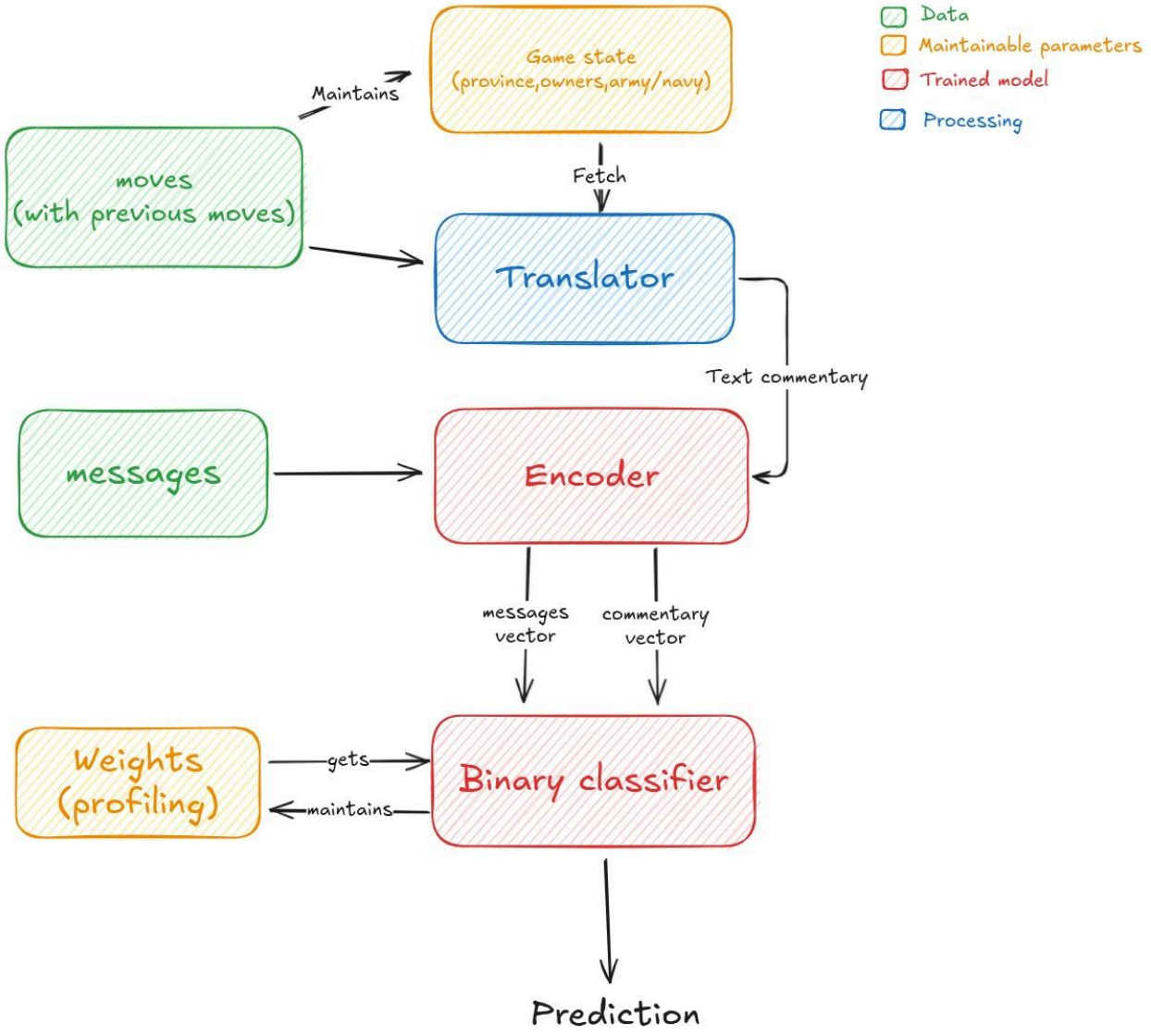
Report: A Novel Architecture for Deception Detection in Diplomacy

Executive Summary



This report details a novel neural architecture, the **Context-Infused Profiling Network (CIP-Net)**, designed specifically for detecting deception in the strategic game of Diplomacy. Traditional NLP models often fail to capture the nuanced interplay between language, game state, and player psychology inherent in complex negotiations. Our proposed solution addresses this gap by creating a multi-modal system that learns to identify discrepancies

between a player's stated intentions (their messages) and their demonstrated actions (their moves).

The core innovations of CIP-Net are:

1. A **Strategic Move Translator** that converts raw game-state data and move histories into natural language commentary, grounding linguistic analysis in factual game events.
2. A **Dual-Stream Transformer Encoder** that separately processes conversational text and strategic commentary, creating distinct vector representations for what is *said* versus what is *done*.
3. A **Player Profiling Classifier** that incorporates learnable embeddings for each player, allowing the model to adapt to individual deceptive styles and tendencies.

By explicitly modeling the relationship between communication and action, and by profiling individual players, CIP-Net is fundamentally designed to detect betrayal, misdirection, and broken promises—the hallmarks of deception in Diplomacy.

---

1. Introduction

The challenge of detecting deception in Diplomacy messages requires more than standard text classification. Deception in this domain is not merely a linguistic phenomenon but a strategic one, deeply embedded in the context of the game state, player relationships, and historical actions. A lie is often not a simple falsehood but a promise that is intended to be broken, a misrepresentation of one's strategic position, or a statement designed to build false trust before a betrayal.

The provided dataset, containing 17,289 messages from 12 games, offers rich data including messages, sender/receiver labels, and detailed game context (

game_score, seasons, years, previous moves)[1]. Our proposed architecture, CIP-Net, is designed to leverage this full spectrum of information, creating a model that understands not just the language of Diplomacy, but the game itself.

2. Proposed Architecture: Context-Infused Profiling Network (CIP-Net)

As illustrated in the provided diagram, the CIP-Net architecture consists of three novel,

interconnected modules designed to process linguistic, strategic, and player-specific data streams.

2.1 Game State and Move Translator

This is the foundational innovation of the architecture. Instead of treating game moves as simple categorical data, this module translates them into a narrative textual format.

- **Input:** Historical game state data, including previous moves, unit positions (province, owner, army/navy), and player scores. This information is available in the dataset's Moves folder and message metadata[2].

- **Process:** The Translator is a rule-based or simple sequence-to-sequence model that generates "Text Commentary" describing the strategic situation. For example:
  - A move A MUN -> RUH becomes: *"Germany moves an army from Munich to Ruhr."*
  - A sequence showing a broken promise might be translated as: *"In Spring 1903, Russia promised to support Austria in Galicia. In Fall 1903, Russia moved to attack Austria in Galicia, breaking the agreement."*
- **Output:** A stream of text commentary that represents the factual, on-the-board reality of the game.

2.2 Dual-Stream Encoder

The core of the network features two parallel Transformer-based encoders (e.g., BERT or RoBERTa) to process the distinct forms of text.

1. **Linguistic Stream:** This encoder processes the raw conversational messages between players. Its purpose is to capture the nuances of negotiation, persuasion, tone, and explicit promises.
   - **Input:** Raw text from the messages field.
   - **Output:** A message vector, Vmsg, representing the linguistic content.
2. **Strategic Commentary Stream:** This encoder processes the output from the Move Translator. Its purpose is to understand the objective strategic context and the history of player actions.
   - **Input:** "Text Commentary" generated by the Translator.
   - **Output:** A commentary vector, Vcom, representing the game-action context.

2.3 Player Profile-Weighted Classifier

This final module integrates the outputs of the encoder with player-specific information to make a final prediction.

- **Input:**
  1. The message vector (Vmsg).
  2. The commentary vector (Vcom).
  3. **Player Profile Embeddings:** Learnable weight vectors unique to each player (e.g., 'italy', 'germany'). These are retrieved based on the speakers and receivers fields from the dataset[3]. These embeddings act as a "profiling" mechanism, capturing individual player styles.

- **Process:** The vectors are concatenated and passed through a series of dense layers. The player embeddings modulate the classification process, allowing the model to weigh features differently based on who is speaking and who is listening. For example, a promise from a historically treacherous player (as learned by their embedding) might be treated with more suspicion.
- **Output:** A binary prediction (0 or 1) corresponding to the sender_labels (truthful vs. deceptive).

---

3. Rationale for Deception Detection

The CIP-Net architecture is fundamentally tailored to deception detection for the following reasons:

- **Detecting Discrepancy:** Deception in Diplomacy often manifests as a divergence between words and actions. By encoding both streams separately, the model can directly compare the linguistic content (V_{msg}) with the strategic reality (V_{com}). A message containing a promise of alliance can be contrasted with a commentary vector describing troop movements that threaten that same ally. This discrepancy is a powerful signal for deception.
- **Grounding Language in Reality:** The Move Translator grounds the model's understanding. It prevents the model from being swayed by purely persuasive language that has no basis in the game's reality. A player might speak of peace and friendship, but the commentary vector will reflect aggressive troop build-ups on a border, providing a factual counter-narrative.
- **Modeling Player Psychology:** Deception is not uniform; different players lie in different

ways. The Player Profile Embeddings allow the model to learn player-specific "tells." For example, Player A might become overly polite when lying, while Player B might make grandiose, unrealistic promises. The model can learn these individual patterns, making it far more effective than a one-size-fits-all approach. This directly addresses the multi-player interaction dynamics of Diplomacy.

---

4. Data Handling and Training

- **Data Source:** The primary data source will be the provided .jsonl files [4][4][4], supplemented by the

  Moves folder mentioned in the README.md[5].

- **Input Preparation:**
  - The messages field will be tokenized for the Linguistic Stream Encoder.
  - The game state fields (game_score, game_score_delta, seasons, years, and historical moves) will be fed into the Move Translator to generate commentary for the Strategic Commentary Stream[6].
  - The speakers and receivers fields will be used to look up the corresponding player embeddings[7].

- **Training:** The model will be trained end-to-end using a **Binary Cross-Entropy** loss function. The target variable is the sender_labels field, which indicates the ground truth for deception[8]. The evaluation will be based on

  Accuracy and **Macro F1-Score**, as stipulated by the hackathon requirements.

5. Conclusion and Novelty

The CIP-Net architecture represents a novel approach that goes beyond standard classification. Its fundamental innovation lies in its explicit modeling of the relationship between language and strategic action. By translating game moves into a linguistic format and comparing them against player messages within a player-aware framework, it is uniquely positioned to identify the discrepancies that form the very essence of deception and betrayal

in Diplomacy. This context-infused, player-profiled approach is designed not just to classify text, but to understand the strategic intent behind it.