動画に対する否定的コメント・フィルタリングにおける few-shot 学習の評価

Evaluation of Few-Shot Learning in Filtering Negative Comments posted to Videos

三田寺　聖†　　宇津呂武仁††

† 筑波大学 理工学群 工学システム学類　〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学 システム情報系 知能機能工学域　〒 305-8573 茨城県つくば市天王台 1-1-1
理化学研究所 革新知能統合研究センター　〒 103-0027 東京都中央区日本橋 1-4-1

あらまし　本論文では，動画に対する否定的コメントに対して，大規模言語モデルを用いたフィルタリングを行う．大規模言語モデルを用いたコメントフィルタリングにおいては，現在最も性能が優れていると考えられる ChatGPT を用いても，本来肯定的であるはずのコメントを否定的であると判定する過剰フィルタリングが多数発生してしまう．この過剰フィルタリングは，あらかじめ少数の分類例を提示する few-shot 学習によって減少させることが可能である．本論文では，過剰フィルタリングを減少させることができる効果的な few-shot 学習の手法について検討し，評価を行った．その結果，ChatGPT を用いて内容に即したコメント・クラスターを作成し，これをもとに few-shot 学習の例示コメントを選定することによって，効果的に過剰フィルタリングを減少させることが可能であることが分かった．またこのとき，各クラスターの概要やコメント数を参考にして例示コメントを選定する最適手法を示す．

キーワード　LLM，テキスト分類，極性分類，動画コメント

1　はじめに

YouTube 等の動画サイトには，日々多種多様な動画がアップロードされており，これらの動画には視聴者から多くのコメントが寄せられている (例: 図 1)．その一部には動画の内容や出演者に対する否定的なコメントも散見され，投稿者や視聴者にとって不都合となることも多くある．本論文は，このような動画に対する否定的コメントに対して，大規模言語モデルを用いたフィルタリングを行うことを目的とする．

本論文では，否定的コメントをフィルタリングするモデルとして ChatGPT[1] を用いる．ChatGPT は，現在最も性能が優れていると考えられる大規模言語モデルである．しかし，ChatGPT を用いて否定的コメントをフィルタリングしても，動画内の文脈を読み取ることができずに，人間の感覚による分類と異なる誤フィルタリングが発生してしまう．この誤フィルタリングは，ChatGPT に対して few-shot 学習を行うことによって，改善することが可能である．

動画に対する否定的コメントのフィルタリングタスクにおいて，誤フィルタリングのうち大部分を占めるのが，本来肯定的であるはずのコメントを否定的であると判定する過剰フィルタリングである．そこで本論文では，一度 ChatGPT を用いて zero-shot 学習によるフィルタリングを行い，否定的であると予測されたコメントのみを対象とし，few-shot 学習によって再予測を行う．

few-shot 学習の際に用いる例示コメントは再予測対象コメントの中から数例を選定し，人手で分類を付与する．この際選定する例示用コメントは，類似するコメントが多数存在し，かつ肯定的である可能性の高いコメントである方がよい．そこで，ChatGPT でコメントのクラスタリングを行い，コメント数の多いクラスター内から最も肯定的であるコメントを選定することで，予測結果の改善により効果的な例示を実現する．

2　関連研究

感情分析タスクの代表的な手法として，評価表現辞書を用いた手法が考えられる．これは，個々の単語や表現に対して，あらかじめ肯定的・否定的の度合いを表す評価値を付与しておき，この値をもとにして文章全体の感情分析を行うという手法である．評価値を決定する辞書の作成手法については様々な研究が行われているが，その代表例として，共起情報を利用した手法に関する研究 [17] が挙げられる．この研究では，代表的な肯定単語として “excellent”，代表的な否定単語として “poor” を設定し，それぞれの単語との共起しやすさをもとに，各単語の評価値を決定している．しかしながらこの手法は，一文中に肯定的・否定的な要素が混在する場合や，未知語に弱いといった欠点が存在する．

一方で近年では，感情分析タスクには ChatGPT をはじめとする様々な大規模言語モデルが用いられている．感情分析に大規模言語モデルを用いた本論文の関連研究として，様々な感情分析タスクにおける複数の大規模言語モデルの性能を評価

1：https://chat.openai.com

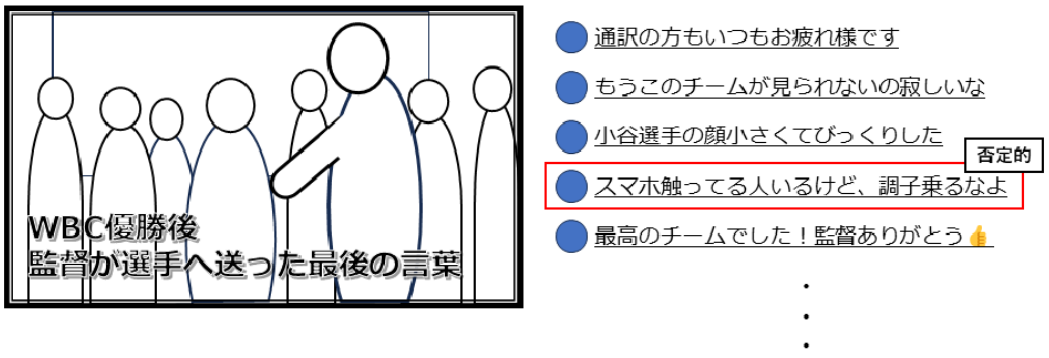

図 1　動画に寄せられるコメントの例

$$
\begin{aligned}
\text{適合率 (肯定的)} &= \frac{\left|「\text{予測} = \text{肯定的}」の集合 \cap 「\text{参照} = \text{肯定的}」の集合\right|}{\left|「\text{予測} = \text{肯定的}」の集合\right|} \\
\text{再現率 (肯定的)} &= \frac{\left|「\text{予測} = \text{肯定的}」の集合 \cap 「\text{参照} = \text{肯定的}」の集合\right|}{\left|「\text{参照} = \text{肯定的}」の集合\right|} \\
\text{適合率 (否定的)} &= \frac{\left|「\text{予測} = \text{否定的}」の集合 \cap 「\text{参照} = \text{否定的}」の集合\right|}{\left|「\text{予測} = \text{否定的}」の集合\right|} \\
\text{再現率 (否定的)} &= \frac{\left|「\text{予測} = \text{否定的}」の集合 \cap 「\text{参照} = \text{否定的}」の集合\right|}{\left|「\text{参照} = \text{否定的}」の集合\right|} \\
\text{分類精度} &= \frac{\left|「\text{予測} = \text{肯定的}」の集合 \cap 「\text{参照} = \text{肯定的}」の集合\right| + \left|「\text{予測} = \text{否定的}」の集合 \cap 「\text{参照} = \text{否定的}」の集合\right|}{\left|「\text{参照} = \text{肯定的}」の集合\right| + \left|「\text{参照} = \text{否定的}」の集合\right|}
\end{aligned}
$$

図 2　評価尺度として用いる適合率・再現率・分類精度の定義

した研究 [19] が存在する．この研究では `gpt-3.5-turbo` [8]，`text-davinci-003` [8]，`Flan-T5` [2]，`Flan-UL2` [15] の 4 つの大規模言語モデルが比較されており，いずれのタスクでも `gpt-3.5-turbo`，`Flan-UL2` の 2 つのモデルで特に高い精度が見られた．また，Instruction Tuning を行い一般目的の大規模言語モデルを金融感情分析に適応させ，金融感情分析の精度を向上した研究 [18] も関連研究として挙げられる．

プロンプトを活用して感情分析の精度向上を行った関連研究としては，プロンプトに背景情報を追加した際のヘイトスピーチ検出の精度を，複数の大規模言語モデルで評価した研究 [9] が挙げられる．この研究では比較対象となる大規模言語モデルとして `gpt-3.5-turbo`，`text-davinci-003`，`Flan-T5` の 3 つが用いられ，プロンプトやデータセットによってその精度に違いが見られた．また，段階的なプロンプトによって，言外の感情の分析精度を向上した研究 [3] や，感情の推論と推論結果の信頼性評価を 2 種類のモデルで二重に行い，感情分析の精度を向上した研究 [14] なども存在している．一方，本論文では，few-shot 学習に用いる例示コメントの選定に焦点を当て，例示コメントの選定にも大規模言語モデルを活用している点が，上記の研究と異なる点である．

few-shot 学習に用いる例示テキストの選定に関する研究としては，データセットから適切な例示コメントの選定を行い，大規模言語モデルを用いてプロンプトの生成を行った研究 [13] が存在する．この研究ではまず，kNN 検索や，fine-tuning によってテキスト分類タスクに特化させたモデルを用いて，入力テキストに対して最適となる例示テキストをデータセットから選定する．次に，ここで選定したテキストをもとに，大規模言語モデルを用いて手がかりとなるキーワードや推論の過程を出力し，few-shot 学習に用いるプロンプトを作成する．これを用いて感情分析やカテゴリ分類を行うことにより，大幅な精度向上を達成している．この研究では，あらかじめ分類ラベルが付与されたデータセットの中から，適切な例示コメントの選定を行っているのに対して，本研究では分類対象とするコメント群の中から例示コメントを選定し，その場で分類ラベルの付与を行っている，という点が相違点となっている．

近年では，動画コメントを活用し様々なタスクを解く研究が，複数行われている．動画コメントを感情分析に用いた関連研究として，動画コメントを数値ベクトルに変換し，機械学習によって炎上動画の検出を行った研究 [10] が存在する．この研究では，収集した動画コメントデータを用いてモデルの学習を一から行い，分類器を作成することで，炎上動画の分類を試みている．また，ルールベースや BERT，`gpt-3.5-turbo` など様々な手法で動画コメントから感情推定を行い，各手法の精度を比較した研究 [4] [5] も存在する．この研究では感情を 7 つに分類して感情値を 7 次元ベクトルとして表し，各手法によって推定された動画ごとの感情値を，人手で付与した感情値と比較することによって，各手法の精度を測っている．この他に，楽曲動画に寄せられたコメントから歌声の印象を分類した研究 [1] や，動画コメントを含む動画のメタデータから動画の再生回数予測を行った研究 [6]，動画コメントをもとに動画の評価項目ごとのスコアを算出することで，タイトルや説明文に表れない動画の特徴をユーザーに提示し，検索を可能にした研究 [16] [11] [12]

なども存在し，動画コメントは動画内容の分析を行うタスクにおいて幅広く活用されている．

3 動画およびコメントのデータセット

本論文では，YouTube が提供する Data API[2]を用い，YouTube 上の複数の動画から動画データ，およびコメントデータの収集を行った．対象とする動画は，コメント内容の多様性を確保し，なおかつ分析コストが大きくなりすぎないように，収集時点でのコメント数が 300 件以上，1 万件以下である動画に限定した．また，動画内容の多様性を確保するため，動画投稿者によって設定された動画カテゴリが偏ることのないように，動画の収集を行った．結果として 248 動画からコメントを収集した．

次に，動画コメントを 1 つずつ目視し，コメントの内容が肯定的・否定的のいずれであるかの分類を，各コメントに対して人手で付与した．ここで付与した人手分類を参照データとすることで，評価に用いるコメントデータセットを作成した．

本論文ではこのデータセットの中から，動画カテゴリ，全コメントにおける否定的コメントの割合等を考慮し，表 1 に示す 9 動画に対して評価を行った．

4 ChatGPT を用いたコメントのフィルタリング

本節では，ChatGPT を用いてコメントをフィルタリングする手法，およびその性能について述べる．なお，本研究では GPT-4 シリーズ [7] のモデルである `gpt-4-1106-preview` を利用する． 4.1 節では，例示を行わずに zero-shot 学習でフィルタリングを行う手順とその性能について述べる． 4.2 節では， 4.1 節で行った zero-shot 学習によるフィルタリングの結果をもとに，例示を用いた few-shot 学習によってフィルタリング精度を向上させる方法について述べる．

4.1 zero-shot 学習によるフィルタリング

zero-shot 学習によるフィルタリングでは，3 節で収集したコメントを用い，動画ごとにフィルタリングを行う．ChatGPT にコメントを 1 つずつ入力して感情分析を行い，肯定的・否定的のいずれかに分類した結果を出力させる．あらかじめ人手で付与した参照分類と出力結果が一致した場合正答，そうでない場合誤答とする．入出力の流れ，および予測結果の具体例を図 3 に示す．図 3 において評価尺度として用いた適合率，再現率，分類精度の定義を図 2 に示す．

zero-shot 学習によるフィルタリングによる予測結果では，肯定的であると予測されたコメントの適合率が高い一方で，否定的であると予測されたコメントの適合率が低くなる傾向にある．これは，ChatGPT を用いた zero-shot 学習によるフィルタリングでは，本来肯定的であるコメントを過剰にフィルタリングしてしまい，否定的であると予測されたコメントに誤分類が多いことが原因である．そこで次に，対象を zero-shot 学習によるフィルタリング時に否定的であると予測されたコメントに限定して再度フィルタリングを行うことによって，フィルタリング精度の改善を目指す．

4.2 否定判定コメントに対する few-shot 学習での再予測

few-shot 学習によるフィルタリングでは， 4.1 節で行った zero-shot 学習によるフィルタリングにおいて否定的であると予測されたコメントのみを対象とし，フィルタリングを行う．手順としてはまず，フィルタリング対象のコメントの中から数例のコメントを選定する．本研究では 10 コメントを例示コメントとして選定する．次にこれらのコメントを目視で判定し，肯定的・否定的のいずれかのラベルを適切に付与する．これを例示として用いて few-shot 学習を行い，再度分類した結果を出力させる．入出力の流れ，および再予測結果の具体例を図 4 に示す．

few-shot 学習によるフィルタリングによる再予測の結果，過剰にフィルタリングされるコメントが減少し，否定的コメントの再現率を大きく下げることなく，肯定的コメントの再現率を大きく改善することができる．しかし，これは例示コメントとして適切なコメントを人手で適切に選定したときの結果であり，実際には多数のコメントの中から自動で適切な例示コメントを選定することが必要である．ここで，全 n 個のコメントから 10 個の例示コメントを選定するパターン数は ${}_nC_{10}$ となる．本論文で評価を行った 9 動画では，表 1 に示すように，例示コメントを選定する対象となる “予測-否定的” コメントが最小でも 95 件存在する．そのため，例示コメントを選定するパターン数は最小でも ${}_{95}C_{10} = 1.01 \times 10^{13}$ となり，非常に大きくなってしまう．次節ではこれらのパターンの中から，フィルタリングをできるだけ多く改善させるために適切な例示コメントの選定手法について，検討を行う．

5 ChatGPT による例示コメントの選定

few-shot 学習によるフィルタリングで否定判定コメントの再予測を行う際，適切な例示コメントの条件として，「類似するコメントが多数存在する」ことが挙げられる．本節では，この条件について述べるとともに，条件を満たすような例示用コメントを，ChatGPT を用いて適切に選定するための手法について提案し，評価する． 5.1 節では，類似するコメントが多数存在するコメントを選定する手法として，ChatGPT を用いた類似コメントのクラスタリングについて述べる． 5.2 節では， 5.1 節で作成したコメント・クラスターをもとに，各クラスターから最適な例示コメントを選定する手法について述べる． 5.3 節では，提案手法について，実際に評価を行った結果を記述する．

5.1 コメントのクラスタリング

few-shot 学習による再予測では，例示コメントの内容を参考に再予測が行われるため，例示コメントと類似したコメントのフィルタリング結果が多く改善する．そのため，例示コメントでより多くの予測結果を改善するためには，類似するコメントが多数存在するコメントを選定することが必要である．

2：https://developers.google.com/youtube/v3

表 1　評価を行った動画

動画番号	動画タイトル	動画カテゴリ	コメント数				
			全体	人手分類結果 (参照データ)		zero-shot 学習結果 (予測データ)	
				肯定的	否定的	肯定的	否定的
動画 1	【応援ありがとう】栗山監督が最後に 選手たちに贈ったメッセージ	Sports	406	370	36	311	95
動画 2	映画『アバター：ウェイ・オブ・ウォーター』 本予告編【異次元の”没入型”映像体験】 12 月 16 日（金）劇場公開	Film & Animation	372	323	49	233	139
動画 3	【４人】間違いなく仲間に疑われる おもしろボードゲーム【お邪魔者】	Gaming	359	348	11	241	118
動画 4	【衝撃】九州→四国を最短で結ぶ ””隠されたルート””で移動してみた！	Travel & Events	379	342	37	223	156
動画 5	新型クラウンは今までと別モノ?! 土屋圭市が 試乗して驚愕！ 沢すみれも感激！ 工藤貴宏が徹底解説	Autos & Vehicles	313	213	100	130	183
動画 6	「コミケのために日本に来た」世界が注目の “コミケの世界”　ツボにはまる!? コミケの意外な楽しみ方 【N スタ解説】｜ TBS NEWS DIG	News & Politics	463	361	102	197	266
動画 7	【神速】誰も追いつけない”イカ速 3.9 スパッタリーヒュー”がチート すぎるんだがwwww【スプラトゥーン 3】	Gaming	409	408	1	304	105
動画 8	「Re:Unknown X」を弾いてみた 【東方ダンマクカグラ】	Music	423	419	4	292	131
動画 9	ペンギンの赤ちゃんの反抗期	Pets & Animals	361	340	21	265	96
平均	—	—	387.2	347.1	40.1	244.0	143.2

類似コメントが多数存在するコメントを選定するための手法として，ChatGPT を用いたコメントのクラスタリングを行う．これは，同じクラスターに分類されるコメントは類似性が高く，各クラスターに分類されたコメントの数が，類似コメントの多さを表す指標となるためである．例示コメントの選定対象となるコメントを ChatGPT に一度に全て入力し，クラスタリングを行うと，クラスターが細分化されてしまい，類似コメント数を見るのに適切なクラスタリングが行われない．そのため，以下の手順でクラスタリングを行う．まず，例示コメントを選定する対象となる，zero-shot 学習によるフィルタリングにおいて否定的であると予測されたコメントを，全て ChatGPT に入力し，コメント内容をもとに分類を行った際の各分類の概要を出力させる．次に，出力された概要を各クラスターの項目名として，それぞれのコメントがどのクラスターに分類されるのか出力を行う．これにより作成されたクラスターをもとに，コメント数の多いクラスターから例示コメントを選定することによって，類似コメントが多数存在するコメントの選定が可能となる．コメント・クラスタリングを行う際の入出力の具体例を，図 5 に示す．

5.2　コメント・クラスターを利用した例示コメント選定

続いて，5.1 節で作成したコメント・クラスターをもとに，ChatGPT を用いて few-shot 学習に用いる例示コメントの選定を行う．手順としては，1 つのクラスターに分類されたコメントを全て ChatGPT に入力し，あらかじめ決められた基準にしたがって，各クラスターから例示コメントを選定し出力する．入出力の具体的な流れの例を，図 6 に示す．

ここで重要となるのが，各クラスターからどのような基準にしたがって例示コメントを選定するか，という点である．本研究では，例示コメントを選定する手法として，以下の 2 つの要素を考える．

1 つ目の要素は，1 クラスターあたりの選定コメント数のように設定するかという点である．例示コメント数を一定とした場合，1 クラスターあたりの選定コメント数を少なくすることで，多くのクラスターからコメントを例示することが可能となる．一方で，1 クラスターあたりの例示コメント数が少ないと，規模の大きいクラスターの特徴をとらえきることが難しい可能性も考えられる．本研究では，4.2 節で述べた通り，10 コメントを例示コメントとして選定する．そこで，選定元となるクラスターと例示コメントの数を以下のように設定した，2 つの手法について評価を行う．

few-shot 学習の例示コメントの選定元クラスター数，および，選定する例示コメント数

- 10 クラスターから 1 コメントずつ選定
- 5 クラスターから 2 コメントずつ選定

2 つ目の要素は，few-shot 学習に用いる例示コメントとし

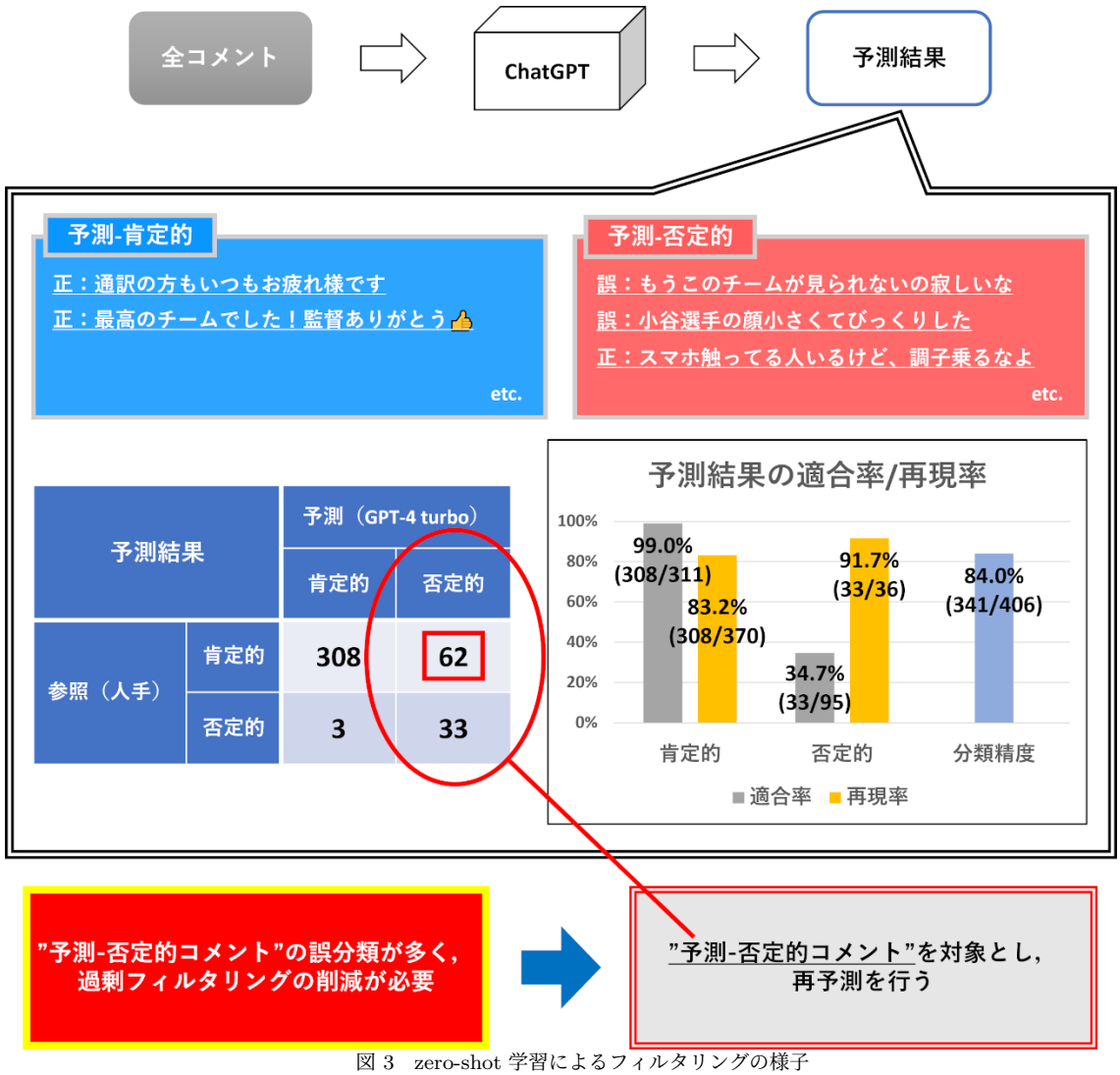

図 3　zero-shot 学習によるフィルタリングの様子

て，どのようなコメントを選定対象とするかという点である．few-shot 学習による再予測の対象となるコメントは全て，予測時点で否定的であると予測されたコメントであるため，正分類された実際に否定的であるコメントよりも，誤分類された本来肯定的であるコメントを例示として用いたほうが，より多くの予測結果を改善することができると考えられる．一方で，コメント・クラスターには，分類されるコメントの主な内容が否定的であるクラスターも存在するため，各クラスターにおける肯定的なコメントがそのクラスターの特徴を適切に表しているとは限らない．そこで，各クラスターから選定する例示コメントとして，以下の 2 つの手法を考え，評価を行う．

選定対象とする few-shot 学習の例示コメント

- 肯定的コメントを選定
- 代表コメントを選定

5.3　評 価 結 果

5.2 節で示した 2 つの要素を組み合わせ，計 4 手法に対して評価を行う．評価結果を表 2，および，表 3 に示す．ここでは，zero-shot 学習時に否定的であると予測されたコメントを，「“予測-否定的” コメント」と定義する．これは，few-shot 学習による再予測の対象となるコメントである．また，ero-shot 学習時

表 2　各コメント・フィルタリング手法における分類精度 (%)(マクロ平均)

(a) zero-shot 学習，および無作為に選定した例示コメントを用いた few-shot 学習

手法	全コメント内	“予測-否定的” コメント内
zero-shot 学習	72.7	23.9
few-shot 学習 (無作為 5 回の平均値)	83.6	57.4

(b) コメント・クラスターをもとに選定した例示コメントを用いた few-shot 学習

全コメント内 “予測-否定的” コメント内	肯定的コメントを選定	代表コメントを選定
10 クラスターから 1 コメントずつ選定	82.1 52.6	83.4 56.6
5 クラスターから 2 コメントずつ選定	**83.8** 58.1	83.3 **58.6**

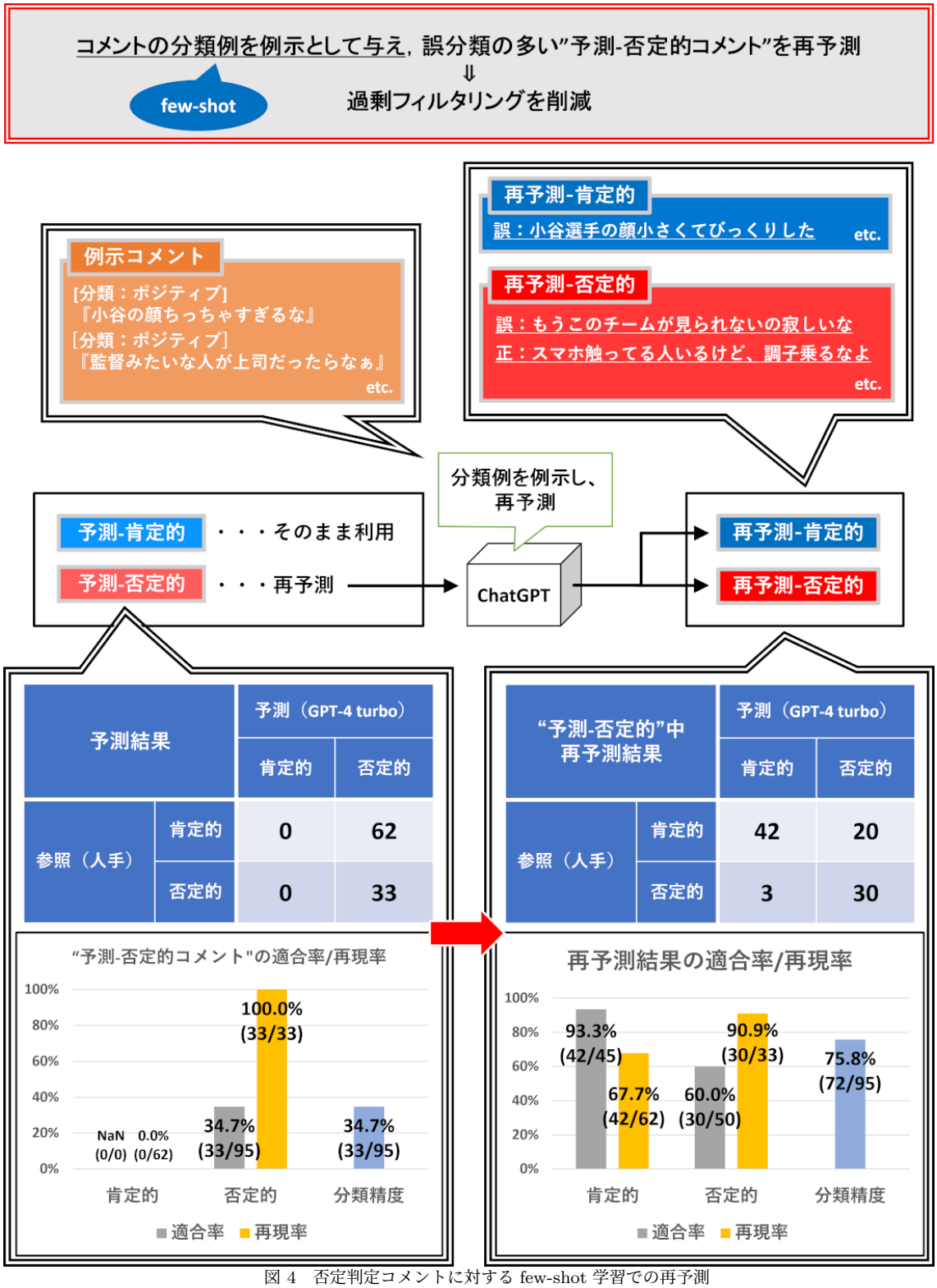

図 4　否定判定コメントに対する few-shot 学習での再予測

に肯定的であると予測されたコメントを「“予測-否定的” コメント」に加えた，1 動画における全てのコメントを，「全コメント」と定義する．本論文では，これら 2 種類のコメント集合内での分類精度を評価の基準とする．評価のベースラインとする，zero-shot 学習，および無作為に選んだ 10 コメントを例示した few-shot 学習の評価結果を表 2(a) に示す．また，4 手法それぞれにおいて 2 種類のコメント集合内での分類精度を算出した結果を表 2(b) に示す．さらに，評価を行った 9 動画それぞれの評価結果を表 3 に示す．なお，肯定的コメントを選定する手法と代表コメントを選定する手法では，選定した 10 コメントのうち，最大で 5 個，平均して 2 個程度のコメントが同一のものとなった．

表 2 より，few-shot 学習によってフィルタリングを行ったときの分類精度は，いずれも zero-shot 学習時の分類精度を大き

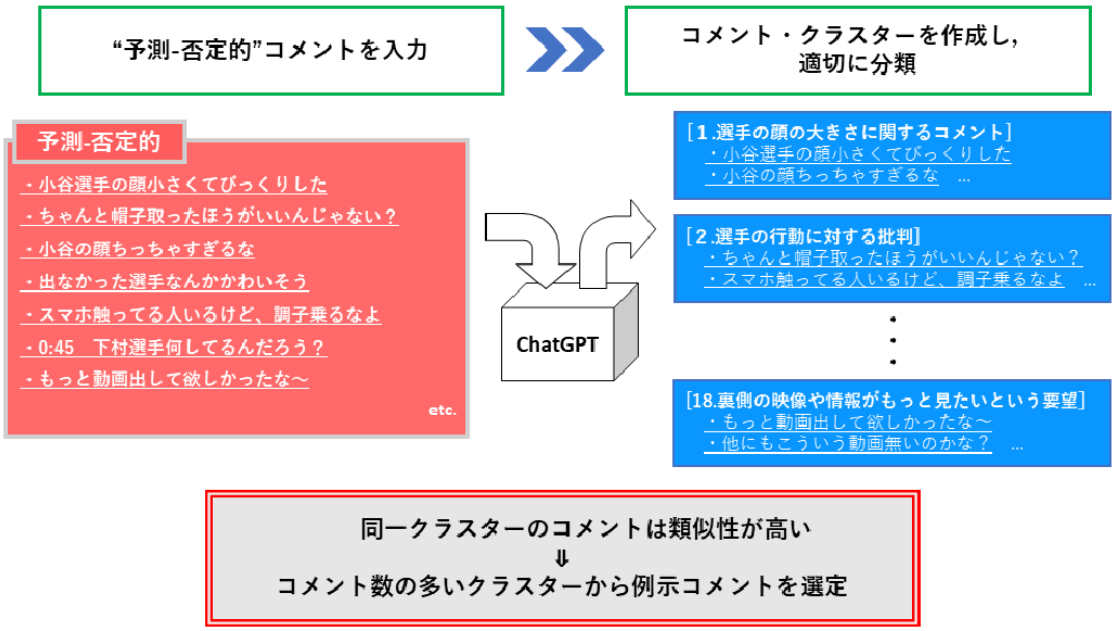

図 5　コメントのクラスタリング

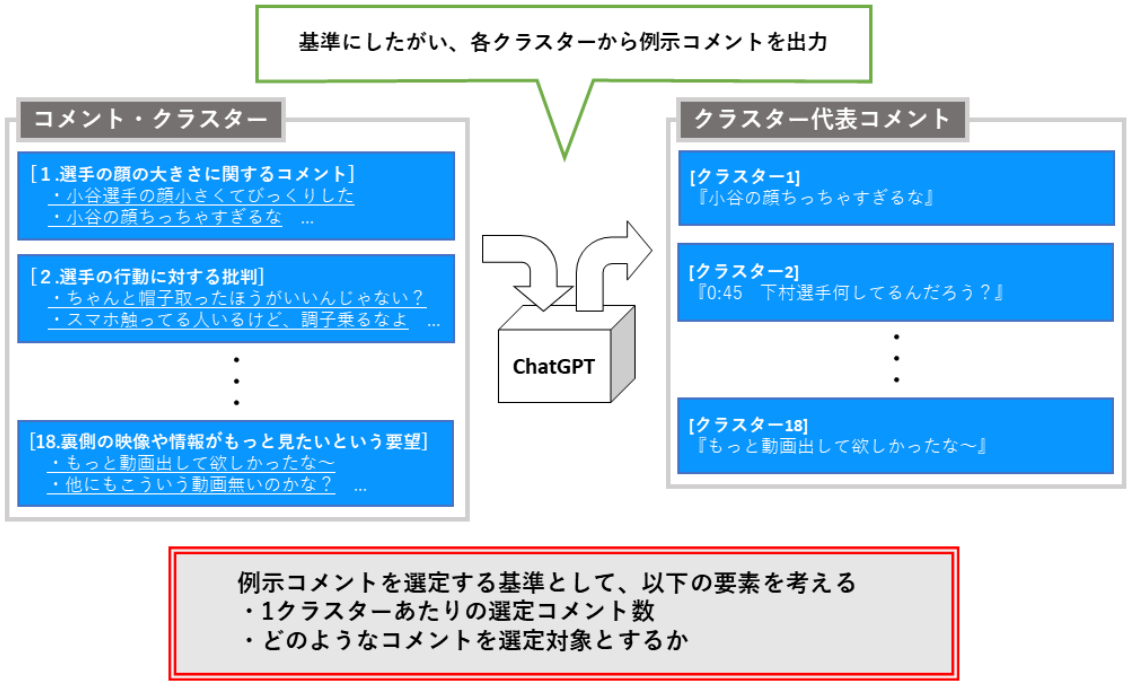

図 6　コメント・クラスターを利用した例示コメント選定

く上回っており，本論文における few-shot 学習を用いた手法が，コメント・フィルタリングにおいて有効であることが確認された．

10 クラスターから 1 コメントずつ選定を行う手法では，few-shot 学習の例示コメントとして肯定的コメント，代表的コメントのいずれを選定した場合でも，正解率のマクロ平均値が無作為に選定した 10 コメントを例示した際を下回った．特に，肯定的コメントを例示した場合では，無作為に選定した 10 コメントを例示した際の分類精度を上回ったものは 9 動画中 1 動画のみであった．一方で，5 クラスターから 2 コメントずつ選定を行う手法では，肯定的コメントを選定した場合，全コメント内，“予測-否定的” コメント内のいずれの分類精度の平均値も，無作為に選定したときの分類精度の平均値を上回った．特に，全コメント内での分類精度は 83.8%となり，コメント・クラスターを用いた例示コメント選定手法 4 つの中で，最も高い結果となった．また，5 クラスターから 2 コメントずつ，代表コメントを選定する手法においては，全コメント内の分類精度は無作為選定時を下回ったものの，“予測-否定的” コメント内の分類精度は 58.6%と無作為選定時を上回り，4 手法の中で最も高くなった．

表 3　各コメント・フィルタリング手法における動画ごとの分類精度 (%)

全コメント内 “予測-否定的” コメント内	zero-shot 学習	few-shot 学習			
		無作為に選定 (5 回の平均値)	コメント・クラスターをもとに選定		
				肯定的コメント を選定	代表コメント を選定
動画 1	84.0 34.7	89.1 56.4	10 クラスターから 1 コメントずつ選定	88.9 55.8	86.9 47.4
			5 クラスターから 2 コメントずつ選定	92.1 69.5	**92.4** **70.5**
動画 2	75.3 34.5	81.6 51.5	10 クラスターから 1 コメントずつ選定	80.9 49.6	80.1 47.5
			5 クラスターから 2 コメントずつ選定	**82.0** **52.5**	78.2 42.4
動画 3	70.2 9.3	88.1 63.9	10 クラスターから 1 コメントずつ選定	86.9 60.2	**93.0** **78.8**
			5 クラスターから 2 コメントずつ選定	90.0 69.5	92.5 77.1
動画 4	68.1 23.1	**80.9** **54.4**	10 クラスターから 1 コメントずつ選定	76.3 42.9	73.1 35.3
			5 クラスターから 2 コメントずつ選定	79.4 50.6	76.5 43.6
動画 5	71.2 53.0	73.1 56.2	10 クラスターから 1 コメントずつ選定	**75.1** **59.6**	72.8 55.7
			5 クラスターから 2 コメントずつ選定	**75.1** **59.6**	72.5 55.2
動画 6	62.4 36.5	74.0 56.7	10 クラスターから 1 コメントずつ選定	73.0 54.9	**77.8** **63.2**
			5 クラスターから 2 コメントずつ選定	72.8 54.5	69.3 48.5
動画 7	74.6 1.0	91.2 65.9	10 クラスターから 1 コメントずつ選定	87.8 52.4	**93.4** **74.3**
			5 クラスターから 2 コメントずつ選定	88.3 54.3	91.9 68.6
動画 8	70.0 3.1	87.1 58.3	10 クラスターから 1 コメントずつ選定	83.9 48.1	**88.2** **61.8**
			5 クラスターから 2 コメントずつ選定	87.0 58.0	87.2 58.8
動画 9	78.1 19.8	87.1 53.8	10 クラスターから 1 コメントずつ選定	86.1 50.0	85.0 45.8
			5 クラスターから 2 コメントずつ選定	87.3 54.2	**89.5** **62.5**

以上より，コメント・クラスターを用いた例示コメント選定手法においては，1 クラスター当たりの例示コメント数を複数とした方が，分類精度が高くなることが分かった．これは，1 つのクラスターに含まれるコメント全体の特徴を，1 コメントの例示のみでは表現することが難しかったためであると考えられる．また，選定対象とするコメントに関しては，肯定的コメント，代表コメントのいずれであっても，大きな変化は見られなかった．

6 おわりに

本論文では，動画に対する否定的コメントのフィルタリングを，few-shot 学習を用いて効果的に行う手法を提案した．ChatGPT を用いて一度コメント・フィルタリングを行い，否定的であると予測されたコメントをクラスタリングしたのち，クラスターごとに例示コメントを選定し再予測を行うことで，フィルタリング精度を大きく向上させることができた．

クラスターごとに例示コメントを選定する手法としては，1 つのクラスターから複数のコメントを選定することが効果的であることが分かった．また，選定するコメントは，クラスター内における肯定的なコメント，代表コメントのいずれであっても，精度向上に大きな差異は見られなかった．

今後の課題としては，さらに多くの動画を用いて few-shot 学

習に関する評価を行い，各手法における精度の向上と動画の特徴との関係について調査を行いたい．また，例示コメントのより効果的な選定手法についての検討も進めたい．

謝　　辞

本研究は科研費 21H00901 の助成を受けたものである

文　　献

[1] 陳沢璨, 熊本忠彦. J-POP における楽曲動画コメントを用いた歌声印象軸の構築. 第 16 回 DEIM フォーラム論文集, 2024.
[2] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
[3] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, and T.-S. Chua. Reasoning implicit sentiment with chain-of-thought prompting. In *Proc. 61st ACL*, pp. 1171–1182, 2023.
[4] 菅野祐希, 坂野遼平. YouTube 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. 第 15 回 DEIM フォーラム論文集, 2023.
[5] 菅野祐希, 坂野遼平. 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, 2024.
[6] 川上大凱, 鈴木優. 動画の属するコミュニティ情報を考慮した動画再生回数予測手法の提案. 第 16 回 DEIM フォーラム論文集, 2024.
[7] OpenAI. GPT-4 technical report, 2023.
[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Proc. 36th NeurIPS*, pp. 27730–27744, 2022.
[9] S. Roy, A. Harshvardhan, A. Mukherjee, and P. Saha. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Proc. Findings of EMNLP*, pp. 6116–6128, 2023.
[10] 堺雄之介, 竹内幹太, 伊東栄典. コメントを利用した炎上動画検出に関する検討. 情報処理学会研究報告, Vol. 2021–ICS–203, No. 9, pp. 1–5, 2021.
[11] 笹原彰斗, 山口史弥, 上田真由美, 中島伸介. 直観的動画検索システムにおける評価項目別スコアリングの精度向上. 第 15 回 DEIM フォーラム論文集, 2023.
[12] 笹原彰斗, 山口史弥, 上田真由美, 中島伸介. 直感的動画検索における単語分散表現を用いた評価項目別スコアリング手法. 第 16 回 DEIM フォーラム論文集, 2024.
[13] X. Sun, X. Li, J. Li, F. Wu, S. Guo, T. Zhang, and G. Wang. Text classification via large language models. In *Proc. Findings of EMNLP*, pp. 8990–9005, 2023.
[14] X. Sun, X. Li, S. Zhang, S. Wang, F. Wu, J. Li, T. Zhang, and G. Wang. Sentiment analysis through LLM negotiations. *arXiv preprint arXiv:2311.01876*, 2023.
[15] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, H. S. Zheng, D. Zhou, N. Houlsby, and D. Metzler. UL2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2023.
[16] 鴇田和士, 上田真由美, 中島伸介. 動画に対する評価項目別スコアを用いた直感的動画検索システム. 第 14 回 DEIM フォーラム論文集, 2022.
[17] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. 40th ACL*, pp. 417–424, 2002.
[18] B. Zhang, H. Yang, and X.-Y. Liu. Instruct-finGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*, 2023.
[19] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*, 2023.