Deception Detection - Project Proposal

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.

1 Problem Definition

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 High-Level Plan

The timeline for our project work is as follows:

Table 1: Combined Project Timeline

Pre Mid Evaluation Submission	
Timeline	**Tasks**
26 Feb – 9 March	Topic selection and finalization.
9 March – 23 March	Working on the baseline codes, understanding the research paper, and exploring the available GitHub codebase.
Post Mid Evaluation Submission	
Timeline	**Tasks**
24 Mar – 15 April	Developing novel models and comparing them with the baseline models.
13 April – 15 April	Final presentation and report preparation.

3 Approach

3.1 Bag Of Words Baseline

Deception detection uses a Bag-of-Words model combined with game-specific power features and logistic regression. It loads JSONL dialogues containing messages, deception annotations, and game score deltas. Messages are preprocessed by converting to lowercase, tokenizing via spaCy, and replacing numbers with a placeholder. Text is transformed into a bag-of-words representation using CountVectorizer, with binary indicators from power score thresholds appended. The logistic regression model is trained with balanced class weights, evaluated using accuracy and F1 scores, with feature importances visualized, and the model and vectorizer saved.

3.2 Context LSTM + Power Baseline

The Context LSTM + power baseline employs a Hierarchical LSTM that integrates the game score delta as a power feature. It starts with a 200-dimensional embedding layer followed by a bidirectional LSTM (hidden size 100) with max pooling to generate 200-dimensional message vectors. These vectors are processed by a unidirectional LSTM (hidden size 200) to capture conversational context, with the game score delta concatenated as a scalar before classification. Regularization is applied via a 0.2 dropout rate and a weighted loss (pos_weight=10) to address class imbalance. The model is optimized using Adam (learning rate 0.003, gradient clipping at max norm 1), trained for 80 epochs with batch size 1. Also, a ReduceLROnPlateau scheduler is used for optimization.

3.3 Future Actions and Additional Model Approaches

Future work will explore advanced transformer-based models (e.g., BERT, RoBERTa) and hybrid approaches that combine deep contextual embeddings with game-specific features. Attention mechanisms and Graph Neural Networks may also be considered to capture relational dynamics in conversations, while ensemble methods will be investigated to improve overall robustness.

Moreover, additional approaches could include incorporating unsupervised learning techniques for anomaly detection, leveraging multi-task learning to jointly model deception and related behavioral cues, and experimenting with reinforcement learning to adaptively refine deception detection strategies during gameplay. These directions aim to enhance model performance and provide deeper insights into the underlying patterns of deceptive communication.

4 References

- Link to the dataset
- Reference for the baseline model implementations
- Link to dataset convokit