BAN443: Transforming Business with AI: The Power of Large Language Models

Candidate numbers: 6, 16, 37, 49

December 2024

Abstract

This paper investigates gender dynamics in the Norwegian debate program "Debatten". Due to prompt size constraints discovered late in the semester, the analysis is limited to ten episodes, as we were unable to reproduce additional data. Using subtitles from these episodes, the study employs a large language model (LLM) to identify participants' genders and analyze behavioral dynamics in same-gender and mixed-gender interactions. The research evaluates dimensions such as argument clarity and relevance, engagement style, and power use to uncover patterns influenced by gender.

The findings reveal that same-gender interactions, particularly female-to-female (F-to-F), score higher in all dimensions, showcasing clearer communication and more collaborative dynamics. Mixed-gender interactions exhibit lower and more varied scores, particularly in power use, suggesting potential challenges in achieving communication balance. Male-to-male (M-to-M) interactions rank second in overall scores but display a more competitive style compared to F-to-F dynamics.

Regression analysis shows minimal, statistically insignificant differences between same-gender and mixed-gender pairs. While same-gender pairs slightly outperform mixed-gender pairs in clarity and power use, the results emphasize that factors such as rhetorical skill and debate strategy could play a larger role. Power use emerged as the most predictive dimension, reflecting its significant influence on debate outcomes.

This study highlights the role of gender in shaping interaction patterns in debates while emphasizing the impact of individual and contextual factors. Future research could explore larger datasets, longitudinal analyses, and variations across cultural or rhetorical contexts to provide a deeper understanding of gendered communication in public discourses.

Contents

List of Figures

List of Tables

1 Introduction

Debates form the backbone of democratic communication, enabling an examination of societal dynamics and power structures. While gender dynamics have been extensively studied in other settings, such as workplaces and educational environments, there is some lack of research exploring how gender influences interactions in live debates. This gap underscores the importance of investigating how communication styles, power dynamics, and engagement strategies vary across gender configurations in such contexts. Building on this premise, this paper seeks to contribute to the limited body of research by analyzing interactions in the Norwegian debate program "Debatten."

This study utilizes subtitles from ten episodes of "Debatten" to analyze interaction patterns. A large language model (LLM) is employed to identify participants, infer their gender, and assess behavioral dynamics within same-gender and mixed-gender pairings. Key dimensions, including argument clarity, engagement style, and power use, are systematically evaluated to uncover patterns that illustrate how gender shapes the flow of debate.

1.1 Research Question

This paper examines gender dynamics in the Norwegian debate program "Debatten," with a focus on participant interactions. Initially, the study investigated whether moderators treated male and female participants differently; however, the findings revealed no significant differences. Consequently, the focus shifted to participant interactions. By analyzing male-to-male, female-to-female, male-to-female, and female-to-male interactions, this study aims to uncover variations in communication styles, engagement strategies, and power dynamics across same-gender and mixed-gender pairings. Against this backdrop, the research addresses the following question:

How do interaction patterns differ between same-gender and mixed-gender pairs in debates?

2 Literature Review

This section reviews prior research examining the influence of gender on communication styles and interpersonal interactions. Carol Kinsey Goman's article, "Is Your Communication Style Dictated by Your Gender?" (Goman, 2016), provides a foundational analysis of gendered communication in workplace settings. Goman highlights that men typically adopt direct and assertive communication styles, prioritizing independence and status, whereas women emphasize collaborative approaches and relationship building. These differences, shaped by social expectations, can sometimes lead to gender misunderstandings. Goman argues that understanding these differences can help improve communication and create more inclusive workplaces.

In the study "Gender Interactions in Online Debates: Look Who's Arguing with Whom," Allan Jeong and Gayle V. Davidson-Shivers explore the use of gender interactions within academic online debates (Jeong & Davidson-Shivers, 2003). Graduate students participated in team-based debates on a discussion board. The study revealed that women were less inclined to challenge other women, potentially contributing to lower participation rates in debates. Conversely, men demonstrated equal interaction rates with both genders but showed a preference for debating other men. On the other hand, men interacted equally with both genders but were more likely to engage in debates with other men. Male-to-male interactions also generated 36% more messages than female-to-female exchanges, showing that conflict tends to drive more discussion in these settings.

Although existing literature offers valuable insights into gendered communication and debate patterns, significant gaps remain, particularly regarding live debates. For instance, Goman's analysis focuses primarily on workplace settings and does not extend to how gendered communication styles play out in public debates. Similarly, Jeong and Davidson-Shivers' study focuses on online academic debates and doesn't account for the unique dynamic of live debates, the direct, in-person power use between participants, which are shaped by physical presence, tone, and body language.

In addition, much of the existing research looks at gender interactions in a broad sense without delving into specific contexts such as debates. For instance, few studies explore how the structured nature of debates, the role of moderators, or the goals of politicians influence these interactions. Our research aims to fill this gap by studying how gender shapes the interaction dynamics in debates, specifically in the Norwegian show "Debatten". We plan to analyze interaction patterns between same-gender and mixed-gender pairs to uncover differences in communication styles, engagement levels, and power use.

3 Data

In this section, we discuss the process of data retrieval, provide a description of the dataset, and explain the steps involved in preparing the data for analysis.

3.1 Illustration Description

The figure provides a visual representation of the process for retrieving, processing, and analyzing subtitle files from NRK TV using an automated flow.

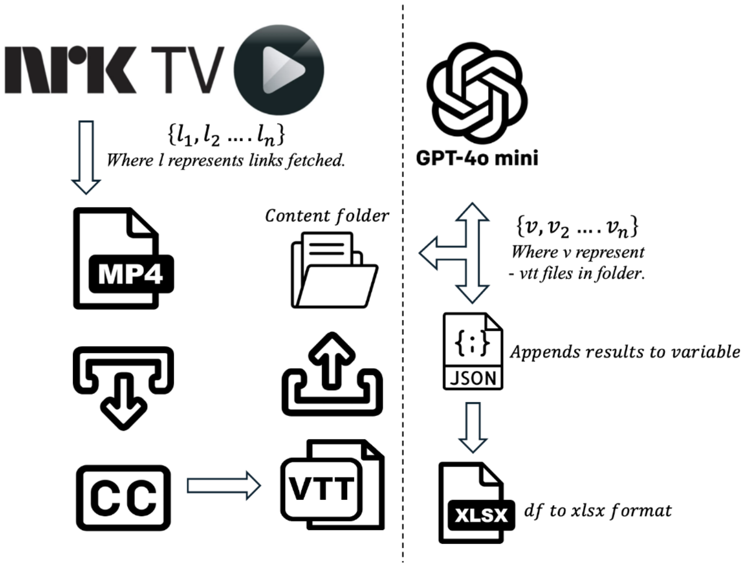
NRK TV
$\{l_1, l_2 \dots . l_n\}$
Where l represents links fetched.
MP4
Content folder
CC
VTT
GPT-4o mini
$\{v, v_2 \dots . v_n\}$
Where v represent - vtt files in folder.
JSON
Appends results to variable
XLSX
df to xlsx format

Figure 1: Dataflow

On the left side, the workflow begins by fetching links $(l_1, l_2, \dots, l_n)$ to episodes from NRK TV. These links are processed to identify associated video files in MP4 format. The subtitles (closed captions, CC) are extracted and saved as `vtt` (Web Video Text to Track File) files. These files are then organized into a content folder, preparing them for further analysis.

On the right side, the `vtt` files $(v_1, v_2, \dots, v_n)$ stored in the content folder are analyzed using the GPT-4o Mini module. This module processes the textual content of the `vtt` files and appends the results into a structured JSON format. The JSON data is then converted into tabular format and exported as an Excel file (`xlsx`) for easy visualization and further analysis.

3.2 Data Sources

The dataset used in this study was obtained from multiple sources, including public video platforms provided by NRK and custom API responses. The dataset's structure was inspired by a guest lecture held by Norges Bank Investment Management (NBIM) on September 17, 2024, which offered valuable insights into systematic data collection and organization. This lecture emphasized the importance of structured data formats, such as JSON, for efficient handling and analysis of complex datasets. Guided by these principles, we organized interaction metrics and participant data into JSON and subsequently tabular formats for further analysis.
Due to prompt size constraints discovered late in the semester, the analysis is limited to ten episodes, as we were unable to reproduce additional data.

Initially, we planned to use our custom API to select the most recent episodes. However, we found that many entries labeled as "debates" were in fact interviews, and there was a significant gender imbalance among participants. Consequently, we manually selected 10 episodes comprising approximately 18,000 lines of dialogue.

To extract subtitles, we employed the yt-dlp package, which downloads subtitle files in vtt format directly from the NRK TV platform. These files were stored in a designated directory for subsequent analysis without any additional format conversion.

3.3 Data Description

Each episode is stored in a separate data file, with each file containing the full transcript of that episode. The dataset is structured such that each record corresponds to a single spoken line. Every line is paired with a timestamp marking the moment it was spoken, along with the full content of the sentence.

Example data in WEBVTT-format:

00:00:00.080 –> 00:00:04.080
Opptak av simultanteksting

00:00:05.520 –> 00:00:11.240
De er erneringsfysiologer, hobbykokker, leger og helsepersonell.

00:00:11.320 –> 00:00:17.200
Gjennom sosiale medier og bokhandlere påvirker matinfluenserne -

00:00:17.280 –> 00:00:20.880
- hva vi burde og hva vi ikke burde spise.

Figure 2: Example of Subtitle Transcription with Timestamp

3.4 Data Handling and Processing

The process of handling and processing data involved several key steps to ensure the dataset was ready for analysis.

Subtitle files were downloaded using the yt-dlp tool with a custom configuration designed to meet specific requirements. Subtitles were extracted in the vtt format while skipping video files, and all files were systematically stored in a designated directory. The script was built to handle potential errors and log issues during the download process, ensuring a robust workflow. The downloaded vtt files were analyzed directly without requiring additional conversions.

JSON data along with episode data, were transformed into tabular formats using Python's pandas library. This transformation resulted in two distinct dataframes.

Data cleaning and integration ensured consistency and usability of the dataset. In our case, there were no missing values; however, if any had been present, they would have been handled manually. Variables with inconsistent naming conventions, such as discrepancies in "Argument Clarity and Relevance" labels, were standardized to improve uniformity. Finally, all processed dataframes were exported as separate Excel files for secure storage and further exploration, facilitating additional analysis as needed.

4 Methodology

This section outlines methods for analyzing debate transcripts, focusing on prompt refinement and scoring gendered interaction patterns.

To analyze the data, we utilized generative AI, specifically OpenAI's ChatGPT-4o-mini, a type of large language model (LLM). This approach aligns well with the research objectives, as LLMs are effective tools for classifying and scoring text (Korinek, 2023). ChatGPT-4o-mini was chosen for its optimal balance of cost-efficiency, scalability, and performance, making it particularly suitable for processing large volumes of debate transcripts in a resource-efficient manner (G. OpenAI, 2024). Despite its lower cost, GPT-4o-mini delivers robust results, outperforming many other compact models on key benchmarks, thereby ensuring high-quality analysis for this project.

4.1 Design of the prompt

The prompt design was inspired by lecture materials, combining all variables into a single, comprehensive prompt rather than separating them into individual prompts. While separate prompts might have enhanced precision, testing showed that a unified prompt provided sufficient accuracy for the task.

To enhance the model's understanding, multiple examples were included in the prompt. This few-shot prompting approach was chosen for its improved accuracy over zero-shot prompting, while also being more efficient and cost-effective than fine-tuning.

4.1.1 Identify the speaker

Speakers in debate transcripts were identified using dialogue analysis, contextual cues, and conversational markers. Each speaker was assigned a unique identifier, such as "Debater 1," "Debater 2," or "Expert 1," and their gender (M/F) was inferred based on linguistic patterns, ensuring real names were excluded to maintain anonymity.

To achieve this, the prompt utilized linguistic cues like gendered pronouns (e.g., "he," "she") and contextual references (e.g., "as a father," "she's the director"). It also incorporated analysis of conversational style, language use, and word choice to provide additional insights into gender tendencies. This structured approach ensured accurate differentiation of speakers while adhering to a neutral and anonymized framework, supporting reliable analysis.

4.1.2 Argument for Using a 1-9 Likert Scale

In our prompt, we implemented a 1-9 scoring scale with 5 as the neutral midpoint. This scale was selected for its balanced range, allowing for clear differentiation between negative and positive performance levels without overly large intervals. The odd-number structure ensures a neutral midpoint, accommodating cases where evaluations are neither strongly positive nor negative.

The use of a 1-9 Likert scale is well-supported in research, which highlights its advantages in reducing response bias and capturing subtle differences in attitudes or perceptions Kusmaryono, Wijayanti, and Maharani (2022). The increased granularity of higher-point scales improves data precision and reliability, as demonstrated by enhanced validity and reliability in psychological and educational assessments Preston and Colman (2000); Warmbrod (2014). For this study, the scale provides the necessary resolution to evaluate nuanced dynamics in debate interactions effectively.

4.1.3 Temperature

The temperature setting of 0.4 was selected to balance coherence and diversity in text generation for analyzing debate transcripts. In GPT models, temperature controls output randomness, with higher values (e.g., 0.7) generating more creative responses and lower values (e.g., 0.2) yielding more focused outputs (D. F. OpenAI, 2023).

Through testing, temperature 0.4 emerged as optimal, providing structured yet flexible outputs. It avoided oversimplifying complex dialogues while minimizing irrelevant randomness, making it suitable for identifying speakers, categorizing interactions, and scoring engagement. This setting supported precise and consistent analysis, aligning with the study's objectives.

4.2 Parameters

To study gender interactions in debates, we analyze four types: females interacting with females (F-to-F), males with males (M-to-M), females with males (F-to-M), and males with females (M-to-F). Each interaction is defined as one participant addressing a statement to another, distinguishing between F-to-M and M-to-F for directional clarity.

This framework enables a systematic examination of communication patterns both within and across gender groups, offering deeper insights into gendered dynamics during debates.

4.3 Categories

We analyze interactions across three categories: Argument Clarity and Relevance, Engagement Style, and Use of Power. These categories are based on insights from Carol Kinsey Goman's study (Goman, 2016), which highlights how gender differences in communication styles can influence clarity, engagement, and power dynamics.

Focusing on these dimensions allows us to systematically evaluate the strengths and weaknesses of gendered communication patterns, aligning with the study's goal of understanding interaction dynamics in debates.

4.3.1 Argument Clarity and Relevance

The Argument Clarity and Relevance category evaluates how clearly arguments are presented and their alignment with the topic. Based on Goman's study (Goman, 2016), gendered communication patterns often differ, with men tending toward directness and women favoring a more exploratory style. This category helps determine whether such patterns persist in same-gender and mixed-gender interactions during debates.

Scoring for this category ranges from 1 to 9. Lower scores (1-4) are assigned to arguments that are vague, poorly structured, or lack relevance. Mid-level scores (5) indicate arguments that are moderately clear and generally relevant. High scores (6-9) are given to arguments that are exceptionally clear, well-structured, highly relevant, and supported by strong evidence or examples. This structured approach ensures consistent evaluation of argument quality across interactions.

4.3.2 Engagement Style

The Engagement Style category evaluates how participants respond to their opponents' arguments and present counterarguments. Goman's study (Goman, 2016) suggests that communication styles may vary, with women often emphasizing collaboration and men adopting a more assertive approach. This category examines whether these tendencies are reflected in debate interactions.

Scoring focuses on the quality of engagement. Limited or irrelevant engagement with weak counterarguments is rated in the lower range (1-4). Mid-level scores (5) indicate basic engagement with main points and adequate but shallow counterarguments. High scores (6-9) are assigned to participants who demonstrate excellent engagement, effectively identifying weaknesses in their opponents' arguments and delivering well-structured, strong counterarguments. This framework ensures a consistent evaluation of interaction quality across debates.

4.3.3 Power Use

The Power Use category evaluates how participants manage conversational control and balance during interactions. This includes behaviors such as interruptions, acknowledgment of others' points, word choice, and the equitable distribution of speaking time. Goman's study (Goman, 2016) suggests that communication styles may differ, with men often exhibiting assertive behaviors and women emphasizing inclusivity. This category examines whether these tendencies appear in debates and how they influence same-gender and mixed-gender interactions.

Scoring for Power Use reflects conversational balance and respect. Low scores (1-4) indicate frequent interruptions, dismissive behavior, or dominance. Mid-level scores (5) suggest occasional power displays with some balance. High scores (6-9) are assigned for respectful, balanced participation that includes acknowledgment of others and minimal interruptions. This framework provides a consistent measure of how participants navigate power dynamics in debates.

4.3.4 Provide the API Model with Examples

To improve accuracy, we implemented a few-shot prompting approach, providing the model with carefully curated examples. As shown in Table 1, these examples were drawn from a test episode of "Debatten," excluded from the final sample set, to align with the category descriptions developed for scoring Clarity and Relevance.

To validate the model's performance, we randomly sampled outputs and cross-checked them with the original episode to ensure consistency and detect potential hallucinations. This iterative process refined the examples, providing the model with clear benchmarks for assigning scores across variables and ensuring reliable outputs.

Score Range	**Category Description**	**Example**
1–4 (Negative)	Unclear or illogical arguments. Poor relevance to the topic.	"Krisesentrene, politisakene, de bulmer over. Ta realiteten innover deg."
5 (Neutral)	Moderately clear arguments with basic logical structure. Generally relevant to the topic.	"Det Norge jeg kom til har endret seg. I dag er det noen som ikke engang har kommet hit ennå, de har en annen kultur og forventer at Norge skal tilpasse seg."
6–9 (Positive)	Clear, logically structured arguments. Highly relevant to the topic with strong supporting evidence.	"De vestlige verdiene står på spill. Demokrati. Ytringsfrihet. Religionsfrihet. Likestilling, synet på homofile. Dette kolliderer med manges praksis."

Table 1: Argument Clarity and Relevance Categories with Examples

5 Results

In this section, we present our findings and discuss what the results mean in relation to our research question.

5.1 Gender and Role Distribution per Episode

To analyze the composition of participants in each episode, we included two dimensions: gender (the proportion of women and men) and roles (experts and debaters).

Gender Distribution per Episode: By analyzing the gender distribution for each episode, we observe that there is always at least one participant from each gender. This ensures the possibility of measuring all interaction types within the dataset. The graph shows that some episodes exhibit a predominance of one gender, which may lead to an unequal distribution of data across interaction types. However, the scoring method accounts for this by utilizing a weighted average to ensure fairness in the analysis.

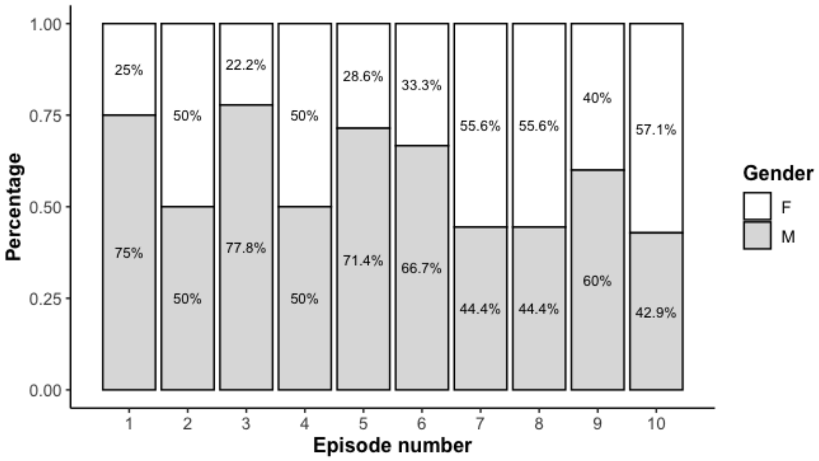

Figure 3: Gender distribution per episode

Role Distribution per Episode: The graph illustrates the distribution of debaters and experts across episodes, providing a clear view of the number of participants in each role for every episode. It is important to emphasize that the scores generated by the LLM were based solely on the debaters and did not take the experts into account, as specified in the prompt.

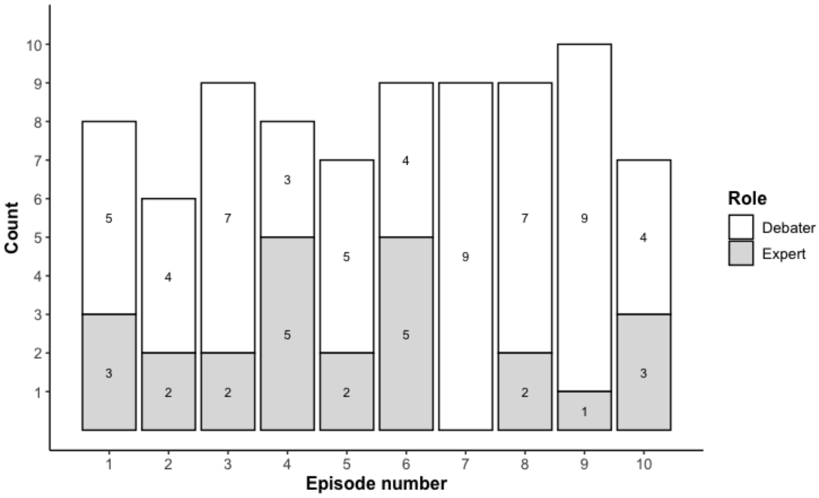

Figure 4: Role distribution per episode

5.2 Final scoring

Overall, the final scores for each interaction type are relatively consistent, with all scores clustering around an average of 6 points. The data suggests a pattern where same-gender interactions tend to receive higher scores compared to mixed-gender interactions. Mixed-gender interactions show the lowest average score, at 6.05 points, indicating a potential challenge in communication or alignment between genders. F-to-F interactions achieve the highest score of 6.58, suggesting a stronger rapport or alignment when women interact with each other. Similarly, M-to-M interactions rank second, with an average score of 6.12, reinforcing the idea that same-gender dynamics may facilitate smoother communication or engagement.

To understand these differences, it is necessary to examine each category in greater detail. By conducting a more granular analysis, we can determine whether these variations in scores are statistically significant. Such an exploration will help identify the underlying factors driving these differences, providing a clearer understanding of how interaction type influences scoring outcomes.

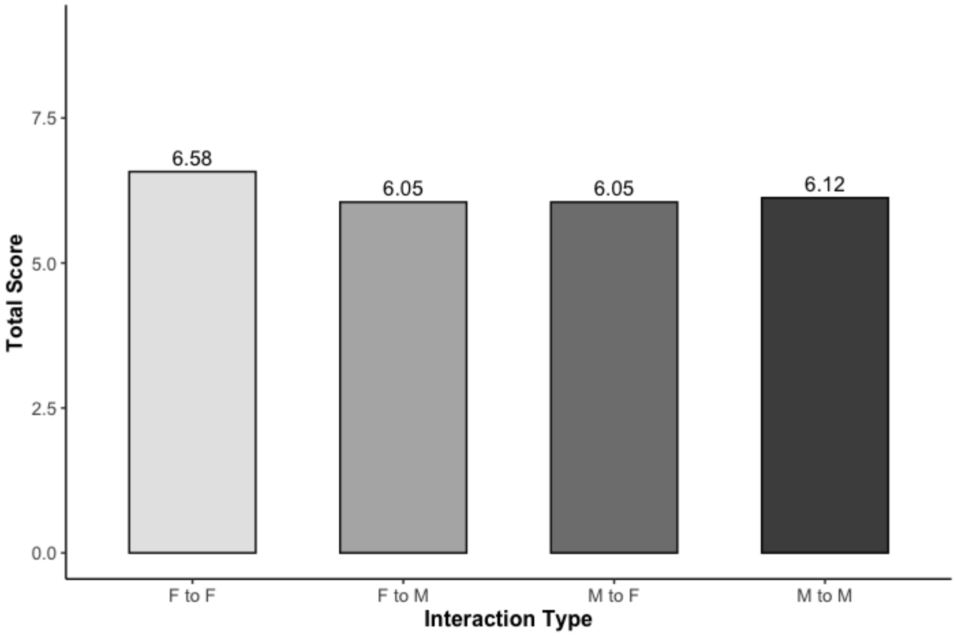

Figure 5: Final Scoring for each interaction pair

5.3 Average Scores per Interaction Type

To analyze how different interaction types perform across dimensions, we have created a bar plot that shows the average scores for each interaction type. These scores provide insights into how gender combinations influence the quality and dynamics of interactions in debates.

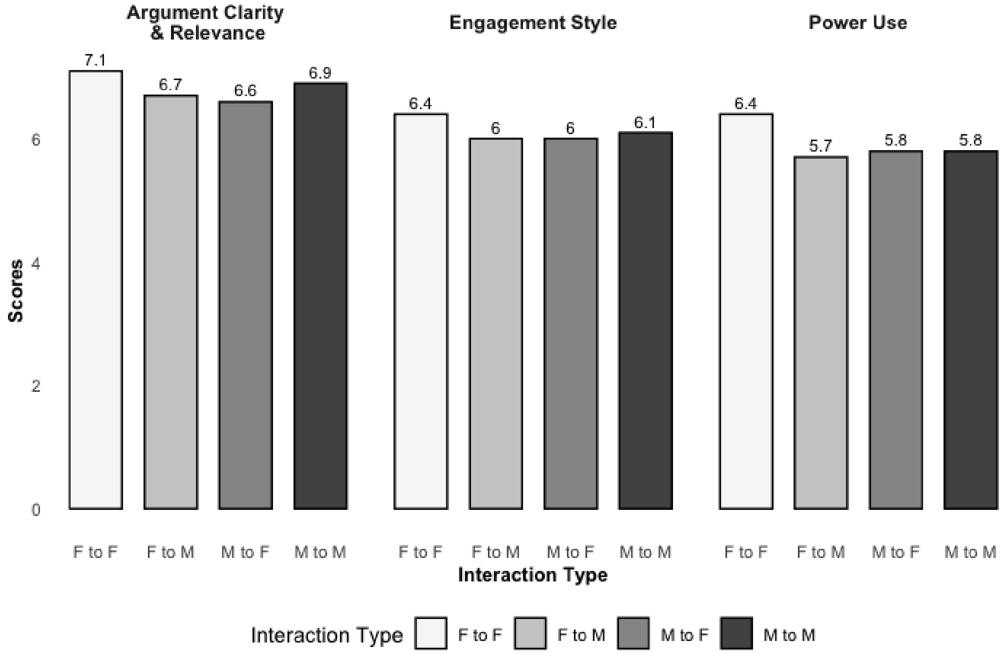

Figure 6: Average scores per Interaction Type

F-to-F interactions scored the highest across all three dimensions. In Argument Clarity & Relevance, they achieved an average score of 7.1, showing that women present clear and relevant arguments when talking to each other. Their Engagement Style shows strong collaboration and responsiveness, while their Power Use is balanced and respectful. These results suggest that women interacting with other women communicate effectively, likely due to a more cooperative dynamic.

F-to-M and M-to-F interactions had lower and more varied scores. In Argument Clarity & Relevance, F-to-M scored 6.7, and M-to-F scored 6.6. For Engagement Style, both types scored around 6.0, showing moderate engagement. In Power Use, these interactions had the lowest scores, suggesting less balanced power dynamics compared to same-gender interactions.

M-to-M interactions also scored lower than F-to-F interactions. They scored 6.9 for Argument Clarity & Relevance, 6.1 for Engagement Style, and 5.8 for Power Use. While these scores are only slightly lower, they may reflect a more competitive and less cooperative style compared to interactions between women.

In summary, F-to-F interactions scored the highest in all dimensions, suggesting more collaborative and balanced discussions. Mixed-gender interactions had lower and more varied scores, especially in Power Use, pointing to potential communication challenges. These findings can help further analyze gender dynamics in debates.

5.4 Regression Analysis

Our regression analysis investigates how same-gender and mixed-gender pairs interact in debates. We focused on three key areas: Argument Clarity and Relevance, Engagement Style, and Power Use.

Property	Argument C & R	Engagement Style	Power Use
Coefficients			
Intercept	3.315 (0.949)	0.599 (0.866)	1.233 (0.725)
Clarity and Relevance	-	0.146 (0.130)	0.160 (0.111)
Engagement Style	0.231 (0.206)	-	0.576 (0.107)
Power Use	0.339 (0.236)	0.771 (0.144)	-
Same gender (=1)	0.174 (0.193)	-0.071 (0.154)	0.150 (0.131)
Regression Statistics			
Multiple R	0.558	0.761	0.781
R Square	0.311	0.580	0.610
Adjusted R Square	0.254	0.545	0.577
Standard Error	0.583	0.463	0.400
Observations	40	40	40

Table 2: Regression results for the categories

Same-gender pairs scored slightly higher in Clarity and Relevance by 0.174 points compared to mixed-gender pairs. However, this difference is minimal and not statistically significant. The model explains 31.1% of the variation in Clarity and Relevance.

For Engagement Style, same-gender pairs scored 0.071 points lower than mixed-gender pairs. This variation is also small and not statistically significant. The model accounts for 58.0% of the variation in Engagement Style.

In Power Use, same-gender pairs scored 0.150 points higher than mixed-gender pairs. As with the other metrics, this difference is minor and not statistically significant. The model explains 61.0% of the variation in Power Use, which is the highest among the three categories.

Overall, the differences between same-gender and mixed-gender pairs are small and not statistically significant across all three aspects. The model for Power Use fits the data best, followed by Engagement Style and Clarity, and Relevance. Interestingly, stronger Power Use appears to be associated with stronger arguments and more engaging interaction styles.

5.5 Explanatory Power Based on Adjusted R^2

The adjusted R^2 values provide insights into the explanatory power of the models. The model for Power Use has the highest adjusted R^2 at 0.577, closely followed by Engagement Style with a value of 0.545. In contrast, the model for Clarity and Relevance has a lower adjusted R^2 of 0.254, suggesting that this dimension is more complex and less directly influenced by the variables in the model. This indicates that power balance and engagement are more measurable and predictable aspects of debates than the clarity and relevance of arguments. The low explanatory power for Clarity and Relevance may imply that this dimension is influenced by factors not captured in the regression model, such as the content of arguments and participants' rhetorical skills.

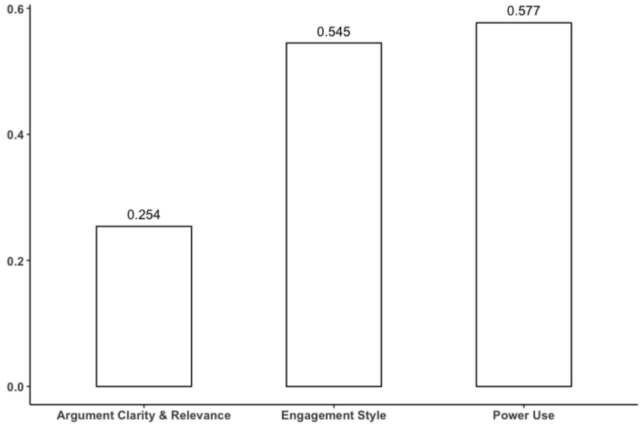

Figure 7: Role distribution per episode

5.6 Positive Correlations Between Categories

The regression results indicate positive correlations between dimensions. For example, increased Power Use is positively correlated with stronger Clarity and Relevance. Similarly, the model shows a strong correlation between Engagement Style and Power Use. These correlations suggest that participants who demonstrate better control over the conversation and actively engage with opponents' arguments also appear more persuasive. This underscores the importance of interaction dynamic, where power balance and engagement not only contribute to better interaction but also support the quality of argumentation.

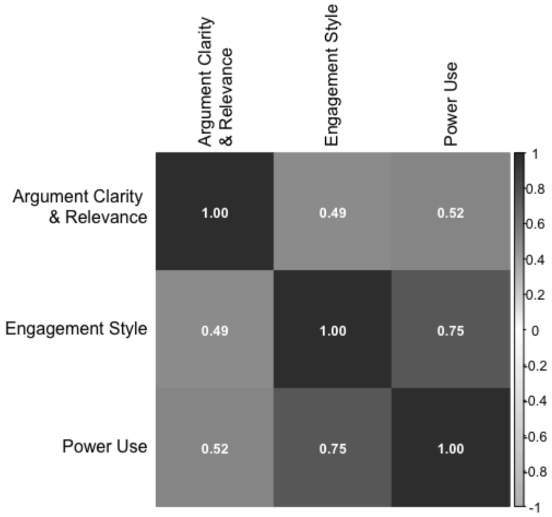

Figure 8: Correlation between the categories

6 Conclusion

6.1 Summary of Findings

This study examined how interaction patterns differ between same-gender and mixed-gender pairs in debates, analyzing 10 episodes of the NRK program "Debatten". Using an automated workflow and GPT-4o mini for text analysis, the study evaluated four interaction types: F-to-F, M-to-M, F-to-M, and M-to-F across three dimensions: Argument Clarity and Relevance, Engagement Style, and Power Use.

The results showed marginal differences between same-gender and mixed-gender pairs, with same-gender pairs scoring slightly higher in Clarity and Relevance 0.174 points higher and Power Use 0.150 points higher. However, these differences were small and not statistically significant, indicating that gender configurations alone have a limited influence on debate performance. Mixed-gender interactions exhibited consistent scores across all dimensions, reflecting comparable levels of engagement and interaction quality. Power Use emerged as the most predictive dimension, with an adjusted R^2 of 0.577, followed by Engagement Style 0.545, and Clarity and Relevance 0.254. These findings suggest that factors beyond gender, such as individual rhetorical skills or debate strategies, play a larger role in shaping interaction quality.

6.2 Strengths and Limitations

The study presents several key strengths. First, the use of an automated pipeline for subtitle extraction, data processing, and analysis ensured consistency and scalability. Second, the integration of advanced AI tools such as GPT-4o-mini enabled detailed scoring and analysis of interaction dynamic. Third, the focus on multiple dimensions provided a comprehensive understanding of debate performance across gender configurations.

Despite these strengths, the study had limitations. The dataset, limited to 10 episodes, reduced statistical power and generalizability. The reliance on coded scores, while systematic, may not fully capture the complexities of conversational dynamic or account for contextual factors like rhetorical style or cultural norms. Additionally, the study did not include longitudinal data, which could provide insights into how interaction patterns evolve over time or across debates.

6.3 Future Research

Future studies should address these limitations by using larger datasets that include more episodes and varied debate formats. Incorporating qualitative methods, such as examining patterns of interaction, could provide deeper insights into how interactions unfold. Exploring factors like rhetorical strategies, cultural differences, and audience reactions would also help to better contextualize the findings.

Longitudinal research tracking participants across multiple debates could provide a better understanding of how experience and familiarity influence interaction patterns. Comparative studies involving international debates or other public forums could explore how gendered communication patterns vary across cultural and political contexts. Finally, examining audience perceptions of debate dynamic could offer practical insights into how interaction styles influence public opinion and debate outcomes.

References

Goman, C. K. (2016). *Is your communication style dictated by your gender.* Retrieved 2024-10-31, from https://www.forbes.com/sites/carolkinseygoman/2016/03/31/is-your-communication-style-dictated-by-your-gender/

Jeong, A., & Davidson-Shivers, G. V. (2003). Gender interactions in online debates: Look who's arguing with whom. In *The annual meeting of the american educational research association, chicago.*

Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, *61*(4), 1281–1317.

Kusmaryono, I., Wijayanti, D., & Maharani, H. R. (2022). Number of response options, reliability, validity, and potential bias in the use of the likert scale: A literature review. *International Journal of Educational Methodology*, *8*(4), 625-637. Retrieved from https://doi.org/10.12973/ijem.8.4.625

OpenAI, D. F. (2023). *Cheat sheet: Mastering temperature and top_p in chatgpt api.* Retrieved 2024-11-25, from https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api/172683

OpenAI, G. (2024). 4o mini: advancing cost-efficient intelligence, 2024. *URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence.*

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(1), 1-15. Retrieved from https://doi.org/dbcr2g

Warmbrod, J. R. (2014). Reporting and interpreting scores derived from likert-type scales. *Journal of Agricultural Education*, *55*(5), 30-47. Retrieved from https://doi.org/10.5032/jae.2014.05030

Appendix

System Prompt and Role Definition

The interaction analysis system is designed to evaluate dialogue content using a structured prompt. Below is the system role and instructions as implemented in the automated analysis workflow:

```
role="system",
content="""
    You are an expert in language analysis and interaction analysis in
       discussions, with a focus on identifying all unique speakers in
       dialogue.
    Your task is to analyze the entire text provided, which may be in
       Norwegian or English
"""
```

User Prompt for Interaction Analysis

The user prompt provides a detailed framework for analyzing and categorizing debate interactions. Below is the structured template used:

```
role="user",
content=f"""
1. Identify and label each unique speaker in the transcript based on
   their dialogue content, contextual cues, or conversational markers.
```

```
    2. Assign a unique identifier to each speaker ("Debater 1", "Debater 2,"
        or "Expert 1" etc.) and determine their gender (M/F) when possible
       based on their name. Do not include their real name in the output.
    3. Include all participants present in the text and ensure no speaker is
        omitted. Do not limit the identification to a predefined number of
       examples.
    4. Analyze variations in language usage, word choice, or conversational
       patterns in the transcript to differentiate between speakers.
    5. Comprehensive Interaction Analysis: Evaluate and score the
       interactions between all debaters, not just a sample. This analysis
       should include every interaction and response in the provided text to
        ensure an accurate and holistic assessment of how each participant
       interacts with others.

Scoring Categories:
   -Argument Clarity and Relevance:
         Negative (1-4):
           "Unclear or illogical arguments. Poor relevance to the topic."
           Example: "Krisesentrene, politisakene, de bulmer over. Ta
              realiteten innover deg."
         Neutral (5):
           "Moderately clear arguments with basic logical structure.
              Generally relevant to the topic."
           Example: "Det Norge jeg kom til har endret seg. I dag er det
              noen som ikke engang har kommet hit enn , de har en annen
              kultur og forventer at Norge skal tilpasse seg."
         Positive (6-9):
           "Clear, logically structured arguments. Highly relevant to the
              topic with strong supporting evidence."
           Example: "De vestlige verdiene st r p  spill. Demokrati.
              Ytringsfrihet. Religionsfrihet. Likestilling, synet p
              homofile. Dette kolliderer med manges praksis."
   -Engagement style:
         Negative (1-4):
           "Limited or no engagement with opponent's arguments. Weak or
              irrelevant counterarguments."
           Example: "Jeg velger  ignorere det du sier. Det m  v re lov
                 kose seg."
         Neutral (5):
           "Basic engagement with main points. Adequate counterarguments,
              but lacking depth or consistency."
           Example: "Jeg mener at for  skape dialogen, kan vi ikke ta
              utgangspunkt i at all muslimsk ungdom er p  n m te. Men
              se nyansene."
         Positive (6-9):
           "Excellent engagement with opponent's arguments. Precise
              identification of weaknesses. Strong, well-formulated
              counterarguments."
           Example: "Vi har studier som viser det Dankel p peker, men det
              betyr ikke en sammenheng. Poenget er at vi ikke kan se p
              totalt inntak, men undergrupper og prosesseringsgrad innad i
              dem."
   -Power Use:
         Negative (1-4):
           Description: Frequent interruptions or complete disregard for
              other participants' contributions. Little or no
              acknowledgment of opposing arguments. The language may be
              overly assertive or dismissive, even without vocal tone cues.
           Example from transcript: "Du generaliserer en generasjon. Syv
              skoler ... Hvor mange elever?"
         Neutral (5):
           Description: Moderate power use with occasional dismissal of
              opposing points, but without excessive dominance. The
           argumentation shows some balance, and the responses hint at
           collaboration or acknowledgment.
        Example from transcript:
          Debater 1: "Synet p  homofile, dette har vi sett i klasserom
             og i forskning. Elever med muslimsk bakgrunn kommer
             d rligst ut ved holdningsunders keler."
          Debater 2: "Men vi m  se det i historisk perspektiv. Det var
             forbudt i Norge for 50  r siden."
      Positive (6-9):
        Description: Respectful and balanced interaction. Opposing
           points are explicitly addressed and built upon. Collaboration
            is evident, with a willingness to consider different
           perspectives in the discussion.
        Example from transcript:
          Debater 1: "Jeg skremmer ikke samfunnet. Samfunnet blir skremt
              av det de ser. Muslimene tar ikke debattene. Det er
             manglende likestilling."
          Debater 2: "Jeg er glad du tar opp utfordringene, men m ten
             vi gj r det p  kan skape mer dialog. Kanskje vi kan sette
              et felles m l for integrering."

6. Use an analysis model: Analyze how gender patterns may influence
   treatment in interactions, while controlling for other variables.
7. Scoring: All scoring should be in the interval 1-9, where 1 is the
   lowest and 9 is the highest.
8. Final Scoring: Combine scores across all categories for each
   participant to assign a final score reflecting their overall
   performance and engagement style in the full debate.

Analyze the interactions between all participants and categorize them
   into gender-based combinations (M to M, F to M, F to F, M to F).
   Ensure the output is presented in JSON format as follows:
{{
    "episode_name": "<generated_name>",
    "participants": [
        {{
            "role": "<role>",
            "gender": "<M/F>",
            "background": "(party affiliation, organization, etc.)"
        }}
    ],
    "interaction_analysis": {{
        "M_to_M": {{
            "Argument Strength": <number>,
            "Summary Argument Strength": "<summary>",
            "Engagement Style": <number>,
            "Summary Engagement Style": "<summary>",
            "Power Use": <number>,
            "Summary Power Use": "<summary>",
            "Total score (average of all)": <number>,
            "Total summary": "<summary>"
        }},
        ...
    }},
    "overall_trends": "<description>",
    "final_score": <number>
}}

Analyze the text below and categorize participants as   Debater   1,
     Debater   2,      Expert   1,    etc., not their real name. Evaluate
   how they interact with each other in all gender combinations (M to M,
    F to M, F to F, M to F). Include all participants and do not include
    the moderator in the analysis. In English only.
```