Enhanced Multimodal Emotion Lines Dataset (E-MELD)
Team 5

Guy Elovici[1*] Noa Magrisso[2†] Amit Avitan[3‡] Omer Idgar[4§]

1. Introduction

Understanding group conversation dynamics is a critical area of study in computational linguistics. Existing datasets like MELD [2] primarily focus on emotions and sentiments but lack information about the roles participants play within conversations. This limitation hinders the ability to analyze interactions beyond the surface level of what is being said, restricting insights into how participants shape and influence group discussions.

Our project focuses on expanding conversational datasets by introducing a **speaker role** feature to address this gap. We rely on the role coding scheme proposed in [6], which classifies speakers into distinct emergent roles based on their behaviors and contributions. This scheme assumes that while a speaker's role may shift over time, their role remains relatively stable within short conversational windows. Each role reflects specific actions and behaviors, defined as follows:

- **Protagonist:** A speaker who dominates the conversation, drives the discussion, asserts authority, and takes a personal perspective.
- **Supporter:** A cooperative speaker who encourages, demonstrates acceptance and offers both technical and relational support.
- **Neutral:** A passive participant who accepts ideas from others without actively contributing to or challenging the discussion.
- **Gatekeeper:** A speaker who moderates and facilitates communication, ensuring the conversation flows smoothly and inclusively.
- **Attacker:** A speaker who challenges others, expresses disapproval and undermines the contributions of fellow participants.

*Email: guyelov@post.bgu.ac.il; ID: 208726935
†Email: noamagri@post.bgu.ac.il; ID: 206934978
‡Email: avamit@post.bgu.ac.il; ID: 211425426
§Email: idgaro@post.bgu.ac.il; ID: 323040758

By enriching datasets with these speaker roles, we aim to enable a more nuanced analysis of group conversations. This enhancement provides insights into not only the content of speech but also the social structure and interaction dynamics that govern group communication. This contribution supports advanced conversational modeling and improves the interpretative capabilities of AI models in understanding complex social behaviors.

1.1. Limitations of Naive LLMs in Conversational Analysis

The main limitation we assume of using Naive LLMs for this approach is that Naive LLMs analyze speaker roles based only on text but fail to account for crucial social dynamics like relationships and centrality, leading to oversimplified interpretations.

Conversation Example:

- **Rachel:** "I'm planning to quit my job and move to another city."
- **Monica:** "Are you sure that's the best decision?"
- **Phoebe:** "If that's what feels right, go for it."

Naive Approach (Without Social Context): Prompt to LLM: "*Given the following conversation, identify the role of each speaker.*"

LLM Output: *Rachel: Protagonist (initiating discussion). Monica: Attacker (criticizing Rachel's decision). Phoebe: Supporter (offering encouragement).*

Enhanced Approach (With Social Context): Prompt to LLM: *"Rachel and Monica are best friends. Rachel has high centrality in this social network. Monica is highly supportive of Rachel but sometimes raises concerns out of care. Using this information, identify the role of each speaker in the following conversation."*

LLM Output: *Rachel: Protagonist (leading the discussion). Monica: Supporter (expressing concern rooted in care for Rachel's well-being). Phoebe: Neutral (passively affirming Rachel's autonomy).*

The main difference is that naive LLM Relies only on the text, misinterpreting Monica's role as critical or negative, while the enhanced model leverages social context to correctly identify Monica's role as supportive and concerned.

2. Related Work

Speaker role classification has garnered increasing attention in recent years, reflecting broader trends in computational sociolinguistics and dialogue system design. Early work often relied on handcrafted features, such as turn-taking patterns and politeness strategies [3]. More recent efforts leverage deep learning models trained on conversation corpora, applying context-aware architectures [7]. Techniques from dialogue act classification [5] and conversation disentanglement [1] similarly underscore the importance of contextual cues—such as speaker intent, turn-taking, and emotional tone—to accurately recognize roles like "leader," "follower," or "supporter" within group interactions. Additionally, work on multi-participant chat dialogues has shown that including metadata (e.g., speaker's background, speaking time, and emotional state) can significantly boost performance in role or stance detection[4]. Our approach adds to this literature by integrating comprehensive dialog-level and metadata-driven features, demonstrating their complementary benefits even when speaker identities are anonymized.

3. Dataset: MELD

The dataset used in this study is the Multimodal EmotionLines Dataset (MELD) [2]. MELD contains 13,708 dialogues and 137,000 utterances annotated with seven emotion categories (e.g., anger, joy, sadness) and three sentiment categories (positive, neutral, negative). The dataset includes multimodal annotations spanning text, audio, and visual signals, enabling comprehensive analysis of multi-party conversations. By preserving turn-taking structures and speaker information, MELD supports the study of linguistic and vocal interaction patterns.

3.1. Dialogue Length Distribution

We analyzed the distribution of dialogue lengths in MELD to understand its structure. Figure 1 shows that most dialogues range between 5 and 14 utterances, with a median length of 9 and a maximum of 24. This variability in dialogue lengths highlights the dataset's diversity and supports its suitability for studying multi-party conversations.

4. Proposed Method

Our method enhances MELD by incorporating speaker roles and constructing dialogue-specific social context, resulting in the E-MELD dataset (as shown in Figure 2). The approach consists of three key steps:

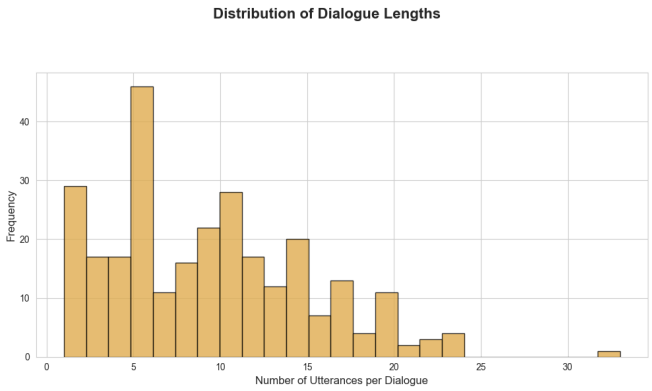

Figure 1. Histogram of dialogue lengths in MELD.

1. **Role Definition and Feature Extraction**: Speaker roles are defined using a predefined scheme [6] (e.g., protagonist, supporter), and extracted based on conversational patterns. Social context is collected for each pair of speakers in the dialogue. Metrics such as exchange frequency, emotional alignment, and duration of response are derived to quantify conversational dynamics.

2. **Role Assignment Using LLMs**: A fine-tuned large language model (LLM) assigns speaker roles by processing textual excerpts and speaker context. Prompts are iteratively refined, and human annotators validate a subset of assignments to ensure alignment with the defined roles.

3. **Dataset Integration and Validation**: Validated role annotations are integrated into the E-MELD dataset. This enriched dataset captures nuanced speaker roles and interaction patterns, supporting advanced analysis of group conversations.

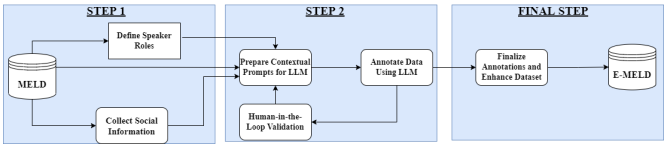

Figure 2. Reminder of our proposed pipeline for enhancing MELD with speaker roles.

5. Approaches Examined

To evaluate the effectiveness of speaker role classification, we tested three approaches, each leveraging different levels of context and metadata. The design of prompts tailored to these approaches was critical for guiding the model's understanding of the task.

5.1. Approach Descriptions

1. **Sentence-Only Approach (Naive Baseline)**: This method classifies sentences independently, using only the content of the sentence. It provides no dialogue-level context or metadata, serving as a simple baseline for comparison.

2. **Dialogue-Aware Approach (Full-Dialogue Baseline)**: Here, each sentence is classified using the full dialogue context, enabling the model to capture conversational dynamics and interactions between speakers.

3. **Context-Enriched Dialogue-Aware Approach (Contextual Baseline)**: This builds on the dialogue-aware approach by including metadata to enhance context. Examples include:

    - **Response Duration**: Average time between responses.
    - **Word and Letter Counts**: Indicators of communication style.
    - **Sentiments and Emotions**: Commonly expressed sentiments and emotions (e.g., joy, sadness).

    This enriched context provides nuanced insights into speaker behavior, improving classification accuracy.

5.2. Prompt Design

Prompts were designed to progressively increase task complexity, corresponding to the level of context used. Each prompt began with a shared introduction defining the task and speaker roles:

> *"You are an expert in analyzing conversations and assigning speaker roles. The following dialogue is taken from various contexts. Your task is to assign roles to each speaker based on their utterances and their role in the overall conversation."*

Dialogue Context: For dialogue-aware approaches, the entire dialogue context was included, helping the model understand conversational flow and speaker interactions:

> *"Here is the context of the dialogue:*
> *Sr No. 1, Mark (2.25s): 'Why do all your coffee mugs have numbers on the bottom?'*
> *Sr No. 2, Rachel (6.76s): 'Oh, that's so Monica can keep track. She'll say, 'Where's number 27?!'""*

Contextual Enrichment: The contextually enriched approach added metadata summaries, providing details on interactions and overall communication patterns:

> *"Interaction summary for Rachel and Mark:*
> *Avg Response Duration: 1.87s*
> *Sentiments: negative (40%), positive (40%)..."*

Concluding Instructions: Prompts concluded with standardized output instructions to ensure clarity and consistency:

> *"Identify the role and provide justifications for each speaker. Include 'Sr No.', 'Speaker', 'Role,' 'Justification,' and the length of each dialogue utterance."*

This structured prompt design—from a simple sentence-based baseline to dialogue-aware and context-enriched methods—equipped the model to handle increasingly complex speaker role classification tasks.

5.3. Speaker Hashing in Role Classification

To assess potential bias toward speaker identities in role classification, we implemented a hashed speaker approach, replacing actual names with generic labels (e.g., Speaker A, Speaker B). This anonymization preserves dialogue context while preventing reliance on speaker-specific patterns.

Applied to the Full-Dialogue Baseline (Approach 2) and Contextual Baseline (Approach 3), this method tests the impact of removing speaker identity. Approach 2 is expected to be more affected, as it lacks metadata to compensate for missing names, whereas Approach 3 is anticipated to remain robust, leveraging contextual cues like response durations and sentiment.

This approach isolates contextual influence from speaker identity, examining whether role classification is driven by conversational behavior rather than associations with individual speakers.

5.4. Evaluation Methods

Our evaluation focuses on verifying the accuracy and consistency of speaker role assignments through both quantitative metrics and a human annotation task. The role assignment accuracy will be assessed by comparing LLM-assigned roles with a gold-standard subset, using metrics such as accuracy, precision, recall, F1-score, and inter-annotator agreement (e.g., Cohen's Kappa). These metrics will also be applied to evaluate the LLM's speaker role predictions against our annotations.

The task of manual annotations will involve four annotators (the project members) and will focus on the test set, which is not used during the prompt engineering stage.

The annotation process will provide a human benchmark for role assignments, enabling a direct comparison to determine whether our LLM surpasses human performance in accurately identifying speaker roles. By incorporating these evaluation metrics, we aim to rigorously assess the quality of the LLM's predictions relative to human annotations.

5.5. Results

To assess the classification accuracy of speaker roles, we compared the results of the three approaches against our manual annotations, which serve as the gold standard for speaker role identification. We anticipated that the third approach—the Contextual Enriched Dialogue-Aware Approach—would best capture the sentence's context.

5.5.1 Comparative Analysis of Results

Inter-Annotator Agreement (Kappa Score Analysis) Before evaluating the model predictions, we assess the consistency of manual annotations using Cohen's Kappa Score, which measures inter-annotator agreement.

The obtained Kappa score of 0.613 suggests a moderate-to-substantial level of agreement among annotators. While this confirms that the majority of labels are reasonably reliable, the score also indicates that some subjectivity or disagreement exists in role assignments. This may arise from cases where speaker intent is ambiguous, requiring deeper context to determine the exact role. Given this level of agreement, we consider the majority annotations as a valid reference for evaluating model performance. However, some role labels might inherently contain variability, which could affect classification accuracy, especially in borderline cases where context is complex.

Model Performance Comparison We evaluated the alignment of three different approaches with the majority of our annotations. Additionally, we examined the two variations in which speaker names were hashed.

Figure 3 compares the performance of five classification approaches. The **Naive Baseline (Approach 1)**, which classifies sentences independently, performs the worst (**0.275**), highlighting its inability to capture conversational context. The **Full-Dialogue Baseline (Approach 2)**, incorporating full dialogue context, improves accuracy to **0.404**, demonstrating the benefit of considering conversational flow. However, the **Hashed version (Approach 2 Hashed)**, where speaker names are replaced with non-indicative labels, drops to **0.332**, suggesting that speaker identity contributes to role classification. The **Contextual Baseline (Approach 3)**, which integrates metadata like response duration and sentiment, achieves the highest accuracy (**0.514**), emphasizing the value of contextual cues. The **Hashed version (Approach 3 Hashed)** slightly lowers accuracy to

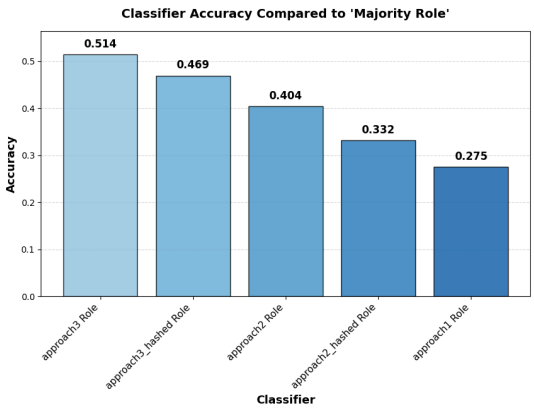

Figure 3. Approaches Predictions vs Majority of Annotators

0.469, indicating that while metadata helps, speaker identity still plays a role. These results confirm that Approach 3 (Contextual Baseline) without hashing is the most effective, demonstrating that combining dialogue context with speaker information and metadata leads to the best role classification performance.

Role Distribution by Emotion (Best Accuracy Approach) We present the distribution of speaker roles by emotion using the Contextual Baseline (Approach 3), as it achieved the highest accuracy (0.514). This analysis highlights how role classification aligns with emotional expression when leveraging metadata and dialogue context.

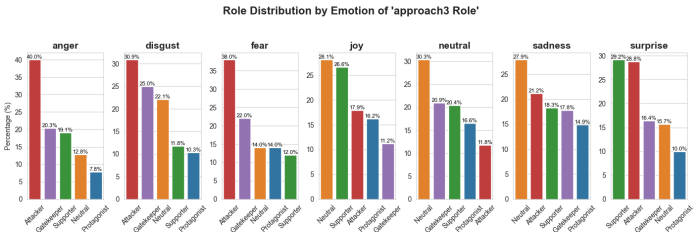

Figure 4. Role (3rd Approach) Distribution by Emotion

The Contextual Enriched Dialogue-Aware Approach effectively captures role distributions across emotions (Figure 4).The distribution of negative emotions reveals a strong association with the Attacker role, particularly in expressions of anger, fear, and disgust. This pattern suggests that individuals classified as attackers tend to engage in emotionally charged, confrontational discourse, reinforcing their role in adversarial interactions. While sadness is not as strongly concentrated in a single role as anger or fear, Neutral and Attacker roles show a slightly higher presence, suggesting that sadness is more broadly distributed but not exclusively linked to confrontational roles.

In contrast, the distribution of surprise is notable, as both

supporters and attackers display a heightened presence. This aligns with the nature of surprise as an emotion that can arise in both positive and negative contexts—supporters may express surprise in response to unexpected favorable developments, whereas attackers might react strongly to unforeseen challenges or threats. These findings highlight the intricate relationship between role classification and emotional expression, emphasizing the significance of contextual factors in discourse analysis.

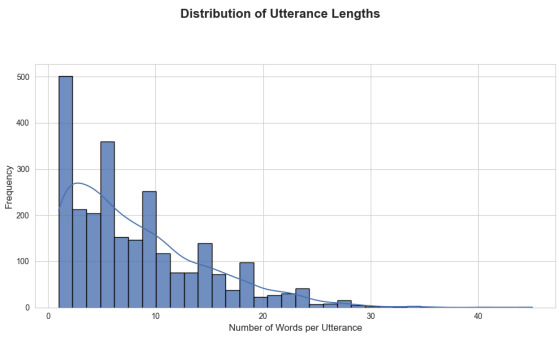

Figure 5. Utterance Length Distribution

Influence of Utterance Length on Accuracy **Utterance Length Distribution** The distribution of utterance lengths, as depicted in Figure 5, indicates that the majority of utterances in the dataset are short, with the highest frequency occurring between 1–6 words. As the length increases, the frequency of utterances decreases significantly. This skewed distribution highlights the predominance of brief conversational exchanges in the dataset, reflecting natural patterns in group dialogues where shorter utterances are more common than longer, detailed responses.

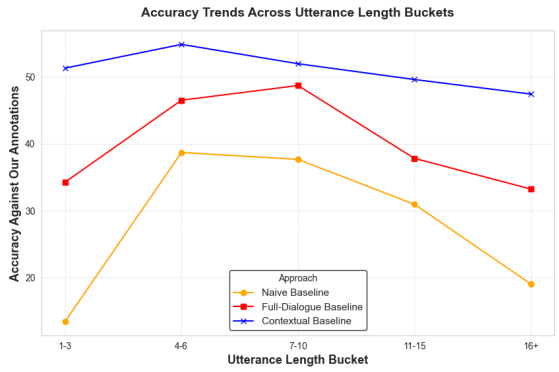

Figure 6. Accuracy Across Utterance Lengths

Accuracy Across Utterance Lengths Figure 6 evaluates the role classification accuracy of the three approaches across different utterance length buckets. The results reveal that for short utterances (1–6 words), accuracy is generally lower across all approaches, particularly for the Naive Baseline, which struggles to correctly classify speaker roles when relying solely on individual sentences. This limitation underscores the challenge of inferring role dynamics without additional dialogue context.

However, there is a clear improvement from the Naive Baseline to the Full-Dialogue Baseline and then to the Contextual Baseline. While the Naive Baseline achieves low accuracy for short utterances, the Full-Dialogue Baseline benefits from considering the broader conversation, leading to a noticeable performance gain. The Contextual Baseline further enhances accuracy, demonstrating its ability to correctly classify roles even when the utterance alone provides minimal information.

A clear performance progression is observed from the Naive Baseline to the Full-Dialogue Baseline and then to the Contextual Baseline, highlighting the impact of incorporating dialogue structure and contextual cues. Additionally, while all approaches experience a decline for long and complex utterances (16+ words), the Contextual Baseline remains the most stable across varying utterance lengths, further reinforcing the value of integrating broader conversational context.

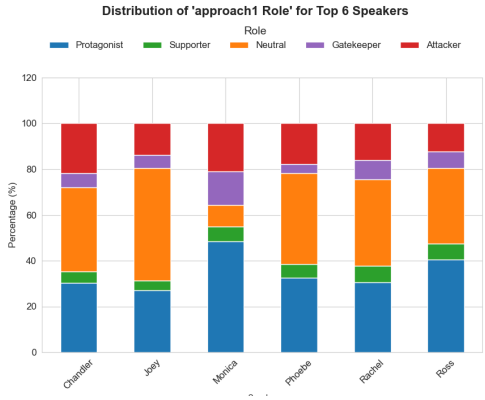

Figure 7. Role Distribution in the Naive Baseline (Approach 1)

Speaker Corr - Change name Figure 7 presents the role classification distribution for the Naive Baseline (Approach 1). In this approach, Protagonist and Neutral roles are overrepresented due to the model's inability to incorporate conversational context. Without dialogue awareness, many utterances are misclassified as Protagonist, as they appear to introduce new topics, or as Neutral, reflecting the model's limited understanding of role dynamics.

Figure 8 compares the Full-Dialogue Baseline (Approach 2) and the Contextual Baseline (Approach 3), both with and without hashed speaker names. In Approach 2, role distributions shift when speaker identities are removed, indicating a reliance on speaker-specific patterns. In contrast, Approach 3 remains stable across hashed and non-

Figure 8. Role Distribution Across Full-Dialogue and Contextual Baselines (Approaches 2 & 3, Regular vs. Hashed)

hashed versions, demonstrating that contextual information alone is sufficient for accurate classification. Overall, Approach 2 reduces the randomness of Approach 1, while Approach 3 further refines role assignments, achieving structured and consistent predictions even in anonymized settings.

5.5.2 Case Study: Phoebe's Utterance Analysis

Figure 9 presents Dialogue 33, which involves three different speakers, with the corresponding Speaker Role classifications produced by each approach for each utterance, along with our manual annotations. The utterances in this dialogue are numbered Sr. 330-340.

The analysis of the dialogue in this figure reveals that, while some utterances were consistently classified across all approaches (highlighted in green), discrepancies emerged in several instances. Notably, the Contextual Baseline most closely aligned with our manual annotations, correctly capturing subtle shifts in speaker roles, except for a single misclassification.

Speaker	Utterance	Naive Baseline	Full-Dialogue Baseline	Contextual Baseline	Our Annotation
Phoebe	Well, alright, we already tried feeding her, changing her, burping her, oh try this one!	Protagonist	Protagonist	Protagonist	Protagonist
Phoebe	Go back in time and listen to Phoebe!	Protagonist	Neutral	Attacker	Attacker
Monica	Alright here's something, it says to try holding the baby close to your body and then swing her rapidly from side to side.	Protagonist	Supporter	Gatekeeper	Gatekeeper
Rachel	Ok.	Neutral	Neutral	Neutral	Neutral
Monica	It worked!	Protagonist	Supporter	Protagonist	Supporter
Rachel	Oh oh no just stopped to throw up a little bit.	Neutral	Attacker	Attacker	Attacker
Rachel	Oh come on, what am I gonna do, its been hours and it won't stop crying.	Protagonist	Attacker	Attacker	Attacker
Monica	Umm, she Rach, not it, she.	Gatekeeper	Gatekeeper	Gatekeeper	Gatekeeper
Rachel	Yeah, I'm not so sure.	Neutral	Neutral	Neutral	Neutral
Monica	Oh my god, I am losing my mind.	Protagonist	Attacker	Attacker	Attacker

Figure 9. Speaker role classifications in Dialogue 33 across different approaches.

A detailed examination of Phoebe's statement, "Go back in time and listen to Phoebe!" (Sr. 331), underscores the strengths of the Contextual Baseline. Video analysis of this scene reveals Phoebe delivering the line in a forceful tone, indicative of an "Attacker" role. The Naive Baseline misclassified this as "Protagonist," reflecting its general bias toward classifying assertive utterances as such, justified by the reasoning that Phoebe initiates and leads the conversation. The Full-Dialogue Baseline labeled it "Neutral," suggesting the comment lacked significant conversational impact. Only the Contextual Baseline accurately identified the adversarial tone, demonstrating its effectiveness in integrating social and sub-contextual cues, supported by the justification that Phoebe's overall communication style includes emotionally charged responses.

5.5.3 Case Study: Video Model Analysis

Figure 10. An example of output of a LLM which received both the prompt and the correspond MP4 video of the conversation.

We wanted to explore further Dialogue 33 by examining whether incorporating video data would allow a model (in this case, the LLaVA-Video model) to more accurately interpret and annotate social speaker roles—potentially offering richer insights than text alone. The hypothesis was that the visual cues (body language, facial expressions, etc.) might help the model discern interpersonal dynamics more effectively. From the model's output(10, we can see that it describes the two individuals in the scene with the roles of a "protagonist" (Phoebe) and a "supporter" (Monica). In contrast, our annotations for those same individuals as having the roles of "attacker" and "gatekeeper."

One possible reason for this discrepancy is that the model, which was provided with the video, inferred a cooperative dynamic—seeing one person offering suggestions and the other person responding positively—leading it to label them as "protagonist" and "supporter." By contrast, the model that received the full conversation and social context (without relying on video cues) was able to more accurately detect the underlying tension or conflict in their interaction, which aligns better with the "attacker" and "gatekeeper" annotation. This suggests that, in certain scenarios, textual and contextual cues from the conversation itself can provide more reliable indicators of social roles than video or surface-level observations alone.

5.5.4 Case Study: Hashed Speakers Analysis

To investigate potential bias towards speaker identities in our LLM-based speaker role classification, we implemented

a hashed speaker approach. We hypothesized that removing speaker names would negatively impact performance, particularly for the Dialogue-Aware approach, as speaker identification provides a valuable cue for tracking conversational turns and relationships. The Contextual Enriched Dialogue-Aware approach, however, was expected to be less susceptible to this change due to the presence of additional metadata (response durations, sentiment, emotion) which could partially compensate for the missing speaker information. Specifically, we anticipated a drop in Kappa score from approximately 0.4 to 0.32 for the Dialogue-Aware (Hashed) condition, and a smaller decrease from approximately 0.5 to 0.46 for the Contextual Enriched Dialogue-Aware (Hashed) condition. Our experimental results confirmed these hypotheses. The Dialogue-Aware approach exhibited a substantial performance decline, underscoring the importance of explicit speaker identification. While the Contextual Enriched Dialogue-Aware approach also showed a reduction in performance, the magnitude was smaller, suggesting that the supplementary contextual features mitigated the impact of anonymizing speaker identities. This finding highlights the benefits of incorporating diverse contextual information, as it not only enhances overall performance but also improves the model's robustness and reduces its dependence on potentially biased or superficial cues like speaker names. Furthermore, the differential impact of hashing on the two approaches provides evidence that the model is indeed learning to associate patterns of conversational behavior with specific roles, rather than simply relying on pre-existing knowledge about individual characters.

6. Conclusions and Discussion

This study explored enhancing speaker role classification in group conversations by leveraging contextual information and metadata within the MELD dataset. Our findings demonstrate the limitations of naive LLMs that rely solely on textual content, as they often misinterpret social dynamics and speaker roles. The Contextual Enriched Dialogue-Aware Approach, which incorporates dialogue-level context and metadata, significantly outperformed both the Sentence-Only and Dialogue-Aware approaches. This highlights the importance of integrating diverse contextual cues, such as response duration, sentiment, and emotion, for accurate speaker role identification.

The hashed speaker analysis further underscores the model's ability to learn complex conversational patterns beyond superficial cues like speaker names. While anonymizing speakers impacted performance, the Contextual Enriched Dialogue-Aware Approach exhibited greater robustness due to its reliance on richer contextual features. This suggests that the model effectively associates behavioral patterns with speaker roles, rather than relying on pre-existing knowledge about individual speakers.

However, our study also revealed challenges in classifying longer utterances, indicating the need for improved methods to capture extended contextual dependencies. Additionally, the case study involving video data suggests that relying solely on visual cues may not always provide the most accurate representation of social dynamics, as textual and contextual information can offer deeper insights into speaker roles.

Future research directions may include exploring advanced techniques like transformer-based models to address challenges in longer utterances and investigating the interplay between visual, textual, and contextual cues for a more comprehensive understanding of speaker roles in group conversations. Furthermore, expanding the dataset with more diverse social contexts and metadata could enhance the model's generalizability and robustness in real-world scenarios.

In conclusion, our study demonstrates the potential of leveraging contextual information and metadata to enhance speaker role classification in group conversations. By incorporating diverse cues and moving beyond naive LLMs, we can achieve a more nuanced understanding of social dynamics and speaker roles, paving the way for more sophisticated conversational AI models.

7. Reflection

Throughout this project, we discovered the complexities of speaker-role classification in multi-party dialogues. At the beginning, our attempts to load and run a LLM for speaker annotations from HuggingFace yielded outputs that were frequently noisy, inconsistent, and difficult to interpret or incorporate into downstream analyses. This lead us to us the Ollama framework, which offered a more controlled environment and yielded more coherent, stable annotations. Even with improved stability, refining prompts and metadata inputs remained a highly iterative process, revealing how sensitive classification results can be to subtle changes in context or feature engineering.

On the human side, manually labeling speaker roles proved both enlightening and challenging. The role definitions—ranging from "Protagonist" to "Attacker"—inevitably overlapped in certain scenarios, particularly when speakers displayed multiple conversational functions within a single utterance. Disagreements among annotators underscored the subjective elements in interpreting intentions and emotional tones. Nonetheless, these annotations served as a critical benchmark, guiding our algorithmic improvements and providing essential feedback for fine-tuning the modeling approach.

Ultimately, this project highlighted that no single strategy suffices for high-quality role classification. Instead, leveraging comprehensive dialogue context, carefully se-

lected metadata (such as average duration and sentiment frequency), and iterative model adaptation led to more consistent and reliable outputs. In this way, the limitations we encountered became stepping stones for methodological rigor and innovation.

8. Work Split

We divided the work so that each team member undertook core tasks without overlap (as shown in Table 1), while still collaborating on shared responsibilities.

1. **Omer** implemented the baseline and the second approach, led the hashed utterances scenario, and created prompts for these methods.

2. **Amit** developed the third (social context-enriched) approach, designed the annotation framework, and crafted prompts tailored to that approach.

3. **Guy** conducted the comprehensive dataset exploration, performed majority role-tagging evaluation, and summarized approach comparisons.

4. **Noa** collated and preprocessed the results of the the three examined approaches, carried out in-depth evaluation of the third approach, and performed additional case studies (including Phoebe's analysis and video-based LLM).

In addition, **Amit** and **Noa** collaboratively worked on the related work review and speaker-role types selection, while **Omer** and **Guy** investigated and implemented how to use LLMs locally via OLLaMA. Finally, all four of us collaborated on the speaker-role annotation to maintain consistent labeling throughout the project.

Task	**Done By**
Baseline and second approach implementation	Omer
Hashed utterances scenario	Omer
Prompts for baseline/hashed scenario	Omer
Development of the third (social context-enriched) approach	Amit
Annotation framework design	Amit
Prompts for the third approach	Amit
Comprehensive dataset exploration	Guy
Majority role-tagging evaluation	Guy
Summarizing approach comparisons	Guy
Approaches results collection	Noa
In-depth evaluation of the third approach	Noa
Phoebe's analysis & video-based case studies	Noa
Related work & speaker-role types selection	Amit, Noa
Local LLM prompt investigation & implementation	Omer, Guy
Annotation of speaker roles	All Team Members

Table 1. Division of Work Among Our Team

References

[1] Micha Elsner and Eugene Charniak. Disentangling chat. *Computational Linguistics*, 36(3):389–409, 2010.

[2] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[3] Rick Rienks and Dirk Heylen. Dominance detection in meetings using easily obtainable features. In *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction (MLMI)*. Springer, 2006.

[4] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[5] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Marie Meteer, and Carol van Ess-Dykema. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, 2000.

[6] Massimo Zancanaro, Bruno Lepri, and Fabio Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, ICMI '06, page 28–34, New York, NY, USA, 2006. Association for Computing Machinery.

[7] Zhaojiang Zhang, Siqi Sun, Michel Galley, Yiming Chen, Chris Brockett, Xiangyang Gao, and Jianfeng Liu. Modeling multi-turn conversation with deep utterance aggregation. *arXiv preprint arXiv:1806.09102*, 2018.