Project Proposal: Hybrid Approach for Deception Detection in Diplomacy Communications

Amartya Singh
2022062
amartya22062@iiitd.ac.in

Anish
2022075
anish22075@iiitd.ac.in

Adarsh Jha
2022024
adarsh22024@iiitd.ac.in

Abstract

In this proposal, we share our plan to build a deception detection system tailored for the Diplomacy game dataset. Our system aims to improve upon the existing Context LSTM+Power baseline (58.1% F1) by leveraging recent advances in transformer models, graph neural networks, and logical reasoning through forward/backward chaining. The ultimate goal is to achieve a robust and explainable deception detector with a target Macro-F1 score in the 65-70% range.

1 Problem Definition

In the game of Diplomacy, players engage in strategic communication where deception is a key factor. Our objective is to accurately detect deceptive messages using a dataset of over 17K messages, with lies comprising only about 5% of the data. The challenge is to overcome class imbalance and the subtle nature of deceptive language, while also modeling complex social dynamics (e.g., alliances and power differentials).

2 High-Level Plan

Our approach comprises the following major components:

- **Data Preprocessing:** Tokenize messages, extract contextual features, and construct a player interaction graph based on game metadata.
- **LLM-Modeling:** Fine-tune a transformer model (e.g., LLaMA, DeepSeek, or Qwen) for message-level deception detection.
- **Graph Neural Network Integration:** Incorporate a GNN module to model player relationships and power dynamics, complementing text-based features.
- **Trust Calibration and/or Logical Reasoning Module:** Integrate historical deception patterns to adjust prediction probabilities and improve interpretability, or some methods like forward or backward chaining to assess argument consistency and bias.

3 Approach

Our system will combine deep language models with structured reasoning:

1. **Transformer Backbone:** We will use an open-source transformer (via Hugging Face) to encode individual messages along with their conversational context, Eg- Llama, Deepseek, Qwen, etc.
2. **GNN for Social Dynamics:** A graph neural network will be used to capture inter-player interactions and the evolving power dynamics during the game.
3. **Ensemble and Trust Calibration and/or Reasoning Module:** By combining outputs from the transformer, GNN, and reasoning modules, and calibrating using a trust score based on historical behavior, maybe also using techniques to check for argument consistency, our ensemble will output a final deception probability.

Limitations

Due to the time constraint, any reinforcement learning approach would not be feasible. Instead, our focus will be on establishing a strong pipline with the hybrid transformer+GNN+reasoning architecture and thorough error analysis.

Note - The evaluation metric seems to be accuracy so we will focus on that too, aside from our main metric of Macro F1 score.