Project Report: Deception Detection in Diplomacy

Team: PARADOX
Final Model: Context-Aware Multi-Vector Ensemble Framework (CMVEF)
Result: Macro F1-Score of **0.6407** [1]

1. Introduction: The Challenge

The Esya '25 Deception Detection Challenge presented a unique and complex problem: to build a **novel architecture** for identifying deception in messages from the strategic game of Diplomacy. The core constraint, "Create, Don't Just Apply!", explicitly forbade the use of standard, off-the-shelf classification models, pushing for genuine innovation. The evaluation was based on **Accuracy** and, critically for the imbalanced dataset, **Macro F1-Score**.

Our team, PARADOX, embarked on this challenge with an ambitious vision for a deep learning architecture. This report chronicles our journey, detailing our initial proposal, the strategic pivot to a more pragmatic approach, the architecture of our final successful model, and the rationale behind the changes made along the way.

---

2. The Initial Vision: The Context-Infused Profiling Network (CIP-Net)

Our initial proposal, the **Context-Infused Profiling Network (CIP-Net)**, was born from the core philosophy that deception in Diplomacy is not just about the words used, but about the **discrepancy between words and actions**. We proposed a novel, end-to-end neural architecture designed to learn this interplay automatically.

The proposed CIP-Net consisted of three core modules:

- **Game State and Move Translator:** This module was designed to process raw game data of moves in the game (given in the data set repository) and translate it into natural language "Text Commentary". This would create a factual, textual representation of the game's reality to contrast with a player's claims.

- **Dual-Stream Transformer :** The model would process two text streams in parallel: one for the player's conversational messages and another for the machine-generated "Text Commentary." By comparing the resulting vectors, the network would learn to spot discrepancies—the very essence of a broken promise.

- **Player Profile-Weighted Classifier:** The final stage would concatenate these two vectors and feed them into a classifier. Crucially, this classifier would incorporate **learnable embeddings** for each player, allowing it to "profile" individual tendencies and weigh evidence differently based on who was speaking and who was listening.

This end-to-end, multi-modal architecture was designed to be fundamentally novel by directly modeling the relationship between language, action, and individual player psychology within a single deep learning framework.

3. The Strategic Pivot: From End-to-End to Feature Engineering

While the CIP-Net was a powerful concept, its implementation as a single, end-to-end trainable model presented significant engineering challenges. We realized that a more pragmatic and modular approach could retain the spirit of our initial proposal while allowing for more robust development and experimentation. This led to our pivot from a monolithic architecture to a **multi-stage feature engineering pipeline** coupled with an ensemble of models.

Why We Changed Our Approach:

- **Tackling Class Imbalance:** The extreme imbalance (~95% truthful messages) was the single greatest technical hurdle. We realized that a multi-pronged approach using specialized techniques (SMOTEENN, Focal Loss, Weighted Samplers) applied to different model types would be more effective than relying on a single model to solve it.
- **Engineering Robustness & Practicality:** Building and fine-tuning a custom dual-stream transformer network from scratch is a high-risk, time-intensive task. Deconstructing the problem into a feature-engineering pipeline allowed for modular development, independent testing of each component, and greater overall stability.
- **Preserving the Core Idea:** Our goal was not to abandon our initial vision, but to find a better way to implement it. We successfully deconstructed the core concepts of CIP-Net into the components of CMVEF.

The core ideas of CIP-Net were not abandoned but were instead implemented as distinct, powerful steps in our new framework:

- **The "Move Translator" became a Data Augmentation Step:** The concept of translating game state into text was realized through a preprocessing script (create_commentary.py). This script generated a commentary.csv file ,from the moves data given, containing natural language summaries of each player's strategy for a given turn, effectively grounding the linguistic data in game reality, just as we had envisioned.

- **The "Dual-Stream Encoder" became a Feature Extraction Step:** Instead of training two large Transformer models from scratch, we used a highly optimized Sentence Transformer, to generate separate numerical embeddings for the messages and the newly created commentary. This captured the semantic meaning of both "what was said" and "what was done" as distinct feature vectors.

- **The "Player Profiling" became Historical Feature Engineering:** The goal of modeling individual player tendencies was achieved by creating a rich set of behavioral and historical features. We engineered features to track a player's historical "lie rate," the similarity of their current message to recent ones, and their specific truthfulness history with the message's receiver. This captured player reputation and patterns without the complexity of training custom embeddings within the model.

This pivot allowed us to maintain the innovative principles of our proposal while building a more robust and ultimately more successful solution.

---

4. The Final Architecture: CMVEF

Our final solution, the **Context-Aware Multi-Vector Ensemble Framework (CMVEF)**, is a comprehensive pipeline that integrates multiple vectors of information.

A. Comprehensive Feature Engineering

The cornerstone of our success was a diverse and powerful set of features designed to provide a 360-degree view of each message:

1. **Semantic & Lexical Features:**
   - **Sentence Embeddings:** 384-dimension vectors for messages and commentaries captured from the Sentence Transformer.
   - **TF-IDF SVD:** 128-dimension vectors capture the importance of specific keywords and phrases[13].

2. **Linguistic & Heuristic Features:**
   - **Promise Features:** Binary flags for promise-related language (e.g., "support," "I will")[14].
   - **Linguistic Cues:** Counts of negations, message length, punctuation, and linguistic "hedges" like "maybe" or "perhaps".

3. **Behavioral & Historical Features:**
    - **Player History:** A player’s lie rate and the receiver's trust rate over the last K seasons.
    - **Recent Message Similarity:** The cosine similarity between a current message and the player's last few messages to detect shifts in tone or intent.
    - **Speaker-Receiver Reputation:** The historical truthfulness rate between a specific speaker and receiver.
    - **Game Context:** Numerical features like year, season, and game score were also included and scaled.

B. The Stacked Ensemble Model

Given the highly imbalanced nature of the dataset (~95% truthful messages), a simple classifier would fail. Our solution employed a stacked ensemble of two diverse base models whose predictions were fed into a final meta-classifier.

1. **Base Model 1: LightGBM:** A powerful gradient-boosted decision tree model, excellent for the tabular feature data we created. To combat imbalance, the training data for this model was resampled using **SMOTEENN**, a sophisticated technique that synthesizes new minority class examples (lies).

2. **Base Model 2: Multi-Layer Perceptron (MLP):** A deep neural network designed to capture complex, non-linear relationships in the features. This model addressed imbalance using a **WeightedRandomSampler** (to show the model more lies) and a **Focal Loss** function (to focus learning on difficult-to-classify examples).

3. **Meta-Classifier: Logistic Regression:** The probabilities from both the LightGBM and MLP models were used as inputs to train a final, simple Logistic Regression model. This "stacking" approach allowed the meta-model to learn the optimal way to weigh the predictions from its two very different base models, leading to a more accurate and robust final decision.

---

5. Results and Conclusion

The effectiveness of our approach was validated by the final results on the test set, where the stacked ensemble significantly outperformed each of its constituent base models.

Model	Macro F1-Score
Stacked Meta-Model (Final)	**0.6407**
LightGBM (Base)	0.6307
MLP (Base)	0.5625

The journey from the conceptual CIP-Net to the implemented CMVEF demonstrates a successful adaptation to real-world modeling challenges. By deconstructing our initial end-to-end vision into a modular pipeline of data augmentation, rich feature engineering, and advanced ensembling techniques, we stayed true to our core philosophy: that detecting deception requires a deep understanding of context, history, and the intricate dance between words and actions. The CMVEF stands as a novel and powerful framework that effectively met and exceeded the demands of the challenge.

6. Real-World Impact & Viability : A Scalable Framework

The evolution from CIP-Net to CMVEF significantly enhances the project's real-world impact and viability.

- **Modularity and Maintainability:** The CMVEF pipeline is far more scalable and maintainable than a single, monolithic model. New feature sets can be added, and individual models can be updated or replaced without redesigning the entire system.
- **Broader Applications:** The core principle of CMVEF—contrasting stated intent (messages) with observed behavior (commentary) and historical context—is highly applicable to other critical domains. This framework could be adapted for:
    - **Financial Fraud Detection:** Analyzing communications versus transaction histories.
    - **Content Moderation:** Identifying malicious actors in online communities based on their stated intent versus their posting patterns.
    - **Enterprise Risk Management:** Analyzing corporate statements against operational data to detect misleading information.

Our final solution is not just a game-playing model; it is a blueprint for a robust contextual analysis framework.