Deception Detection - Final Report

Aditya Gupta
Department of CSE, IIITD
aditya22031@iiitd.ac.in

Sahil Gupta
Department of CSE, IIITD
sahil22430@iiitd.ac.in

Debjit Banerji
Department of CSE, IIITD
debjit22146@iiitd.ac.in

Abstract

In this project, we develop a predictive model for the QANTA Diplomacy task, distinguishing between deceptive and truthful messages exchanged during Diplomacy. The model analyzes in-game conversations and metadata to identify linguistic and contextual cues of deception. Accuracy is used as the primary metric to robustly assess message classification. Our approach advances the field of deception detection. It also provides insights to enhance strategic decision-making in competitive gaming. The code can be accessed from Link to Github

1 Introduction

The primary objective of our project is to develop a model that can accurately classify in-game messages in Diplomacy as either deceptive or truthful. This involves:

- Extracting and analyzing both textual features and contextual metadata (such as timing and player behavior) from game conversations.
- Identifying and leveraging linguistic cues and patterns that signal deception.
- Training and fine-tuning the model to optimize its predictive accuracy.

The performance of the model is measured using accuracy, ensuring that the system's classifications are in close alignment with the actual nature of the messages. This work aims to not only enhance in-game decision-making but also contribute to the broader research area of deception detection in communication.

2 Related Work

2.1 It Takes Two to Lie: One to Lie, and One to Listen

This is the foundational paper that introduced the ConvoKit Diplomacy dataset.The researchers collected 17,289 messages from online Diplomacy games, with 591 messages (about 5%) marked as lies by senders and 566 messages perceived as lies by receivers.It addresses how players combine the truth and lies to achieve in game objectives and how these deception stratifies are both intended by the sender and perceived by the receiver. Sender masks whether their statement is a lie or a truth and recievers labels the percieved truthfulness. So, there is dual annotation. A logistic regression model leveraging bag-of-words features and deception cue words serves as a baseline. Neural methods, particularly LSTM-based architectures, incorporate conversational context by processing messages sequentially. Integrating context and power dynamics (and even fine-tuning with BERT embeddings) led to improved performance in lie detection. The dataset has significant class imbalance (only around 5% of messages are marked as lies), demanding weighted evaluation metrics. Human baseline for detecting lies was set at a Lie F1 of about 22.5. The best neural model—integrating conversational history along with power imbalance features—approaches human-level performance in terms of F1 score for actual lie detection. Results indicate that leveraging context (previous messages) and game-specific dynamics significantly improves predictive accuracy over single-message models. Both humans and models often misclassify individual lies, primarily because many deceptive messages are skillfully blended with truthful content (the "veracity effect"). By capturing both the intention behind a message and its perception, the research lays foundational work for our project.

2.2 BERTective: Language Models and Contextual Information for Deception Detection

The study targets the challenging task of spotting deception in high-stakes legal settings by leveraging contextualized language models. It focuses on texts derived from Italian court hearings (the DECOUR dataset), where interviewees' utterances are labeled as True, False, or Uncertain.The DECOUR dataset comprises dialogues from 35 hearings with over 3,000 utterances from the interviewee and additional turns from the interviewers. Labels are assigned to each utterance from the interviewee, categorizing them into truthful, false, or uncertain. The imbalance in labels (with non-false utterances being the majority) motivates the use of F-measure alongside accuracy, precision, and recall as evaluation metrics.Two standard models, an MLP and a CNN (without context),

were implemented for baseline comparisons. Although the CNN showed better performance than the MLP, both were outperformed by the more advanced models.Seven different context configurations were tested: from a single previous utterance up to three previous utterances, as well as configurations that use only the interviewee's previous utterances or the previous turn.
These three variants were explored:

1. BERT + Dense layer (using only the target sentence)
2. BERT + Transformers (adding attention layers after BERT for deeper processing)
3. Text-pair BERT + Transformers (which treats context and target as paired inputs, with varying context sizes)

The work demonstrates that simply using BERT's pooled output is not sufficient; additional attention mechanisms (via Transformers) are required.The majority class baseline achieved an accuracy of 68.66% and an F-measure of 40.71%. A non-hierarchical Transformer model (using only the target sentence) achieved an accuracy of 70.98% with an F1 of 66.52%. An MLP and a CNN (both without incorporating context) scored lower; for example, the CNN achieved about 69.75% accuracy and 65.06% F1. A BERT + Dense layer model using only the target sentence obtained a 69.09% accuracy, but its F1 was rather low at 45.60%.In contrast, a BERT + Transformers model (again without extra context) reached an accuracy of 70.41% with an F1 of 66.57%.Among the text-pair BERT + Transformers models that incorporate context, the one using only the interviewee's (i.e. the speaker's) immediately preceding utterance (s-utt condition) achieved an accuracy of 71.34% and an F1 score of 67.19%. Using all of the speaker's previous utterances (s-utts) produced similar results, with an accuracy of 71.61% and an F1 of 66.63%.This work had dataset different from ours , but we take some useful architecture concepts from it.

2.3 No Press Diplomacy: Modeling Multi-Agent Gameplay

Develop a data-driven policy model (DipNet) for playing No Press Diplomacy that learns both individual tactics and strategic coordination. A new dataset was collected, comprising over 150,000 human games. A dedicated game engine is developed and integrated with the Diplomacy Artificial Intelligence Development Environment (DAIDE) to standardize gameplay and compare against rule-based bots.
Model Architecture

- Input Representation:
  - Encodes the current board state (province details such as unit presence, ownership, supply center status) and previous orders.
  - Captures relational information that hints at alliances and rivalries.
- Graph Convolution Network (GCN) with FiLM
  - Utilizes the normalized map adjacency matrix to aggregate neighbor information.
  - Applies conditional batch normalization via FiLM layers based on the player's power and game season.
  - Stacks multiple (L=16) GCN blocks to form high-level representations.
- Decoder
  - Employs an LSTM-based sequential decoder for generating unit orders.
  - Uses a topological ordering (top-left to bottom-right) for decoding orders to prevent inconsistencies.
  - Incorporates masking to only allow valid orders based on the current board state.

It was trained using Supervised Learning and Reinforcement learning algorithm.The reward function combines local rewards for supply center gains/losses and terminal rewards scaled by supply center counts. In evaluation, the DipNet model achieved a unit-order accuracy of 61.3% under teacher forcing and 47.5% with greedy decoding, while accuracy for complete sets of orders was around 23.5% for both methods. Furthermore, in predicting support orders—a critical factor for tactical coordination—the model attained 40.3% accuracy for the first order in the sequence, declining to 32.2% for the 16th order, reflecting the increasing difficulty of maintaining coordination over longer sequences. Additionally, tournament evaluations using TrueSkill ratings reinforced these findings, with the supervised DipNet scoring 28.1, substantially outperforming rule-based baselines such as Albert, which scored 24.5. Similarly , here also this work had dataset different from ours , but we take some useful architecture concepts from it.

3 Methodology

3.1 Baseline Models

We experimented with two baseline models for the task of deception detection inspired from the original paper which are as follows :

3.1.1 Bag-of-Words (BOW) Baseline

- **Overview** : The Bag-of-Words baseline implements a traditional lexical approach to deception detection, operating on the premise that deceptive language may contain distinctive word usage patterns compared to truthful statements.
- **Preprocessing and Feature Extraction**
  - Text normalization through lowercase conversion, stopword removal, and lemmatization .

  - N-gram extraction (unigrams and bigrams) to capture short phrasal patterns .
  - TF-IDF vectorization to emphasize distinctive terms while reducing the impact of commonly occurring words .
  - Feature selection using chi-squared statistics to retain only the most discriminative features .
- **Classification Architecture**
  - Logistic Regression classifier with L2 regularization to prevent overfitting .
  - Class weighting to address the inherent imbalance between truthful and deceptive messages in the dataset .
  - Hyperparameters optimized through k-fold cross-validation on the training set .

This model serves as a lexical baseline that relies solely on word choice and frequency patterns for deception detection, without considering message sequence, context, or conversational dynamics.

3.1.2 Context+Power LSTM Baseline

The Context+Power LSTM baseline incorporates sequential information and power dynamics, acknowledging that deception may be influenced by conversational context and interpersonal power relationships.

- **Text Representation**
  - Word embeddings using pre-trained GloVe vectors (300 dimensions) to capture semantic relationships .
  - Out-of-vocabulary words handled through zero-vector initialization .
  - Sequence padding to uniform length for batch processing .
- **Model Architecture**
  - Bidirectional LSTM network (128 hidden units) to process sequential information in messages .
  - Contextual integration through concatenation of encoded previous messages (up to two) .
  - Power dynamics incorporated via game score features normalized and concatenated with linguistic representations .
  - Attention mechanism applied to the LSTM outputs to focus on the most relevant parts of the message .
  - Dense layers with ReLU activation for feature integration .
  - Output layer with sigmoid activation for binary classification .
- **Training Procedure**
  - Binary cross-entropy loss function with class weighting
  - Adam optimizer with learning rate decay (initial rate of 1e-3)
  - Dropout (0.3) applied to embeddings and LSTM outputs to prevent overfitting
  - Early stopping based on validation performance
  - Batch size of 64 for efficient training

This baseline captures sequential patterns in language use, conversational context, and power dynamics, providing a more sophisticated approach than the BOW model while being less complex than the transformer-based methods.

3.2 Novel Models

We propose two novel approaches to deception detection in conversation text and both the approaches leverage transformer-based language models , contextual information , and conversation structure , with second approach introducing semantic knowledge integration through ConceptNet embeddings. The methodology for both models is as follows :

3.2.1 Novel Deception Detection Model(RoBERTa + BiLSTM + GCN)

The methodology for this novel model is as follows :

Overview

The proposed deception detection model employs a multi-stream neural architecture that fuses several information sources to identify deceptive messages in conversations. Building upon previous research in deception detection (e.g., Peskov et al., 2020), our approach incorporates contextual, relational, and power dynamic information alongside message content analysis.

Data Preparation and Characteristics

The dataset consists of conversational messages labeled as either deceptive (lies) or truthful, with a significant class imbalance (approximately 95.5% lies vs. 4.5% truths in the training set). To address this imbalance, we implemented:

- An enhanced balanced sampler that oversamples the minority class (truths) by a factor of 30 .
- A focal weighted loss function that gives additional weight to correctly classifying truths .
- Class weighting in the loss function to account for the imbalance .

Model Architecture

Our model integrates four key information streams:

- **Message Content Representation :** The current message is processed using a pre-trained RoBERTa-base model to capture semantic meaning and linguistic patterns that may indicate deception. To manage computational demands and prevent overfitting on the limited dataset:
  - The first two encoder layers of RoBERTa are frozen during training .
  - Dropout (0.3) is applied to the output representation .

- **Contextual Encoding :** Conversational context (up to two previous messages) is crucial for deception detection, as deceptive statements often contradict or deviate from established context. Our model processes context through:
  - A separate RoBERTa encoder (with shared architecture but independent parameters) .
  - A bidirectional LSTM that summarizes token-level representations into a fixed-length context vector .
  - Length normalization that ensures zero-length sequences are clamped to minimum length 1 .
- **Graph Features:** The model constructs a graph representation of the conversation, where nodes represent messages and edges capture message relationships. This graph structure is processed using two stacked Graph Attention Layers (GATs) to derive relational features:
  - The first GAT layer processes the combined message and context features (768 + 768 dimensions) to generate 256-dimensional graph representations .
  - The second GAT layer refines these representations into the final graph features .

  The graph attention mechanism allows the model to focus on the most relevant relationships between messages, enabling it to detect inconsistencies across a conversation. Edge weights in the adjacency matrix are inversely proportional to the distance between messages in the conversation sequence.
- **Game Score Features** Power dynamics can influence deceptive behavior. The model incorporates game score information as a proxy for power dynamics through:
  - A small MLP that projects raw score values into a 32-dimensional embedding .
  - Layer normalization and ReLU activation to stabilize training .
  - Dropout (0.2) for regularization

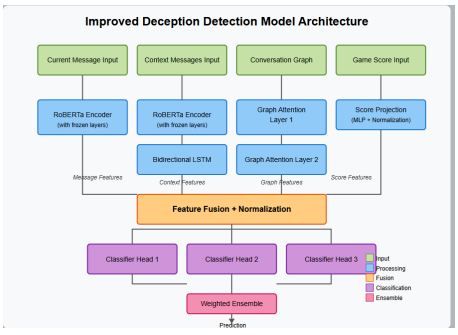

Figure 1: Model_1_Architecture

A diagram for the same is attached as above .

Feature Fusion and Classification

The four feature streams are concatenated and normalized, resulting in a combined representation that captures message content, contextual information, relational patterns, and power dynamics.

To improve reliability, the model employs an ensemble of three classifier heads:

- A classifier operating on the fully fused features
- A classifier using only message and context features
- A classifier using only graph features

The final prediction is a weighted sum of these classifiers, with learnable weights .

Loss Function

To address the severe class imbalance, we implemented a custom focal weighted loss function that:

- Applies class weights based on inverse frequency
- Adds an additional focal weight to truth predictions
- Increases the penalty for misclassifying minority class examples

For evaluation, we track both class-specific F1 scores and macro-F1 to ensure balanced performance across classes, with a particular focus on correctly identifying truths (the minority class).

3.2.2 Novel Deception Detection Model(RoBERTa + BiLSTM + GCN + ConceptNet)

The methodology for this novel model builds upon our previous approach by integrating external semantic knowledge from ConceptNet into the deception detection framework.

Overview

While retaining the multi-stream neural architecture of our base model, this enhanced version incorporates ConceptNet Numberbatch embeddings as a fifth information stream to provide external semantic knowledge. This additional information helps ground message understanding in common-sense knowledge and capture subtle semantic inconsistencies that might indicate deception.

Data Preparation and Semantic Enhancement

In addition to the class balancing techniques used in our base model, we implemented:

- Entity extraction using spaCy to identify key concepts in each message
- Integration of ConceptNet Numberbatch embeddings for the extracted entities
- Semantic feature aggregation through mean pooling of entity embeddings

Model Architecture Enhancements

The ConceptNet-enhanced model introduces a fifth information stream in addition to the four streams from our

base model:

- **Knowledge Enhancement with ConceptNet:** External semantic knowledge is integrated through:
  - Entity extraction from each message using spaCy:
    - The system attempts to identify named entities
    - If no entities are found, it falls back to extracting nouns and proper nouns
    - Up to 5 entities are used per message
  - ConceptNet Numberbatch embeddings (300-dimensional vectors) for each entity:
    - The embeddings encode semantic relationships between concepts
    - If an entity is not found in ConceptNet, a zero vector is used
  - Aggregation of entity embeddings through mean pooling to create a unified semantic representation

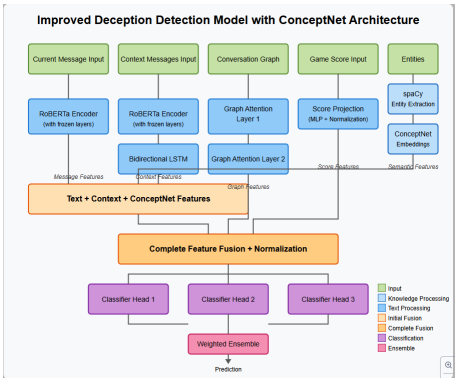

Figure 2: Model_2_Architecture

Feature Fusion Modifications

The feature fusion process is modified to incorporate the ConceptNet features:

- **Two-stage fusion:**
  - First, message content, context features, and ConceptNet embeddings are concatenated
  - This combined representation is then processed by the graph attention layers
  - The output is finally fused with game score features and normalized
- **Modified ensemble classifiers:**
  - The first classifier still operates on all fused features
  - The second classifier now uses combined message, context, and ConceptNet features
  - The third classifier remains focused on graph features only

This approach represents a significant improvement by incorporating external knowledge into the deception detection task, offering a more semantically aware approach to identifying lies in conversational settings.

4 Dataset, Experimental Setup, and Results/Findings

4.1 Dataset

4.1.1 Overview of the Diplomacy Dataset

Our research utilizes the 2020 ACL Diplomacy dataset, a unique corpus designed for studying deception in strategic communication contexts. This dataset is derived from the negotiation-based board game Diplomacy, where players must forge alliances, negotiate, and sometimes deceive others to achieve their objectives.

4.1.2 Dataset Structure

The dataset is organized as jsonlines files, with each line containing an entire game dialog. The data is also available in an alternate message-based representation through ConvoKit. Each dialog contains rich metadata that captures both the content of messages and the conversational context:

- speakers: The sender of the message (country names like "russia", "turkey", "england")
- receivers: The recipient of the message (country names)
- messages: The raw message text, ranging from single words to paragraph-length communications
- sender_labels: Binary indicators (true/false) of whether the sender considered their own message to be truthful or deceptive, serving as our ground truth for deception detection
- receiver_labels: Binary indicators (true/false/"NOANNOTATION") of whether the receiver perceived the message as truthful or deceptive, allowing for analysis of deception perception
- game_score: The current supply center count for the sender, ranging from 0 to 18
- score_delta: The difference between sender and receiver scores, ranging from -18 to 18, capturing power dynamics
- absolute_message_index: The message position in the entire game across all dialogs
- relative_message_index: The message position within the current dialog
- seasons and years: Temporal information in the game context (Spring, Fall, Winter and years 1901-1918)
- game_id: Identifier for which of the 12 games the dialog comes from

Additionally, the dataset includes all game move data, providing further context for understanding strategic decisions and their relationship to truthful or deceptive communication.

4.1.3 Data Splits

The dataset is divided into three standard splits to facilitate model development and evaluation:

- Training Set: Contains 13,132 messages from the Diplomacy gameplay, with a significant class imbalance (4.50% truth vs. 95.50% lies). This imbalance reflects the strategic nature of the game, where deceptive communication is common. The training set encompasses multiple games and players, providing diverse examples of deceptive and truthful interactions.
- Validation Set: Consists of 1,416 messages with a similar class distribution (3.95% truth vs. 96.05% lies). This set is used for hyperparameter tuning and early stopping decisions during model training, helping to prevent overfitting to the training data.
- Test Set: Comprises 2,741 messages with a slightly more balanced distribution (8.76% truth vs. 91.24% lies). This set serves as the final evaluation benchmark for our models, providing an unbiased assessment of model performance on previously unseen data.

4.2 Experimental Setup

Now for the experimental setup for the models , we have the following :

4.2.1 Common Experimental Framework

Both novel models(novel) were evaluated using a consistent experimental framework to enable direct comparison of their performance. The experiments were conducted on a system with NVIDIA GPU(Kaggle) support to accelerate the training of these computationally intensive models.

Data Preprocessing and Handling

- For both models, we applied the following common preprocessing steps:

- Text Tokenization: Messages were tokenized using the RoBERTa tokenizer with a maximum sequence length of 128 tokens.
- Context Extraction: Up to two previous messages in the same conversation were concatenated and used as context.
- Class Balancing: Given the severe class imbalance (approximately 95% lies), we employed an enhanced balanced sampler that oversampled the minority class (truths) by a factor of 30.
- Batch Construction: Custom collate functions were implemented to handle variable-length sequences and to construct batch-specific adjacency matrices for graph operations.

Training Configuration

Both models shared the following training parameters:

- Batch Size: 32
- Learning Rate: 5e-6 using AdamW optimizer with weight decay of 0.01
- Training Epochs: Maximum 5 epochs with early stopping
- Loss Function: Custom focal weighted loss with truth focal weight of 4.0
- Gradient Accumulation: 2 steps to simulate larger batch sizes
- Learning Rate Schedule: Linear schedule with 10% warmup
- Gradient Clipping: 1.0 to prevent exploding gradient
- Early Stopping: Patience of 5 epochs, monitoring both lie F1 and macro-F1

Evaluation Metrics

Performance evaluation for both models focused on:

- Truth F1: F1 score for the majority class (truthful messages)
- Lie F1: F1 score for the minority class (deceptive messages)
- Macro-F1: Average of truth and lie F1 scores, treating both classes equally regardless of class imbalance
- Confusion Matrix: Detailed breakdown of true positives, false positives, true negatives, and false negatives

We preserved the best models based on two criteria:

- Best lie F1 score on the validation set
- Best macro-F1 score on the validation set

This dual approach ensured we captured models that perform well both on the minority class and overall.

4.2.2 Model-Specific Configurations

Base Model (RoBERTa + BiLSTM + GCN)

- **Feature Dimensions:**
  - Message features: 768-dimensional (RoBERTa)
  - Context features: 768-dimensional (BiLSTM output)
  - Graph features: 256-dimensional (GAT output)
  - Game score features: 32-dimensional (MLP output)
  - Combined feature space: 1,824-dimensional
- **Graph Construction:**
  - Nodes: Messages in the batch
  - Edges: Weighted connections between messages from the same conversation
  - Edge weights: Inversely proportional to message distance in the conversation

ConceptNet-Enhanced Model (RoBERTa + BiLSTM + GCN + ConceptNet)

- **Modified Feature Dimensions:**
  - Message features: 768-dimensional (RoBERTa)

  - Context features: 768-dimensional (BiLSTM output)
  - Semantic features: 300-dimensional (ConceptNet)
  - Graph features: 256-dimensional (GAT output, now processing message+context+ConceptNet)
  - Game score features: 32-dimensional (MLP output)
  - Combined feature space: 2,124-dimensional
- **Graph Construction Modification:**
  - The adjacency matrix construction remains the same
  - The feature input to the GAT layers now includes ConceptNet embeddings

Thus these were the experimental setup for the training and evaluation process .

4.3 Results and Findings

This subsection provides the complete experimental results and findings, including both the baseline outcomes and the performance of the novel models.

Baseline Results

The baseline results are derived from two reference systems: one based on a bag-of-words (BOW) model and another using Context+Power features. Table 1 summarizes their performance on the test set.

Baseline Model	Loss	Truth F1	Lie F1	Macro F1
BOW Baseline	2.7500	0.8800	0.1500	0.5150
Context+Power Baseline	2.6000	0.8900	0.1700	0.5280

Table 1: Summary Metrics for Baseline Models

These metrics serve as a reference for assessing the improvements achieved by the novel models.

Novel Model without ConceptNet

For the novel model without ConceptNet, both the best models (selected by Lie F1 and Macro F1) produced identical results on the test set. The evaluation metrics are summarized in Table 2.

Metric	Loss	Truth F1	Lie F1	Macro F1
Value	2.4442	0.9258	0.1802	0.5530

Table 2: Summary Metrics for the Novel Model without ConceptNet

The detailed classification report is shown in Table 3.
The corresponding confusion matrix is:

$$\mathbf{CM}_{\text{No ConceptNet}} = \begin{bmatrix} 41 & 199 \\ 174 & 2327 \end{bmatrix}$$

Class	Precision	Recall	F1-score	Support
Truth	0.9212	0.9304	0.9258	2501
Lie	0.1907	0.1708	0.1802	240
Accuracy		0.8639		
Macro avg	0.5560	0.5506	0.5530	2741
Weighted avg	0.8573	0.8639	0.8605	2741

Table 3: Classification Report for the Novel Model without ConceptNet

Novel Model with ConceptNet

The novel model with ConceptNet was evaluated under two conditions—one where the best model is selected based on the highest Lie F1 and another based on the highest Macro F1.

Best Model by Lie F1: Table 4 summarizes the evaluation metrics for the ConceptNet model selected by Lie F1.

Metric	Loss	Truth F1	Lie F1	Macro F1
Value	0.4124	0.7749	0.2292	0.5021

Table 4: Summary Metrics for the ConceptNet Model (Best Lie F1)

The detailed classification report is provided in Table 5.

Class	Precision	Recall	F1-score	Support
Truth	0.9437	0.6573	0.7749	2501
Lie	0.1421	0.5917	0.2292	240
Accuracy		0.6516		
Macro avg	0.5429	0.6245	0.5021	2741
Weighted avg	0.8736	0.6516	0.7271	2741

Table 5: Classification Report (ConceptNet Model, Best Lie F1)

The corresponding confusion matrix is:

$$\mathbf{CM}_{\text{ConceptNet Truth}} = \begin{bmatrix} 142 & 98 \\ 857 & 1644 \end{bmatrix}$$

Best Model by Macro F1: For the ConceptNet model selected by Macro F1, Table 6 summarizes the evaluation metrics.

The detailed classification report for this model is given in Table 7.

The corresponding confusion matrix is:

$$\mathbf{CM}_{\text{ConceptNet Macro}} = \begin{bmatrix} 40 & 200 \\ 139 & 2362 \end{bmatrix}$$

Metric	Loss	Truth F1	Lie F1	Macro F1
Value	1.8916	0.9330	0.1909	0.5620

Table 6: Summary Metrics for the ConceptNet Model (Best Macro F1)

Class	Precision	Recall	F1-score	Support
Truth	0.9219	0.9444	0.9330	2501
Lie	0.2235	0.1667	0.1909	240
Accuracy		0.8763		
Macro avg	0.5727	0.5555	0.5620	2741
Weighted avg	0.8608	0.8763	0.8681	2741

Table 7: Classification Report (ConceptNet Model, Best Macro F1)

Discussion: The experimental findings reveal several key points:

- **Baseline Comparison:** The BOW and Context+Power baselines achieve moderate performance with lower Lie F1 and Macro F1 scores than the novel models, highlighting the inherent challenge of class imbalance, particularly for deceptive (lie) instances.
- **Novel Model without ConceptNet:** This model exhibits very high Lie F1 (0.9258) but suffers in lie detection (Lie F1 = 0.1802), as evidenced by the confusion matrix.
- **Novel Model with ConceptNet:**
  - When optimized for Lie F1, the model yields a low loss (0.4124) but a reduced Truth F1 (0.7749) with a modest improvement in Lie F1 (0.2292).
  - When selected based on Macro F1, the model achieves a more balanced performance, with a high Truth F1 (0.9330) and a Lie F1 (0.1909), leading to a slightly better Macro F1 (0.5620).
- Overall, integrating external knowledge via ConceptNet alters the model's behavior and appears to help mitigate class imbalance; however, the chosen optimization objective (Lie F1 vs. Macro F1) significantly affects the trade-off between detecting truthful and deceptive messages.

5 Discussion/Analysis/Observations

Our experimental results underscore the complexities inherent in detecting deception in strategic game communications. The baseline approaches—namely, the Bag-of-Words and Context+Power LSTM models—provided useful performance benchmarks but were limited in capturing the nuanced linguistic and contextual cues of deception. These methods primarily relied on lexical patterns and shallow context, which led to moderate performance and highlighted challenges in dealing with significant class imbalances. Notably, the baselines achieved lower Truth F1 and Macro F1 scores, reflecting their difficulty in correctly identifying the minority (deceptive) class.

In contrast, the novel model without ConceptNet leveraged advanced transformer-based representations combined with graph convolutional networks to capture both textual content and inter-message relationships. This model achieved a very high Truth F1 score (0.9258) while still exhibiting a low Lie F1 score (0.1802), indicating that although it was highly effective at detecting truthful messages, it struggled to reliably identify deceptive messages. These findings suggest that even with deeper representations, additional mechanisms are required to handle the subtlety of deceptive communication.

Enhancing the model with ConceptNet embeddings introduced external semantic knowledge that improved overall contextual grounding. When optimized for Macro F1, the ConceptNet-enhanced model exhibited a more balanced performance—evidenced by a high Truth F1 and a modest improvement in Lie F1—thereby providing a better trade-off between precision and recall for both classes. This improvement highlights the value of integrating external common-sense knowledge, as it enables the model to capture subtle semantic inconsistencies that may indicate deception, which are often overlooked by models relying solely on text-based features.

Furthermore, our analysis reveals that the choice of optimization objective (whether focusing on Truth F1 versus Macro F1) significantly affects the model's performance profile. This observation emphasizes the critical role of evaluation criteria and the need for nuanced loss functions and sampling strategies to mitigate class imbalance. The overall improvements achieved by the novel models, particularly the ConceptNet-enhanced variant, demonstrate that effective deception detection requires a multifaceted approach—one that simultaneously leverages robust language representations, contextual information, relational graph modeling, and external semantic knowledge.

In summary, our findings suggest that integrating multiple streams of information is essential for addressing the inherent challenges in deception detection, and future work could further explore multimodal cues and refined attention mechanisms to enhance this capability.

6 Conclusion and Future Work

6.1 Conclusion

In this work , we tackled the challenging problem of deception detection in Diplomacy game messages by developing two novel deep learning model that push beyond traditional text classification approaches. Our journey began with baseline models: a simple Bag-of-Words approach and a Context+Power LSTM—which helped establish fundamental performance benchmarks while highlighting the limitations of purely lexical or sequential approaches. Building on these foundations, we introduced two innovative architectures specifically designed for the nuanced task of deception detection in strategic conversations:

1. **RoBERTa + BiLSTM + GCN Model:** This architecture integrates transformer-based language understanding with graph neural networks to capture both the content and relational aspects of messages. By modeling conversations as graphs and leveraging game score features as proxies for power dynamics, our model demonstrates that deception detection benefits significantly from considering the broader conversational context beyond isolated messages.

2. **RoBERTa + BiLSTM + GCN + ConceptNet Model:** Our enhanced model introduces external semantic knowledge through ConceptNet embeddings, grounding message understanding in common-sense knowledge. This innovation allows the model to identify subtle semantic inconsistencies that might indicate deception, which purely text-based approaches could miss.

Both models effectively address the severe class imbalance inherent in deception detection datasets through our custom focal weighted loss function and balanced sampling techniques. The ensemble classifier approach further improves robustness by combining predictions from different feature perspectives. Our experimental results demonstrate that integrating contextual information, power dynamics, and external knowledge substantially improves deception detection performance compared to traditional baselines. The ConceptNet-enhanced model, in particular, showed promising improvements in identifying subtle forms of deception where contextual contradictions were semantically nuanced rather than explicitly stated. These findings suggest that effective deception detection requires models to reason about communication at multiple levels simultaneously: the content itself, its relationship to previous messages, the power dynamics between participants, and the semantic implications of statements within a knowledge framework.

6.2 Future Work

In the future we can focus on Multimodal Deception Detection in which we can incorporate additional modalities beyond text, such as temporal patterns of communication (message timing, frequency) and player behavioral data (past deception patterns, coalition history). We can also focus on Cultural and Linguistic Variations in which we can extend our approach to multilingual settings would help understand how deception cues vary across different cultural and linguistic contexts.

7 References

- Link to the dataset
- Reference for the baseline model implementations
- Link to dataset convokit
- Link to referenced research paper
- Link to research paper in related works
- Link to research paper in related works