一般論文

Peer-Reviewed

オンライン動画共有サービスにおける視聴者感情の推定

菅野　祐希[a]・坂野　遼平[b]

近年のインターネットやスマートフォンの普及により，動画共有サービスに多くの人がアクセスするようになった．YouTube や TikTok に代表される動画共有サービスは，現代において様々な人の行動や選択に大きな影響力を持つようになっている．それにより動画共有サービスはビジネスやマーケティングの場としても活用されるようになった．視聴者が動画を閲覧することでどのような感情を得るかという情報は，視聴者とマーケターの両方において有益となる．本稿では，オンライン動画共有サービス上にアップロードされた動画に付けられたコメントから，動画の視聴者に引き起こされる感情の推定を行う手法を提案する．BERT 及び数種類の大規模言語モデル (LLM) を用い，動画コメントを利用した各モデルの感情推定に関する能力の違いを明らかにする．提案手法では，7 種類の感情の強さを成分とした 7 次元ベクトルにより感情を表現し，推定を行う．実験の結果，100 件のコメントを使用して精細な感情の強度を推定する場合には LLM が優位であることが分かった．また，10 件の少ないコメント件数から最も強い感情を推定する場合，BERT が高いスコアを示した．

キーワード：オンライン動画サービス，感情推定，BERT，大規模言語モデル

Estimating Viewer Emotions Using Comments on Online Video-sharing Services

Yuki Kanno[a] and Ryohei Banno[b]

The rapid proliferation of internet access and smartphones has significantly increased engagement with video-sharing platforms like YouTube and TikTok. These platforms have emerged as powerful influencers of individual behavior and societal trends, particularly in the domains of marketing and consumer decision-making. Understanding the emotional responses elicited in viewers during video consumption holds substantial value for content creators and marketers. This study presents a novel approach for estimating viewer emotions based on user-generated comments associated with online videos. By leveraging Bidirectional Encoder Representations from Transformers (BERT) and various large language models (LLMs), the emotion estimation capabil-

[a] 工学院大学大学院 工学研究科, Graduate School of Engineering, Kogakuin University
[b] 一橋大学大学院 ソーシャル・データサイエンス研究科, Graduate School of Social Data Science, Hitotsubashi University
本論文は言語処理学会年次大会で発表された同一著者による研究 (菅野，坂野 2024) に加え，使用したモデルと評価手法を増やす形で発展させたものである．

ities of different models are systematically compared. Emotions are represented as a seven-dimensional vector, where each component reflects the intensity of a specific emotional state. Experimental evaluations reveal that LLMs perform better in estimating nuanced emotional intensities when a larger set of comments (e.g., 100) is available. Conversely, BERT performs better in identifying the predominant emotion when only a limited number of comments (e.g., 10) are analyzed. These findings highlight the complementary strengths of different models and offer practical insights into emotion estimation from social media data.

Key Words: *Online Video Services, Emotion Estimation, BERT, Large Language Models*

1 はじめに

近年，動画共有サービスの普及が進んでいる．調査によって，インターネットの利用時間の内訳で，最も長い時間利用されているのが動画共有サービスであることが分かっている (総務省 2024)．動画共有サービスの中でも特に影響力を持つものとして，YouTube[1]やニコニコ動画[2]，TikTok[3]等が挙げられる．

動画共有サービスは，現在マーケティングの場としても扱われるようになっている (YouTube 2023; OXFORD ECONOMICS 2022)．テレビを見ない代わりにインターネットを利用する傾向にある若年層にも効率的に商品を紹介することができ，企業等が著名なインフルエンサーに商品紹介を依頼して報酬を支払うという事例も多数存在する．

動画媒体は人間の感情に大きな影響を与えることが分かっている (Gross and Levenson 1995)．また，動画サイトにおける広告やテレビ CM に代表される，動画を用いたマーケティングにおいて視聴者の感情をコントロールすることで商品に対する見方を変えられる事が実証されている (大友 他 2010)．こうした事例から，動画視聴者がどのような感情を動画から受け取ったかという情報は，非常に有用な情報であると言える．例えば，ある商品の広告を，その商品の購買意欲を促進する上で適した感情を誘起する動画コンテンツに差し込むといった応用が考えられる．また，動画をアップロードするクリエイターにとっても感情の変化を目的としたコンテンツ制作の際の重要な参考情報となるほか，メンタルヘルスの観点からも有益な情報となりえる．この他，感情情報の応用例としては，特定の感情を与える動画を検索可能とするシステムや，視聴動画と類似した感情を与える動画の推薦システムといったものが挙げられる．

図 1 に感情による動画推薦システムの概要を示す．感情を入力として，データベース内で動画に結びつけられた感情ラベルと比較を行い，入力された感情を発生させる動画を出力する．

[1] https://www.youtube.com
[2] https://www.nicovideo.jp
[3] https://www.tiktok.com

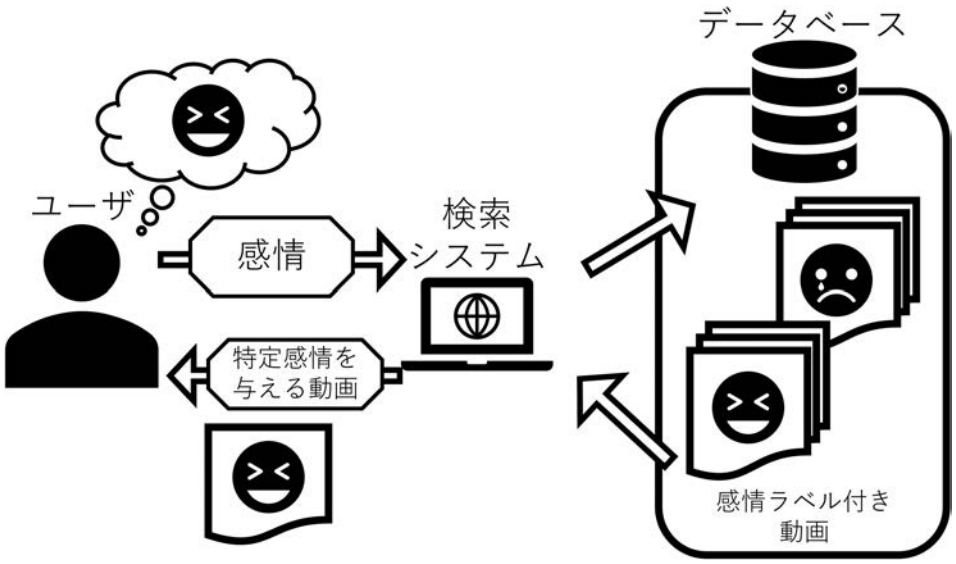


図 1　感情による動画検索・推薦システム概要

このように動画と感情を結び付けた応用例が考えられる.

感情を推定する手法についてはこれまで様々な研究がなされてきている．視聴者の表情等, 生体情報を用いて感情を測定する手法 (Soleymani et al. 2012; Suhaimi et al. 2020) であったり，視聴者にアンケートを取る手法などが挙げられる (Drigas and Papoutsi 2021)．これらを用いることで動画から得られた感情を推定することは可能である．しかし，動画コンテンツの数に対するスケーラビリティが問題となる．一例として，YouTube では 2016 年時点で 20 億本以上の動画がアップロードされている (Tiernan 2016)．動画共有サービスの利用者数が増加傾向にある中，各サービスにおける動画コンテンツの数は一層増加していると考えられる．先述のような手法では，数多くの動画に対して視聴者が得た感情を判別することはコストや時間の面で難しい．動画コンテンツそのものの映像や音声等の情報から視聴者に与える感情を推定することは容易ではなく, 我々が知る限りでは高精度な推定を可能とする手法は実現されていない.

本研究では，世界的に最もよく知られた動画共有サービスの一つである YouTube を対象とし，コメント文を用いて動画が視聴者に与える感情を推定する手法を提案する．YouTube 動画に付加されているコメントは, 視聴者が感じ取ったことを含んでおり，視聴者の感情をある程度反映していると考えられる. そこで，YouTube 動画コメントを利用することで，視聴者が得た感情を言語モデルを用いて抽出し, 推定する．近年では大規模言語モデルが発達し様々なタスクに応用されつつあるが，本稿で提案している視聴者コメントに基づく動画視聴者の感情推

定については精度や推定の傾向は明らかではない．そこで，BERT 及び 5 種類の大規模言語モデル (LLM) を用い, 当該タスクにおける推定能力の差異を明らかにする．具体的には，深層学習モデルとして BERT (Devlin et al. 2019) を，大規模言語モデルとして GPT (3.5-turbo, 4, 4-turbo, 4o) (Brown et al. 2020; OpenAI 2024)，Claude (3.5 Sonnet, 3 Sonnet, 3 Opus) (Anthropic 2024a, 2024b)，Gemini (1.5 Pro, 1.5 Flash) (Gemini Team Google 2024a, 2024b)，Deepseek (DeepSeek-AI 2024)，Llama-3-ELYZA-JP-8B (Llama team 2024; ELYZA, Inc. 2024) を使用している．

本論文の貢献は以下のとおりである．

- オンライン動画サービスのコメントを用いて，言語モデルにより視聴者の詳細な感情を 7 次元ベクトルとして推定する手法を提案する．
- 評価に用いる正解データの作成に関する予備実験として，100 名の実験協力者によるアンケートを実施し，正解データの作成に必要なアンケート回答件数を検証する．
- YouTube の実データを用い，1 種類の深層学習モデル及び 5 種類の大規模言語モデルによる差異を明らかにする．7 次元のベクトルの推定精度及び，最大感情の推定精度という 2 つの側面から評価を行う．

2　関連研究

機械学習を用いた感情推定の既存研究について，本研究と関連が深いものを以下に述べる．

2.1　ルールベース・深層学習による感情推定

深層学習以前に存在した，文章による感情推定の手法としてルールベースによる研究が存在する (Lee et al. 2010)．出現する単語や文章の形から感情の推定を行うものであり，現在主流である BERT や LLM と比較すると，柔軟な推定が不可能であるため性能が低い傾向にある．

2017 年に深層ニューラルネットワークモデルとして提案された Transformer (Vaswani et al. 2023) アーキテクチャは，2018 年に BERT のモデル構造として採用されてから多くの場面において利用されることになった．機械翻訳 (Zhu et al. 2020) やテキスト分類 (Qasim et al. 2022)，文書要約 (Grail et al. 2021) などの NLP タスクに応用されている．NLP タスクの一つに挙げられる感情推定に関しても，BERT を始めとした深層学習モデルを用いた手法 (Acheampong et al. 2021) が提案されている．

サイバー虐待検知のために文章内における攻撃性の検知を行うシステムの提案 (Malte and Ratadiya 2019) においては，エンコーダーに対し訓練用データセットを入力することによって各単語の複数の Attention 特徴を学習することにより文章内の攻撃性の分類を可能とした．X（元 Twitter）上のデータの感情分析及び感情認識へのアプローチ (Chiorrini et al. 2021) にも BERT

が使用された．事前学習済みモデルに対し，Fine-tuning を行うことによって感情の分類を行う手法が提案されている．

2.2 LLM を用いた感情推定

Transformer による事前学習を発展させ，大規模なテキストコーパスを用いた GPT が提案された．BERT との違いとして，Transformer のデコーダを利用している点が挙げられる．数億から数千億のパラメータによる学習を用いることにより，事前学習や Fine-tuning 無しでも高い言語処理能力を持つ．後続のモデルとして Gemini や Claude，Llama などが発表され，現代において LLM の選択肢は幅広い．大量のテキストを学習した LLM がどのような言語処理能力を有するかは不透明な点も多く，LLM による判断や推定の特徴を評価する研究が多くなされている．

LLM がコーパスから獲得している社会集団に対する偏見や感情について評価を行う研究 (田中 他 2024) や，LLM が特定状況において想起する感情について，人間との比較を行っている研究 (Huang et al. 2024) などが存在する．それらの研究において，LLM が状況に応じて感情の強度やポジティブ・ネガティブを変化させること，LLM が持つ特定の社会集団に対する感情がすでに社会調査によって得られている人間の感情傾向と一致すること等が示されている．ここから，LLM は人間に近い感性や考えを，コーパスから獲得している可能性が示唆されている．

様々な言語のコーパスから学習を行っていることから，英語などのメジャーな言語のみならず，話者の少ない言語においてもタスクを遂行出来る．LLM と少数話者の言語に特化させたモデルとの感情推定能力の比較を行った研究 (Rathje et al. 2024) では，Fine-tuning を行っていないにも関わらず，多くの特定言語特化モデルに性能が匹敵したことが示された．

Hama らの研究では，会話における感情認識に LLM を応用している (Hama et al. 2024)．LLM に会話文を入力して感情を予測させる 1-Step プロンプティング，及び会話文における発話の意味を問う質問の生成と回答を LLM 自身に行わせたうえで感情を予測させる Multi-Step プロンプティングを提案している．MELD 及び IEMOCAP データセットを用い，7 種類の感情のマルチラベル分類を行う評価により，Multi-Step プロンプティングが感情推定精度を向上し得ることが示されている．

薛らの研究では，Plutchik の感情の輪に存在する感情の有無を LLM により測定する手法を提案している (薛，小林 2024)．この研究では，LLM に分析対象テキストとプロンプトを与え，48 種類の感情によるマルチラベル分類を行っている．また，得られた感情を 8 次元の感情ベクトルに変換する手法も示されており，各成分は 0 から 3 の 4 段階の値を取る．

LLM による感情推定は，AI コンパニオンアプリケーション[4,5]による人間との対話等に応用されている．こうしたアプリケーションによる LLM のメンタルヘルス等への活用は，LLM が

[4] https://www.yuna.io
[5] https://www.luzia.com/en

人間の感情や感性を理解することへの社会的ニーズの一端を示している．

本節で言及した既存研究に対する本研究の主な特徴としては，感情推定の対象の違い，データセットの違い，比較検証範囲の違い，出力の違いなどが挙げられる．例えば，Hama らの研究は会話文における発話の感情推定を想定し，MELD 及び IEMOCAP データセットを用いて 1 種類の LLM を BERT ベースの SPCL と比較評価している．また，薛らの研究では，SNS 投稿等の単一文章の感情推定を想定し，WRIME データセットを用いて 4 種類の LLM について評価を行っている．これらに対し，本研究では動画視聴者の感情推定を目的としており，動画と紐づく複数のコメント文から感情推定を行う．Fine-tuning 及び評価にはクラウドソーシング及び YouTube Data API を用いて収集・構築したデータセットを用いており，BERT 及び 5 種類の LLM についてバージョンやパラメータの違いによる計 15 のバリエーションで幅広い比較検証を行っている．また，本研究では，7 種類の感情の強度を成分とする感情ベクトルを出力しており，Hama らの研究における 7 種類のマルチラベル分類や，田中らの研究におけるポジティブ・ネガティブ・ニュートラルの分類に基づく分析とは異なっている．

2.3　動画視聴者の感情推定

動画の視聴者の感情を推定する研究について述べる．ルールベースや機械学習を用いた視聴者感情推定の手法 (Yi et al. 2003; Yasmina et al. 2016) が存在している．これら手法では，動画のコメントを用いてルールベースや機械学習により感情の推定を行っている．

また，視聴者の生体データから感情を探る手法 (Soleymani et al. 2012) も提案されている．目の動きや反応等をもとに，活発さを示す覚醒度及び快適さを示す価値観の観点から感情を推定している．動画共有サイトにおける炎上に着目した研究 (堺 他 2021) では，YouTube において動画コメントのネガティブさを評価指標として，動画が炎上しているか否かの推定を行っている．BERT を用いた手法によって動画コメントの感情を推定している．

以上のように様々な研究がなされているものの，本研究のようにオンライン動画サービスのコメントを用い，深層学習モデルや大規模言語モデルにより視聴者の詳細な感情を同一の基準を用いて推定することは行われていない．必然的に，近年多数のバリエーションが生じている大規模言語モデル間の推定能力の差異も明らかではない．

本研究では，前述の課題に対し，BERT 及び LLM を活用した感情推定手法を提案し，モデル間の感情推定能力の差異を明らかにする．オンライン動画サービスにおける視聴者コメントは，スラングを含んだ短文が多いことから，BERT を始めとする事前学習済みモデルより前の自然言語処理技術では，特徴量エンジニアリングに大きな労力を要する問題がある．また，感情推定の精度面も限定的と考えられ，実際に我々の過去の検証ではルールベースによる推定は精度が BERT に劣ることが確認されている (菅野，坂野 2023)．松林らの研究 (松林 他 2016) においても，崩れた単文が多い点で動画コメントと類似している SNS の投稿を対象とし，ランダムフォ

レストによる 5 種類の感情への分類を試みた結果，一般的な投稿文を用いた検証にて妥当と評価されたものは検証データ中 35.5%に留まった事が報告されている．また，前述の Soleymani らの研究のような，視聴者の生体情報等に基づく感情推定手法も存在するが，膨大な量の動画コンテンツに対して個別に感情推定を行うことはコスト等の面で現実的ではない．これらを踏まえ，本研究では BERT 及び LLM を用いた感情推定を行う．LLM を用いた感情推定の既存研究に対する差異は，2.2 節末尾に述べたとおりである．

なお，本研究では多数の動画コンテンツの感情を推定し，マーケティングや推薦等に適用することを想定している．各動画の感情推定を比較的少量のコメントで行うことが出来れば，実応用時のコストを低減でき，かつコメント数が少ない動画にも適用できることから，入力するコメントは全件ではなく 100 件あるいは 10 件としている．詳細については以降の章で述べる．

3 提案手法

本節では，推定を行う感情について述べ，次に BERT による感情推定手法について述べる．最後に大規模言語モデルを利用した感情推定手法について述べる．

3.1 推定を行う感情について

本研究では，推定を行う対象の感情について，日本語感情表現辞書 (SOCIOCOM Social Computing Laboratory) より，以下の 7 種類の感情に設定している．

- sad（悲しみ）
- anxiety（不安）
- angry（怒り）
- disgust（嫌悪感）
- trust（信頼）
- surprise（驚き）
- joy（喜び）

これらの 7 種類の感情を，各成分 0 以上 1 以下の 7 次元の単位ベクトルとして扱う（以下，“感情ベクトル” とする）．以降に述べる各手法では，コメントを入力とし，感情ベクトルを出力とする．

感情推定においては，2.3 節に挙げた関連研究 (田中 他 2024) のように，感情極性を想定して分類や強度推定を行う事例が多く見られる．これに対し，本研究では 7 次元の感情ベクトルを用いる事で，より詳細な感情の推定を図る．

詳細な感情の推定は様々な応用において有用と考えられる．一例として，マーケティングにおいては特定の感情を誘起する動画コンテンツをユーザに推薦することで，ユーザの購買意欲

を高められる可能性がある．その際，感情がポジティブ・ネガティブの 2 種類となっていると，不適切な感情を誘起してしまう恐れがある．例えば，視聴者が被災者の置かれている状況への共感をチャリティーグッズ販売に結び付ける必要がある場合，動画が分類される悲しみではなく，同じネガティブ感情である怒りや嫌悪感を誘起してしまうケースや，技術書販売において信頼の誘起が有効である場合に，同じポジティブ感情である喜びを誘起してしまうケース等が考えられる．本研究における 7 感情ベクトルは，こうした応用において有用となることが期待できる．

感情の分類については，様々なモデルが存在している．代表的なものとして，Ekman による分類や，Plutchik の感情の輪が挙げられる．日本語感情表現辞書は，広く認知された感情分類である Plutchik の感情の輪をベースとしている．クラウドソーシングによって様々な体験に関する文章を収集したデータセットである EPISODE BANK を用いており，文章と感情を結び付けたうえで，各感情中の単語の出現頻度に基づいて単語ごとの感情値を定めたものである．日本語感情表現辞書の 7 感情を用いる事で，これら単語ごとの感情値を活用して感情推定を行うアプローチとも比較することが出来，多面的な検証が可能となることから，本研究ではこの 7 感情を採用した．なお，単語ごとの感情値の活用については，実際にルールベースの感情推定手法の検証を行い，発表済みである (菅野，坂野 2023).

実応用においては，7 次元感情ベクトルの全体を用いるケースや，最大感情を抽出して用いるケースが考えられる．例えば，ユーザが視聴している動画と類似する感情を得られる動画を推薦するシステムの場合，7 次元ベクトルの類似度が高い動画を推薦対象とすることが考えられる．また，ユーザが指定する感情を得られる動画を提示するシステムの場合，7 次元ベクトルの中で最も強く表れている成分をその動画の感情ラベルとし，指定感情とマッチする感情ラベルを持つ動画を提示することが考えられる．こうした応用例を想定し，5 章では，7 次元ベクトルの推定精度評価及び最大感情の推定精度評価を行っている．

3.2　BERT による感情推定手法

大まかな流れとして，事前学習済みの BERT に対し，感情のラベルが付加された動画のコメントを用いて Fine-tuning を行う．特定感情とコメントの間にある関係性を学習させることによって BERT を感情分類を行う分類器として用いる．その後，推定を行いたい動画のコメントを入力し，出力として感情ベクトルを得る．

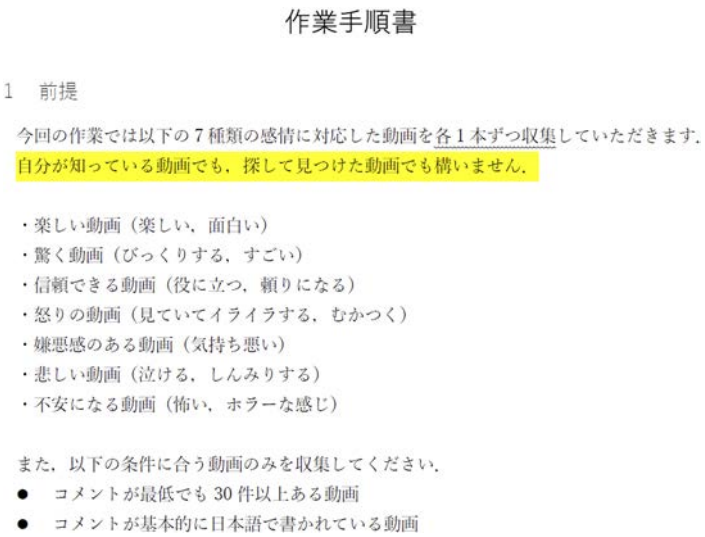
本研究では，東北大学乾・鈴木研究室が公開している日本語事前学習済み BERT モデル[6]を用いる．当該モデルは日本語版の Wikipedia コーパスによって学習が行われている．感情ごとのコメントの傾向を学習させるために，Fine-tuning を行う．Fine-tuning で用いる動画のコメン

[6] https://github.com/cl-tohoku/bert-japanese?tab=readme-ov-file

トは，クラウドソーシング[7]を用いて収集した．150 名の実験協力者を対象に，視聴した際に 3.1 節で示した 7 つの感情を得られる動画を 1 本ずつ提示させた．これにより，各感情につき 150 本ずつ，合計 1,050 本の動画の ID を得た．図 2 に動画 ID 収集時の作業手順書の内容を，図 3 に入力用フォームを示す．

各動画から YouTube Data API[8]を用いて，人気順で上位 100 件のコメントを収集した．コメントを収集し，テキストファイルへと変換する際に絵文字・記号・英字・数字の削除を行っている．人気順とは，コメントについたコメント（子コメント）や，コメントについたいいねの数量に基づいてコメントの人気を評価し，順序付けしたものである．本実験では，コメントの取得を全て人気順で行っている．

収集の際に提示された感情でコメントにラベルを付与した．例えば，sad（悲しみ）の感情を得る動画として提示された動画 150 本のコメントにはそれぞれ sad（悲しみ）のラベルを付与した．こうして得られたラベル付きのコメントデータを用い，Fine-tuning を行った．Fine-tuning によって出力される感情ベクトルの値が変化するため，Fine-tuning と感情ベクトルの出力をセットとして，10 回行った平均を BERT の出力結果としている．



作業手順書

1　前提

今回の作業では以下の 7 種類の感情に対応した動画を各 1 本ずつ収集していただきます．
自分が知っている動画でも，探して見つけた動画でも構いません．

・楽しい動画（楽しい，面白い）
・驚く動画（びっくりする，すごい）
・信頼できる動画（役に立つ，頼りになる）
・怒りの動画（見ていてイライラする，むかつく）
・嫌悪感のある動画（気持ち悪い）
・悲しい動画（泣ける，しんみりする）
・不安になる動画（怖い，ホラーな感じ）

また，以下の条件に合う動画のみを収集してください．
- コメントが最低でも 30 件以上ある動画
- コメントが基本的に日本語で書かれている動画

図 2　動画 ID 入力作業手順書

[7] https://crowdworks.jp
[8] https://developers.google.com/youtube/v3

図 3　動画 ID 入力用フォーム

3.3　大規模言語モデルによる感情推定手法

LLM では追加の学習を行わず，コメントからどのような感情を推定することが出来るかをプロンプトを用いて判別させる．大まかな流れとして，大規模言語モデルに推定を行いたい動画のコメントを入力する．1 コメントずつの入力となり，プロンプトによって指定されたフォーマットによって感情値の返答が行われる．最終的に，返答された感情の値の平均を取ることによって，感情ベクトルを出力する．プロンプトを以下に示す．

プロンプト 1　感情推定プロンプト

```
"あなたには入力した動画のコメントから，動画が視聴者に与える感情の推定を行っていただきます．"
"条件として，感情は必ず次に示す7つの選択肢から選択してください．"
"[悲しい],[不安],[怒り],[嫌悪],[信頼],[驚き],[嬉しい]"
"感情の強さに応じて，0,1,2,3,4の5段階で評価してください．"
"感情が強く含まれる場合は大きい数字，感情が含まれない場合は小さい数字で評価してください．"
"今日は新発売のケーキを買った．最高の一日だ．"
"[悲しい:0][不安:0][怒り:0][嫌悪:0][信頼:0][驚き:1][嬉しい:4]"
"愛猫が亡くなってしまった．泣きながら一日眠ってしまった．"
"[悲しい:4][不安:2][怒り:0][嫌悪:0][信頼:0][驚き:2][嬉しい:0]"
```

GPT や Llama, Claude は，入力の例と理想の返答について設定することが可能である．Gemini に関しては入力例と返答例を送信することが不可能であるため，プロンプト 1 の下 4 行の例示を行っていない．例示として使用した文章は作成した物であり，文章の感情値は実験協力者 5 名にアンケートを取った平均である．

4　正解データの作成に関する予備実験

提案手法の評価においては，実験協力者に対し動画から得られる感情についてアンケートを実施し，結果を平均することで正解データを作成する．正解データを作成するために十分なアンケートの回答件数を明らかにするために，予備実験を行った．動画の中で最も強く得られる感情について，何件のコメントがあれば安定して取得可能かについて評価している．

3 本の動画について，クラウドソーシングを用いて 100 名の実験協力者に対して動画を視聴させ，7 つの感情の強度について 1 から 5 の 5 段階でアンケートを行った．図 4 に本研究で用いた感情評価用のアンケートフォームを示す．

手法として，100 件のアンケート結果を平均した値を基準値とし，100 件の中から一定件数をランダムに抽出して平均した値と比較する．この作業を 50 回繰り返す．基準値から値が大きく離れている場合はアンケートの件数として適さない．基準値と値が近い場合，特に最も強い感情が一致している場合はアンケートの件数として適していると考えられる．表 1 に抜き出し件数ごとの誤差の大きさの平均と最大感情の一致率を，表 2 に抜き出し件数ごとの一定以上の誤差を示した件数を示す．

誤差大きさ平均は，基準値と抜き出し平均のベクトル成分の誤差を示す．基準値と抜き出し平均において，同じ動画の同じベクトル成分から誤差の絶対値を求める．以上の作業を全ての動画，ベクトル成分に対して行う．最終的にそれらの平均を算出する．

最大感情一致率は基準値と抜き出し平均のベクトルを比較し，最も値の高い成分が一致した割合を示している．動画 1 から 3 において，基準値と抜き出し平均の感情ベクトルを比較し，最も高い成分が一致するかを判定する．最終的にそれらの割合を算出する．

誤差数は基準値と抜き出し平均のベクトル成分の誤差を，誤差の絶対値の大きさごとにカウントしたものである．誤差大きさ平均と同様に，同じ動画の同じベクトル成分から誤差の大きさを求め，それらを集計した．±0.05 の場合，誤差が −0.05 以上 0.05 以下である数をカウントしている．

結果として，アンケート結果を 10 件以上抜き出すことで最大感情の一致率は 95%を越え，コメントを 20 件以上集めることで 99%を越えることが分かった．また，30 件以上集めることで 100 件のコメントを集めた時と同一の最大感情一致率を示すことが分かった．誤差の大きさは，10 件集めることで 0.236 となり，0.25 を下回る．20 件集めることで 0.154 となり 0.2 を下回る

図 4　感情評価用アンケートフォーム

ことが分かる．アンケート結果の成分ごとの誤差に関しては，30 件以上収集することで，100 件収集した時の成分と比較して，±0.5 以上の誤差が発生しなくなる．

以上の結果より，10 名以上のアンケート回答を平均化することで，95%以上の確率で 100 名のアンケート回答と同様の最大感情を定められることが分かった．多くの動画について正解データを作成する場合，この結果に基づいてアンケート回答数を設定することで実験コストを抑えることが出来る．以降の実験では，5.1 節において 3 本の動画について 100 名のアンケート回答から正解データを作成して用い，5.2 節においては 30 本の動画について 1 本あたり 10 名のアンケート回答から正解データを作成して用いている．

抜き出し件数	誤差大きさ平均	最大感情一致率
3	0.442	0.873
5	0.332	0.913
7	0.294	0.940
10	0.236	0.953
20	0.154	0.993
30	0.120	1.000
40	0.091	1.000
50	0.076	1.000
75	0.044	1.000

表 1　アンケート件数についての評価結果（誤差大きさ）

抜き出し件数	誤差数					
	±0.05	±0.1	±0.15	±0.3	±0.5	±0.5 以上
3	89	69	48	238	243	363
5	75	140	85	285	256	209
7	141	54	149	288	232	186
10	126	178	119	318	216	93
20	244	201	185	285	117	18
30	300	214	199	294	43	0
40	381	293	178	182	16	0
50	467	304	151	125	3	0
75	699	280	62	9	0	0

表 2　アンケート件数についての評価結果（誤差件数）

5　評価実験

本節では，提案手法の評価実験について説明する．二つの評価を実施した．一つ目は 100 件のコメントから 7 次元感情ベクトルを推定する精度の評価であり，二つ目は少数のコメント（10 件）から最大感情を推定する精度の評価である．BERT 及び 5 種類の大規模言語モデルについて，これら二種類の観点で評価を行い，言語モデル間の感情推定能力の比較を行う．

本研究では，以下のモデル及びバリエーションについて評価を行っている．

- BERT
  - 350 file 256 token, 350 file 512 token, 1,050 file 256 token, 1,050 file 512 token
- GPT
  - 3.5-turbo, 4, 4-turbo, 4o
- Deepseek
- Gemini
  - 1.5 Flash, 1.5 Pro

- Claude
  - 3 sonnet, 3 opus（類似度評価のみ），3.5 sonnet
- Llama-3-ELYZA-JP-8B

BERT は，Fine-tuning の際のパラメータによって 4 種類の推定モデルを用意している．Fine-tuning に用いたテキストファイルの数 (350 file, 1,050 file) と Fine-tuning の入力 Token 数 (256 token, 512 token) の組み合わせである．前者は学習に用いた動画の本数，後者は 1 動画ごとのコメント量と対応している．

5.1　7 次元感情ベクトルの推定精度評価

7 次元感情ベクトルの推定精度評価では，推定を行いたい動画のコメント 100 件を入力とし，感情ベクトルを出力とし，正解データとの類似度を算出することで推定精度の評価を行う．評価に用いた動画は以下の 3 本である．

- 動画 1: 【衝撃】火山にゴミを捨てて処理する場合に起こること（VIENCE バイエンス）[9]
- 動画 2: back number - 手紙 (full)[10]
- 動画 3: 貫禄ありすぎて、父親と間違われる引きこもり生徒【ジェラードン】[11]

各動画について，4 章で示した 100 名のアンケート回答を用い，平均・正規化して単位ベクトルとしたものを正解データとした．アンケートでは 1 から 5 の 5 段階としたが，感情が得られない場合の値を 0 とするため，平均の際は 0 から 4 の 5 段階として扱った．各モデルから出力された感情ベクトルと正解データをコサイン類似度により比較し，その数値を評価の基準とした．

表 3 に各動画の感情ベクトルの正解データを示す．表 3 に示されている通り，動画 1 は信頼，驚き，喜びの 3 つの感情において高い数値を示しており，複数の感情が同時に表れやすい動画であることが示されている．動画 2 では悲しみが突出して高く，信頼と喜びも他の感情と比較してやや高い事が分かる．動画 3 では喜びの感情が最も高く，驚きの感情も高い事が示されている．全ての動画において感情の傾向が分かれていることが分かる．表 4 に評価結果を示す．

3 つの動画において最も高いスコアを示したモデルは，動画 1 が GPT-4 の 0.9950，動画 2 が

感情	sad	anxiety	angry	disgust	trust	surprise	joy
動画 1	0.1487	0.2588	0.1197	0.1352	0.5465	0.6083	0.4577
動画 2	0.7051	0.0673	0.0972	0.0747	0.4709	0.2791	0.4285
動画 3	0.0621	0.1002	0.1623	0.1599	0.2458	0.5489	0.7566

表 3　正解データの感情ベクトル

[9] https://youtu.be/uJBHO2NXvxQ
[10] https://youtu.be/woRV5VxJDkU
[11] https://youtu.be/OtJvFyqqeo0

Model	BERT				GPT			
	350/256	350/512	1050/256	1050/256	3.5-turbo	4	4-turbo	4o
動画 1	0.8340	0.8061	0.7621	0.7376	0.9796	0.9950	0.9592	0.9925
動画 2	0.8609	0.8539	0.9035	0.8885	0.8997	0.9427	0.9349	0.9231
動画 3	0.8896	0.8359	0.9189	0.8901	0.9522	0.9397	0.9877	0.9509
Model	Gemini		Claude			Deepseek	ELYZA-llama	
	1.5-flash	1.5-Pro	3 sonnet	3 opus	3.5 sonnet			
動画 1	0.7720	0.9486	0.9582	0.9856	0.9775	0.9867	0.9397	
動画 2	0.8196	0.9458	0.9456	0.9421	0.9354	0.9154	0.9202	
動画 3	0.8487	0.9789	0.9804	0.9611	0.9672	0.9408	0.9774	

表 4　言語モデルごとの 7 次元感情ベクトルと正解データのコサイン類似度

Gemini-1.5-Pro の 0.9458，動画 3 が GPT-4-turbo の 0.9877 となっている．LLM については，ほとんどが 0.9 以上のスコアを示しており，0.95 を越えるモデルも多い．BERT は LLM と比較してスコアが低く，その中では 350 file 250 token で学習されたものは性能が比較的高く，1,050 file 256 token で学習されたものは動画 2 と動画 3 において高い性能を示している．

動画 1 に関しては複数の感情が高く，joy や sad などの比較的分かりやすい感情以外が高いという特徴がある．GPT-4 及び GPT-4o は複雑な成分の感情に対して強力であるという特徴がある．それ以外の LLM は 0.9500 程度が多く，複雑な感情の推定能力は一歩劣ると考えられる．

動画 2 に関しては全体的に 0.9500 以下になるものが多い．これはコメントの持つ感情の値と正解データに差が存在するためであると考えた．本動画は両親への家族愛を叙情的に書いた歌である (森 2015)．コメントを行う人は基本的に家族や家族愛に対する言及をすることが多い．そのため，コメントから直接感情推定を行うことによって得られる感情は，主に trust（信頼）が主になる傾向が見られた．また，sad（悲しみ）はあまり高い数値を示していない．しかし，クラウドソーシングによって集めた視聴者が感じた感情は sad（悲しみ）が中心である．そのため，1 回見ただけの視聴者とコメントを行う層が食い違うことによって類似度の低下を招いたと考えられる．

動画 3 に関しては，基本的に 0.95 以上を示すモデルが多い．joy（喜び）が分かりやすい感情であるためだと考えている．しかし，最も高いスコアでも 0.9900 を越えていない．喜びが最も高く驚きが次いで高いことを推定したモデルは多いが，喜びと驚きのベクトルの成分の比率を高い精度で一致させたモデルが少ないためであると考えた．また，正解データにわずかに含まれている angry（怒り）と disgust（嫌悪感）の感情が，LLM の推定においては極めて少なく推定されている（付録 A）．そのずれによってわずかに低いスコアを示していると考えられる．

5.2　最大感情の推定精度評価

最大感情の精度評価では，推定を行いたい動画のコメント 10 件を入力とし，感情ベクトルを出力とする．クラウドソーシングを用いて 10 名の実験協力者から 3.1 節で示した 7 つの感情を得られる動画を収集し，その中から各感情の動画が 4 から 5 本になるようにランダムに選定した 30 本の動画について感情推定を行った．表 5 に各感情ラベルが付与された動画の本数を示す．

各動画について，クラウドソーシングを用いて 100 名の実験協力者に，30 本の動画からランダムに選定した 3 本の動画を視聴させ，7 つの感情の強度について 1 から 5 の 5 段階評価でアンケートを行った．各動画につき 10 名のアンケート結果が得られ，平均して正規化し，単位ベクトルとしたものを正解データとした．5.1 節と同様に，感情が得られない場合の値を 0 とするため，平均の際は 0 から 4 の 5 段階として扱った．各モデルから出力された感情ベクトルの中で最も数値の大きいベクトル成分と，正解データで最も数値の大きいベクトル成分を比較し，一致した動画の数を評価の基準としている．

表 6 に結果を示す．正解数は 30 本の動画中で最大感情が正解データと一致した数，正解率はその割合，コサイン類似度平均は 30 本の動画ごとのコサイン類似度の平均を表している．

最もスコアの高いモデルは，BERT の 1,050 file 256 token で Fine-tuning を行ったモデルとな

感情	sad	anxiety	angry	disgust	trust	surprise	joy
動画数	4	4	5	4	4	4	5

表 5　選定動画のラベル

	BERT				
	350/256	350/512	1,050/256	1,050/512	
正解数	17	19	22	21	
正解率	57%	63%	73%	70%	
コサイン類似度平均	0.8652	0.8619	0.8831	0.8786	
	GPT				Deepseek
	3.5-turbo	4	4-turbo	4o	
正解数	14	18	17	17	21
正解率	47%	60%	57%	57%	70%
コサイン類似度平均	0.7843	0.8354	0.8321	0.8155	0.8245
	Gemini		Claude		ELYZA-
	1.5-flash	1.5-Pro	3 sonnet	3.5 sonnet	llama
正解数	17	21	20	16	17
正解率	57%	70%	67%	53%	57%
コサイン類似度平均	0.6699	0.8015	0.8480	0.8258	0.7758

表 6　モデルごとの最大感情の推定結果

り，正解数は 22 で正解率は 73%となった．次に性能が高いモデルとして，BERT の 1,050 file 512 token モデル，Gemini-1.5-Pro と Deepseek となった．ランダムに感情を選択した場合，正解率は 1/7（約 14%）となるため，ランダムに感情を選択した場合よりも性能が高く，推定する能力は全てのモデルにおいて存在していることが分かる．BERT は Fine-tuning の際に，入力可能な Token に上限がある．そのため，学習用のコメントを 1 動画につき 1 つのテキストファイルとして入力している関係上，入力されているコメントは人気順で上位のコメントに限られる．そのため，上位 10 件のコメントのみを使用した最大ベクトルの一致率評価において高い性能を示した可能性がある．それに対し，LLM は多数のコメントを平均することによって高い性能を持っていたが，少ないコメントではコメント一つ一つの持つ感情のバラつきを精細に感じ取ることで平均した際に想定していない感情が最も高くなる可能性を持つ．

詳細な結果（付録 C）から分かることとして，全体的に正答率の高い感情として，joy（正答率 80%），trust（正答率 73%）が挙げられる．正答率の低い感情としては，sad（正答率 45%），disgust（正答率 39%）が挙げられる．

BERT は LLM と比較してネガティブ感情 (sad, anxiety, angry, disgust) の正解率が高い傾向にある．LLM の正答率が低い感情において高い正解率を示していることから，LLM 自体がネガティブな感情の推定能力が低い可能性を示唆している．BERT は学習によって，ネガティブな動画のコメントが持っている雰囲気や傾向を上手く取得できていることが分かる．しかし，ポジティブな感情 (trust, surprise, joy) に関しては LLM の正答率の方が高い．そのため，Fine-tuning による学習を行った BERT と LLM は反対の性質を持っている．

6 分析

5 章の評価結果より，各言語モデルの感情推定能力を考察する．

6.1 BERT の感情推定性能

BERT に関しては，基本的に 1,050 file 256 token のものが高い性能を示していた．すなわち，より多くの動画について，より少ないコメントを学習したモデルが高い性能を示している．少ないコメントが良い結果を示している点については，YouTube のコメントを人気順で取得している影響が考えられる．人気順とは，コメントに付けられる「いいね」評価の高さに応じた順番である．[12] 「いいね」が多くついている，ユーザから共感を集めたコメントが上位となる．「いいね」が少ないコメントは，動画と無関係なものや内容が薄い物が含まれる場合があり，結果的に 256 token が 512 token を上回ったと考えられる．また，学習に用いたテキストファイル数

[12] https://support.google.com/youtube/answer/9482556?hl=ja

が多いほうがより高い精度での推定が可能であることが結果から考えられる．

LLM と比較すると，コメントそのものが持つ感情ではなく，特定の感情を与える動画とコメントの対応を学習しているという点で大きな違いがある．そのため，コメントから直接感情を感じることが出来ない場合に高い性能を示す．

正解率評価では，嫌悪感や怒りなどのネガティブな感情を与える動画に関しても推定を行った．それらの感情を人に与える動画のコメントは，直接的な感情がコメントから読み取ることが出来ないと考えられる．そのため，動画の与える感情とコメント内容の対応を学習した BERT が高い正解率を示した．

しかし，BERT の短所として感情推定の出力結果のベクトル成分の最低値が 0.2500 から 0.3500 程度になることが挙げられる（付録 A）．LLM の場合，最低値は 0.1000 未満になる．動画を見て得られない感情は 0 に近い値になるためである．しかし BERT の場合は得られない感情でもわずかに高い数値を示してしまう．そのため，感情ベクトルの類似度評価において低い値を示したと考えられる．

6.2　LLM の感情推定性能

全体を通して，Gemini-1.5-Flash を除いた LLM の感情推定の性能は高い．しかし，後発のモデル程性能が高いというわけではない．Claude 3 Sonnet と 3.5 Sonnet を比較すると，動画 1 の精度は上昇しているが，それ以外の精度はわずかに減少している．GPT に関しても，4 と 4o は動画 1 において性能が高く，対して 4-turbo は動画 3 において性能が高い．

LLM はコメントの入力件数を増加させることで動画視聴者の感情推定能力を向上させることが可能であることが分かる．コメントが少ない場合，GPT のモデルでは悲しみや嫌悪感，不安の正解数が低い．

また，LLM のコストと性能が必ずしも一致している訳ではない．表 7 に各大規模言語モデルの入出力 1Mtoken（100 万 token）における 2024 年 11 月におけるコストを示す．GPT-4 や Claude 3 Opus などの高コストなモデルが，必ずしも GPT-4o や Claude 3 Sonnet より高いスコアを示しているわけではないことが分かる．

6.3　BERT と LLM の違いについて

本研究では，BERT と LLM のデータ入力の形式が異なっている．これはデータ入力をモデルごとに行いやすい形式へ変換して行っているためである．データ入力の形式が異なることによる影響が存在すると考えられる．

BERT への入力は 100 件あるいは 10 件のコメントを結合したテキストファイルによって行われている．入力トークン数の制限によって，末端のコメントが読み込まれない場合がある．コメントは人気順で結合しており，上位には人目を惹く意外性のあるコメント等が入る場合があ

Model	入力	出力
GPT-3.5-turbo	0.50	1.50
GPT-4	30.00	60.00
GPT-4-turbo	10.00	30.00
GPT-4o	2.50	10.00
Gemini-1.5-flash	0.08	0.30
Gemini-1.5-pro	1.25	5.00
Claude 3 Sonnet	3.00	15.00
Claude 3 Opus	15.00	75.00
Claude 3.5 Sonnet	3.00	15.00
Deepseek	0.14	0.28

表 7　モデルの 1Mtoken あたりのコスト（単位：$）

る一方，感情を表現した素朴なコメントは下位となる場合がある．動画から得られる感情を直接的には含んでいないコメント等が上位にあった場合，その影響が相対的に大きくなる可能性がある．学習時にもトークン数により上位コメントを相対的に多く学習しており，動画の感情の傾向と上位のコメントの傾向を学習することで，正解数評価において，入力コメントそのものが持つ感情ではなく，入力コメントが付きやすい動画の感情を推定できたと考えられる．

それに対し LLM は 1 件ずつコメントを入力する形式である．100 件中の下位のコメントが削られるといったことなく，感情の推定が行われる．これにより，精度評価においては，人気下位であるものの視聴者感情が書かれたコメントを含むことで LLM が高いスコアを示すことに寄与した可能性がある．それに対し，正解数評価では入力コメント数は 10 件であり，上位のコメントのみで感情推定を行っている．動画と関係が薄いが多くいいねの付いたコメントや，動画の与える感情とは異なる感情の表現を含むコメントが 10 件の内多くを占めた場合，精度が低くなると考えられる．

また，正解率評価における BERT と LLM の性能の違いについては，特にネガティブな感情において性能差が見られた．これは，上記でも述べている通り，コメントそのものが持つ感情ではなく，特定感情の動画につきやすいコメントの傾向を BERT が学習していることが要因だと考えられる．BERT の Fine-tuning において用いたデータセットは，動画ごとに視聴者が最も強く感じた感情のラベルが付加されている．そのため，コメントそのものが持つ感情と動画の感情ラベルが異なる場合がある．表 8 に，実際に BERT の Fine-tuning に用いた動画コメントから抜き出した，ネガティブな動画感情とコメントから読み取れる感情が異なる例と，そのコメントの LLM (GPT-4o) による感情推定結果を示す．これらのコメントはテキストファイルの中でも上位に位置するコメントであり，トークン数 256 token でも読み込まれる．このように，動画の持つ感情とコメントから読み取れる感情が異なる事例も存在する．そのため，コメント

動画の感情	コメント	LLM による感情推定結果
悲しみ	この簡単な絵でここまで 表せるこの人がすごい	[悲しい:0][不安:0][怒り:0][嫌悪:0] [信頼:0][驚き:2][嬉しい:2]
不安	こうゆう絵を描く人って才能あるよな	[悲しい:0][不安:0][怒り:0][嫌悪:0] [信頼:3][驚き:1][嬉しい:2]

表 8　動画感情とコメント感情の差異

から読み取れる感情だけでは視聴者感情を上手く推定出来ない動画において，動画とコメントの傾向を学習した BERT による推定が有効に働くと考えられる．

6.4　動画視聴回数による影響について

5.1 節で述べたように，動画の視聴回数が視聴者感情，ひいては本研究における評価結果に影響を与えている可能性がある．本研究ではクラウドソーシングを用いて視聴者感情を収集し，正解データとして用いている．クラウドソーシングでは視聴する動画を我々が指定しているため，各作業者にとっては初回視聴となる動画が多かったものと考えられる．すなわち，正解データはおおむね初回視聴者が得る感情を反映しているはずである．

一方，オンライン動画サービスにおいて人気順で上位のコメントを残すのは，いわゆるヘビーユーザであり，感情推定に用いているコメントは当該動画を繰り返し視聴しているユーザによるものである可能性がある．動画の内容によっては，視聴回数によって得られる感情が変化する場合があり，動画 2 において正解データが悲しみの感情を強く示していたのに対し，LLM がコメントから信頼を強く読み取った事は，その影響を受けている可能性が考えられる．

ただし，本研究が想定する応用においては，この点の影響は限定的と考えられる．5.1 節で示したように，3 件の動画に対して多くの LLM では 0.9000 以上の結果が得られており，中には 0.9900 を超えているものもある．例えば，Gemini 1.5-Pro や Claude 3 sonnet，Claude 3 Opus は動画 3 件共に 0.9400 を超えている．コサイン類似度のベースラインとして，ランダムな感情ベクトル同士のコサイン類似度を 100 回計算したところ，平均は約 0.764 となった．ジャンルの異なる 3 件の動画について，いずれもベースラインと比べて高いコサイン類似度が得られていることから，上記の影響が存在するとしても正解データとコメントに基づく感情ベクトルとの乖離は限定的と考えられる．

本研究ではマーケティングやコンテンツ推薦への応用を想定しており，感情推定の精度が高いことが好ましくはあるものの，正解データとの完全な一致を必要とするものではなく，そもそも視聴者によって得られる感情に多少の幅があるものであり，上記の影響は限定的と考えられる．また，繰り返し視聴されにくいジャンルの動画（時事ニュース等）であれば，コメントは初回視聴者によるものとなりやすく，上記の影響は生じづらいという想定の元，提案手法を

適用可能と考えられる．

なお，初回視聴における感情の推定が求められる場合には，コメントから初回視聴によるものを選別して感情推定を行うアプローチが考えられる．YouTube Data API においては，ユーザ個別の動画視聴回数は得られない事から，コメント内容に基づく選別が必要となる可能性がある．こうしたアプローチについては，本研究ではスコープ外としているが，今後検討すべき課題のひとつである．

7　おわりに

本研究では，オンライン動画共有サービスにおける動画視聴者の感情を自然言語処理によって推定する手法を提案し，実際の視聴者の感情と提案手法による推定結果を比較することによって評価を行った．深層学習モデルとして Fine-tuning を行った BERT を，大規模言語モデル (LLM) として GPT-4 を始めとした複数の LLM を使用し，これらモデル間の感情推定能力の差異を明らかにした．

精細な感情推定の能力を評価する感情ベクトル類似度評価では，GPT-4 や GPT-4-turbo, Gemini-1.5-Pro を始めとした 2024 年に発表された最先端のモデルにおいて高い性能を示した．対して BERT のスコアは比較的低い値を示した．

少数のコメントから最も強い感情を推定する能力を評価する最大ベクトルの一致率評価では，全ての LLM を越えて Fine-tuning を行った BERT が最も高いスコアを示す結果となった．次に Gemini-1.5-Pro や Deepseek が高いスコアを示した．また，BERT はネガティブな感情の正解率が高く，LLM はポジティブな感情の正解率が高い傾向が示された．

既存手法と比較し，提案手法では感情を 7 次元ベクトルで表現することにより，ポジティブ・ネガティブのみではない，詳細な推定が可能であり，複数の感情の強度を数値化できる．これにより，強い感情と弱い感情のみでなく，表出するが強くない感情などを捉えることも可能である．また，視聴者の生体情報等でなく，動画のコメントを入力として推定を行うため，数多くの動画に対して推定が可能である．

また，本研究では複数の LLM による同一の手法による感情推定を行った．これによって，各 LLM モデルの持つ感情推定能力にどのような差が存在するかについて結果を示すことが出来た．

7.1　今後の課題

今後の課題として，BERT の Fine-tuning の方法やデータの前処理方法に工夫を加えることで，推定能力の向上を図ることを考えている．また提案手法による感情推定を応用し，動画の検索や類似動画の推薦を行うシステムを検討していきたい．

また，提案手法の応用に際し，異なる複数の感情が共に強く表れるような動画において問題

が生じる恐れがある．賛否両論であるトピックを扱う動画や，過激な内容を含む動画では，ポジティブ・ネガティブ両方のコメントが多く書き込まれる場合がある．こうした動画は，関連動画の推薦システムへの応用等において問題となり得る．例えば，喜びの感情を得られる動画を求めているユーザに推薦した動画が，そのユーザにとっては怒りの感情を誘起するものであるといった状況が生じる可能性がある．ナイーブな対処方法としては，7 次元ベクトルとして出力が得られる点を活かし，複数の感情が共に強く表れる場合には推薦対象から除外するといった対応が考えられるが，推薦可能な動画コンテンツが限定されてしまう問題もある．こうした，異なる感情を強く与え得る動画の扱いについては，今後の課題である．

時間経過が及ぼす影響についても今後検討するべき事項である．本研究ではコメントを人気順で取得し感情推定を行っているが，動画の公開初期のコメントと，時間が経過した後のコメントでは，社会情勢等の変化により，異なる感情を内包している可能性がある．ひとつの動画内においても，長時間の動画の場合，視聴する箇所によって得られる感情が大きく異なる状況が考えられる．こうした，時間経過が及ぼす影響に対し，コメントの投稿日時やコメントが指し示す動画内の時間的位置を考慮した，感情推定についても検討したい．

この他，Plutchik の感情の輪をはじめとする各種感情分類モデルによる違いや，感情が発生しない場合を示すニュートラル感情の有無による影響を検証することも考えられる．本研究では 7 次元の感情ベクトルにおいて各要素が 0 に近い値を取ることで間接的にニュートラルな状態を表現可能であるが，実験用の動画 URL 収集等においてはニュートラル感情の存在は考慮していない．これを考慮することで，より精細な感情推定を行うことを検討したいと考えている．

参考文献

Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). “Transformer Models for Text-based Emotion Detection: A Review of BERT-based Approaches.” *Artificial Intelligence Review*, **54** (8), pp. 5789–5829.

Anthropic (2024a). “Introducing Claude 3.5 Sonnet Anthropic.” `https://www.anthropic.com/news/claude-3-5-sonnet`.（2024 年 11 月 20 日閲覧）.

Anthropic (2024b). “Introducing the Next Generation of Claude Anthropic.” `https://www.anthropic.com/news/claude-3-family`.（2024 年 11 月 20 日閲覧）.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). “Language Models are Few-Shot Learners.” *arXiv*

preprint arXiv:2005.14165.

Chiorrini, A., Diamantini, C., Mircoli, A., and Potena, D. (2021). "Emotion and Sentiment Analysis of Tweets using BERT." In *EDBT/ICDT Workshops*, pp. 1706–1711.

DeepSeek-AI (2024). "DeepSeek LLM: Scaling Open-Source Language Models with Longtermism." *arXiv preprint arXiv:2401.02954*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "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. Minneapolis, Minnesota. Association for Computational Linguistics.

Drigas, A. and Papoutsi, C. (2021). "Nine Layer Pyramid Model Questionnaire for Emotional Intelligence." *International Journal of Online & Biomedical Engineering*, **17** (7), pp. 123–142.

ELYZA, Inc. (2024).「GPT-4」を上回る日本語性能の LLM「Llama-3-ELYZA-JP」を開発しました | ELYZA, Inc. `https://note.com/elyza/n/n360b6084fdbd`.（2024 年 11 月 20 日閲覧）, [ELYZA, Inc. (2024).「GPT-4」wo Uwamawaru Nihongoseino no LLM「Llama-3-ELYZA-JP」wo Kaihatsushimashita | ELYZA, Inc. https://note.com/elyza/n/n360b6084fdbd.].

Gemini Team Google (2024a). "Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context." *arXiv preprint arXiv:2403.05530*.

Gemini Team Google (2024b). "Gemini: A Family of Highly Capable Multimodal Models." *arXiv preprint arXiv:2312.11805*.

Grail, Q., Perez, J., and Gaussier, E. (2021). "Globalizing BERT-based Transformer Architectures for Long Document Summarization." In Merlo, P., Tiedemann, J., and Tsarfaty, R. (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1792–1810, Online. Association for Computational Linguistics.

Gross, J. J. and Levenson, W. R. (1995). "Emotion Elicitation Using Films." *Cognition and Emotion*, **9**, pp. 87–108.

Hama, K., Otsuka, A., and Ishii, R. (2024). "Emotion Recognition in Conversation with Multistep Prompting Using Large Language Model." In *Social Computing and Social Media: 16th International Conference, SCSM 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part I*, pp. 338–346.

Huang, J. T., Lam, M. H., Li, E. J., Ren, S., Wang, W., Jiao, W., Tu, Z., and Lyu, M. R. (2024). "Emotionally Numb or Empathetic? Evaluating How LLMs Feel Using EmotionBench." *arXiv preprint arXiv:2308.03656*.

菅野祐希，坂野遼平 (2024)．オンライン動画サービスにおける BERT 及び GPT-3.5 を用いた視聴者感情の推定．言語処理学会第 30 回年次大会論文集, pp. 1067–1072．[Y. Kanno and R. Banno (2024). Onrain Doga Sabisu ni okeru BERT oyobi GPT-3.5 wo Mochiita Sichosha Kanjo no Suitei. Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing, pp. 1067–1072.].

菅野祐希，坂野遼平 (2023)．YouTube 動画コメントを用いた動画の感情推定—ルールベース及び BERT の精度比較—．In *DEIM forum 2023*, 4a-6-1．[Y. Kanno and R. Banno (2023) YouTube Douga Komento wo Mochiita Doga no Kanjo Suitei—Rurubesu oyobi BERT no Seido Hikaku—. DEIM forum 2023, 4a-6-1.].

Lee, S. Y. M., Chen, Y., and Huang, C.-R. (2010). "A Text-driven Rule-based System for Emotion Cause Detection." In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 45–53.

Llama team (2024). "The Llama 3 Herd of Models." *arXiv preprint arXiv:2407.21783*.

Malte, A. and Ratadiya, P. (2019). "Multilingual Cyber Abuse Detection using Advanced Transformer Architecture." In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pp. 784–789. IEEE.

松林圭，五味京祐，古川和祈，松尾祐佳，松原良和，日諸マルセロ優次，中村拓哉，山下晃弘，松林勝志 (2016)．Twitter 上に投稿された文章に基づく感情推定法とその応用に関する検討．第 78 回全国大会講演論文集, 2016, pp. 79–80. 情報処理学会．[K. Matsubayashi et al. (2016). Twitter jo ni Toko sareta Bunsho ni Motozuku Kanjo Suiteiho to Sono Oyo ni Kansuru Kento, Proceedings of the 78th National Convention of IPSJ, 2016, pp. 79–80.].

森朋之 (2015). back number「手紙」インタビュー—音楽ナタリー 特集・インタビュー. `https://natalie.mu/music/pp/backnumber05`．（2024 年 11 月 20 日閲覧）．[T. Mori (2015). buck number "Tegami" Intabyu—Ongaku Natari Tokushu・Intabyu. https://natalie.mu/music/pp/backnumber05.].

大友章司，竹島久美子，広瀬幸雄 (2010). 感情状態が商品広告の情報処理方略に及ぼす影響について．人間環境学研究, **8** (2), pp. 123–132. [O. Syouji et al. (2010). The Effects of Mood State on the Information Processing of Advertisement—The multiple functions of information—. Journal of Human Environmental Studies, 8(2), pp. 123–132.].

OpenAI (2024). "GPT-4 Technical Report." *arXiv preprint arXiv:2303.08774*.

OXFORD ECONOMICS (2022)．YouTube Impact Report 2022 年日本における YouTube の経済的，文化的，社会的影響．テクニカル・レポート，YouTube. `https://kstatic.googleusercontent.com/files/cd1dc8580f0aeca037253f88a5b68f179821b077568823f113f0f94a63afd89ceea0ea3fc92053cd386736f5758ab151142c80dd04a6f8cdb0d238dc46754964`（2024 年 11 月 20 日閲覧）．[(2022) YouTube Impact Report 2022nen Nihon ni Okeru

YouTube no Keizaiteki, Bunkateki, Shakaiteki Eikyo. Tekunikaru Ripoto, YouTube. https://kstatic.googleusercontent.com/files/cd1dc8580f0aeca037253f88a5b68f179821b077568823f113f0f94a63afd89ceea0ea3fc92053cd386736f5758ab151142c80dd04a6f8cdb0d238dc46754964].

Qasim, R., Bangyal, W. H., Alqarni, M. A., and Ali Almazroi, A. (2022). "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification." *Journal of Healthcare Engineering*, **2022** (1), p. 3498123.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C. E., and Bavel, J. J. V. (2024). "GPT is an Effective Tool for Multilingual Psychological Text Analysis." *Proceedings of the National Academy of Sciences*, **121** (34), p. e2308950121.

堺雄之介，竹内幹太，伊東栄典 (2021). コメントを利用した炎上動画検出に関する検討. 情報処理学系研究報告, **Vol.2021-ICS-203** (9), pp. 1–5. [Y. Sakai et al. (2021) A Study of Flaming Comment Detection using Text based Machine Learning. IPSJ SIG Technical Report, 2021-ICS-203 (9), pp. 1–5.].

SOCIOCOM Social Computing Laboratory. 日本語感情表現辞書 JIWC-Dictionary. `https://sociocom.naist.jp/jiwc-dictionary/`.（2024 年 11 月 20 日閲覧）. [Nihongo Kanjo Hyogen Jisho JIWC-Dictionary. https://sociocom.naist.jp/jiwc-dictionary/.].

Soleymani, M., Pantic, M., and Pun, T. (2012). "Multimodal Emotion Recognition in Response to Videos." *IEEE Transactions on Affective Computing*, **3** (2), pp. 211–223.

総務省 (2024). 令和 5 年度情報通信メディアの利用時間と情報行動に関する調査報告書. `https://www.soumu.go.jp/menu_news/s-news/01iicp01_02000122.html`.（2024 年 11 月 18 日閲覧）. [Soumusyou (2024). Reiwa 5 Nendo Joho Tushin Media no Riyo Jikan to Joho Kodo ni Kansuru Chousa Hokokusho https://www.soumu.go.jp/menu_news/s-news/01iicp01_02000122.html.].

Suhaimi, N. S., Mountstephens, J., and Teo, J. (2020). "EEG-Based Emotion Recognition: A State-of-the-Art Review of Current Trends and Opportunities." *Computational Intelligence and Neuroscience*, **2020** (1), p. 8875426.

田中邦朋，笹野遼平，武田浩一 (2024). 大規模言語モデルに含まれる社会集団間の感情の抽出. 言語処理学会第 30 回年次大会発表論文集, pp. 1715–1720. [K. Tanaka et al. (2024). Daikibo Gengo Moderu ni Fukumareru Shakai Shudankan no Kanjo no Chushutsu. Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing, pp. 1715–1720.].

Tiernan, R. "YouTube's 2 Billion Videos, 197M Hours Make it an 'Immense' Force, Says Bernstein." `https://www.barrons.com/articles/youtubes-2-billion-videos-197m-hours-make-it-an-immense-force-says-bernstein-1462978280`. BARRON's.（2024 年 11 月 20 日閲覧）.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and

Polosukhin, I. (2023). “Attention Is All You Need.” *arXiv preprint arXiv:1706.03762*.

薛沛欽，小林亜樹 (2024). 大規模言語モデルによる細かな感情推定手法. DEIM2024 第 16 回データ工学と情報マネジメントに関するフォーラム. [P. Xue and A. Kobayashi (2024). Daikibo Gengo Model ni yoru Komakana Kanjou Suitei Shuhou, DEIM 2024 16th Forum on Data Engineering and Information Management].

Yasmina, D., Hajar, M., and Hassan, A. M. (2016). “Using YouTube Comments for Text-based Emotion Recognition.” *Procedia Computer Science*, **83**, pp. 292–299.

Yi, J., Nasukawa, T., Bunescu, R., and Niblack, W. (2003). “Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques.” In *3rd IEEE International Conference on Data Mining*, pp. 427–434.

YouTube (2023). YouTube Japan Blog: YouTube の経済的影響と収益化の新しい方法. `https://youtube-jp.googleblog.com/2023/10/impact-report-2022-low-tier-fan-funding.html`. （2024 年 11 月 20 日閲覧）. [YouTube Japan Blog: YouTube no Keizaiteki Eikyou to Shuekika no Atarashii Hoho. https://youtube-jp.googleblog.com/2023/10/impact-report-2022-low-tier-fan-funding.html.].

Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., Zhou, W., Li, H., and Liu, T.-Y. (2020). “Incorporating BERT into Neural Machine Translation.” *arXiv preprint arXiv:2002.06823*.

付録

A　ベクトル成分の類似度評価における詳細な出力

各モデルが出力した詳細な感情ベクトルの成分を示す．表 9 に動画 1 の成分を，表 10 に動画 2 の成分を，表 11 に動画 3 の成分を示している．

B　最大ベクトルの一致率評価の詳細な出力

各ラベルごとのコサイン類似度の値を示す．表 12 に LLM のコサイン類似度，表 13 に BERT のコサイン類似度を示す．

C　感情ごとの正解数

表 14 に，動画の感情ラベルごとの各モデルの正解数を，表 15 に動画 30 本に対する正解数と正解率を示す．

動画	感情						
	sad	anxiety	angry	disgust	trust	surprise	joy
BERT256-50	0.2613	0.4541	0.3655	0.3971	0.5164	0.3850	0.1392
BERT512-50	0.2550	0.4980	0.3722	0.4037	0.5029	0.3417	0.1255
BERT256-150	0.1784	0.5816	0.2734	0.5094	0.3940	0.3585	0.1094
BERT512-150	0.2247	0.5701	0.3147	0.5155	0.3856	0.3185	0.0978
GPT-3.5turbo	0.0143	0.3286	0.1429	0.1286	0.5143	0.6822	0.3536
GPT-4	0.0833	0.2900	0.0800	0.1333	0.5099	0.6499	0.4499
GPT-4-turbo	0.0693	0.3233	0.1451	0.1880	0.3101	0.7159	0.4750
GPT-4o	0.0993	0.3590	0.0802	0.1260	0.5499	0.5919	0.4315
Deepseek	0.0811	0.3411	0.1370	0.2125	0.4669	0.6458	0.4221
Gemini-1.5-flash	0.1342	0.2788	0.0774	0.1239	0.2788	0.1601	0.8829
Gemini-1.5-Pro	0.1540	0.3363	0.0446	0.1378	0.3039	0.5875	0.6361
Claude-3-Sonnet	0.0760	0.4318	0.1159	0.1839	0.3838	0.6996	0.3518
Claude-3-Opus	0.1388	0.3502	0.0852	0.2082	0.4449	0.6563	0.4228
Claude-3.5-Sonnet	0.0840	0.2834	0.0490	0.1189	0.3918	0.7066	0.4932
Llama-ELYZA	0.0184	0.1349	0.1410	0.1135	0.3250	0.8034	0.4446

表 9　ベクトル成分の類似度評価出力（動画 1）

Model	感情						
	sad	anxiety	angry	disgust	trust	surprise	joy
BERT256-50	0.6312	0.3186	0.3508	0.3593	0.2430	0.2957	0.3184
BERT512-50	0.6549	0.3463	0.3674	0.3115	0.2055	0.2762	0.3173
BERT256-150	0.6036	0.3672	0.1953	0.2078	0.2329	0.3285	0.5073
BERT512-150	0.7296	0.3765	0.2398	0.2498	0.2253	0.2449	0.3087
GPT-3.5turbo	0.3625	0.2200	0.0250	0.0200	0.6525	0.2725	0.5650
GPT-4	0.5295	0.2599	0.0219	0.0146	0.6290	0.1773	0.4736
GPT-4-turbo	0.4605	0.2638	0.0336	0.0144	0.5804	0.3070	0.5348
GPT-4o	0.4667	0.2639	0.0171	0.0122	0.6818	0.2224	0.4447
Deepseek	0.4630	0.3451	0.0400	0.0189	0.6334	0.2609	0.4419
Gemini-1.5-flash	0.4403	0.0341	0.0000	0.0000	0.3072	0.0034	0.8430
Gemini-1.5-Pro	0.5392	0.1849	0.0231	0.0154	0.5572	0.1232	0.5905
Claude-3-Sonnet	0.6077	0.3379	0.0227	0.0199	0.4998	0.1647	0.4885
Claude-3-Opus	0.5470	0.2049	0.0163	0.0140	0.6099	0.0978	0.5261
Claude-3.5-Sonnet	0.4946	0.2065	0.0072	0.0096	0.6218	0.1537	0.5498
Llama-ELYZA	0.4124	0.1749	0.0125	0.0083	0.6019	0.2541	0.6102

表 10　ベクトル成分の類似度評価出力（動画 2）

Model	感情						
	sad	anxiety	angry	disgust	trust	surprise	joy
BERT256-50	0.2308	0.2597	0.4037	0.3822	0.2633	0.3532	0.6133
BERT512-50	0.1820	0.2317	0.5353	0.4141	0.2030	0.3234	0.5562
BERT256-150	0.1673	0.1878	0.4332	0.3096	0.2819	0.3513	0.6711
BERT512-150	0.1656	0.1691	0.4792	0.3587	0.2773	0.3299	0.6325
GPT-3.5turbo	0.0034	0.0570	0.0637	0.0838	0.5031	0.5400	0.6641
GPT-4	0.0328	0.0558	0.0624	0.0624	0.5418	0.4728	0.6862
GPT-4-turbo	0.0199	0.0828	0.0663	0.0994	0.3379	0.5731	0.7321
GPT-4o	0.0316	0.1228	0.0491	0.0632	0.4983	0.4491	0.7263
Deepseek	0.0183	0.1149	0.0392	0.0496	0.5248	0.5300	0.6528
Gemini-1.5-flash	0.0071	0.0107	0.0498	0.0391	0.2488	0.0782	0.9632
Gemini-1.5-Pro	0.0877	0.1196	0.1157	0.1555	0.2273	0.3829	0.8615
Claude-3-Sonnet	0.0239	0.0796	0.0716	0.1432	0.4097	0.5410	0.7120
Claude-3-Opus	0.0784	0.0914	0.0620	0.1339	0.4897	0.5126	0.6791
Claude-3.5-Sonnet	0.0160	0.0319	0.0128	0.0415	0.3926	0.5202	0.7564
Llama-ELYZA	0.0059	0.0059	0.0267	0.0445	0.2760	0.5668	0.7745

表 11　ベクトル成分の類似度評価出力（動画 3）

Model	ラベル							
	sad	anxiety	angry	disgust	trust	surprise	joy	平均
GPT-3.5-t	0.7935	0.7582	0.7613	0.7448	0.8679	0.8292	0.8890	0.7843
GPT-4	0.8995	0.7196	0.7887	0.7774	0.8756	0.8252	0.9060	0.8354
GPT-4-t	0.8851	0.7944	0.8015	0.8103	0.8482	0.8178	0.8908	0.8321
GPT-4o	0.8830	0.7364	0.8301	0.7836	0.8908	0.7972	0.9047	0.8155
deepseek	0.8999	0.8173	0.8139	0.8261	0.7882	0.7967	0.8717	0.8245
Gemini-1.5-flash	0.8004	0.5858	0.6592	0.5532	0.7354	0.5899	0.7652	0.6699
Gemini-1.5-Pro	0.8947	0.7105	0.7956	0.7826	0.7985	0.7823	0.8462	0.8015
Claude-3-sonnet	0.9280	0.8450	0.8357	0.8111	0.8417	0.8029	0.8889	0.8480
Claude-3.5-Sonnet	0.8984	0.7504	0.8165	0.7881	0.8564	0.8142	0.8995	0.8258
Llama-ELYZA	0.7600	0.7407	0.7434	0.7010	0.8229	0.8118	0.8829	0.7758
感情平均	0.8643	0.7459	0.7846	0.7578	0.8326	0.7867	0.8745	

表 12　感情ラベルごとのコサイン類似度 (LLM)

Model	ラベル							
	sad	anxiety	angry	disgust	trust	surprise	joy	平均
BERT50/256	0.9017	0.9190	0.9094	0.9329	0.7535	0.8154	0.8243	0.8652
BERT50/512	0.9157	0.9068	0.9161	0.9061	0.7520	0.8184	0.8180	0.8619
BERT150/256	0.9405	0.9490	0.9157	0.9430	0.7279	0.8478	0.8580	0.8831
BERT150/512	0.9262	0.9537	0.8957	0.9040	0.7473	0.8576	0.8657	0.8786
感情平均	0.9210	0.9321	0.9092	0.9215	0.7452	0.8348	0.8415	

表 13　感情ラベルごとのコサイン類似度 (BERT)

Model	正解数						
	動画ラベル						
	sad	anxiety	angry	disgust	trust	surprise	joy
GPT-3.5-t	0	1	4	0	3	3	3
GPT-4	1	1	4	1	4	3	4
GPT-4-t	2	1	3	0	3	3	5
GPT-4o	1	2	2	1	4	2	5
deepseek	2	2	4	2	3	3	5
BERT50/256	3	1	4	2	2	2	3
BERT50/512	3	3	4	2	3	1	3
BERT150/256	3	4	4	3	2	3	3
BERT150/512	2	4	4	3	2	3	3
Gemini-1.5-flash	2	2	2	1	3	2	5
Gemini-1.5-Pro	2	2	4	3	3	2	5
Claude-3-sonnet	2	3	3	2	3	3	4
Claude-3.5-Sonnet	1	2	2	2	3	3	3
Llama-ELYZA	1	1	4	0	3	3	5
平均	1.7857	2.0714	3.4286	1.5714	2.9286	2.5714	4

表 14 感情ごとの正解数

Model	正解数	正解率
GPT-3.5-t	14	47%
GPT-4	18	60%
GPT-4-t	17	57%
GPT-4o	17	57%
deepseek	21	70%
BERT50/256	17	57%
BERT50/512	19	63%
BERT150/256	22	73%
BERT150/512	21	70%
Gemini-1.5-flash	17	57%
Gemini-1.5-Pro	21	70%
Claude-3-sonnet	20	67%
Claude-3.5-Sonnet	16	53%
Llama-ELYZA	17	57%
平均	18.3571	61%

表 15 合計正解数と正解率

略歴

菅野　祐希：2023 年工学院大学情報学部情報通信工学科卒業．同年同大学大学院工学研究科情報学専攻に進学，現在に至る．自然言語処理の研究に従事．

坂野　遼平：2010 年北海道大学工学部情報エレクトロニクス学科卒業．2012 年同大学大学院情報科学研究科修士課程修了．2018 年東京工業大学大学院情報理工学研究科博士課程修了．2012 年日本電信電話株式会社入社．2018 年東京工業大学情報理工学院研究員．2020 年工学院大学情報学部情報通信工学科助教，2022 年より准教授．2024 年一橋大学大学院ソーシャル・データサイエンス研究科准教授，現在に至る．IoT システム，オーバレイネットワーク等の研究に従事．博士（理学）（2018 年 9 月，東京工業大学）．ACM, IEEE, IEICE, IPSJ 各会員．

（2024 年 11 月 25 日　受付）
（2025 年 2 月 28 日　再受付）
（2025 年 4 月 3 日　採録）