Deception Detection: QANTA Diplomacy

Aditya Sahai (MT24009), Shreyas Rajendra Gore (MT24087), Sharad Jain (MT24132)

Abstract

The Deception Detection project or QANTA Diplomacy project is a classification project where in-game messages between players are to be identified as either deceptive or truthful. Goal is to develop model that uses conversational text and metadata to predict deception in the message, with dataset of 17,249 messages. The model is then evaluated using accuracy. It has many implications in the field of NLP, decision making tasks and game theory.

1 Problem Definition

The project task is to develop model that detects deceptive messages in communication between players. The task involves analyzing text and metadata to classify deceptive or truthful messages.

2 High-Level Plan and Approach

1. **Preprocessing**
   - Messages are converted to lowercase, special characters-extra spaces are removed, vocabulary is made using frequent, embeddings and tokenization is used.
   - Dataset is split into training, validation and test sets, weighted samples are used to handle class imbalance in deception detection.
2. **Model Selection And Optimization**
   - Choosing a baseline: Initial baseline model used is Bidirectional LSTM with Attention Mechanism.
   - Features: Contextual text understanding, dynamic attention, integration of metadata features, focal loss for class imbalance.
   - Optimization: Hyperparameter tuning, regularizations (dropout, weights), loss functions (Focal loss, wighted corss entropy) and learning rate scheduling (ReduceLRonPlateau)

3 Baseline Folder Contents

- **Working Code**: Jupyter Notebook containing data preprocessing, model training, and evaluation.
- **Relevant Data**: Preprocessed datasets with labels.
- **Initial Results**:
  - Significant improvement with threshold adjustments but class imbalance challenge remains.

Metric	Baseline
Accuracy	0.6702
F1-score	0.4781

Table 1: Initial Results Summary

4 Next steps and improvement

For improvements require:

- Feature engineering: Exploring domain specific linguistic features.
- Model: Exploring more advanced architectures and ensemble methods.
- Augmentation: Implementing Data Augmentation Strategies.