RAG による芸能人の話題集約及びその経歴の良否判定

土田陸斗 [1] 横山響 [2] 宇津呂武仁 [3]
[1] 筑波大学 理工学群 工学システム学類
[2] 筑波大学大学院 システム情報工学研究群 知能機能システム学位プログラム
[3] 筑波大学 システム情報系 知能機能工学域
s2110466_@_u.tsukuba.ac.jp s2320808_@_u.tsukuba.ac.jp
utsuro_@_iit.tsukuba.ac.jp

概要

本論文の目的は，先行研究 [9] の情報探索の問題点の解決と，大規模言語モデルによる良否判定の能否を判明させることである．その手法として，まずウェブページの芸能人に関する記事から対象の芸能人に言及している文を収集する．この際，記事を全て人が確認することは困難であるため，その収集を大規模言語モデルの ChatGPT で行う．次に，集めた文章を ChatGPT によって内容でカテゴリ分けし，カテゴリ名を付ける．ここで付けた名前を芸能人の観点と呼ぶことにする．この芸能人の観点について，先行研究 [9] の手法との対応付けや，大規模言語モデルによる経歴の良否判定を行う．

1 はじめに

本論文では芸能人に関する情報を基に，芸能人の経歴の良否を大規模言語モデル (LLM) の ChatGPT に判定させることを目的としている．正確な判定のために，対象となる芸能人に関する多くの情報が必要になる．

本論文の先行研究 [9] では，X のポストから芸能人の評価対象に関する感想の収集・集約を行っている．芸能人の評価対象とは，感想の対象となっている事柄である．対して，本論文ではポストではなくウェブページから芸能人の観点を抽出し，詳細な情報を集約するという手法を新たに提案している．さらに，先行研究 [9] の評価対象・感想・理由と，本論文の観点・話題の対応や，芸能人の観点に対して ChatGPT による良否の判定を行う．

2 関連研究

芸能人の情報を扱う研究は，1 章で述べている先行研究 [9] 以外に，マイクロブログにおける芸能人と感想との関係を決定する研究 [4] や，マイクロブログから芸能人の評価対象に対する感想を抽出する研究 [8] がある．しかし，これらは過去に問題を起こした芸能人は対象としておらず，そのような芸能人を対象としたときに発生する問題点を本論文では解決している．

また，本論文では検索拡張生成 (RAG) [3] を使用している．これは外部から得られる情報を参照することで LLM のハルシネーションを抑え，出力を安定させるものである．RAG に関連する研究には，検索強化型言語モデル [6] や，検索付き言語モデルによる信頼性・適応性・帰属可能の向上 [1] などがある．また，ChatGPT に関する研究には，エンティティリンキング [5] や対話分析 [2]，抽出型要約 [10] などがある．

さらに，本論文の重要な特徴として，LLM による芸能人の経歴の良否の判定があるが，LLM と法律に関する研究には，LLM の効果を法の分野で検証した研究 [7] などがある．

3 X 上のポスト・ウェブページを情報源とする芸能人への感想とその理由の集約 [9]

本論文の先行研究として，1 章に挙げた，芸能人の感想を表すポストの理由集約における検索拡張生成の評価 [9] がある．そこでは，まず X のポストから LLM の ChatGPT を用いて芸能人の感想を含むポストを収集している．ここで収集されたポストから，芸能人の何に対しての感想であるかを芸能人の評価対象として抽出する．この時抽出された「芸能人の評価対象+感想」の理由を RAG [3] を用いて収集・集約している．図 1 にあるように「芸能人の評価対象+感想」をキーワードとしてウェブページを検索し，出てきたウェブページを ChatGPT に与

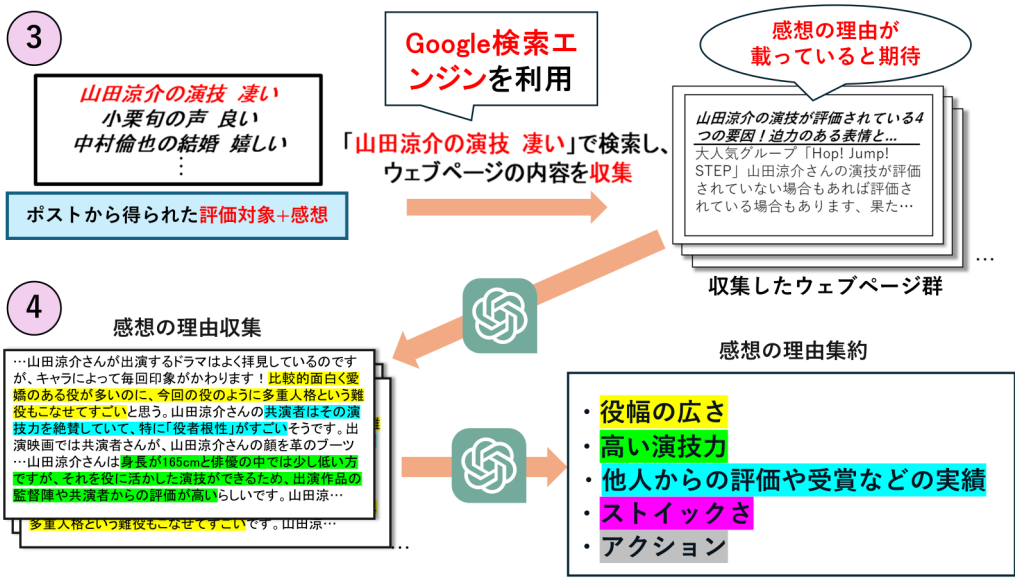

図 1　先行研究 [9] で行っている評価対象・感想・理由の収集・集約

え，その内容から感想の理由を探している．ここで，RAG [3] を使用する理由は，ポストには，感想はあってもその理由まで詳細に書かれていることが少ないからである．そこで，外部のウェブページから情報を取ってくる必要がある．

しかし，先行研究 [9] で扱っていた芸能人は，過去に問題を起こしていない人のみであった．そこで，より様々な芸能人を対象に収集・集約を行った結果，過去に問題を起こしている芸能人に関するポストには評価対象が明確ではなく，抽出が困難であった．つまり，先行研究の手法は過去に問題の無い人に対しては有効だが，過去に問題のある芸能人に対して有効ではないと考えられる．

以上の問題点を受けて，本論文では過去に悪い行いをした芸能人も対象に入れること，その際にポストに明確に描かれていない芸能人の評価対象を抽出すること，を目指す．

4　ウェブページを情報源とする芸能人の話題集約

3 章で述べた，先行研究 [9] の問題点を解決するために本論文では，ポストから評価対象を抽出せず，ウェブページから芸能人のあらゆる評価対象を取得する手法を提案する．

4.1　ウェブページの収集

初めに「芸能人の名前」で検索してウェブページを収集する．先行研究 [9] では検索のキーワードを「芸能人の評価対象+感想」としていたので，ポストの情報と関連したウェブページしか取得することが出来なかったが，このキーワードの変更により，幅広い情報を取得することが可能になる．

4.2　芸能人に対する観点の抽出および話題集約

次に，集めたウェブページ上の大量の文章から，対象の芸能人に言及している文を収集し，内容ごとにカテゴリ分けを行う．その際にカテゴリ名として，芸能人の「何の話題」であるかを ChatGPT に判断してもらい，名付けさせる．ここで生成されたカテゴリ名が「芸能人の観点」となる．最後に，芸能人の観点に関する様々な話題を集約する．

ここまでの一連の作業は全て ChatGPT で行う．使用する ChatGPT のモデルは gpt-4o であり，後に ChatGPT を使用している場面がいくつかあるが，全て同様のモデルを使用している．対象とする芸能人の例として「フワちゃん」でウェブページの収集から観点の抽出まで行った様子を図 2 に示す．

4.3　評価

4.2 節で ChatGPT が行う作業について，人手で芸能人の観点の抽出と話題の集約を行った結果と比較して評価する．ChatGPT と人手の両方で観点抽出と話題集約を行い，ChatGPT の観点・話題が共に人手と同じであるか，もしくは類似する内容の場合は一致するとした．つまり，観点が人手と同じでも話題が異なる場合は一致するとしないが，本研究ではそ

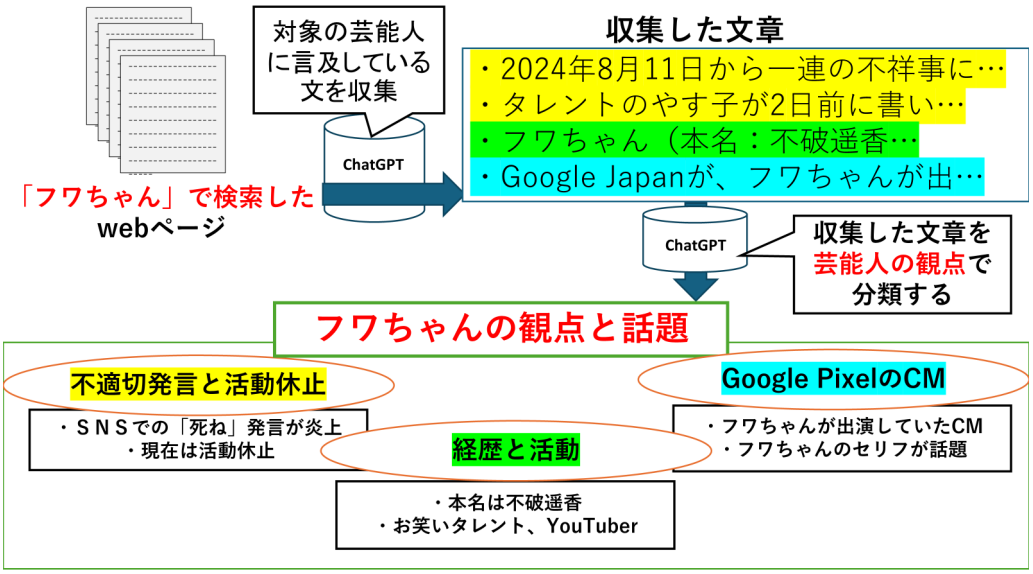

図 2　フワちゃんを例とした観点の抽出と話題の集約

のような観点と話題は見られなかった．

また，ある芸能人 c に対して人手の観点・話題の集合を R(c)，ChatGPT の観点・話題の集合を S(c) としたとき，再現率・適合率を以下のように定義する．つまり，ChatGPT の観点・話題が 5 つ，人手の観点・話題が 6 つ，一致した観点・話題が 4 つの場合，再現率は 4/6，適合率は 4/5 となる．

$$\text{再現率} = |R(c) \cap S(c)|/|R(c)|,$$

$$\text{適合率} = |R(c) \cap S(c)|/|S(c)|$$

芸能人の観点の抽出と話題集約に関する評価は，表 1 に示す．再現率はおよそ 6 割程度，適合率は 9 割近い結果になり，正確に抽出と集約が行えていることが分かる．

表 1　「抽出した観点・集約した話題」の人手評価

芸能人名	再現率	適合率
フワちゃん	0.75 (6/8)	1.00 (6/6)
ピエール瀧	0.53 (8/15)	1.00 (8/8)
中丸雄一	0.64 (9/14)	0.90 (9/10)
宮迫博之	0.60 (6/10)	1.00 (6/6)
槇原敬之	0.82 (9/11)	0.90 (9/9)
山田涼介	0.78 (7/9)	1.00 (7/7)
小栗旬	0.50 (5/10)	1.00 (5/5)
綾野剛	0.71 (10/14)	0.83 (10/12)
二宮和成	0.63 (5/8)	0.83 (5/6)
菊池風磨	0.86 (6/7)	1.00 (6/6)
マクロ平均	0.68	0.95

4.4　先行研究による感想・理由集約結果との対応付け

本論文の手法で抽出された観点や集約された話題は，先行研究 [9] で抽出される評価対象・感想・理由と重複するものや，本論文の手法でのみ抽出されるものもある．そこで，再び ChatGPT を用いて，先行研究の「芸能人の評価対象+感想+感想の理由」と，本論文の手法の「芸能人の観点+話題」を対応付けさせ，重複している観点・話題及び評価対象・感想・理由を調べた．具体例を用いた様子を付録の図 4 に示し，対応付けの全体の結果が表 2 である．加えて，各芸能人ごとの詳細な結果は表 6 に示している．表 2 を見ると，ChatGPT の行う対応付けの精度はかなり高く，正確に行えていることが分かる．また，手法 [9] では重なっている数が多いのに対し，本論文の手法では重なっている数は少ないことから，手法 [9] でのみ得られる評価対象・感想・理由は少なく，本論文の手法でのみ得られる観点・話題が多く，より幅広い情報探索を行えていることが分かる．

5　芸能人の観点・話題の良否判定

続いて，4.2 節で抽出された観点に対して良否の分類を行う．この分類の目的は，法律の知識に特化して学習していない大規模言語モデルでも，細かい悪の分類を正確に行えるか判明させることである．ChatGPT に判定させる具体的な分類項目は次の 5 つである．

表 2　ChatGPT による観点・話題と評価対象・感想・理由の対応付けの評価結果 (10 人の芸能人の平均)

—	手法 [9]	本論文の手法
二手法間で**重なった** 観点・話題 / 評価対象・感想・理由の数	5.67 人手評価の精度: 88.20 (%)	2.00 人手評価の精度: 80.56 (%)
二手法間で**重ならなかった** 観点・話題 / 評価対象・感想・理由の数	2.67 人手評価の精度: 41.11 (%)	5.83 人手評価の精度: 91.67 (%)
ChatGPT による観点・話題 / 評価対象・感想・理由の数	8.33	7.83

表 3　10 人の芸能人を対象とした良否判定の評価結果

—	—	ChatGPT での判定結果					
		善	法的に悪	倫理的に悪	評判が悪い	善悪関係なし	計
人手での判定結果	善	2	0	0	0	0	2
	法的に悪	0	2 / 槇原・瀧	0	0	0	2
	倫理的に悪	0	1 / 宮迫	2 / フワ・中丸	0	0	3
	評判が悪い	0	0	0	1 / 宮迫	0	1
	善悪関係なし	1	0	1	0	61	63
	計	3	3	3	1	61	71

- 善
- 法的に悪
- 倫理的に悪
- 評判が悪い
- 善悪と関係なし

特に，悪の 3 つの分類について，「法的に悪」は明確に法律に違反して問題となった人，「倫理的に悪」は法律には違反していないが倫理的に問題がある言動をして世間から批判された人，「評判が悪い」は特に悪いことをしていないが周りからの評判が良くない人である．これらの細かい違いを ChatGPT が判定できるかの調査を行う．

5.1　判定手順

判定する際に ChatGPT に渡す入力情報は，4.2 節で抽出した芸能人の観点と集約した話題である．フワちゃんを例とすると，抽出された観点の 1 つである「不適切発言と活動休止」について，集約した話題も合わせて ChatGPT に入力する．判定の結果，フワちゃんの「不適切発言と活動休止」は「倫理的に悪」となる．

5.2　評価

5.1 節で行われた ChatGPT の判定がどれほど正確であるかについて，人手の結果と比較を行う．各芸能人から抽出される観点と集約された話題に対して，人手で良否の判定を行った結果と ChatGPT の結果が一致しているか調べ，再現率・適合率を計算した．この時，1 つの観点に対し 1 つのラベルを付ける作業のため，再現率と適合率の分母は同数になる．各芸能人ごとの評価を表 4，10 人の結果をまとめた混合行列を表 3 に示す．表 3 では，特に悪の 3 つの項目に分類された芸能人の名前を表記している．どの芸能人も問題を起こして芸能界を休止した人であり，正確な判定が行えていると言える．

表 4　芸能人ごとの良否判定の評価結果

芸能人名	再現率・適合率
フワちゃん	1.00 (6/6)
ピエール瀧	1.00 (8/8)
中丸雄一	1.00 (9/9)
宮迫博之	0.67 (4/6)
槇原敬之	1.00 (9/9)
山田涼介	0.86 (6/7)
小栗旬	1.00 (5/5)
綾野剛	1.00 (10/10)
二宮和成	1.00 (5/5)
菊池風磨	1.00 (6/6)

6　おわりに

本論文では先行研究 [9] で過去に問題を起こした芸能人を対象としたときに発生する問題点を解決するための新たな手法を提案した．提案手法に対し，人手での結果との比較や先行研究 [9] との対応付けを行った．

さらに，芸能人の観点に対し，ChatGPT がどこまで芸能人の良否を人が行うように判定できるのかについても調査を行った．その結果，問題を起こした芸能人に対しては，悪と判定することが出来ており，その細かい分類も区別できていた．

謝辞

本論文は，一部，科研費 21H00901，電気通信普及財団 2023 年度研究調査助成，弥生株式会社共同研究の支援を受けたものである．

参考文献

[1] A. Asai, Z. Zhong, D. Chen, P. W. Koh, L. Zettlemoyer, H. Hajishirzi, and W. T. Yih. Reliable, adaptable, and attributable language models with retrieval. https://arxiv.org/abs/2403.03187, 2024.

[2] S. E. Finch, E. S. Paek, and J. D. Choi. Leveraging large language models for automated dialogue analysis. In **Proc. 24th SIGDIAL**, pp. 202–215, 2023.

[3] P. Lewis, E. Perez, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In **Proc. 34th NeurIPS**, pp. 483–498, 2020.

[4] Y. Nozaki, K. Sugawara, Y. Zenimoto, and T. Utsuro. Tweet review mining focusing on celebrities by MRC based on BERT. In **Proc. 36th PACLIC**, pp. 757–766, 2022.

[5] R. Peeters and C. Bizer. Using ChatGPT for entity matching. **arXiv preprint arXiv:2305.03423**, 2023.

[6] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham. In-context retrieval-augmented language models. https://arxiv.org/abs/2302.00083, 2023.

[7] V. Shaurya, Z. Atharva, D. Somsubhra, S. Anurag, B. Upal, N. Shubham Kumar, G. Shouvik, R. Koustav, and G. Kripabandhu. LLMs – the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on Indian court cases. In **Findings of EMNLP**, pp. 12451–12474, 2023.

[8] K. Sugawara and T. Utsuro. Developing a dataset for mining reviews in tweets focusing on celebrities' aspects. In **Proc. 7th ABCSS**, pp. 466–472, 2022.

[9] 横山響, 宇津呂武仁. 芸能人への感想を表す X 上のポスト集約およびウェブ検索・RAG によるその理由の集約. 言語処理学会第 31 回年次大会論文集, 2025.

[10] H. Zhang, X. Liu, and J. Zhang. SummIt: Iterative text summarization via ChatGPT. In **Findings of EMNLP**, pp. 10644–10657, 2023.

A　手法 [9] による評価対象の抽出

3 章で述べた，先行研究 [9] による評価対象の抽出で発生する問題を表しているのが図 3 である．図 3 にはフワちゃんを例に，過去に問題を起こした人のポストには明確な評価対象が示されていないことを表している．

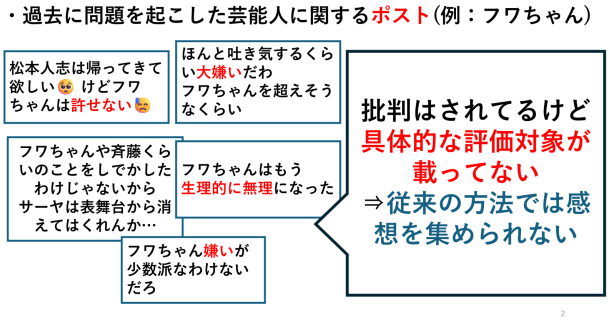

図 3　フワちゃんを例とした評価対象の抽出

B　対象となった芸能人リスト

表 5 は本論文で対象とした芸能人のリストである．芸能人は 10 人おり，そのうち 5 人が先行研究 [9] で元々対象となっていた人であり，過去に問題の無い人である．残りの 5 人が本論文で新たに対象となった人であり，過去に問題があり，活動を休止していた人となる．

表 5　対象とする芸能人

手法 [9] から対象の人	本論文から対象となった人
山田涼介	フワちゃん
二宮和也	ピエール瀧
菊池風磨	中丸雄一
小栗旬	槇原敬之
綾野剛	宮迫博之

C　手法 [9] との対応付けの具体例

図 4 は 4.4 節で行っている対応付けの様子を示したものであり，表 6 は，その結果の具体例である．フワちゃんの場合，手法 [9] でのみで得られる評価対象・感想・理由はなく，逆に本論文の手法でのみ得られる観点・話題が多くあることから，4.4 節で述べたように，本論文の手法の方が幅広い情報を得ることが出来ると分かる．

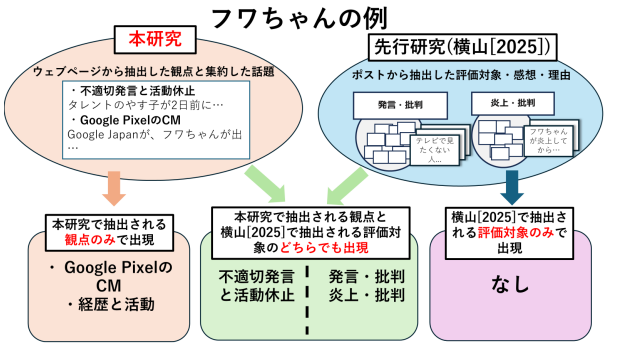

図 4　観点・話題と評価対象・感想・理由の対応付けの例

D　芸能人の経歴の良否判定

図 5 は，5.1 節で行っている，フワちゃんを対象としたときに抽出された観点の 1 つである「不適切発言と活動休止」に対して良否判定の図である．芸能人の観点に対して，詳細な情報を参考にして 5 つの分類のどこに当てはまるかを ChatGPT が判断している．

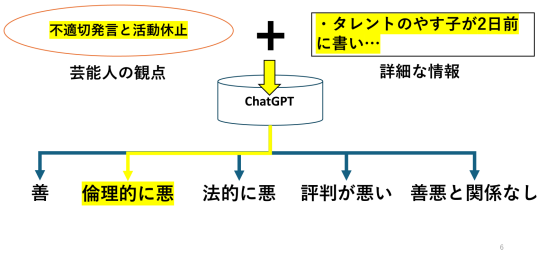

図 5　フワちゃんの観点・話題の良否判定

表 6　観点・話題と評価対象・感想・理由の対応付け

芸能人名	フワちゃん	
	手法 [9]	本論文の手法
二手法間で**重なった** 観点・話題 / 評価対象・感想・理由	(発言・批判，暴言・批判，炎上・批判，炎上・共感，嫌い・共感，嫌い・批判，嫌われている・共感，消えた・批判)	不適切発言と活動休止
二手法間で**重ならなかった** 観点・話題 / 評価対象・感想・理由	なし	経歴と活動，言語能力と学歴，交友関係，ファッションと影響，メディア出演