Deception Detection In Diplomacy

Dasari Sai Harsh
2022144
IIIT-Delhi
dasari22144@iiitd.ac.in

Avi Sharma
2022119
IIIT-Delhi
avi22119@iiitd.ac.in

Parth Sandeep Rastogi
2022352
IIIT-Delhi
parth22352@iiitd.ac.in

Abstract

Deception detection is a critical yet underexplored task in multi-agent strategy games like *Diplomacy*, where players often engage in calculated misdirection. Building upon prior datasets that capture labeled deceptive intent, we propose two complementary models to address core limitations in current approaches: (1) **HiS-Attention**, a transformer-based multimodal fusion architecture that integrates message content, strategic game state, and historical dialogue; and (2) **LieDetectorGAT**, a graph attention network that explicitly models interplayer interactions enriched with linguistic deception cues and power dynamics. Through detailed error and speaker-level analyses, we demonstrate the importance of contextual modeling.

1 Introduction

Deception is a core aspect of human communication, deeply intertwined with trust, strategic intent, and social dynamics. From diplomatic negotiations to social media discourse, identifying deceptive behavior — and distinguishing it from honest misjudgment or exaggeration — remains a critical challenge. Computationally modeling deception requires both nuanced contextual information and a robust signal of ground truth intent, both of which are scarce in real-world datasets.

The *Diplomacy* board game presents a compelling sandbox for studying long-term strategic deception. In their landmark study, *It Takes Two to Lie* (Peskov et al., 2020), the authors construct a deception-labeled dataset grounded in this game, where players must form and break alliances over a sequence of turns. Messages exchanged between players are annotated in real time by both senders (actual intent) and receivers (perceived deception), offering a dual-perspective view of lying within a competitive, goal-driven setting.

However, the ACL Diplomacy paper itself highlights several limitations and under-explored directions. First, many existing models suffer from contextual blind spots, as they operate on isolated messages and overlook the critical interaction history between players — a key cue for detecting shifts in trust or alliance breakdowns. Second, there is the issue of sparse deception signals: only around 5% of the messages are labeled as lies, which creates a significant class imbalance and demands more robust indicators for deception detection. Third, while the authors incorporate linguistic markers from a deception lexicon — including categories such as claims, subjectivity, and temporal discourse cues — these features are often hardcoded and brittle, limiting their generalizability across different games or player styles. Finally, the paper points to a lack of cross-player interaction modeling: deception in Diplomacy is rarely player-agnostic, and accurately identifying it often requires recognizing inconsistencies across multiple interactions — for example, when promises made to one player are covertly violated in messages to another.

Moreover, the original analysis confirms that lies tend to be verbose, apologetic, or hyper-specific ("support X into Y"), while truths are often terse and factual. This underscores the **central role of linguistic cues**, especially those rooted in discourse, subjectivity, and game context, in surfacing deceptive patterns. Beyond surface words, deception is embedded in how players evolve relationships, reveal or withhold intentions, and leverage rhetorical structure to persuade — or mislead — their counterparts.

1.1 Contributions

We summarize our key contributions as follows:

- **We propose LieDetectorGAT**, a graph-based deception detection model that constructs dynamic player interaction graphs, where mes-

sages form edges enriched with BERT embeddings, power differentials, and lexicon-derived linguistic cues. It leverages multi-head edge-aware attention to model message-level deception within game-specific social structures.

- **We propose HiS-Attention (Historical-Structured Attention)**, a transformer-based architecture that fuses three streams of information — the current message, structured game metadata, and multi-turn conversational history — using cross-modal attention to generate deception-aware representations at the message level.

2 Related Work

Deception detection in negotiation and multi-party game settings has recently gained significant attention, particularly with the introduction of datasets and models tailored to complex, strategic environments. Peskov et al. (2020) introduced the Diplomacy Deception dataset, which consists of over 17,000 annotated messages exchanged between players in the game *Diplomacy*. Their work is notable for its dual annotation scheme, capturing both the sender's intent and the receiver's perception of truthfulness. The authors explored a range of models, from logistic regression using linguistic and power-dynamic features to neural architectures such as LSTM and BERT-based models that incorporate conversational context and in-game metadata. Their findings demonstrate that combining message content with game context significantly improves deception detection, approaching human-level performance.

Building on this foundation, several recent studies have explored deception in similarly complex interactive settings:

- **Wongkamjan et al. (2024)**: In their recent work, the authors address deception detection in Diplomacy negotiations by combining logical forms of agreements, reinforcement learning-based value functions (derived from the Cicero agent), and BERT-based classifiers. Their approach aligns closely with ours in integrating game state, negotiation context, and semantic features.

- **Fu (2019)**: In their thesis ,on deception in the online Mafia game Fu utilized LSTM, BiLSTM, and CNN models to analyze player conversations. This study underscored the importance of conversational context and linguistic markers in detecting deception within adversarial game environments.

- **Kumar et al. (2021)**: Their work on deception in the game *The Resistance* employed dynamic interaction networks to model evolving group interactions. By representing these interactions as dynamic graphs, they demonstrated the utility of graph-based structures—echoing our use of graph attention networks for modeling Diplomacy's social complexity.

Across the literature, contextual modeling has emerged as a central requirement for effective deception detection. Models that incorporate long-range dependencies—whether through attention mechanisms or graph structures—consistently outperform those relying on isolated message representations, as they better capture evolving player relationships and conversational nuance. In parallel, graph-based approaches are becoming increasingly popular in multi-agent and strategic environments, where the ability to model inter-player dynamics and role-specific interactions provides a strong structural prior. Linguistic cues also continue to play a valuable role, with many studies demonstrating performance gains when integrating semantic and lexical features derived from domain-specific deception lexicons. Despite the widespread adoption of attention mechanisms in NLP more broadly, they remain relatively underexplored in Diplomacy-focused deception models. In particular, the application of multi-head attention and graph attention networks has received limited attention in this space, pointing to a significant gap that our proposed methods aim to address.

3 Methodology

3.1 Multimodal Transformer Based Attention Fusion

To improve upon the baseline architecture proposed in the QANTA Diplomacy Deception Detection task, we replaced the concatenating features and then passing the feature vector strategy used in the original model with a Learnable Query Attention Pooling mechanism. Our model encodes the text message using BERT, the game board state using a custom state encoder, and the dialogue history

using a transformer. Instead of simply concatenating these features before classification, we treat the three modality vectors as a sequence and apply multi-head attention to allow inter-modality interactions. A learnable query vector is then used to perform attention pooling over these outputs, producing a single fused representation. This allows the model to dynamically weigh each modality based on context, leading to more adaptive and informative fusion compared to static concatenation.

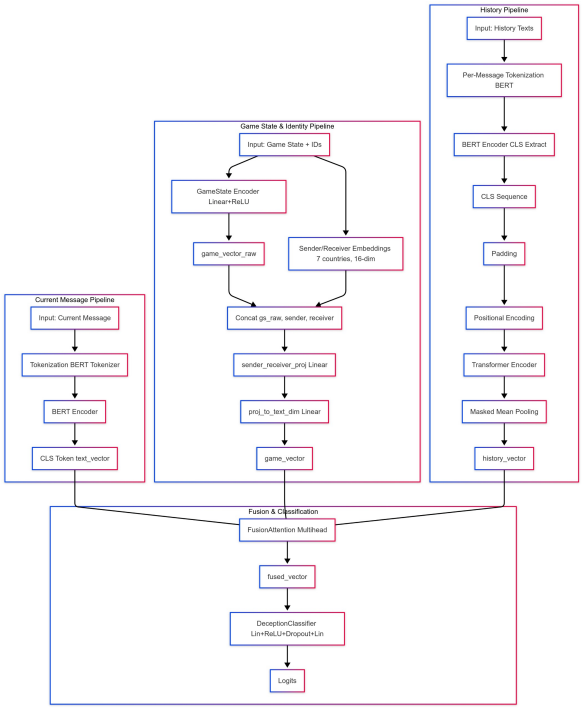

Figure 1: Overview of our multimodal approach

- **Message Encoding:** The current message is encoded using a pre-trained BERT model to capture rich contextual and semantic information. BERT's bidirectional transformer layers are particularly effective in identifying linguistic cues that may signal deception.

- **Game State Encoding:** Game-specific features (e.g., current score, score delta, season, year) are encoded using a feedforward neural network. Categorical features such as season and year are one-hot encoded, while numerical features are passed as-is. This representation provides strategic context to the model.

- **Sender and Receiver Embeddings:** Each player (country) is assigned a learnable embedding vector. These embeddings help the model account for player-specific strategies and interpersonal dynamics.

- **Conversational History Encoder:** The last 10 messages from the sender to the receiver in that game are encoded using BERT and passed through a transformer encoder with positional encodings. This models the evolving conversational tone and interaction patterns between players over time.

- **Fusion Attention (Learnable Query Attention Pooling):** To replace static concatenation of modality features, we treat the three encoded vectors — from the text, game state, and history — as a short sequence and apply multi-head self-attention to allow inter-modality interactions. A learnable query vector then performs attention pooling over these outputs, producing a single fused representation. This dynamic weighting allows the model to contextually prioritize the most relevant modality, leading to more robust deception classification.

- **Classification Head:** The final fused representation is passed through a feedforward classifier to predict the likelihood that the message is deceptive.

This architectural enhancement enables the model to better model cross-modal dependencies and adaptively weigh different sources of information, outperforming simple late fusion(concatenation) and making it more suited for the complexities of strategic deception in Diplomacy.

3.2 Graph Attention Network with Linguistic Cues

Our final model, **LieDetectorGAT**, is a graph-based neural architecture designed to detect deception in the game of Diplomacy. It explicitly models the interaction graph between players and leverages both semantic and strategic cues embedded within message exchanges. This version builds upon our earlier implementation with key improvements that ensure safe and leakage-free feature design.

3.2.1 Game-Level Graph Construction

Each Diplomacy game is modeled as a directed graph $G = (V, E)$, where:

- Nodes V represent unique players (countries) in the game, initialized with 7-dimensional one-hot encodings to indicate player identity.
- Directed edges $(u, v) \in E$ represent individual messages sent from player u to player v.

Messages are sorted chronologically by their `absolute_message_index`, ensuring that only past information is available for computing features.

3.2.2 Liguistically Enriched Edge Features

For each edge (i.e., message), we construct a feature vector using the following components:

- **BERT Embeddings:** The semantic content of the message is encoded using the `[CLS]` token from a frozen `BERT-base-uncased` model, producing a 768-dimensional contextual representation.
- **Deception Lexicon Features:** A 10-dimensional vector is computed using word counts from a psychological deception lexicon (2015 Diplomacy Lexicon), covering categories such as *claims*, *justifications*, and *temporal cues*.
- **Pre-Message Strategic Metadata:**
  - **Running Lie Count:** The number of lies told by the sender up to (but not including) the current message. This serves as a proxy for deception propensity without inducing label leakage.
  - **Average Pre-Message Score Delta:** The average change in the sender's power (score delta) from earlier messages only, capturing long-term power dynamics without using future knowledge.
  - **Game Context Features:** Season (categorical), year bucket (discretized), and normalized message index are included to capture message timing within the game context.

$$\underbrace{BERT}_{768}; \underbrace{Lexicon}_{10}; \underbrace{LieCount}_{1}; \underbrace{ScoreDelta}_{1}; \underbrace{Meta}_{3}$$

These features are concatenated to form a 783-dimensional vector per edge, which is passed through an edge encoder MLP.

3.2.3 Graph Attention and Edge-Aware Learning

We employ a multi-layer, multi-head Graph Attention Network (GATv2Conv) to process the message graph. Node embeddings are updated iteratively with attention weights modulated by edge features:

$$h_v^{(l+1)} = GATv2Conv(h_u^{(l)}, h_v^{(l)}, e_{uv})$$

This mechanism allows the model to capture not just who talks to whom, but also the salience of what is being said based on deception-related features.

3.2.4 Deception Classification

For each message (edge), we predict a deception label using a 2-layer MLP that takes as input:

- Sender node embedding h_u
- Receiver node embedding h_v
- Edge feature embedding e_{uv}

The MLP outputs a scalar logit, which is converted to a deception probability using the sigmoid function. We use `BCEWithLogitsLoss` with a class-weighted positive loss to mitigate the heavy class imbalance ($\sim$5% lies).

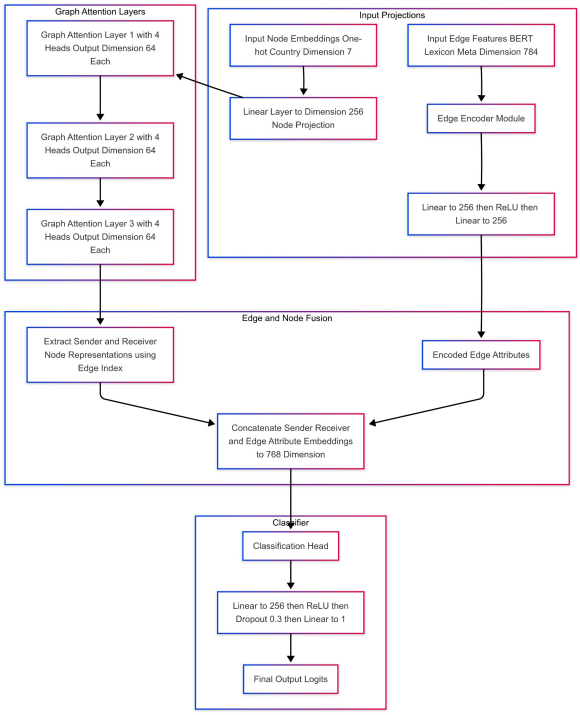

Figure 2: Overview of our graph-based LieDetectorGAT model for deception detection.

4 Dataset, Experimental Setup, and Results/Findings

4.1 Dataset Description

The dataset used in this work is sourced from the QANTA Diplomacy Deception Detection task, designed 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.

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**

This class imbalance presents a challenge in training robust models and highlights the importance of using evaluation metrics such as *Lie-F1* and *Macro F1*, which are sensitive to minority class performance.

4.2 Preprocessing

Before feeding the data into our model, we apply a series of preprocessing steps to clean and enrich the message content and structure. These steps help reduce noise and improve the quality of input data.

1. **Text Cleaning**
   Each message undergoes the following transformations:
   - **Lowercasing**: All text is converted to lowercase to ensure uniformity.
   - **URL Removal**: Any links (e.g., starting with `http://`, `https://`, or `www.`) are removed as they do not contribute to the linguistic features.
   - **Special Character and Emoji Removal**: All characters that are not alphanumeric or whitespace are stripped out.
   - **Whitespace Normalization**: Multiple consecutive whitespaces are collapsed into a single space, and leading/trailing whitespaces are trimmed.
2. **Conversation History Augmentation**
   To provide context to each message, we incorporate the history of past communications between the same sender and receiver within the same game. Specifically:
   - For each message, we identify its game id, the sender and the receiver
   - Using this combination as a key—(game_id, sender, receiver)—we track the chronological list of cleaned messages exchanged between this pair in that particular game.
   - For every new message, we retrieve the last k messages (with k set to 10) that the same sender has previously sent to the same receiver in that game.
   - This list of up to k messages is attached as the `history` field to the current message, providing prior conversational context.
   - After assigning the history, the current cleaned message is added to the sender-receiver's conversation history for future reference.

4.3 Experimental Setup

4.3.1 His-Attention

We experimented with a series of attention-based architectures that use BERT embeddings for the current message, game state encodings, and varying forms of contextual message history. All models included game state features throughout. The models differ in how they incorporate context and fuse multimodal information:

- **Model 1 – No Context (Concat Fusion):** Used only the BERT embedding of the current message and game state features, fused via concatenation.
- **Model 2 – 5-Message Context + BiLSTM + Mean Attention:** Incorporated the last five sender-receiver messages, processed via BiLSTM. Fused with game state using multi-head attention and mean pooling. Achieved **53.8** macro F1.

- **Model 3 – 10-Message Context + Transformer + Query Attention:** Used last ten messages with a Transformer encoder for history. Fused using multi-head attention with query pooling. Best performance with **57.38** macro F1.

Training Configuration and hyperparameters: The best model was trained for 10 epochs using a batch size of 16 and Adam optimizer with learning rate 1×10^{-5}. A weighted binary cross-entropy loss was used to handle class imbalance (`pos_weight` $= 21.22$). An adaptive learning rate scheduler was applied with patience of 1 and decay factor 0.5. Gradients were clipped with a max norm of 1.0. Weights for the best model are uploaded here.

4.3.2 LieDetectorGAT

Training Configuration and hyperparameters: LieDetectorGAT is trained for 20 epochs using the AdamW optimizer with a fixed learning rate of 3×10^{-4}. To address the class imbalance (only $\sim$5% of messages are labeled as lies), we use a `BCEWithLogitsLoss` objective with a positive class weighting factor computed from the ratio of true to false labels. Gradient clipping is applied with a maximum norm of 1.0 to improve training stability. Each game instance is represented as a graph, and the model processes one graph per batch (i.e., batch size of 1) using a PyTorch Geometric `DataLoader`. Moreover, we dont shuffle the datasets and sort them in order during preprocessing to preserve the conversational nature of the data. We use the deception lexiconn provided under utils in the dataset for linguistic cues construction. Weights for the best model are uploaded here.

5 Results

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**

Table 2: Performance comparison of attention-based architectures

	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.

6 Error Analysis

6.1 HiS-Attention

6.1.1 Quantitative Analysis

The best-performing architecture incorporated 10 prior messages using Transformer encoding and query-based attention. It achieved a macro F1

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.

score of **57.38** and lie-F1 of **21.98**, outperforming all the baselie models on both metrics. However, the model demonstrates a stark class imbalance in performance,despite using a weighted loss function.moreover, Despite the use of a weighted loss, the model struggles with the minority class, failing to learn subtle cues of deception effectively.

6.1.2 Qualitative Analysis

A review of false positives and false negatives shows that the model often misinterprets diplomatic vagueness and cautious planning as lies. Many truthful but strategically worded messages are flagged as deceptive. In contrast, lies that mimic cooperative or neutral tones are often missed.

Speaker-wise, nations like Germany and France exhibit higher misclassification rates (21.8% and 16.3%, respectively), likely due to complex or indirect language styles. Conversely, Italy and Austria have lower error rates, suggesting clearer or more consistent communication.

Overall, the model appears to overfit to dominant truthful patterns, necessitating more refined modeling of deception cues, possibly with speaker-specific conditioning.

6.2 LieDetectorGAT

6.2.1 Quantitative Analysis

Table 6 presents the confusion matrix for LIEDETECTORGAT. Overall, the model achieves an accuracy of **81.3%**. Notably, it identifies a larger portion of the actual *Lie* class than our best HIS-ATTENTION model (higher recall), but at the cost of more false positives (lower precision).

Table 7 shows the classification report. While the *Truthful* class is predicted with high accuracy (**F1** = 0.893), performance on the *Lie* class remains challenging. Specifically, LIEDETECTORGAT attains a better recall (0.388) relative to the attention-based model but lower precision (0.203), indicating a tendency to over-flag messages as deceptive.

6.2.2 Qualitative Analysis

Speaker-wise Error Rates. Table 8 shows that **Germany** has the highest error rate (32.33%), likely reflecting the model's struggles with more intricate or multi-turn strategic discourse. **Italy** has the lowest error rate (5.15%), possibly due to more consistent or simpler messaging patterns. These findings align with our HIS-ATTENTION speaker-level trends, suggesting that a speaker's rhetorical style can shape detection difficulty.

Similar to the HIS-ATTENTION model, overly detailed or cautious phrasing can be mistaken for deceit; conversely, succinct and friendly lies sometimes appear truthful. Overall, while LIEDETECTORGAT improves recall for deceptive statements by leveraging a more structured, graph-based view of inter-player interactions, it pays the price in lower precision. This reflects a recurring pattern in deception detection systems where more aggressive lie identification often leads to overestimating deceptive intent.

7 Limitations

Despite the strong performance of our models—HIS-ATTENTION and LIEDETECTORGAT—in deception detection within Diplomacy dialogues, several limitations remain. The models often fail to detect subtle deception, particularly when lies are embedded in cooperative or indirect language that lacks overt deceptive cues. Additionally, the reliance on a fixed deception lexicon constrains the models' adaptability to evolving or context-specific language use, leading to both false positives and false negatives. Finally, the use of a fixed contextual window of ten prior messages limits the ability to capture long-range dependencies, which are often critical in the strategic flow of diplomatic conversations.

8 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 adver-

sarial 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.

References

Lisa Fu. 2019. Deception detection in online mafia game interactions. Technical report, Stanford University.

Yihao Guo, Longye Qiao, Zhixiong Yang, and Jianping Zhang. 2023. Fake news detection: Extendable to global heterogeneous graph attention. *Tsinghua Science and Technology*, 28(5):850–862.

Loukas Ilias, Felix Soldner, and Bennett Kleinberg. 2022. Explainable verbal deception detection using transformers. *arXiv preprint arXiv:2210.03080*.

Srijan Kumar, Chongyang Bai, V.S. Subrahmanian, and Jure Leskovec. 2021. Deception detection in group video conversations using dynamic interaction networks. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 15, pages 319–330.

Riccardo Loconte, Chiara Battaglini, Stéphanie Maldera, and Pietro Pietrini. 2025. Detecting deception through linguistic cues: From reality monitoring to natural language processing. *Journal of Language and Social Psychology*, 44(1):3–24.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3811–3824.

Wichayaporn Wongkamjan, Yanze Wang, Feng Gu, Denis Peskoff, Jonathan K. Kummerfeld, Jonathan May, and Jordan Lee Boyd-Graber. 2025. Should i trust you? detecting deception in negotiations using counterfactual rl. *arXiv preprint arXiv:2502.12436*.

Fan Xu, Lei Zeng, Qi Huang, Keyu Yan, Mingwen Wang, and Victor S. Sheng. 2024. Hierarchical graph attention networks for multi-modal rumor detection. *Neurocomputing*, 569:127112.

Shayi Zhang. 2012. Lingcues – a linguistic cues software tool for research in text-based automatic deception detection. Master of science thesis, University of Georgia, Athens, Georgia. Under the direction of Michael A. Covington.