



Deception Detection in Diplomacy

Avi Sharma -2022119

Dasari Sai Harsh -2022144

Parth Sandeep Rastogi -2022352

INTRODUCTION



Deception is a nuanced challenge in human communication, notably in strategic settings like Diplomacy, where players' messages are dual-annotated for both intent and perception. The study reveals that only about 5% of messages are deceptive, and current models face challenges due to sparse signals, isolated context, and inadequate cross-player interaction, despite deceptive messages often being verbose and contextually rich.

LieDetectorGAT: A graph-based model that builds dynamic player interaction graphs where messages are connected via BERT embeddings, power differentials, and linguistic cues, using multi-head edge-aware attention to capture message-level deception within game-specific social structures.

HiS-Attention: A transformer architecture that integrates the current message, game metadata, and multi-turn dialogue history with cross-modal attention to generate deception-aware representations at the message level.

LieDetectorGAT

- 1) **Graph Structure:** Each game = directed graph
- Nodes: Players (7-d one-hot encoding)
 - Edges: Messages ($u \rightarrow v$), sorted by time

- 2) **Edge Features (782-d):**
- BERT [CLS]: Message semantics (768-d)
 - Deception Lexicon: Psychological cues (10-d)
 - Strategic Metadata: Lie count, avg. score delta, season, year bucket, msg index

- 3) **Graph Attention (GATv2):** Multi-layer, edge-aware updates capture "who says what to whom" and how deceptive

- 4) **Classification:**
- 2-layer MLP on sender, receiver, and edge embeddings
 - Outputs deception probability using weighted BCE loss

Models message interactions and deception cues in a structured, game-aware manner.

HiS-Attention

Message: Encoded via BERT for linguistic deception cues.

Game State: Encoded using FFNN; includes score, season, score_delta, etc.

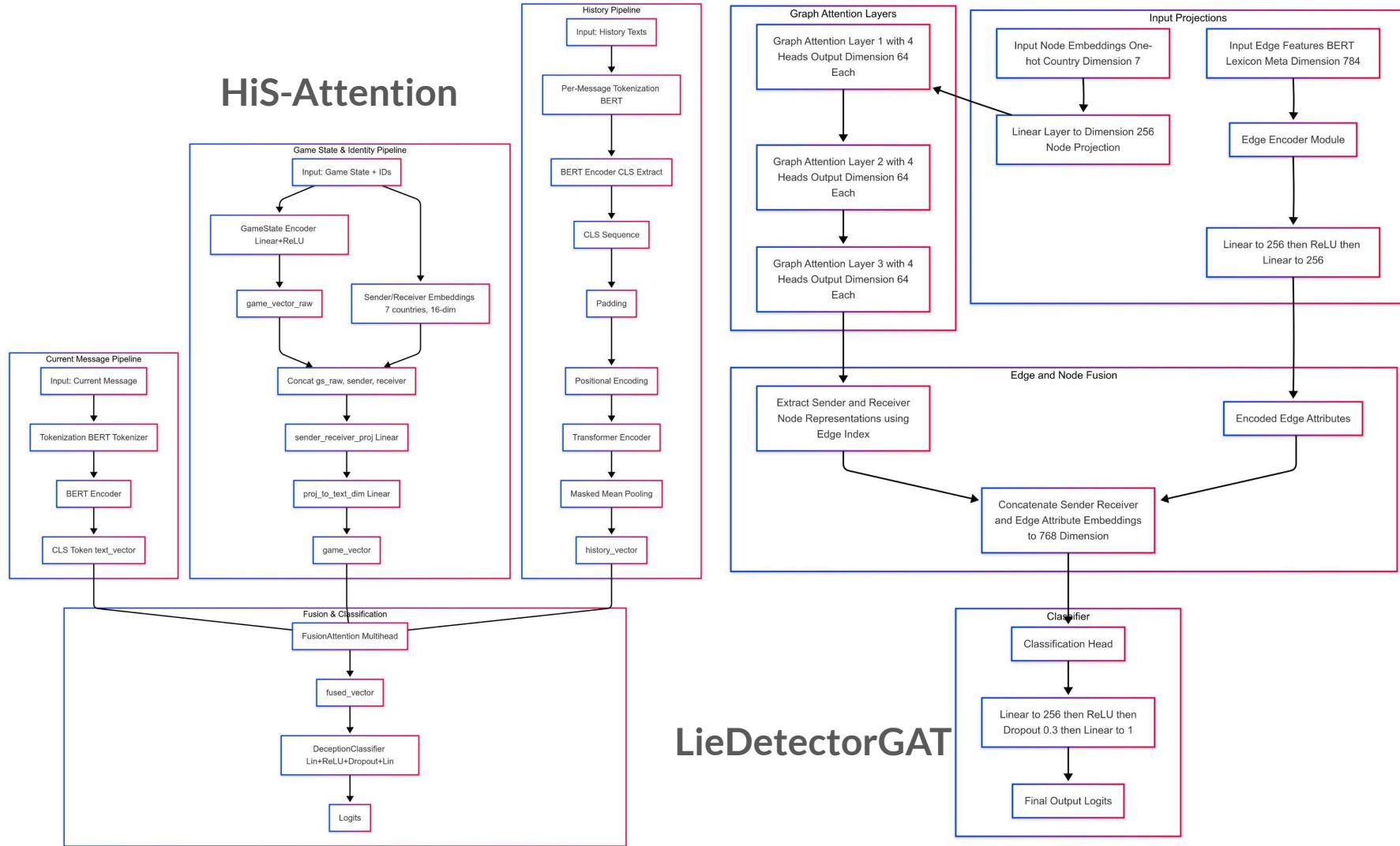
History: Last 10 messages encoded with BERT + Transformer.

Player Embeddings: Learnable vectors per sender/receiver.

Multi-head attention: Replaces feature concatenation with Learnable Quasi-Attention Pooling over all modalities for adaptive, context-aware fusion.

Output: Fused vector passed to classifier for deception prediction.

HiS-Attention



LieDetectorGAT

Dataset Description and Preprocessing



The dataset used in this work is sourced from the QANTA Diplomacy Deception Detection task, de4 signed to predict whether messages exchanged between players in the game Diplomacy are deceptive or truthful. Each sample corresponds to a single message exchanged between two players, along with contextual metadata. In this work, we do not use this field, as we are only concerned with whether the sender actually intended to lie, not how the receiver perceives the message.

PreProcessing

1. Text Cleaning:

- Convert to lowercase
- Remove URLs, emojis, special characters
- Normalize whitespace

2. Conversation History Augmentation:

- For each message, track (game_id, sender, receiver)
- Retrieve last 10 messages from sender → receiver
- Attach as history for contextual understanding
- Update history after each message

→ Enhances input quality and contextual depth for deception detection.

The dataset is imbalanced, with significantly more truthful messages than deceptive ones. Label distributions across data splits are summarized in Table 1.

Table 1: Label distribution across dataset splits

Split	Truthful	Deceptive
Train	12,541	591
Validation	1,360	56
Test	2,501	240
Overall	16,402	887

Experimental Setup

HiS Attention

Model Variants:

- Model 1 No Context (Concat), BERT + Game State, concatenated.
- Model 2: 5-Message Context + BiLSTM + Mean Attention:
History via BiLSTM, fused with Multi-head Attention.
Macro F1: 53.8
- Model 3 – 10-Message Context + Transformer + Query Attention:
History via Transformer, fused with Query Attention Pooling.
Macro F1: 57.38

Training Details: Epochs: 10, Batch Size: 16, Optimizer: Adam, LR = $1e-5$, Loss: Weighted BCE (pos_weight = 21.22), Scheduler: Patience = 1, Decay = 0.5, Gradient Clipping: Max norm = 1.0

LieDetectorGAT


Training Parameters: Epochs: 20, Optimizer: AdamW, Learning Rate: 3×10^{-4} , Loss Function: BCEWithLogitsLoss with positive class weighting to address class imbalance (~5% lies) Gradient Clipping: Maximum norm of 1.0

Data Handling: We used one graph per batch (i.e., one game instance) and employed the PyTorch Geometric DataLoader with shuffling disabled to preserve conversational order. During preprocessing, message sequences were chronologically sorted to maintain dialogue flow. For linguistic features, we utilized the deception lexicon provided in the dataset's utilities.

Model	Macro F1 Score
No Context + Concat Fusion	51
Last 5 Messages + BiLSTM + Mean Attention	53.8
Last 10 Messages + Transformer + Query Attention	57.38

→ **Query Attention + Transformer History yields strongest deception detection performance.**

Results



	Predicted Truth	Predicted Lie
Actual Truth	2328	173
Actual Lie	189	51

Table 3: Confusion matrix of the best HiS-Attention model

Class	Precision	Recall	F1 Score	Support
Truth (0)	0.92	0.93	0.93	2501
Lie (1)	0.23	0.21	0.22	240
Accuracy	0.8679			
Macro Avg	0.58	0.57	0.57	2741
Weighted Avg	0.86	0.87	0.87	2741

Table 4: Classification report of HiS-Attention model

Nation	Number of Samples	Error Rate
Russia	852	12.79
Turkey	441	12.93
England	307	7.17
France	804	16.29
Germany	133	21.80
Italy	97	6.19
Austria	107	7.48

Table 5: Speaker-wise error rates and sample counts of HiS-Attention

	Pred. Truth	Pred. Lie
Actual Truth	2136	365
Actual Lie	147	93

Table 6: Confusion matrix of the LIEDETECTORGAT model on the test set.

Class	Precision	Recall	F1	Support
Truth (0)	0.936	0.854	0.893	2501
Lie (1)	0.203	0.388	0.266	240
Accuracy	0.813			
Macro Avg	0.569	0.621	0.580	2741
Weighted Avg	0.871	0.813	0.838	2741

Table 7: Classification report of LIEDETECTORGAT on the test set.

Speaker	Number of Samples	Error Rate
Austria	107	7.48
England	307	13.68
France	804	19.78
Germany	133	32.33
Italy	97	5.15
Russia	852	19.84
Turkey	441	19.50

Table 8: Speaker-wise error rates for LIEDETECTORGAT on the test set.

Model	Macro F1	Lie F1
Human Baseline	0.581	0.226
Harbringers AL	0.528	0.246
Harbringers + Power AL	0.529	0.237
Harbringers SL	0.459	0.147
Bag-of-Words AL	0.539	0.179
Bag-of-Words + Power AL	0.548	0.198
Bag-of-Words SL	0.518	0.143
Bag-of-Words + Power SL	0.518	0.144
Random AL	0.397	0.149
Random SL	0.384	0.118
Majority AL	0.477	0.000
Majority SL	0.483	0.000
LSTM	0.530	0.122
ContextLSTM	0.535	0.137
HiS-Attention (Ours)	0.570	0.220
LieDetectorGAT (Ours)	0.580	0.266

Table 9: Macro and binary (Lie) F1 for all baselines (AL = Actual-Lie, SL = Suspected-Lie). Bolded entries indicate best performance across all models.

Error Analysis - HiS-Attention



Quantitative Analysis

- Best-performing architecture uses Transformer encoding with 10 prior messages and query-based attention
- Achieved a **Macro F1 score: 57.38** and **Lie-F1: 21.98**
- Outperformed all baseline models
- Despite using a weighted loss function, the model still struggles with the minority (deceptive) class

Qualitative Analysis

- **False Positives:**
 - Diplomatic vagueness and cautious planning misinterpreted as lies
 - Many truthful, strategically worded messages flagged as deceptive
- **False Negatives:**
 - Lies are mimicking cooperative or neutral tones often missed
- **Speaker-Level Observations:**
 - Higher misclassification in nations like Germany (21.8%) and France (16.3%)
 - Lower error rates in Italy and Austria, likely due to clearer communication
- **Overall:**
 - The model tends to overfit to dominant truthful patterns
 - Suggests a need for refined deception cue modeling, possibly with speaker-specific conditioning

Error Analysis - LieDetectorGAT



Quantitative Analysis

- **Overall Accuracy:** Achieved an accuracy of **81.3%**.
- **Deception (Lie) Class Performance:** Recall is **0.388** (higher than the HiS-Attention model), indicating improved detection of lies. Precision is only **0.203**, which shows the model tends to flag more false positives.
- **Truthful Class:** Maintains a high F1 score of **0.893**.
- **Key Insight:** While the model effectively captures more deceptive cases, the trade-off is a reduction in precision due to increased false alarms.

Qualitative Analysis

- **Speaker-wise Error Trends:**
 - **Germany** shows the highest error rate at **32.33%**, suggesting difficulties with complex or multi-turn strategic messaging.
 - **Italy** displays the lowest error rate at **5.15%**, implying that clearer, more consistent communication is easier to classify correctly.
- **Error Patterns:** Overly detailed or cautious phrasing is often misclassified as deception. Lies that mimic friendly or neutral tones are sometimes missed.
- **Overall Observation:** The graph-based approach of LieDetectorGAT enhances the recall of deceptive messages by explicitly modeling inter-player interactions, yet it also overestimates deceptive intent, leading to a higher rate of false positives.



Conclusion and Future Work

While our models advance the state of deception detection in Diplomacy, several promising directions remain open. First, the current model lacks adversarial awareness; future iterations could incorporate counterfactual training, where deceptive messages are paired with hypothetical truthful variants to strengthen representation learning. In addition, leveraging reinforcement learning (RL) techniques and agent-based simulation can enable adaptive decision-making frameworks that better mirror the dynamics of multi-agent deception. Moreover, improved linguistic cue extraction using tools like LIWC could supplement our current lexicon-based approaches by providing more fine-grained semantic and psychological markers. Finally, fine-tuning large language models (LLMs) specifically for deception understanding promises to provide nuanced interpretations of context and subtle cues that may escape traditional methods.