Project Proposal: Deception Detection in Diplomacy

Satwik Garg **Shreyansh Srivastav** **Ashutosh Dwivedi**

satwik22461@iiitd.ac.in, shreyansh22485@iiitd.ac.in, ashutosh22116@iiitd.ac.in

Abstract

Deception is a strategic element in many negotiations and social interactions, particularly in competitive environments. This project aims to explore computational ML and DL models for detecting deception in given human-human text conversations from the game of Diplomacy. We implement various models ranging from simple baselines to advanced deep learning approaches, comparing their performance on detecting both actual lies (sender's intention) and suspected lies (receiver's perception).

1 Problem Definition

This project aims to develop and evaluate machine learning models to detect deceptive communication within Diplomacy game dialogues. The challenge is to classify messages as truthful or deceptive based on both sender intention and receiver perception.

2 Dataset

The dataset consists of messages each annotated with `game_id`, `sender`, `receiver`, timestamp, and a deception label. The deception label tells whether a message is a truth or lie based on the sender's intent.Suspected deception from the receiver's perspective is also captured.

3 High-Level Plan

- **Dataset Exploration:** Explore and analyze message content, game metadata, sender-receiver interactions, and deception labels.
- **Model Development:** Implemented logistic regression, deep learning models (LSTM, BERT), bag of words and context-aware architectures.
- **Evaluation and Optimization:** Measure performance using various metrics such as accuracy, Macro F1 score, Binary/Lie F1 score (F1 for deceptive class), Precision, and Recall.

4 Approach

- **Data Preprocessing:** Tokenize messages, clean text, and encode categorical metadata.
- **Model Implementation and Selection:** Train logistic regression, XGBoost, Bag of Words, Harbinger, LSTM, and LSTM with context-aware mechanisms with baseline model .
- **Evaluation:** Compare against human deception detection baselines and analyze feature importance.

  We evaluate models on two tasks:

  - **Sender Task (Actual Lie Detection):** Predict whether the sender intends to deceive.
  - **Receiver Task (Suspected Lie Detection):** Predict whether the receiver perceives a message as deceptive.

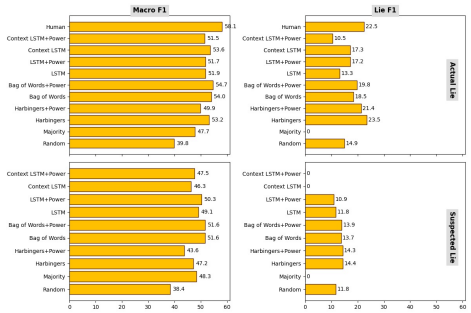

Figure 1: Baseline. Note Context LSTM is still in development

References

[1] Peskov, D., Cheng, B., Elgohary, A., Barrow, J., Danescu-Niculescu-Mizil, C., Boyd-Graber, J. (2020). It Takes Two to Lie: One to Lie, and One to Listen. (pp. 3811-3854).