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

Candidate numbers: 6, 16, ?? ??

December 2024

Abstract

Include a title page with the paper's title and a short abstract/executive summary (max 250 words). The title page does not count toward the page limit.

Most journals allow 100-150 words. Obey this limit now. The main function of the abstract is to communicate the one central and novel contribution, which you just figured out. You should not mention other literature in the abstract. Like everything else, the abstract must be concrete. Say what you find, not what you look for. Here too, don't write "data are analyzed, theorems are proved, discussion is made.."

Contents

1 Introduction

This paper explores how gender dynamics influence interactions in political debates, focusing on the Norwegian show Debatten. Using a large language model (LLM), we will analyze behavioral patterns in participant interactions, categorizing them into same-gender and mixed-gender pairings. Our study aims to understand whether gender influences the flow and tone of the debates and to explore how fair and inclusive the debate platform is for all participants.

The main contribution of this paper is the use of an LLM to systematically analyze interactions in debates. We will work with subtitles from ten episodes of Debatten, using the model to identify unique speakers, determine their gender, and analyze the dynamics of their interactions. These interactions will be assessed based on factors like argument clarity, engagement style, and power use, using a scoring system to identify patterns in communication. We aim to explore how gender affects debate dynamics, uncovering possible biases and differences in how participants interact.

At first, we planned to look at differences in how male and female participants were treated by the moderator. However, our initial findings didn't show any clear gender-based differences in the moderator's behavior. Because of this, we decided to shift our focus to the interactions between participants, which allows for a deeper analysis of gender dynamics in debates. By grouping interactions into male-to-male, female-to-female, male-to-female, and female-to-male categories, we aim to better understand how gender influences communication styles and strategies in high-stakes political debates.

Should we write something more in the introduction? For example what types of analysis we are going to do, etc.

1.1 Research Question

This paper investigates the role of gender dynamics in debate settings, focusing on behavioral interactions between participants. We aimed to use LLM's to analyze and scale the interactions between the participants in "Debatten". Initially, we wanted to research the behavioral difference in moderator interactions between men and females in debate settings. However, our initial findings indicated that there were no notable differences in moderator interactions based on the gender of the participants. Moderator Fredrik Solvang seemed indifferent to men and women. This lack of distinction led us to reconsider our approach. To gain a deeper understanding of gender dynamics within "Debatten," we shifted our focus towards examining the participants' behavioral dynamics by exploring patterns in same and mixed-gender interactions. We had to adjust our original code to implement this. We adjusted the model to be more generalized as we could analyze new participants, allowing for a broader analysis of interaction patterns in pairs of the same gender and mixed gender. Our new approach provides insight into how gender might influence debate flow and participant interactions, allowing us to see if 'Debatten' is an equal place for both men and women.

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

2 Literature Review

There has been a lot of research on how gender influences communication styles or interactions in debates. For example, the article "Is Your Communication Style Dictated by Your Gender?" by Carol Kinsey Goman examines how men and women communicate differently in the workplace (Goman, 2016). Goman explains that men tend to have a more direct and assertive style, often focusing on independence and status. In contrast, women usually prioritize building connections and working collaboratively. These differences, shaped by societal expectations, can sometimes lead to gender misunderstandings. Goman argues that understanding these differences can help improve communication and create more inclusive workplaces.

Another relevant study, "Gender Interactions in Online Debates: Look Who's Arguing with Whom" by Allan Jeong and Gayle V. Davidson-Shivers, looks at how men and women interact in academic online debates (Jeong & Davidson-Shivers, 2003). Graduate students participated in team-based debates on a discussion board. The study found that women were less likely to argue with other women than men, which might explain lower participation rates among women in such debates. 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 these studies provide useful information about gender communication and debate patterns, they don't fully address the questions we want to explore. For example, Goman's work is focused on the workplace and doesn't consider how these gendered communication styles play out in public debates, especially in political contexts. Similarly, Jeong and Davidson-Shivers' study focuses on online academic debates and doesn't account for the unique dynamics of live political debates, such as audience reactions, media coverage, or the power dynamics between participants.

In addition, much of the existing research looks at gender interactions in a broad sense and doesn't explore specific contexts like political 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 flow of interactions in live political debates, specifically in the Norwegian show Debatten. We plan to analyze the differences in interaction patterns between same-gender and mixed-gender pairs to uncover potential biases, inequalities, or distinct strategies used by participants. This will help us better understand how gender influences political debate dynamics and whether these debates provide a fair platform for everyone involved.

3 Data

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 system.

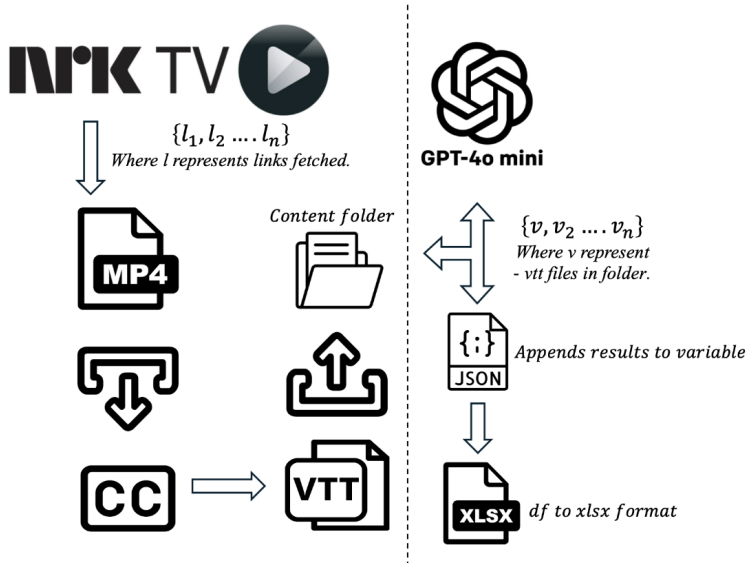

Figure 1: Dataflow

On the left side, the workflow begins by fetching links $(l_1, l_2, \ldots, l_n)$ 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` files. These files are then organized into a content folder, preparing them for further analysis.

On the right side, the `vtt` files $(v_1, v_2, \ldots, v_n)$ 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 and custom API responses. Additionally, the structure of the dataset and methodology were inspired by a guest lecture held by Norges Bank Investment Management (NBIM) on September 17, 2024. This lecture provided valuable insights into the systematic collection and organization of interaction and participant data. The importance of structured data formats were emphasized , such as JSON, for efficient handling and analysis of complex datasets. Inspired by this approach, our study adopted a similar methodology to organize interaction metrics and participant data into JSON and tabular formats for analysis.

To extract subtitles for episodes of interest, the `yt-dlp` package was utilized. The process involves downloading subtitle files in `vtt` format from platforms such as NRK TV. These files were stored in a designated directory for subsequent analysis without any conversion to other formats. The `vtt` files were chosen for their compatibility with modern subtitle extraction workflows.

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
paavirker matinfluenserne -

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

3.3 Data Description

The dataset comprises two key components:

-Scores Data includes detailed metrics related to interactions in each episode, such as argument strength, engagement style, tone variation, and power dynamics. This data provides both summarized and total scores, offering a comprehensive view of interaction patterns and dynamics. These metrics enable both quantitative analysis and qualitative interpretation of the interactions.

-Participants Data provides contextual information about participants in each episode, including their roles, gender, and backgrounds. This information supports demographic analyses and helps in understanding the participants' influence on discussions and outcomes.

The dataset consists of information from 10 episodes, each featuring multiple interactions and participants. Every record is linked to its respective episode name, allowing relational mapping between the two datasets for effective cross-referencing and analysis.

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. The first, the Scores DataFrame, aggregated interaction metrics such as argument strength, engagement style, tone dynamics, and power dynamics, alongside their summaries. The second, the Participants DataFrame, detailed participant roles, gender, and background information, providing valuable demographic context.

Data cleaning and integration ensured consistency and usability of the dataset. Missing or inconsistent entries were either imputed or flagged for exclusion to maintain data integrity. Variables with inconsistent naming conventions, such as discrepancies in "argument strength" 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 Analysis

4.1 Methodology

To explore how interaction patterns differ between same-gender and mixed-gender pairs in political debates, we analyzed episodes from the NRK program Debatten. We began by selecting a sample of 10 episodes, comprising approximately 18,000 lines of dialogue. To analyze our data, we are using generative AI, specifically OpenAI's ChatGPT-4mini, a type of large language model (LLM). This approach is well-suited to our research question, as LLMs is a good tool for classify and score text. We selected ChatGPT-4mini because it is one of the most powerful LLMs currently available to the public according to Korinek (2023). Next, we sent a request through an API call, which generated a response that was then returned via the API.

4.1.1 Sample

(!! CHANGE THIS SINCE WE USED THE OLD PROMPT !!) We selected 10 episodes to ensure a sample size large enough for meaningful analysis. While we initially aimed to include more episodes, resource limitations with the ChatGPT API made us stick with 10 episodes. We selected episodes including both male and female participants to ensure eligibility for the interaction analysis. To achieve this, episodes were manually picked by reviewing the names of the participants to assure individuals from both genders.

4.2 Design of the prompt

To build the prompt, we took inspiration from the lecture slides and decided to combine all the variables into a single, comprehensive prompt rather than creating separate prompts for each variable. While having individual prompts for each variable might have offered more precision, we tested the results from our current approach and determined that a single prompt was sufficient for this task. Our approach to designing the prompt involves incorporating multiple examples to effectively guide the model's understanding of the task. This few-shot prompting technique is preferred because it offers greater accuracy compared to zero-shot examples while being faster and more resource-efficient than fine-tuning.

4.2.1 Identify the speaker

Our method for identifying speakers in a debate transcript relies on analyzing the dialogue, contextual cues, and conversational markers. Each speaker is given a unique identifier, such as "Debater 1" or "Debater 2," and their gender (M/F) is inferred where possible, while ensuring that real names are excluded. To support this, we designed the prompt to focus on identifying gender through linguistic and contextual cues, such as patterns in names, gendered pronouns (e.g., "he," "she"), and contextual references like roles or titles (e.g., "as a father," "she's the director"). The prompt also guides the analysis toward conversational style, language use, and word choice, which can sometimes provide additional insights into gendered tendencies. This structured approach helps distinguish speakers while preserving a neutral and anonymized framework.

4.2.2 Scoring system

In our prompt, we implemented a 1-7 (!!! ENDRE PÅ DENNE !!!) scoring scale, with (!!4!!) serving as the neutral midpoint. This scale was chosen because it offers a balanced range for evaluating negative and positive scores without creating overly large intervals, making it easier to differentiate performance levels. Additionally, using an odd number ensures a clear neutral middle point for cases where the evaluation is neither strongly positive nor strongly negative.

4.2.3 Argument for Using a 1-9 Likert Scale

The selection of a 1-9 Likert scale is supported by research emphasizing the advantages of scales with an odd number of response options and a higher range of points. The odd-numbered structure, particularly one with more than five options, has been demonstrated to enhance both the reliability and validity of data collection instruments. A neutral midpoint, inherent in odd-numbered scales, allows respondents to express neutrality without being forced into polarized choices, reducing potential response bias Kusmaryono, Wijayanti, and Maharani (2022).

Furthermore, scales with broader response ranges, such as 7 or 9 points, offer finer granularity in capturing attitudes or perceptions. This increased granularity helps in identifying subtle differences in responses, leading to more precise and nuanced data analysis. Studies have also shown that scales with a higher number of points tend to yield higher reliability coefficients, enhancing the consistency of the instrument Preston and Colman (2000); Warmbrod (2014). As a result, a 1-9 scale optimally balances the need for detailed responses with the mitigation of central tendency bias and offers a robust framework for psychological and educational assessments.

4.2.4 Temperature

The choice of temperature 0.5 in our application balances coherence and diversity in text generation for analyzing debate transcripts. Temperature in GPT models controls output randomness. Higher temperatures (e.g., 0.7) produce more creative responses, while lower values (e.g., 0.2) give more focused outputs OpenAI (2023).

For analyzing debates and scoring participants, temperature 0.4 helps the model generate responses that are consistent with the data yet flexible enough for nuanced interactions. This balance is important for identifying speakers, categorizing interactions, and scoring engagement. It avoids oversimplifying complex dialogues while reducing irrelevant randomness. Temperature 0.4 thus supports precise and adaptable analysis, meeting the need for both structured scoring and capturing interaction dynamics OpenAI (2023).

4.3 Parameters

To study how gender interactions are in debates, we examine four types of interactions: females interacting with females (F to F), males interacting with males (M to M), females interacting with males (F to M), and males interacting with females (M to F). An interaction refers to one person addressing a statement to another, which is why we distinguish between F to M and M to F interactions. This approach allows us to systematically analyze how people interact within and across gender groups, providing deeper insight into gendered communication patterns during debates.

4.4 Categories

For each parameter, we analyze the interactions across three categories: Argument Clarity and Relevance, Engagement Style and Use of Power. The reason for focusing on these categories is drawn from Carol Kinsey Goman's study Goman (2016), which highlights distinct communication strengths and weaknesses for both males and females. Goman emphasizes how gender differences in communication styles can impact clarity, engagement, and the balance of power in interactions.

4.4.1 Provide the API model with examples

For each category, we provided concrete examples of what constitutes negative, neutral, and positive scoring. These examples are (FROM DEBATTEN OR GENERAK REFERENCES????), ensuring that the API model has clear benchmarks for assigning scores across categories. This structured approach gives the model the guidance it needs to produce consistent and interpretable results.

4.4.2 Argument Clarity and Relevance

(!! WE ALSO NEED TO CHANGE THE SCALE HERE IF WE GO FOR 1-9 INSTEAD OF 1-7 !!) In the category of Argument Clarity and Relevance, we evaluates how clearly and logically arguments are presented and how closely they align with the topic. From Goman's study, one identified weakness for women is a tendency to meander, while a strength for men is being direct and to the point. By focusing on this category, we can examine whether these patterns are consistent when individuals interact with the same or opposite gender in debates. The scoring for this category is as follows: arguments that are vague, poorly structured, or lack relevance are rated in the lower range (1-3). Arguments that are moderately clear, generally relevant, and have a basic logical flow receive a mid-level score (4). The highest scores (5-7) are given to arguments that are exceptionally clear, well-structured, highly relevant, and supported by strong evidence or examples.

4.4.3 Engagement Style

For the category of Engagement Style, we assess how participants interact with their opponent's arguments and respond with counterarguments. According to Goman's study, women tend to focus on collaboration, while men often adopt a more assertive approach. This category allows us to explore whether these tendencies are evident in debates and how they influence interactions with the same or opposite gender. Scoring for this category focuses on the quality of engagement: limited or irrelevant engagement with weak counterarguments is rated in the lower range (1-3). Basic engagement with main points, providing adequate but shallow counterarguments, falls in the mid-range (4). The highest scores (5-7) are reserved for participants who show excellent engagement, clearly identifying weaknesses in their opponent's arguments and delivering strong, well-structured counterarguments.

4.4.4 Power Use

For the category of Power Use, we evaluate how participants manage conversational control and balance in their interactions. This includes examining behaviors such as interruptions, acknowledgment of others' points, word choice, and the equitable distribution of speaking time. According to Goman, men often assert dominance by taking credit, interrupting frequently, and using authoritative language, while women foster inclusivity through

collaboration, building on ideas, and promoting harmony. This category enables us to explore whether these patterns are reflected in debates and how they influence interactions with the same or opposite gender. Scoring for Power Use evaluates conversational balance and respect. Low scores (1-3) reflect frequent interruptions, dismissive behavior, or dominance. Mid-level scores (4) indicate occasional power displays with some balance. High scores (5-7) are given for respectful, balanced participation, with acknowledgment of others and minimal interruptions

4.5 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
"""
```

4.6 User Prompt for Interaction Analysis

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

!! CHANGE THE SCALE SO IT ALIGNS WITH THE REST OF THE TASK !!

```
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 (1-3):
        "Unclear or illogical arguments. Poor relevance to the topic."
        Example: "We should focus on space exploration because I once saw a shooting star.
             It was pretty."
    Neutral (4):
        "Moderately clear arguments with basic logical structure. Generally relevant to
            the topic."
        Example: "Space exploration is important for scientific advancement. It might lead
             to new technologies we can use on Earth."
    Positive (5-7):
        "Clear, logically structured arguments. Highly relevant to the topic with strong
            supporting evidence."
        Example: "Space exploration is crucial for three reasons: First, it drives
            technological innovation, as evidenced by NASA's development of memory foam
            and scratch-resistant lenses. Second, it provides valuable data on climate
            change through satellite observations. Third, it inspires future generations
            of scientists, as shown by the 40% increase in STEM applications following the
             Mars rover landing."

- Engagement Style:
    Negative (1-3):
        "Limited or no engagement with opponent's arguments. Weak or irrelevant
            counterarguments."
        Example: "I disagree with everything you said. My policy is better because it just
             is."
    Neutral (4):
        "Basic engagement with main points. Adequate counterarguments, but lacking depth
            or consistency."
        Example: "You mentioned economic impact, which is important. However, I think my
            plan would be more effective because it focuses on long-term growth."
```

```
    Positive (5-7):
        "Excellent engagement with opponent's arguments. Precise identification of
            weaknesses. Strong, well-formulated counterarguments."
        Example: "Your point about job creation is valid, but it overlooks two critical
            factors. First, the data from the Smith study shows that similar policies have
             led to short-term gains but long-term job losses. Second, it fails to account
             for the technological shifts outlined in the Jones report, which suggests a
            different approach is needed."

- Power Use:
    Negative (1-3):
        "Frequent interruptions, dismissive tone, or lack of acknowledgment of others'
            points. Displays conversational dominance with limited balance in discussion
            flow."
        Example:
        Person A: "I think we need to allocate more funds to healthcare."
        Person B: (interrupting) "No, that's completely wrong. Let's move on to the next
            topic."
    Neutral (4):
        "Moderate power displays. Occasional interruptions or longer speaking turns, but
            overall balanced participation without being overly dismissive."
        Example:
        Person A: "This policy could increase efficiency in transportation."
        Person B: "You make a good point, but I think focusing on public transport would
            be better. Let me elaborate on why that approach works in urban areas."
    Positive (5-7):
        "Demonstrates respectful and balanced use of conversational power. Rare
            interruptions and active acknowledgment of others' contributions. Shares the
            discussion space equitably."
        Example:
        Person A: "Investing in education is crucial for long-term economic growth."
        Person B: "I completely agree with your point about education being essential.
            However,  I d  like to add that we also need to consider vocational training
            to address immediate workforce gaps. What are your thoughts on combining these
             approaches?"

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 1-7, where 1 is the lowest and 7 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>",
            "Tone Dynamics": <number>,
            "Summary Tone Dynamics": "<summary>",
            "Power Dynamics": <number>,
            "Summary Power Dynamics": "<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.
```

```
"""
```

5 Results

5.1 Final scoring

The total scores indicate that interactions between females (F_to_F) have the highest average total score, with a mean value of 6.58. This highlights that female-to-female interactions generally achieve higher scores compared to the other interaction types assessed.

Male-to-male (M_to_M) interactions rank second, with an average total score of 6.12. Although slightly lower than F_to_F interactions, this score demonstrates that male-to-male exchanges also maintain a relatively high level of performance based on the metrics evaluated.

Mixed-gender interactions, specifically female-to-male (F_to_M) and male-to-female (M_to_F), exhibit closely aligned average scores, both recorded at 6.05. While slightly lower than the scores for the same-gender interactions, these scores indicate a comparable level of engagement and interaction quality across mixed-gender exchanges. Together, these findings provide a comprehensive view of how interaction types perform on average, with clear distinctions in total scores across the categories.

5.2 Regression Analysis

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

Property	**Argument Strength**	**Engagement Style**	**Power Dynamics**
Coefficients			
Intercept	3.315 (0.949)	0.599 (0.866)	1.233 (0.725)
Argument Strength	-	0.146 (0.130)	0.160 (0.111)
Engagement Style	0.231 (0.206)	-	0.576 (0.107)
Power Dynamics	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 1: Regression results for Argument Strength, Engagement Style, and Power Dynamics

Same-gender pairs scored slightly higher in Argument Strength by 0.174 points compared to mixed-gender pairs. The difference is small and not statistically significant. The model accounts for 31.1% of the variation in Argument Strength.

In Engagement Style, same-gender pairs scored slightly lower by -0.071 points compared to mixed-gender pairs. This difference is also small and not statistically significant. The model explains 58.0% of the variation in Engagement Style.

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

The differences between same-gender and mixed-gender pairs are small for all three aspects, and none of these differences are statistically significant. The Power Dynamics model fits the data best, followed by Engagement Style, and then Argument Strength. There are interesting relationships between the three aspects; for example, stronger Power Dynamics are linked to stronger Arguments and more engaging styles.

5.3 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). This provides an understanding of how the episode dynamics may be influenced by participant groupings.

Gender Distribution per Episode: We propose using a stacked bar chart to display the proportion of women and men in each episode. This clearly visualizes variations in gender balance across episodes, with the total proportion set to 100% for each episode. Each bar represents one episode, with segments indicating the percentage of women (e.g., 40%) and men (e.g., 60%).

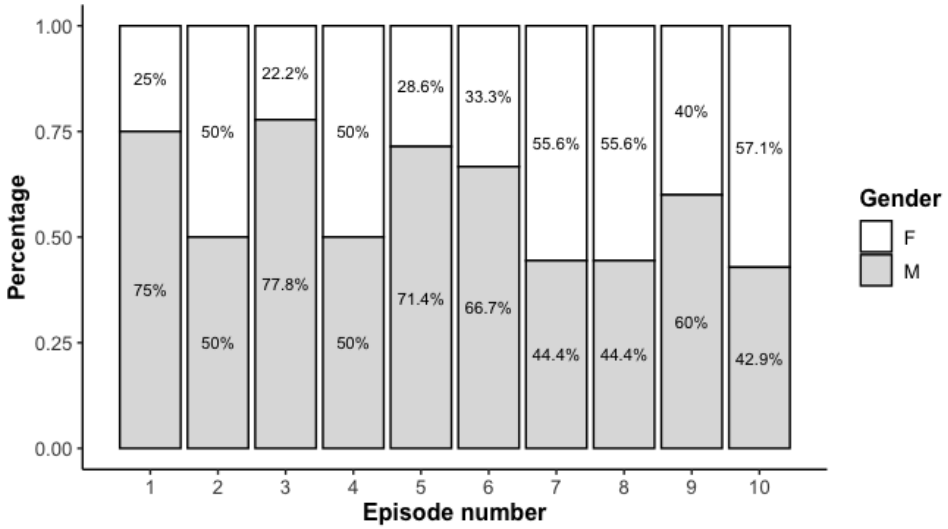

Figure 2: Gender distribution per episode

Role Distribution per Episode: To display the number of debaters and experts in each episode, we recommend a non-stacked bar chart. This provides an absolute overview of how many participants from each role are present in each episode. Each bar represents the total number of participants in an episode, divided into debaters and experts.

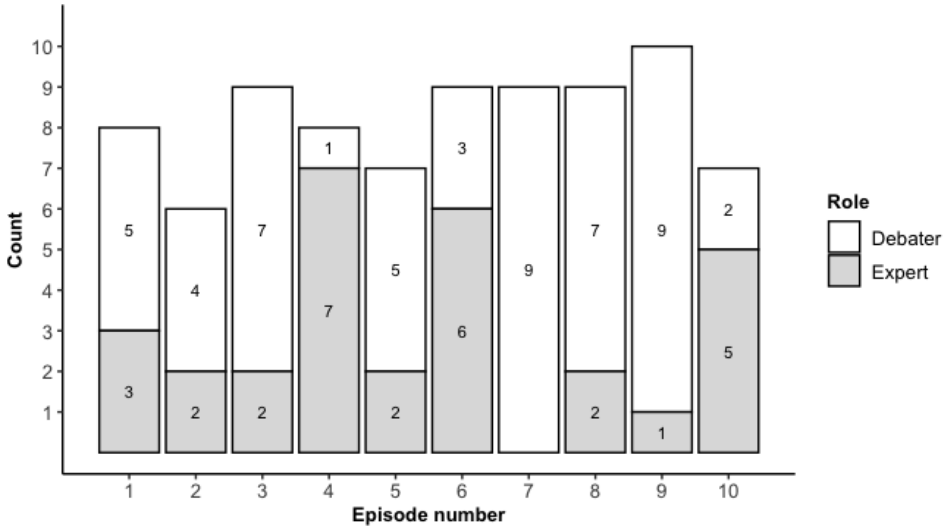

Figure 3: Role distribution per episode

What the Graphs Show:

- **Gender Distribution:** The graph illustrates whether certain episodes have a predominance of male or female participants, potentially indicating an imbalance in interaction dynamics.
- **Role Distribution:** The role distribution graph highlights how the episodes are structured in terms of expertise versus debate perspectives, and whether certain episodes are dominated by experts or debaters.

5.4 Average Scores per Interaction Type

To analyze how different interaction types perform across dimensions, we propose creating a graph that shows the average scores for each interaction type (F_to_F, F_to_M, M_to_F, M_to_M). This provides an overview of whether there are discernible patterns or differences in performance based on gender combinations in interactions.

Type of Graph: We recommend a bar chart where each bar represents one interaction type, and the height of the bar indicates the average total score (*total_score*) for that interaction. This makes it easy to compare the relative performance of the different types.

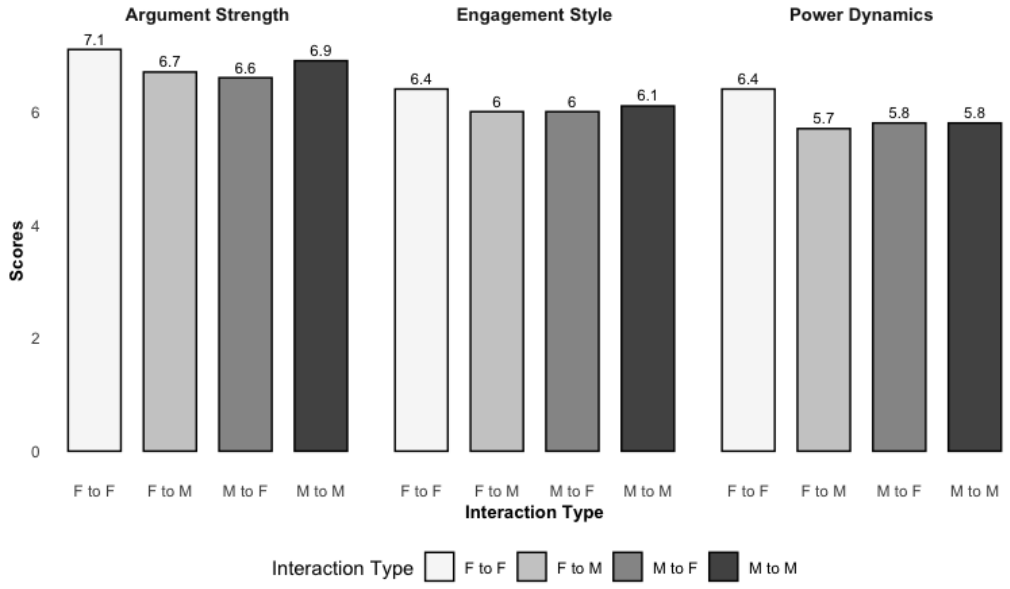

Figure 4: Average scores per Interaction Type

What the Graph Shows:

- The graph will provide a visual indication of which interaction types perform best across dimensions. For example, F_to_F interactions may have higher average total scores compared to other types, potentially reflecting stronger collaboration and engagement among female participants.
- The graph may also reveal whether mixed-gender interactions (F_to_M and M_to_F) show more consistent or less varied performance compared to same-gender interactions.

The graph highlights significant differences or similarities between interaction types and should be included in the discussion section to support interpretations of how gender affects interactions in debates. It can be placed in a section analyzing interaction patterns or performance.

5.5 Small and Non-Significant Differences Between Same- and Mixed-Gender Configurations

Regression analysis shows that differences between same- and mixed-gender configurations, represented by the coefficient for "Same gender," are small and not statistically significant across all three dimensions: *Argument Strength*, *Engagement Style*, and *Power Dynamics*. For instance, the coefficient for *Same gender* is 0.174 for *Argument Strength*, indicating a weak positive relationship but insufficient significance to draw definitive conclusions. This suggests that gender alone is not a decisive factor in the dynamics of political debates. Instead, individual attributes such as rhetorical skills, experience, and knowledge may play a larger role. The results show that gender itself does not create marked differences, influencing dynamics to a lesser degree than previously assumed.

Graph Placement: The graph *gender_differences* can be placed here to illustrate the small differences between same- and mixed-gender configurations for the three dimensions. The graph should be a bar plot displaying the coefficients with error bars for *Argument Strength*, *Engagement Style*, and *Power Dynamics*.

5.6 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 Dynamics* 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 *Argument Strength* 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 strength of arguments. The low explanatory power for *Argument Strength* 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.

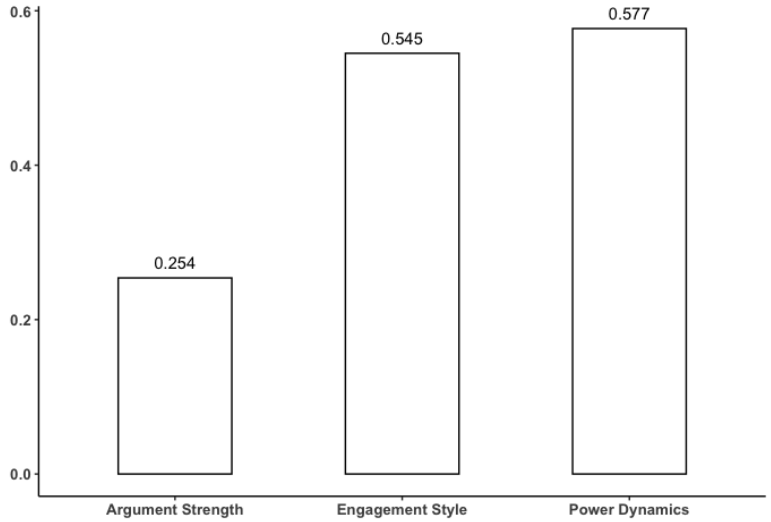

Figure 5: Role distribution per episode

Graph Placement: The graph *r2_ explanatory_ power* can be included here to compare the adjusted R^2 values for the three models. The graph could be a bar chart with each dimension represented and clearly labeled.

5.7 Positive Correlations Between Categories

The regression results indicate positive correlations between dimensions. For example, increased *Power Dynamics* is positively correlated with stronger *Argument Strength* (coefficient = 0.339). Similarly, the model shows a strong correlation between *Engagement Style* and *Power Dynamics* (coefficient = 0.771). 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 dynamics, where power balance and engagement not only contribute to better interaction but also support the quality of argumentation.

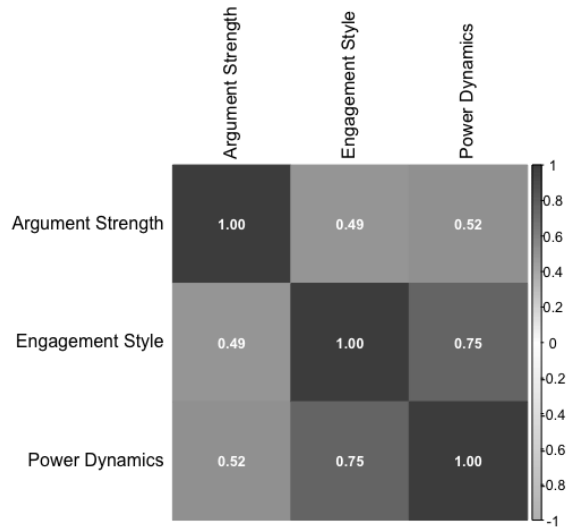

Figure 6: Correlation between the categories

Graph Placement: The graph *correlations* can be placed here to show how the three dimensions relate to

each other. The graph could be a correlation map or scatter plot with trend lines illustrating the positive relationships between variables.

5.8 Marginal Differences Between Same- and Mixed-Gender Configurations

The results indicate marginally higher scores for same-gender configurations in both *Argument Strength* (0.174) and *Power Dynamics* (0.150). Although these differences are not significant, they may suggest that same-gender configurations provide a slight advantage, possibly due to a sense of increased comfort and trust among participants. However, the lack of statistical significance emphasizes that gender alone does not have a strong effect, and other factors likely play a more important role.

Graph Placement: The graph *interaction_ types* can be placed here to show the small differences between same-gender and mixed-gender configurations. The graph could be a bar chart showing average scores for *Argument Strength* and *Power Dynamics* by interaction type.

Start with the main result. Do not do warmup exercises, extensive data description (especially of well-known datasets), preliminary estimates, replication of others' work. Do not motivate the specification that worked with all your failures. If any of this is really important, it can come afterwards or in an appendix.

Present Findings: Report your findings in a clear and logical manner. Use subheadings if necessary to organize results.

Use of Visual Aids: Include figures, tables, and graphs to illustrate your results. Ensure they are clearly labeled and referenced in the text.

Interpretation: Discuss what the results mean in relation to your research question. Highlight any patterns, trends, or significant observations.

Regression, R-test, standard deviation

6 Conclusion

6.1 Summary of Findings

This study examined how interaction patterns differ between same-gender and mixed-gender pairs in political debates, analyzing 10 episodes of the NRK program *Debatten*. Using an automated workflow and GPT-4mini for text analysis, the study evaluated four interaction types: female-to-female (F_to_F), male-to-male (M_to_M), female-to-male (F_to_M), and male-to-female (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 *Argument Strength* (0.174 points higher) and *Power Dynamics* (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 *Argument Strength* (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-4mini enabled detailed scoring and analysis of interaction dynamics. 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 dynamics 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 expanding the dataset to include more episodes and broader debate formats. Incorporating qualitative methods, such as discourse analysis or audience feedback studies, could reveal deeper insights into the dynamics of interaction. Exploring additional variables, such as rhetorical strategies, cultural differences, and audience reception, would help contextualize the findings further.

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 dynamics could offer practical insights into how interaction styles influence public opinion and debate outcomes.

6.4 Closing Remarks

This study demonstrates that while gender configurations have a limited direct impact on political debate performance, interaction qualities such as *Power Use* and *Engagement Style* are central to effective communication. By leveraging innovative methodologies and automated tools, this research provides a foundation for further exploration of gender dynamics in public discourse. Understanding these patterns is essential for fostering more inclusive and impactful communication in high-stakes settings, paving the way for future studies that build on these findings.

Summary of Findings: Recap the main findings of your research succinctly.

Strengths and Limitations: Reflect on the strengths of your study and acknowledge any limitations or challenges faced.

Future Research: Suggest areas for further investigation or how the research could be expanded or improved.

7 References

According to (Goman, 2016), understanding how communication is influenced by gender is essential.

References

Goman, C. K. (2016). *Is your communication style dictated by your gender.* Retrieved 2024-10-31, from https://www.fiskeridir.no/Akvakultur

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

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