Studying the Representation of the LGBTQ+ Community in RuPaul's Drag Race with LLM-Based Topic Modeling

Mika Hämäläinen
Metropolia University of Applied Sciences
Helsinki, Finland
firstname.lastname@metropolia.fi

Abstract

This study investigates the representation of LGBTQ+ community in the widely acclaimed reality television series RuPaul's Drag Race through a novel application of large language model (LLM)-based topic modeling. By analyzing subtitles from seasons 1 to 16, the research identifies a spectrum of topics ranging from empowering themes, such as self-expression through drag, community support and positive body image, to challenges faced by the LGBTQ+ community, including homophobia, HIV and mental health. Employing an LLM allowed for nuanced exploration of these themes, overcoming the limitations of traditional word-based topic modeling.

1 Introduction

The representation of LGBTQ+ identities in mass media is an important area of research to gain a better understanding on what kind of an image of the LGBTQ+ community is broadcast to the public. Media representations contribute significantly to the shaping of public perceptions (McCombs, 2002) and they influence on societal attitudes towards marginalized communities (see German 2017).

Within this context, RuPaul's Drag Race (RPDR)[1] has emerged as a prominent cultural artifact, offering a platform that foregrounds the art of drag and simultaneously engages with themes of gender, sexuality and queer culture (see Chan 2024). The series, which debuted in 2009, has garnered widespread acclaim and critical attention[2], becoming a touchstone for LGBTQ+ representation in mainstream entertainment.

Media and television studies have long studied the role of popular culture in reflecting and shaping societal norms and ideologies (Calvert et al., 2007). Television, as a mass medium, occupies a unique position in the cultural landscape, blending entertainment with implicit and explicit narratives about minority identities (Greenberg et al., 2002). Scholars have argued that television serves as both a mirror and a mold that offers audiences representations which both reflect their lived realities and influence their perceptions of the world (Ott and Mack, 2020). For this reason, it is important to study what kind of a picture of the LGBTQ+ RPDR paints, especially since it is one of the few widespread LGBTQ+ shows that is broadcast globally.

Furthermore, television studies have emphasized the interplay between audience reception and media production in how viewers actively interpret and negotiate the meanings embedded within televised texts (Jensen, 2002). These interpretations are shaped by cultural, historical and personal contexts, and thus they create a complex feedback loop between creators and consumers (see Hagen and Wasko 2000).

Recent advances in large language models (LLMS) provide new opportunities for analyzing large-scale textual data, which makes more detailed topic modeling possible as we no longer need to rely on word-level methods that were in fashion before LLMs. Topic modeling with large language models (LLMs) has emerged as a powerful tool for exploring thematic structures in text corpora (Pham et al., 2023; Kapoor et al., 2024; Invernici et al., 2024).

This study employs LLM-based topic modeling to analyze the representation of the LGBTQ+ community in RuPaul's Drag Race subtitles from seasons 1-16. By analyzing the transcripts of the show, we aim to see how it reflects and represents the LGBTQ+ community. This method helps us explore narrative and in-group attitudes portrayed in the show. Our goal is to contribute to conversations about how media portrays LGBTQ+ identities and

[1] A show produced by World of Wonder

[2] https://www.televisionacademy.com/shows/rupauls-drag-race

Proceedings of the Queer in AI Workshop, pages 1–5

to show how our method can help us understand these representations better.

2 Related work

RPDR is no stranger to scientific study. In this section, we will go through some of the recent body of work that has studied the show.

Edgar (2011) explores how the show frames drag performance through normalization, reinforcing stereotypical ideals of femininity while simultaneously illustrating the complexities of gender as a performative construct. By examining the experiences of key contestants, Edgar highlights how drag performance is judged not only by skill but by adherence to specific gendered expectations, such as natural beauty and the seamlessness of femininity. While the show borrows successful elements from other reality television formats to engage mainstream audiences, this normalization risks reducing drag to entertainment, sidelining its potential to subvert rigid gender binaries.

In the analysis of RuPaul's Drag Race (RPDR) by Brennan (2017), the author explores the interplay of authenticity, competition and consumption within the show, arguing that these dimensions both reflect and complicate the format of reality television. The study examines how authenticity is negotiated through drag queens' performances, revealing tensions between personal identity and constructed personas, while competition emphasizes individuality in a space shaped by neoliberal values and historical marginalization. Additionally, the author critiques the show's commercial underpinnings the highlight the role of branding and consumerism in shaping perceptions of drag culture.

In their article, Strings and Bui (2014) analyze the interplay of race and gender in the third season of RPDR. They argue that while the show challenges traditional notions of gender through drag performance, it enforces rigid racial authenticity, particularly for African American contestants. This duality allows gender to be fluid and performative, while race is treated as fixed and essential, leading to racial stereotyping and tensions among contestants. The authors highlight how these dynamics reflect broader societal patterns, where racial identities are commodified and constrained even within ostensibly progressive queer spaces.

Goldmark (2015) examines the complex interplay between reality television, queer identity and neoliberal ideals through the lens of the show's first season. The analysis critiques how RPDR employs narratives of transformation and success, tying them to aspirational themes of the American Dream. While the show celebrates diversity and individuality, the study highlights the underlying contradictions, particularly its reliance on cultural and linguistic hierarchies that privilege English and U.S. norms. The article also interrogates racial and economic disparities, showcasing how contestants like BeBe Zahara Benet and Nina Flowers symbolize both the potential and limitations of inclusion, complicating the program's portrayal of upward mobility and integration into an idealized U.S. nation.

3 Data

We use the subtitles of RuPaul's Drag Race for seasons 1 to 16 that are available on OpenSubtitles[3]. The seasons vary in length. The shortest season is the first season with only 9 episodes and the longest season is the 13th season with 17 episodes. A typical length of a season is 14 episodes. The corpus size for each episode can be seen in Table 1.

Season	Tokens	Season	Tokens
1	78k	9	104k
2	101k	10	172k
3	129k	11	153k
4	119k	12	165k
5	114k	13	199k
6	103k	14	182k
7	94k	15	155k
8	104k	16	179k

Table 1: Size of each season in tokens

The seasons extend from the first season released in 2009 to the 16th season released in 2024. Season 17 was excluded as it was still not fully released during this study. The judges have changed over the years and some queens have been featured in multiple seasons. The only judge who has appeared in every season is RuPaul himself.

The subtitle files were cleaned from timestamps and only text was retained. Some of the subtitle files were not encoded in Unicode format, which led to some encoding errors. These erroneous characters including some invisible Unicode characters

[3]https://www.opensubtitles.org/en/ssearch/sublanguageid-eng/idmovie-171453

The main LGBTQ+ conversation topics in these subtitles from RuPaul's Drag Race, focusing on Adore Delano's statements, are: 1. **Drag Identity and Performance:** This is central, encompassing Adore's stage name (Adore Delano, its meanings, and the humor around it), the contrast between their drag persona and their birth name (Danny), and the creation of catchphrases and overall performance style. The discussion of "wearing people's stories" through thrift store finds also speaks to the performative and expressive nature of drag. 2. **Family Acceptance and Support:** Adore discusses their relationship with their mother, highlighting a complex dynamic of love, criticism, and ultimately, support. This touches on common themes of family acceptance within the LGBTQ+ community. The mother's overbearing nature is portrayed humorously, but the underlying love is evident. 3. **Gender Expression and Identity:** The statement "I'm not a boy. So rude" directly addresses gender identity and the rejection of gender norms. The discussion of their natural pink hair and the frustration with pageant queens not understanding it speaks to the broader theme of self-expression and challenging beauty standards. 4. **Queer Community and Influences:** Adore mentions Alaska as their favorite queen, indicating a connection to and appreciation for other members of the LGBTQ+ community and the influence they have. While there's humor and self-deprecation throughout, these subtitles reveal key themes relevant to LGBTQ+ experiences, focusing on self-discovery, family dynamics, the performance of identity, and community.,

Table 2: An example output from the LLM

were removed. This did not affect the textual content of the subtitles as they were in English and all English alphabets were encoded consistently across the files. Some of the subtitles included HTML tags such as <*i*> and <*b*>, these tags along with their possible attributes were deleted as well.

4 LLM-based topic modeling

We use Gemini 1.5 Flash (Georgiev et al., 2024) LLM to extract a list of LGBTQ+ related topics for each episode of each season. This is simply done by prompting the model through the API. We use the prompt template shown in Table 3 populated for each episode.

For the following subtitles from RuPaul's Drag Race, give a list of the main LGBTQ+ conversation topics. Subtitles: <Subtitle data>

Table 3: Prompt template used for extracting the topics

The prompt resulted in an analysis of the main LGBTQ+ topics discussed in the episode (see an example in Table 2). Every analysis has a list of topics indicated by a bolded topic title such as ***Coming out and self-expression:*** or ***Body image and eating disorders:***. Each title is followed by a further analysis of the topic. We sampled 5 LLM produced analyses randomly and compared the topics to what was discussed in the episodes. We found the LLM results to be of sufficient quality.

Using the topic titles, we separate each topic along with its description into different strings for each episode. We remove all text in the LLM answers that is not part of a topic description. This way, each episode is now described by a list of topic strings indicating the topic and description.

We use *text-embedding-004* model from Google Gemini API[4] to produce topic embeddings for each topic string. These topic embeddings are used to cluster the topics together with HDBSCAN algorithm (Campello et al., 2013) using UralicNLP Python library (Hämäläinen, 2019).

HDBSCAN does not require a fixed number of clusters, but it will find an optimal number of clusters on its own. We tested with several parameter values for minimum cluster size and found that 10 resulted in a good number of clusters that was still manageable to go through manually.

The algorithm found 43 cluster, which we further combined manually given that several clusters had similar topics but described using different wordings. The titles were often very similar if not identical, but the semantic contents of the descriptions were different enough for the clusters not being merged. We also tried affinity propagation clustering (Frey and Dueck, 2007), but didn't find it producing any better results, for this reason we proceeded to manual merging.

We removed a few topic clusters altogether because they did not deal with LGBTQ+, but were

[4]https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/text-embeddings-api

Topic	Occurrences	Topic	Occurrences
Ageism within the LGBTQ+ community	21	Drag as a form of self-expression and artistry	187
Mental health and resilience	54	Intersectionality (race and class)	27
Sisterhood and community	29	Gender expression and identity	60
Internalized homophobia and self-acceptance	77	Relationships and intimacy	60
LGBTQ+ community and representation	219	Negative body image and beauty standards	65
Representation and Visibility	86	Coming out and self-acceptance	159
The importance of community and family support	100	Positive body image and self-love	129
HIV/AIDS awareness and activism	11	Homophobia and discrimination	45

Table 4: Topic clusters and how often cluster topics appeared in the analyses

rather about the competition itself such as judging, winning and elimination. We also removed clusters related to humor because they were not LGBTQ+ related.

5 Results

The results of the clustering can be seen in Table 4. The topics listed in the table represent the topic clusters and the occurrences indicate how many times the topic was found the LLM analyses for the all the seasons.

The most commonly discussed themes were *LGBTQ+ community and representation*, which refers to being a representative of the LGBTQ+ community, *Drag as a form of self-expression and artistry*, *Coming out and self-acceptance*, *Positive body image and self-love* and *The importance of community and family support*. All in all, the most common themes are either empowering or can be seen as a growth story.

Although not in the list of the most common topics, RPDR also frequently visits negative themes that are typically seen as problematic for LGBTQ+ people such as *Mental health and resilience*, *Internalized homophobia and self-acceptance*, *HIV/AIDS awareness and activism*, *Negative body image and beauty standards* (including body dysmorphia) and *Homophobia and discrimination*. An additional negative topic that is perhaps not as stereotypically seen as an LGBTQ+ problem is *Ageism within the LGBTQ+ community*.

Some of the more positive and less frequent topics include *Sisterhood and community*, *Representation and Visibility*, which means representation of oneself and visibility as a public figure, *Intersectionality (race and class)*, *Gender expression and identity* and *Relationships and intimacy*.

6 Conclusions

Much of the prior work in research on RPDR has taken a rather critical and negative view on the show as evidenced in the related work section. However, if we look at the LGBTQ+ topic clusters found by our method, a different narrative can be perceived. A narrative of hope. Many of the topics are empowering such as how one can use drag to express themselves or how one is representative of a bigger LGBTQ+ community, i.e. one is not alone.

One can perceive hope through the difficult themes such as coming out and it ultimately leading to self-acceptance. And regardless of the bad thing such as homophobia (internalized or externalized) or the unrealistic beauty standards set by the society, one can still overcome them.

Our intention is not to invalidate any of the existing and more critical research. Our study simply revealed another side of the show. Despite of the problems the show has, our NLP approach has shown that the show serves an important puprose as a beacon of hope for LGBTQ+ people and, by discussing difficult themes that many of us queer people can relate to, the show delivers a message to their LGBTQ+ audience that they are not alone with their problems.

In the future, it would be interesting to study how the topics have evolved throughout the series from one season to another. Also, RPDR has been adapted to many other regions and languages. It would also be interesting to study what kind of topics exist in those shows and how comparable they are to the main series.

7 Limitations

When analyzing large amounts of textual data, no method comes without limitations. We, in particular, have always found traditional topic modeling methods quite limited as they operate on word level. LLMs overcome this limitation as they can produce a more thorough and reasoned analysis. As LLMs extend our topic modeling beyond words, they come with their own limitations. LLMs can get an generate listing of topics, but the listing may

not contain all the topics and there might be unknown biases in how the topics are picked by the LLM due to their black box nature.

We used the free version of Gemini API, which means that conducting a similar study does not require big computational resources or a thick wallet. However, this also means that we did not conduct this research with the best state-of-the-art models. Expensive models such as Gemini 2.0 or GPT-4o would have likely been able to extract even more topics from the subtitles. Their embeddings could have also resulted in more accurate clustering results.

References

Niall Brennan. 2017. Contradictions between the subversive and the mainstream: Drag cultures and rupaul's drag race. *RuPaul's Drag Race and the shifting visibility of drag culture: The boundaries of reality TV*, pages 29–43.

Ben Calvert, Neil Casey, Bernadette Casey, Liam French, and Justin Lewis. 2007. *Television studies: The key concepts*. Routledge.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Alex Siu Wing Chan. 2024. Rupaul's drag race: A cultural phenomenon that challenges gender norms and sparks conversations across borders. *Journal of Homosexuality*, 71(8):1863–1866.

Eir-Anne Edgar. 2011. Xtravaganza!": drag representation and articulation in" rupaul's drag race. *Studies in popular culture*, 34(1):133–146.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science*, 315(5814):972–976.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Daniel B German. 2017. The role of the media in political socialization and attitude formation toward racial/ethnic minorities in the us. In *Nationalism, Ethnicity, and Identity*, pages 285–298. Routledge.

Matthew Goldmark. 2015. National drag: The language of inclusion in rupaul's drag race. *GLQ: a journal of Lesbian and Gay studies*, 21(4):501–520.

Bradley S Greenberg, Dana Mastro, and Jeffrey E Brand. 2002. Minorities and the mass media: Television into the 21st century. In *Media effects*, pages 343–362. Routledge.

Ingunn Hagen and Janet Wasko. 2000. *Consuming audiences?: production and reception in media research*. Hampton Press Cresskill, NJ.

Mika Hämäläinen. 2019. Uralicnlp: An nlp library for uralic languages. *Journal of open source software*, 4(37):1345.

Francesco Invernici, Francesca Curati, Jelena Jakimov, Amirhossein Samavi, and Anna Bernasconi. 2024. Capturing research literature attitude towards sustainable development goals: an llm-based topic modeling approach. *arXiv preprint arXiv:2411.02943*.

Klaus Bruhn Jensen. 2002. Media audiences: Reception analysis: mass communication as the social production of meaning. In *A handbook of qualitative methodologies for mass communication research*, pages 135–148. Routledge.

Satya Kapoor, Alex Gil, Sreyoshi Bhaduri, Anshul Mittal, and Rutu Mulkar. 2024. Qualitative insights tool (qualit): Llm enhanced topic modeling. *arXiv preprint arXiv:2409.15626*.

Maxwell McCombs. 2002. The agenda-setting role of the mass media in the shaping of public opinion. In *Mass Media Economics 2002 Conference, London School of Economics*.

Brian L Ott and Robert L Mack. 2020. *Critical media studies: An introduction*. John Wiley & Sons.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Sabrina Strings and Long T Bui. 2014. "she is not acting, she is" the conflict between gender and racial realness on rupaul's drag race. *Feminist Media Studies*, 14(5):822–836.