# Project Mid-Evaluation: Deception Detection in Diplomacy

**Muthuraj Vairamuthu**
IIIT Delhi
muthuraj22307@iiitd.ac.in

**Syam**
IIIT Delhi
syam22528@iiitd.ac.in

**Pratham**
IIIT Delhi
pratham22374@iiitd.ac.in

## Abstract

This proposal outlines our team's effort to complete the Deception Detection in Diplomacy task, where our objective is to classify messages exchanges between players in the game diplomacy as deceptive or truthful. We re-implement the baseline models in "It Takes Two to Lie: One to Lie, and One to Listen." Moreover, we give a high-level plan to improve our predictions.

## 1 Problem Definition

In this game, Diplomacy and deception play a key role in winning. Players communicate with other players and get a tactical advantage by hiding their true intentions. Hence, the problem boils down to analyzing the player's conversation and determining whether a given message is deceptive or truthful. Our dataset consists of JSON-formatted files containing messages and metadata like(sender, receiver, season, year, game score,etc.)and annotations for deception. We have to effectively classify the messages as either deceptive(label 0) or truthful (label 1) whilst handling the class imbalances effectively(only $\sim$5% of the messages are labeled as lies).

## 2 High-Level Plan

- **Harbinger Model:** This is one of the two baseline models we explored; we re-implemented this logistic regression model using TF-IDF features and linguistic features like politeness cues, message structure, emotional tone, etc. The metadata includes sender/receiver identity, game score, and round details.
- **LSTM Model:**The second baseline model is a Py-Torch LSTM-based classifier that uses a tokenized and padded message sequence.The model is further trained with class weights to handle high dataset imbalances
- **Future Work:** 1.We aim to fine tune the present model currenlty context LSTM has the best results in accordance with the paper we ain to improve it by further fine tuning it with various other modular architecture,we are looking at mainly using Neural network based models for achieving better performance,to fine tune it and give better results.

## 3 Approach

Our approach involves re-implementing two baseline models from prior work: the Harbinger model and a context-aware LSTM. The Harbinger model combines TF-IDF features with linguistic features, along with metadata such as sender, receiver, and game phase. In contrast, the LSTM model processes tokenized messages with padded embeddings and incorporates class weights to address class imbalance. Both models are evaluated using accuracy and F1 scores, with future plans to either fine-tune these models or explore more modular architectures to achieve improved performance.

## Limitations

The dataset is highly imbalanced, with far more truthful messages than lies. Following the original paper *It Takes Two to Lie*, we focus on F1 score over accuracy, as it better handles class imbalance and penalizes missed lie predictions more effectively.

## Acknowledgments

This work is inspired by the paper *"It Takes Two to Lie: One to Lie, and One to Listen"*.