MULTI-FACETED DECEPTION DETECTION IN DIPLOMACY

A COMPARISON OF GRAPH-BASED, LLM-ASSISTED, AND RL-BASED APPROACHES

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

GROUP 91

MEMBERS: VARUN BHARTI, VIDHAN, VAIBHAV SEHARA

ROLL NO.: 2022562, 2022568, 2022554

CSE556 : NATURAL LANGUAGE PROCESSING PROJECT

INTRODUCTION & MOTIVATION

- **Problem Statement:**
  - Detect deceptive messages in the complex multi-agent negotiations of the game Diplomacy.
  - Challenges include severe class imbalance (with deceptive messages <5%), subtle linguistic cues, and evolving game dynamics.
- **Motivation:**
  - Enhance deception detection by leveraging not only textual cues but also relational, temporal, and game state information.
  - Combine established baselines with novel methods that capture global context.
- **Evaluation Focus:**
  - **Primary:** Accuracy
  - **Secondary:** Macro F1 and Lie F1 scores to account for the low frequency of deceptive messages.

RELATED WORK & BASELINES

Existing Work:

- Early studies on deception detection using linguistic features and behavioral cues (e.g., Levine et al., 1999).
- Previous work specific to Diplomacy (e.g., Peskov et al., 2020) used sequential models that treat messages in isolation.

Baselines Established:

- Classical machine learning methods (logistic regression, SVM, etc.).
- Deep learning baselines using Bi-RNN and LSTM architectures.

Limitations Identified:

- Inability to model global context and inter-player dynamics adequately.
- Insufficient handling of temporal and power dynamics inherent to Diplomacy.

PROPOSED NOVEL APPROACHES

Overview:

- We introduce three complementary methods to overcome baseline limitations.

Graph-Based Neural Network:

- Represents conversations as heterogeneous graphs combining player and message nodes.
- Incorporates text embeddings (using DistilBERT), one-hot encoded season/year, and normalized game score differences.

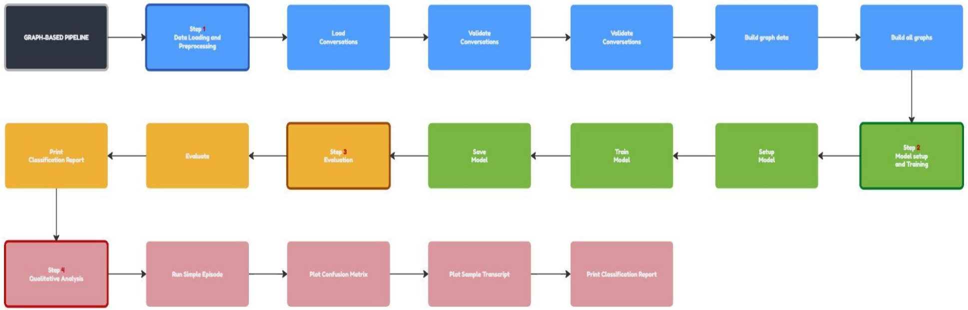
LLM-Assisted Classifier:

- Augments DistilBERT text features with a consistency score derived via LLM queries (using multiple models such as GPT-4O, GPT-4O Mini, O1 Mini, and O3 Mini).
- Captures long-term context and speaker-specific behavioral consistency.

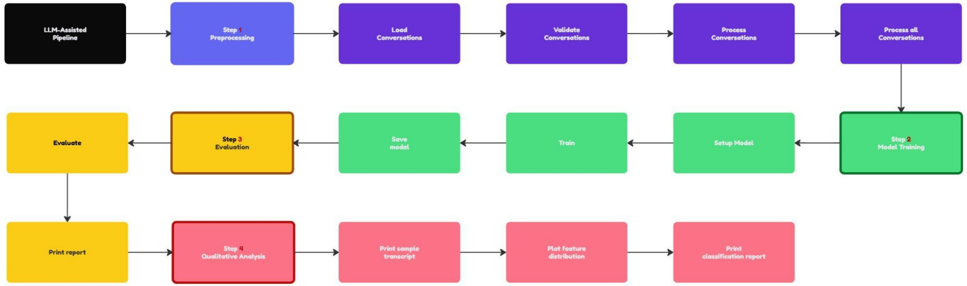
RL-Based Sequential Agent:

- Frames deception detection as a sequential decision-making task.
- Uses Recurrent PPO with a recurrent policy (MlpLstmPolicy) and reward shaping to emphasize the detection of deceptive messages.

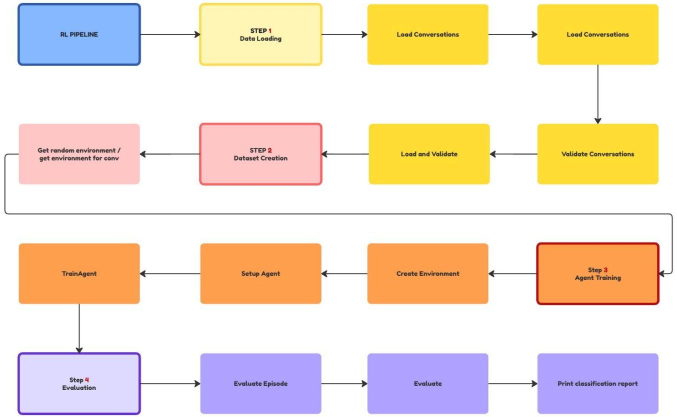
PROPOSED NOVEL APPROACHES



Graph based



LLM based



RL based

EXPERIMENTAL SETUP & QUANTITATIVE RESULTS

Dataset:

- 189 training, 21 validation, and 42 test conversations.
- **Additional statistics:**
  - **Message Length:** count = 13,132; mean ≈ 20.8 tokens; std ≈ 22.27; min = 1; median = 14; max = 294.
  - **Game Score Delta:** mean ≈ 0.07; std ≈ 2.15; min = -14; max = 14.
  - **Label Distribution:** Sender: 12,541 Truth vs. 591 Lie; Receiver: 11,459 Truth vs. 1,673 Lie.
  - **Unique Players:** {austria, england, france, germany, italy, russia, turkey}.
  - **Unique Years:** [1901, 1902, …, 1910].
  - **Unique Seasons:** {Fall, Spring, Winter}.

Evaluation Metrics:

- Accuracy, Macro F1, and Lie F1 (F1 on deceptive messages).

Method	Accuracy	Macro F1	Lie F1
Graph-Based GNN	0.90	0.51	0.95
LLM-Assisted Classifier	0.91	0.64	0.33
RL-Based Agent	0.96	0.49	0.00

QUALITATIVE ANALYSIS

- A transcript from a test conversation shows that nearly all messages were predicted as "Truth" except at a few points where a deceptive message was present (e.g. message 11 and 16 were true "Lie" in ground-truth but misclassified).
- This illustrates that while the relational features and global context capture much of the conversational tone, subtle deceptive cues might be overlooked.
- The conversation graph (see flowchart) visually highlights how player and message nodes interact; misclassifications are often associated with brief or ambiguous messages.

RL-Based Method:

- A sample episode transcript demonstrates that the RL agent correctly predicts most non-deceptive messages, receiving positive rewards (+1) at several steps.
- However, at key timesteps (e.g. steps 7, 8, 9, and 21) when deceptive messages occurred, the agent's consistently incorrect action (predicting "Truth") resulted in steep negative rewards (–10).
- This shows that, despite capturing sequence dependencies through the recurrent policy, the RL setup still struggles to detect deception in rare cases—indicating a need for further reward shaping and refined memory mechanisms.

LLM-Assisted Classifier (Brief Mention):

- Early qualitative findings suggest that including an LLM-derived consistency score helps align predictions with a speaker's typical behavior.
- In some cases, the consistency score provided a meaningful signal for detecting deviations, though occasional neutrality in the score still poses challenges.

CONCLUSION AND FUTURE WORK

Conclusion:

- We present a multi-faceted approach to deception detection in Diplomacy, leveraging three novel methods:
    1. Graph-Based Neural Networks to capture relational and temporal dynamics.
    2. LLM-Assisted Classification that fuses DistilBERT embeddings with LLM-derived consistency scores.
    3. RL-Based Sequential Agents that optimize detection through reward shaping and temporal context.
- Quantitatively, our methods achieve competitive overall accuracy and Macro F1 scores, although challenges remain on the Lie F1 metric—especially for the RL-based approach.

Future Work:

- **Hybrid Integration:** Explore a unified model that combines the strengths of graph propagation, LLM consistency scoring, and sequential decision-making.
- **Reward and Memory Refinement:** Refine the reward function and investigate enhanced recurrent architectures to better capture rare deceptive cues.
- **LLM Fine-Tuning:** Fine-tune LLM components on domain-specific negotiation data to improve the quality of consistency scoring.
- **Feature Fusion Enhancements:** Experiment with deeper fusion layers and additional contextual features (e.g., alliance dynamics, previous game moves) to boost performance on deceptive messages.

THANK YOU

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI