場合には出力の形式も制御したい。後述するようなプロンプトエンジニアリングによって出力の制御をある程度行うことはできるが、ここでは出力形式の安定性について触れておきたい。我々の目的はテキストという非構造化データの構造化である。構造化されたデータは一定の形式に則っている必要がある。Structured Outputs を用いることで指定したJSON スキーマに従って結果が出力される（OpenAI 2024）ため、データの構造化に役立つ。リスクエスト側においてresponse_format というパラメータにJSON スキーマを渡すことで出力形式を制御することができる。

後述するようなプロンプトエンジニアリングに加えて温度パラメータと Structured Outputs を用いた出力形式の制御によって、OpenAI API を介して非英語の文章を英語へと繰り返し、再現性高く翻訳することができるようになる。

1. プロンプトエンジニアリング

GPT などの生成系AI（基盤モデル）に対してプロンプトと呼ばれるテキストを入力することで、様々なタスクを解くことがで

きる（山田ほか、2023: 60）。フレグリーらによると「プロンプトエンジニアリングは、基盤モデルへの理解を深めつつ、タスクやユースケースに基盤モデルを適用することに焦点をあてた」技術のことを指す（Fregly, Barth and Eigenbrode、2024: 16）。典型的なプロンプト構造は「指示、コンテキスト、入力データ、出力インジケーターといった各部分」で構成される（Fregly, Barth and Eigenbrode、2024: 18）。ここで、出力インジケーターは上述した出力形式のことを指し、入力データは解きたいタスクに応じた非構造化データのことを指す。プロンプトエンジニアリングで特に重要な要素は指示とコンテキスト（文脈）である。「指示（instruction）とは、モデルに渡す、モデルに実行させたいタスクを記述したテキストのこと」である（Fregly, Barth and Eigenbrode、2024: 18）。アラマーらによると、「できるだけ具体的に書き、解釈の余地を最小限にすることが重要」である（Alammar and Grootendorst、2025: 167）。要約や分類といった様々なタスクがある中で、どのような処理を行いたいのかを明確にすることで出力の精度を向上させる。

プロンプトの構成要素の中でも特に核心的な要素が文脈である。GPT は文脈からタスクを学習する能力を有している。この文脈は「モデルがタスクやトピックをより深く理解した上で適切に応答するように、モデルに渡す、関連情報や詳細を指し」、「モデルの応答を望ましい出力に導くための一般的なテクニックに、プロンプトと応答のペアの例を、コンテキスト[文脈]情報としてモデルに渡すというもの」がある（Fregly, Barth and Eigenbrode、2024: 19）。この文脈を用いた学習を文脈内学習と呼び、GPT に望ましい出力例を幾つか与えることでタスクの処理能力が向上する。「One-shot プロンプトは 1 つの例、Few-shot プロンプトは 2 つ以上の例を使用」する（Alammar and Grootendorst、2025: 170）。これらの技術と、上述したような不確実性の制御と出力形式の制御の技術を組み合わせることで高い処理能力でデータの構造化を行うことができる。

2. OpenAI API による分類の実例

ここまでで OpenAI API でデータを構造化する際の技術的な議論を行ってきた。以下では実際にそれらの技術を用いた実例を見ていく。例としてマレーシアにおける政治的議論を分類するタスクを考える。これはソーシャルメディア上で行われているニュースの報道やそれらの報道に対するユーザーのコメントを分類するというものである。こうしたタスクを解くことで、ソーシャルメディア上でどのようなトピックがどのような極性を帯びて議論されているかを知ることができ、結果として重要視されている争点を定量的に把握することができる。

ここでは Facebook と Reddit という二つのソーシャルメディアのデータを、Web サイトやソーシャルメディアのデータを収集（スクレイピング）するプラットフォームである Apify にある Facebook Posts Scraper、Facebook Comments Scraper、Reddit Scraper を用いて収集した。合計して 43,546 件ある報道や投稿、およびコメントから 50 件をテストデータとして無作為抽出し、それらに対して人手で 12 個のトピック（民主主義、経済、人種、政治的リーダーシップ、開発、汚職、政治的安定性、治安、行政のパフォーマンス、教育、宗教、環境）に関して 4 つの極性（肯定的、中立的、否定的、言及なし）へとラベル付けした。

プロンプトを適用したモデル（GPT-4o-mini）にこのデータの分類タスクを解かせた。重要な点は二つある。一つ目は明示的なタスクの記述である。このプロンプトでは最初にモデルに果たしてほしい役割を与えた。そして極性が 4 つあることと、分類してほしい 12 個のトピックを明示した。また、極性への分類の仕方をガイドラインとして提供している。二つ目は文脈内学習である。このプロンプトでは三つの例を提示し、どのような場合にどのような分類をすべきかを文脈情報として与えている。

次に分類時に出力形式の指示を、response_format を用いて行った。また、温度パラメータである temperature も 0 にしている。これらの技術的な制御を施したモデルの出力結果を重み付き F1 スコア（Weighted F1）で測定している。精度指標の一つであるアキュラシー（Accuracy）は精度評価で広く使われているが、仮にデータに偏りが見られる場合には精度評価に不向きであり、ゲロンが指摘するようにより優れた精度評価の方法として混同行列がある（Géron、2024）。その混同行列を用いた精度の比較方法として F1 スコアが挙げられる。ゲロンの指摘にもあるように、多数派クラスと少数派クラスに偏りがある場合においては単に多数派クラスを一貫して予測することで高い精度が得られるものの（Géron、2024）、ガーネムらによると「少数派クラスを正確に特定す

ることに苦心する可能性がある」(Ghanem et al., 2023: 2)。これに対してイノホサ・リーらの議論にもあるように (Hinojosa Lee, Braet and Springael, 2024: 3-4)、重み付き F1 では各クラスにおけるクラスの不均衡を考慮しながら、再現率と適合率をバランスした評価を行う。F1 スコアによる性能評価はシルヴァらが開発したブラジルにおけるポルトガル語で記述された政府のデータを分類する GovBERT-BR (Silva et al. 2025) でも用いられている。直感的には精度の方が分かりやすいものの、政治的テキストといったトピックや極性に偏りがあるデータセットでは重み付き F1 の方が性能改善といった観点では適切である。

表 1 は OpenAI API を用いた時の各トピックに対する分類精度を示している。全体的な重み付き F1 スコアは 0.8752 であった。GovBERT-BR の F1 スコアは 0.784 から 0.948 であったため、これが一つのベンチマークであると考えると、今回のようなマルチクラス分類かつマルチラベル分類では比較的高い分類精度であることがわかる。一方で政治的リーダーシップや民主主義、行政のパフォーマンスにおいては 0.7882 や 0.7991 、0.6605 といった比較的低い分類精度が確認された。これらのトピックについて言及があるコメントではマハティール元首相を強く批判するヒシャムディン元副総裁への賛同のように、一方に対しては否定的であるもののもう一方に対しては肯定的といったものが見られた。こうした二つの極性が多くあるトピックで分類精度が低くなったものと考えられる。政治的トピックなどの社会科学の領域における非構造化データでは複数の極性を帯びたデータが時折見られるため、単純に一つの極性へと分類するよりもそうしたデータに対応した極性を考慮した方が良い可能性がある。

表 1 OpenAI API (GPT-4o-mini) における分類精度

トピック	重み付き F1 スコア
民主主義	0.7991
経済	0.9703
人種	0.9640
政治的リーダーシップ	0.7882
開発	0.8860
汚職	1.0000
政治的安定性	0.6899
治安	0.9335
行政のパフォーマンス	0.6605
教育	0.8813
宗教	0.9396
環境	0.9899
全体	0.8752

II BERT のファインチューニング

以下では BERT を用いて上述したものと同様のタスク (マレーシアにおける政治的テキストの分類タスク) を解くことを考える。なお、ファインチューニングを行った BERT のモデル (以降、特に断りがない場合はこれを BERT モデルと呼ぶ) の精度を評価するためのベースラインとして表 2.1 で示した OpenAI API (GPT-4o-mini) を用いる。

OpenAI API ではなく BERT を用いる利点はいくつかあるが、一つ目は再現性である。そしてデータのスケーリングによる精度

向上が挙げられる。BERT ではモデルとトークナイザのバージョンを固定することで同一の入力に対して同一の出力を長期的に維持することができる。一方でOpenAIAPI ではモデル更新の影響を受ける可能性がある。また、上述したように温度（temperature）パラメータを 0 にしても出力結果が微妙に異なることがある。再現性の確保といった点ではBERT の方が扱いやすいという利点がある。

OpenAI API では与えられるプロンプトの長さに上限がある上、大量の例示を与えることで利用時のコストが増大する。BERT では様々な教師データでファインチューニングを行うことで分類精度の更なる向上が期待できる。他方で、このデータのスケーリングによる精度向上には別の問題が生じる。良質なデータを大量に用意するには膨大なコストがかかる。モナークが指摘するように、一般的に「手作業によるデータのラベリングにはコストがかかり、信頼性も低いことが知られており」（Monarch、2023: v）、本報告ではこの問題を解決するための方法として Human-in-the-Loop 機械学習と生成系 AI（GPT-4o-mini）を用いたデータ拡張について見ていく。

1. Human-in-the-Loop 機械学習

BERT を、精読を高くファインチューニングするためには良質なデータセットが必要である。一方で、モスクアイラ・レイによると、ラベル付けがされていないデータの「アノテーションタスクは費用がかかるか時間がかかる」（Mosqueira-Rey, 2023: 3009）。これに対して、Human-in-the-Loop 機械学習の中核的な技術である「能動学習はできる限り少ないラベル付きのインスタンスを用いることで、ラベル付きデータを獲得するコストを最小化することを狙っている」（Mosqueira-Rey, 2023: 3011）。

ウーらが示す通り（Wu et al., 2022）、一般的な自然言語処理における Human-in-the-Loop 機械学習の適用過程は以下のようなものである。まずはテキストのデータセットを用意し、前処理を施した上で学習と推論・予測を行う。そして予測結果の中からサンプルを抽出し、それらのサンプルに対して人間がフィードバックを行う（ラベル付けをする）というものである。このフィードバックループを繰り返すことで効率的に良質なデータセットを作るというものである。

以下では、「人間がアノテーションを行うためにどのデータをサンプリングするかを決定するプロセス」（Monarch、2023: 7）である能動学習を構成するであるランダムサンプリング、不確実性サンプリング、そして多様性サンプリングの内、後者の二要素について議論する。その上でサンプリングだけでは解消しきれないデータの偏りを補正するためのデータ拡張について議論する。

(1) 多様性サンプリング

モナークが議論するように、「多様性サンプリングとは、現状の機械学習モデルにとって未知、あるいは稀なデータを特定するための戦略」（Monarch、2023: 7）である。モスケイラ・レイが述べるように、「トレーニングデータにおいて稀な、ないしは見られない、ラベル付けがされていないものを選択することで問題箇所の解像度を上げる」（Mosqueira-Rey, 2023: 3010）。元のデータセットに偏りがある場合に単純なランダムサンプリングを行うと、抽出されたデータ群にも偏りが反映されうる。その場合には稀な、ないしは元のデータセットでは見られないラベルや極性に対する予測精度が低下し得る。元のデータセットから様々なデータをサンプリングすることでこのような問題に対処することが多様性サンプリングの目的である。

多様性サンプリングには主にモデルベースのサンプリング、クラスタベースのサンプリング、代表点サンプリング、実世界における多様性を考慮したサンプリングという 4 つのサンプリング方法がある（Monarch 2023）。ここではクラスタベースのサンプリング手法として、k 平均法を扱う。k 平均法は指定したクラスタ数 k に基づき、「データセットのクラスタリングを素早く効率的に実行できる単純なアルゴリズムである」（Géron、2024: 240）。このアルゴリズムでは、データセットの中から指定したクラスタ数 k の数だけ重心（セントロイド）を見つけ出す。多様性サンプリングでは外れ値（重心から遠いデータ）が特に重要である。重

心から遠い外れ値を用いることで「クラスタの中で見落とされている可能性がある興味深いデータを特定」する（Monarch、2023: 102）。

表2. BERT のファインチューニング（初回）の精度とベースラインモデルの精度の比較

トピック	重み付き F1 スコア（BERT）	重み付き F1 スコア（ベースライン）
民主主義	0.5769	0.7991
経済	0.8127	0.9703
人種	0.9448	0.9640
政治的リーダーシップ	0.6136	0.7882
開発	0.9323	0.8860
汚職	0.8419	1.0000
政治的安定性	0.0013	0.6899
治安	0.8304	0.9335
行政のパフォーマンス	0.0190	0.6605
教育	0.9815	0.8813
宗教	0.7422	0.9396
環境	0.9815	0.9899
全体	0.6898	0.8752
データ数	405	3（Few-shot 学習）

本報告では関心のある 12 個のトピックに合わせてクラスタ数kを 12 と設定して k 平均法による多様性サンプリングを行う。それにあたってハギングフェース（Hugging Face）で提供されている all-MiniLM-L6-v2 というモデルでテキストの埋め込みを抽出し、それに基づき k 平均法でクラスタリングを行う。そしてこれらのクラスターの中から重心に最も近いデータを 5 件、最も遠いデータを 5 件ずつ抽出した。12 個のクラスターから 10 件ずつ抽出するので 120 件のデータを抽出したことになる。これらのデータに人手でアノテーションをした結果、データに偏りが生じていることがわかった。例えば人種問題に関して肯定的なコメントや環境問題に言及するコメントがほとんど見られなかった。そのため、キーワード検索を用いて追加で 85 件のデータを抽出して追加した（合計で 205 件となった）。このデータは重心と外れ値に偏っている可能性があるので、追加で 200 件をランダムサンプリングして合計 405 件のデータに対して人手でアノテーションを行い、BERT のファインチューニングを行った。

ファインチューニングの結果、全体の重み付き F1 スコアが 0.6898 のモデルとなった（表 2）。このモデルの全体的な重み付き F1 スコアは 0.6898 であった。この結果はベースラインと比べて低いものである。特に政治的安定性や民主主義、政治的リーダーシップにおいて低い精度となっていることがわかる。機械学習ではデータの数と精度は連動する。また、機械学習モデルが上手く識別できないデータに対して人間がフィードバックを与えることで精度向上が見込める。そのため、以降では不確実性サンプリングを用いて BERT モデルが分類できなかったデータを抽出し、人手でアノテーションを行うことでモデルにフィードバックを与える。

(2) 不確実性サンプリング

「不確実性サンプリングとは、機械学習モデルの決定境界付近でラベル付けされていないデータを特定するための戦略」であ

る（Monarch、2023: 7）。「これは現在の訓練済みモデルの下で最も確信度が低いインスタンスを選択する」（Mosqueira-Rey, 2023: 3010）。ここでは、予測確率が 0.8 未満のインスタンスを不確実なデータと定義し、それらのインスタンスをサンプリングして人手でアノテーションを行った。表 2 のように、1 度目の学習時に 200 個のインスタンスを予測し、不確実なデータにアノテーションを行い、2 回目の学習を行った。2 回目以降は同様のループを行った。それらの結果を表 3 で示している。2 回目の学習では全体の精度がやや向上し、重み付き F1 スコアが 0.7134 となった。ここで 600 件のインスタンスを追加で予測し、不確実なものにラベル付けを行った。合計 1,205 件のデータで 3 回目の学習を行ったところ精度が大幅に向上し、スコアが 0.8212 に改善した。特に 2 回目までの課題であった政治的安定性や行政のパフォーマンスで大きな改善が見られている。次に 500 件のデータに不確実性サンプリングを行い、アノテーションを行ったものと、後述するデータ拡張で得た 200 件のデータを追加した合計 1,905 件のデータで 4 回目の学習を行った。結果、精度がやや改善して全体の重み付き F1 スコアが 0.8476 となった。「アノテーションプロセスの開始時は、[ハイパーパラメータのチューニングよりも]学習データを増やすことに集中」する（Monarch、2023: 40）ため、最後に 4 回目の学習で用いたデータとハイパーパラメータのチューニングを用いて最終モデルのファインチューニングを行った。結果は全体の重み付き F1 スコアが 0.8606 とベースラインモデルと近い性能のモデルの構築に繋がった。

表 3 BERT モデルの精度（重み付き F1 スコア）の推移

トピック	2 回目	3 回目	4 回目	最終モデル
民主主義	0.7280	0.6715	0.7631	0.7886
経済	0.7775	0.8822	0.8454	0.9691
人種	0.9448	0.8534	0.8912	0.9047
政治的リーダーシップ	0.6369	0.5915	0.6615	0.6940
開発	0.9323	0.9407	0.9181	0.8716
汚職	0.8959	0.8606	0.8797	0.9701
政治的安定性	0.0183	0.7681	0.8379	0.6593
治安	0.8127	0.7892	0.8003	0.9037
行政のパフォーマンス	0.1032	0.7541	0.7854	0.6914
教育	0.9815	0.9702	0.9375	0.9701
宗教	0.7476	0.8247	0.9027	0.9347
環境	0.9815	0.9481	0.9491	0.9701
全体	0.7134	0.8212	0.8476	0.8606
データ数	605	1,205	1,905	1,905

図 1 は学習のタイミングごとの各トピックにおける予測確率の分布を箱ひげ図で示している。1 回目と 2 回目以降で大きな乖離が見られる。すなわちこれは、1 回目では確信を持てなかった予測結果が多かったのに対して、不確実性サンプリングを行った 2 回目以降ではそうした不確実なインスタンスに対して人間がアノテーションというフィードバックを与えることでモデルが確信を持てるようになったことを示している。2 回目以降で予測確率が低い部分が見られる（例えば治安は予測確率の分布が向上しているのに対して行政のパフォーマンスでは下の裾野が広がっているように見える）。これはインスタンスの数を増やしたことに起因しているものと考えられる。特に 4 回目ではデータ拡張によってそれまで 3 回目まででは稀な、ないしは存在しないデータも学習に使われているため、確信を持ちづらいインスタンスが見られたものと思われる。

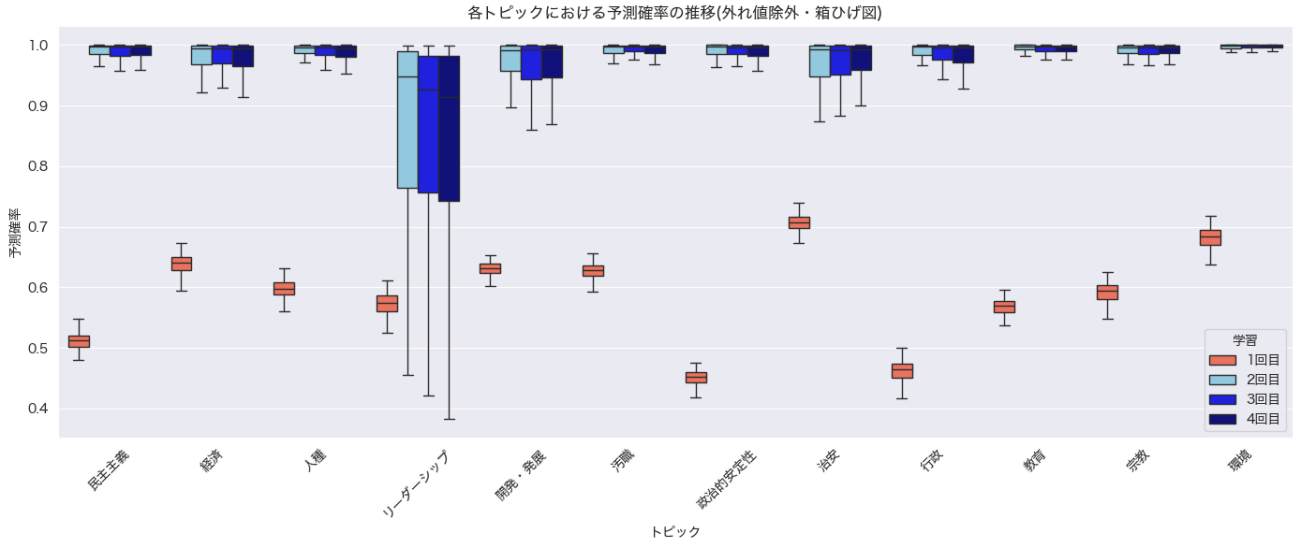

図1 各トピックにおける予測確率の推移（箱ひげ図）

2. データ拡張

これらのサンプリング戦略は有効であるものの、限界がある。サンプリングを行うデータセットが現実世界を全て反映している保証がないためである。ソーシャルメディア上のデータには偏りがある。例えばマレーシアは多民族国家であり、そこでは様々な言語が用いられる。マレー人はマレー語を話し、華人は華語を用いる傾向にある。また、民族横断的に英語が用いられることもある。こうした背景から、言語ごとにユーザーの属性に偏りが生じる。ある言語のグループで重要な争点とされているものが他の言語のグループでは軽視されていることもある。また、ソーシャルメディアごとにもユーザーの属性が異なることが考えられる。そして何より重要なことはソーシャルメディアユーザーと現実世界の国民や有権者が全く同じ属性である保証はないということである。ゆえにあるプラットフォームや媒体で抽出したデータセットをそのまま他のデータセットに適用することは難しい。他のデータセットにも適用可能なデータセットを作成するために人工的にデータを作ることは有効な手立ての一つである。

表 4 では訓練において、人手でアノテーションを行ったデータセットと不確実性サンプリングを行ったデータセットを結合したデータ（$N = 1{,}705$）におけるトピックごとの極性の分布を示している。言及がされているデータに着目すると大半が否定的なものであった。これは抽出対象としたソーシャルメディアのユーザーにおける属性を反映している可能性があり、興味深いものであるものの、否定的な意見が多数派のラベルとなると、データに偏りが発生していることになる。

データの偏りによって少数派のクラスを検出できず、モデルの正常な動作が妨げられることがある（Huyen 2023）。これは既知のデータ、すなわち訓練データでは正常に動作していたモデルが未知のデータ、すなわち実際に運用する上で投入されるデータに対しては正常に動作できないという問題である。このようなクラスの不均衡の問題に対処する方法の一つにデータ拡張やデータオーグメンテーションと呼ばれる「訓練データの量を増やすために用いられる」（Huyen 2023: 111）技術が挙げられる。これには元のデータを少し改変したりノイズを加えたりデータの合成といったものがあるが、今回は OpenAI API を使った人工データの作成を考える。ChatGPT を用いたデータ拡張を行うことでモデルの精度を改善したという報告（Zhao et al. 2023）がありこの発想は汎化性能の向上にも役立つと考えられる。

データの偏りによって少数派のクラスを検出できず、モデルの正常な動作が妨げられることがある（Huyen、2023）。これは既知のデータ、すなわち訓練データでは正常に動作していたモデルが未知のデータ、すなわち実際に運用する上で投入されるデータに対しては正常に動作できないという問題である。このようなクラスの不均衡の問題に対処する方法の一つにデータ拡張やデ

ータオーグメンテーションと呼ばれる「訓練データの量を増やすために用いられる」（Huyen、2023:111）技術が挙げられる。これには元のデータを少し改変したりノイズを加えたりデータの合成といったものがあるが、今回はOpenAI API を使った人工データの作成を考える。ChatGPT を用いたデータ拡張を行うことでモデルの精度を改善したという報告（Zhao, Chen and Yoon, 2023）がありこの発想は汎化性能の向上にも役立つと考えられる。

表4 トピックごとの極性の分布

トピック	肯定的	中立的	否定的	言及なし
民主主義	103	77	94	1,431
経済	54	22	69	1,560
人種	22	23	75	1,585
政治的リーダーシップ	80	90	250	1,285
開発	52	14	30	1,609
汚職	4	7	103	1,591
政治的安定性	17	27	156	1,505
治安	34	18	177	1,476
行政のパフォーマンス	69	66	115	1,455
教育	22	4	13	1,666
宗教	18	31	105	1,551
環境	9	6	28	1,662

特に$N = 1{,}705$件のサンプリングデータでは少数派であった肯定的、ないしは中立的な言説（ニュース報道やコメント）を中心に人工的にデータを作ることを試みた。これらの人工データではソーシャルメディア上のやり取りを反映するため様々な文体を表現するようにしている。

OpenAI API（GPT-4o-mini）を用いて作成したデータにおけるトピックとその極性を表5で示した。今回は肯定的なデータと中立的なデータを重点的に作成した一方で、否定的なデータも少数ながら作成することで人工的に作成されたデータに潜む未観測のバイアスを軽減することを試みた。最終的には上述した$N = 1{,}705$のデータと人工的に作った$N = 200$データを結合し、4回目と最終モデルの学習を行った。結果は上述したように、ベースラインモデルと遜色のない精度となった。

ここまでで、BERT のファインチューニングについて議論してきた。ここではベースラインモデル（GPT-4o-mini）と BERT モデルの精度に関する比較を行う。精度評価の指標は重み付き F1 スコアである。Human-in-the-Loop 機械学習とデータ拡張を用いたことで$N = 1{,}905$という比較的小さなデータセットにおいても、全体的な精度はベースラインモデルと遜色ないものとなった（表6）。各トピックについても類似した精度となっている。一方で、ほとんど全てのトピックでBERT モデルの方がやや低い精度となっている。分類精度が低い場合にはデータの数を増やすことである程度の対応が可能ではあるものの、データを増やした時の精度向上効果が逓減していくことについては上ですでに確認した。特にモデルが分類に苦慮しているトピックを見ていくと政治的リーダーシップや行政のパフォーマンスである。こうしたトピックには一つの文章で賛否両論が含まれることがある。こうした問題に対応するために例えば賛否両論などの別の極性を用意するといった手立ても考えられる。ラベル付けのルールをより現実世界のものに適合したものにすることで、データのスケーリングによってより高い精度向上効果が得られる可能性がある。

表5 データ拡張で生成した人工データの分布

トピック	肯定的	中立的	否定的	言及なし
民主主義	15	10	5	170
経済	18	15	5	162
人種	15	9	3	173
政治的リーダーシップ	16	11	9	164
開発	13	16	5	166
汚職	15	11	7	167
政治的安定性	20	11	7	162
治安	14	10	6	170
行政のパフォーマンス	12	12	4	172
教育	17	8	4	171
宗教	18	13	4	165
環境	15	12	10	163

表6 OpenAI API（GPT-4o-mini）と BERT モデルの精度評価結果の比較

	GPT-4o-mini	BERT（最終モデル）
トピック	重み付き F1 スコア	重み付き F1 スコア
民主主義	0.7991	0.7886
経済	0.9703	0.9691
人種	0.9640	0.9047
政治的リーダーシップ	0.7882	0.6940
開発	0.8860	0.8716
汚職	1.0000	0.9701
政治的安定性	0.6899	0.6593
治安	0.9335	0.9037
行政のパフォーマンス	0.6605	0.6914
教育	0.8813	0.9701
宗教	0.9396	0.9347
環境	0.9899	0.9701
全体	0.8752	0.8606

Ⅳ 大規模言語モデルの利活用に関する結論と課題、そして展望

ここまでで、非構造化データ（テキストデータ）の構造化というタスクを例として、OpenAI API の利活用と BERT のファインチューニングについて見てきた。前者の技術を用いる利点は手軽に精度の高いモデルを出力形式の制御を行いながら、ある程度

の再現性を維持しながら利用できるというものである。そのために重要な技術がいくつかある。プロンプトエンジニアリングではタスクを明示的に示し、例示をモデルに与えることでタスクにモデルを適応させることができる。そして温度のパラメータである temperature を 0 にするとともに Structured Outputs を活用することで出力の多様性や出力形式を制御することについても議論した。こうした技術を用いることで先述した利点をモデルから引き出すことができる一方で、多様性の制御は相対的なものであるという課題が残っている。すなわち、不確実性はある程度制御できるもののその再現性は完全なものではない、という課題である。また、今回用いた GPT-4o-mini のような生成系 AI の利点として少ない例示ですぐにタスクに適用できるというものが挙げられるが、入力できるプロンプトにも限界があるため、例外的なデータや外れ値、決定境界付近にあるデータの処理が難しい。

これに対して BERT ではモデルとトークナイザのバージョンを固定することで同一の入力に対して同一の出力を長期的に維持することができるので、より高い再現性が見込める。また、決定境界付近のデータに関して人間からのフィードバック（すなわち、アノテーション）を与えることで分類に関するパターンをモデルが追加的に学習できる。BERT のファインチューニングでデータの規模を大きくすることで精度の向上が期待できる（ただし、その効果はデータの規模が大きくなるほど逓減することには注意が必要である）。一方で BERT のファインチューニングにおいて課題となるのは良質なデータセットを構築するというものである。人手のアノテーションはコストが高い。更に時間的な制約がある場合、一つ一つのデータをラベル付けすることは難しい。また、サンプリングしたデータに偏りが生じている可能性もある。そこで本報告では Human-in-the-Loop 機械学習という枠組みの下で広くデータをサンプリングしつつ、モデルが判断に苦慮したものに対して集中的に人間がフィードバックを与えるという手段を用いた。また、サンプリング元のデータセットにデータが存在しないという課題に対して、生成系 AI（OpenAI API）を用いてデータ拡張を行うという手段を用いた。これらの技術を活用することで $N = 1{,}905$ という比較的小規模なデータセットでもベースラインモデルと遜色ないモデルのファインチューニングを行うことができた。

これらのことから次のようなことが考えられる。まず、ある程度の再現性という課題が許容できる場合には OpenAI API といったサービスの利用は手軽さに比べた精度の高さという側面で有効であると考えられる。これに対してより再現性の高さやスケーリングによるモデルの精度を求める場合には BERT といった個別のタスクに特化したモデルをファインチューニングすることは有効であると考えられる。

こうした利点から、大規模言語モデルを用いたデータの構造化は社会科学のさまざまなデータの定量的な分析に資するものであると考えられるが、いくつかの課題が残されている。一つはモデルの出力が本当に妥当なものなのかを人間が監視する必要があるというものである。機械学習モデルは繰り返しの処理を高速に実行できる。しかしそれらのモデルが出力した結果が人間の意図通りのものなのかは適切に管理する必要がある。こうした課題の解決策の一つとして伝統的な統計学の枠組みが挙げられる。出力された結果をランダムサンプリングしてそれが意図通りのものなのかを検定・推定してみることで確率的に出力の管理ができる可能性がある。仮に意図しない出力であればプロンプトを変えてみる、モデルのファインチューニングを再度行うといった手立てを打つことができる。全てをコンピュータに任せるのではなく適切なタイミングで人間が介入することで機械学習モデルの利点を引き出すことができる。

また、こうした機械学習モデルを社会科学に適用する際の障壁として技術的なものと学問領域に存するものが考えられる。まず、OpenAI API といった簡易なスクリプトで実行できるものでも Python といったプログラミング言語が動作する環境を用意し、そしてプログラムを書くといった技術的障壁がある。BERT のファインチューニングでは本報告で取り上げた技術に加えて深層学習の知見と、それを実装する知識が必要になる。また、上述したような人間によるモデルの出力結果の管理においても統計学に関する知識が必要になる。

一方で分類対象をどうするかといった課題設定や、データのラベル付けの規則といった側面では問題関心を抱く学問領域に関する十分な知見が必要となる。そのため、単に機械学習やその周辺領域に関する知識を持つだけでは社会科学の研究を行うことは難しい。関心対象への知見を持った上で、当該領域に対して機械学習の知見を応用する必要がある。故に社会科学におけるデ

ータサイエンスの応用では当該領域の専門を持つ者と機械学習や統計学の専門を持つ者が協業する必要がある。機械と人間、そして人間と人間の間に相互作用をもたらすことで初めて大規模言語モデルといった技術を課題解決に応用することができるものと考えられる。

（参考文献）

日本語

Alammar, Jay and Grootendorst, Maarten（2025）、『直感 LLM ―ハンズオンで動かして学ぶ大規模言語モデル入門』（中山光樹訳、原著は2024 年発行）オライリー・ジャパン。

Vajjala, Sowmya., Majumder, Bodhisattwa., Gupta, Anuj and Suranam Harshit（2022）『実践 自然言語処理 ―実世界 NLP アプリケーション開発のベストプラクティス』（中山光樹訳、原著は2020 年発行）オライリー・ジャパン。

岡本正明・八木暢昭・久納源太（2024）、「第5 回 ティックトックの政治化は民主主義を空洞化するのか?」『IDE スクエア – 世界を見る眼』、1-9 ページ。

OpenAI（2020）、『OpenAI API』（2025 年 10 月 7 日に最終アクセス、https://openai.com/ja-JP/index/openai-api/）。

――― （2024）、『API に Structured Outputs を導入』（2025 年 10 月 7 日に最終アクセス、https://openai.com/ja-JP/index/introducing-structured-outputs-in-the-api/）。

Ozdemir, Sinan（2023）、『事例で学ぶ特徴量エンジニアリング』（田村広平・大野真一朗・砂長谷健・土井健・大貫峻平・石山将成訳、原著は2022 年発行）オライリー・ジャパン。

Géron, Aurélien（2024）『scikit-learn、Keras、TensorFlow による実践機械学習 第3 版』（下田倫大・牧允皓・長尾高弘訳、原著は2022 年発行）オライリー・ジャパン。

鈴木貴之編（2023）、『人工知能とどうつきあうか 哲学から考える』勁草書房。

Tunstall, Lewis., Von Werra, Leandro and Wolf, Thomas（2022）、『機械学習エンジニアのための Transformers―最先端の自然言語処理ライブラリによるモデル開発』（中山光樹訳、原著は2022 年発行）オライリー・ジャパン。

Huyen, Chip（2023）、『機械学習システムデザイン―実運用レベルのアプリケーションを実現する継続的反復プロセス』（江川崇・平山順一訳、原著は2022 年発行）オライリー・ジャパン。

Fregly, Chris., Barth, Antje and Eigenbrode, Shelbee（2024）『AWS ではじめる生成 AI―RAG アプリケーション開発から、基盤モデルの微調整、マルチモーダル AI 活用までを試して学ぶ』（久富木 隆一訳、本橋和貴・久保隆宏技術監修、原著は2023 年発行）オライリー・ジャパン。

Monarch, R. M.（2023）、『Human-in-the-Loop 機械学習―人間参加型 AI のための能動学習とアノテーション―』（上田隼也・角野為耶・伊藤寛祥訳、原著は2021 年発行）共立出版。

山田育矢・鈴木正敏・山田康輔・李凌寒（2023）、『大規模言語モデル入門』技術評論社。

英語

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... and McGrew, B. (2023). "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... and Amodei, D. (2020). "Language models are few-shot learners," *Advances in neural information processing systems*, *33*, pp. 1877-1901.

Chinnasamy, S., & Manaf, N. A. (2018). "Social media as political hatred mode in Malaysia's 2018 General Election," *SHS Web of Conferences*, *53*.

Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding," *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.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R. and Huang, J. (2020). "Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?," *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 325-336.

Ghanem, M., Ghaith, A. K., El-Hajj, V. G., Bhandarkar, A., De Giorgio, A., Elmi-Terander, A. and Bydon, M. (2023). "Limitations in evaluating machine learning models for imbalanced binary outcome classification in spine surgery: a systematic review," *Brain Sciences*, *13*(12), 1723.

Grimmer, J., Roberts, M. E. and Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Hinojosa Lee, M. C., Braet, J. and Springael, J. (2024). "Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores," *Applied Sciences*, *14*(21), 9863.

Kasmani, M. F. (2020). "How did people Tweet in the 2018 Malaysian general election: Analysis of top Tweets in# PRU14," *IIUM Journal of Human Sciences*, *2*(1), pp. 39-54.

Kasmani, M. F., Sabran, R. and Ramle, N. (2014). "Can Twitter be an effective platform for political discourse in Malaysia? A study of# PRU13," *Procedia-Social and Behavioral Sciences*, *155*, pp. 348-355.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. and Fernández-Leal, Á. (2023). "Human-in-the-loop machine learning: a state of the art," *Artificial Intelligence Review*, *56*(4), pp. 3005-3054.

Müller-Hansen, F., Callaghan, M. W. and Minx, J. C. (2020). "Text as big data: Develop codes of practice for rigorous computational text analysis in energy social science," *Energy Research & Social Science*, *70*, 101691.

Silva, M. O., Oliveira, G. P., Costa, L. G. and Pappa, G. L. (2024). "GovBERT-BR: A BERT-Based Language Model for Brazilian Portuguese Governmental Data," *Brazilian Conference on Intelligent Systems*, pp. 19-32.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017). "Attention is all you need," *Advances in neural information processing systems*, *30*.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., and He, L. (2022). "A survey of human-in-the-loop for machine learning," *Future Generation Computer Systems*, *135*, pp. 364-381.

Zhang, J., Zhao, Y., Saleh, M. and Liu, P. (2020). "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," *International conference on machine learning*, pp. 11328-11339.

Zhao, H., Chen, H. and Yoon, H. J. (2023). "Enhancing text classification models with generative ai-aided data augmentation," *2023 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pp. 138-145.

Abstract

Large Language Models in Southeast Asian Research Practice

Nobuaki YAGI

This paper examines the practical use of large language models (LLMs) to structure unstructured text in Southeast Asian political research. Prior studies have relied on manual labeling, which creates scalability and consistency challenges for large datasets. This paper presents and compares two approaches on a common task, classifying Malaysian political news and comments into 12 topics with four polarity labels. The first approach employs general-purpose models via the OpenAI API. Through prompt engineering, setting the temperature to 0, and enforcing Structured Outputs in JSON to control diversity and output format, this approach achieves high performance with few in-context examples and delivers a degree of reproducibility, although perfect reproducibility remains elusive. On a test set of 50 items sampled from 43,546 Facebook and Reddit posts, GPT-4o-mini attained a weighted F1 score of 0.8752. The second approach fine-tunes BERT within a Human-in-the-Loop framework. This paper applies diversity sampling using k-means and uncertainty sampling iteratively based on predicted class probabilities to prioritize annotations near the decision boundary, expanding the labeled dataset from 405 to 605, 1,205 and then to 1,705 items. To mitigate class imbalance, particularly the scarcity of positive and neutral instances, this paper augments the data with 200 synthetic examples generated via the OpenAI API. Weighted F1 scores improved from 0.6898 to 0.7134, 0.8212, and 0.8476 across iterations; with hyperparameter tuning, the model reached 0.8606, approaching the performance of GPT-4o-mini. Topic-level analyses indicate persistent difficulty in domains such as political leadership and administrative performance, where single sentences often contain mixed polarity; this paper therefore suggests reconsidering label design (e.g., accommodating mixed polarity) for applied settings. In contexts where some variability in reproducibility is acceptable and rapid, high performance is required, the API-based approach is advantageous. For long-term stable operation, version pinning, and accuracy gains through targeted data growth, fine-tuned BERT is preferable. This paper further argues that human monitoring of outputs, statistical quality control of data, and collaboration between domain expertise and data science are essential and offers a practical roadmap for quantitative analysis in the social sciences using LLMs.