Deception Detection
CSE556: Natural Language Processing Mid Evaluation
Group 91 Project

Varun Bharti
IIIT Delhi
varun22562@iiitd.ac.in

Vidhan
IIIT Delhi
vidhan22568@iiitd.ac.in

Vaibhav Sehra
IIIT Delhi
vaibhav22550@iiitd.ac.in

Abstract

We propose a deception detection system for Diplomacy that predicts whether messages exchanged between players are deceptive or truthful. Our system is evaluated using accuracy, Macro F1, and Lie F1 scores. We implement three baselines—human assessment, classical machine learning, and deep learning models—and report their performance.

Problem Statement

Task Explanation: The QANTA Diplomacy project involves developing a model that predicts whether messages exchanged between players in the game Diplomacy are deceptive or truthful. The model analyzes in-game conversations and associated metadata to make its predictions.

Evaluation Metric: Accuracy is the primary metric; however, given the severe class imbalance (with deceptive messages <5% of the data), Macro F1 and Lie F1 (F1 for deceptive messages) are also used which is inspired from the ACL 2020 paper

Baseline Implementation

Our baseline implementation comprises three main components:

1. Preprocessing Pipeline: Raw dialogs are aggregated into individual messages. Each message is filtered to retain only those with valid sender and receiver annotations. Annotations are converted from booleans or string values to binary labels (0: truthful, 1: deceptive). For classical ML, messages are vectorized using a bag-of-words approach, while for deep learning they are tokenized and padded to a fixed length.

2. Human Baseline: We compare sender annotations (intended truth) with receiver annotations (perceived truth) to compute human performance in terms of Accuracy, Macro F1, and Lie F1.

3. Automated Models:

- **Classical ML:** Six models: Logistic Regression, SVM, KNN, Decision Trees, Random Forest are implemented
- **Deep Learning:** Three recurrent models Bi-RNN, Bi-LSTM, and Bi-GRU are implemented.

Model	**Accuracy**	**Macro F1**	**Lie F1**
Human Baseline	0.884	0.581	0.226
Logistic Regression	0.949	0.862	0.752
SVM	0.985	0.952	0.912
KNN	0.92	0.537	0.116
Decision Tree	0.994	0.98	0.963
Random Forest	0.994	0.982	0.967
Bi-RNN	0.872	0.503	0.076
Bi-LSTM	0.891	0.524	0.106
Bi-GRU	0.889	0.523	0.104

Table 1: Baseline Results (Evaluation metrics as per ACL 2020).

Future Plan

- Incorporate concepts like context and power dynamics as described in the ACL 2020 paper which we believe will help in classifying and improving the accuracy.
- Develop a hierarchical model which will work with combination of LSTM and GRUs with attention mechanisms and transformer-based modules.
- Expand evaluation metrics to include additional measures for deeper insights and getting better analysis of the results.
- Explore the use of large language models in both in-context learning and fine-tuning.

The end goal is to build an end-to-end pipeline that should outperform current baseline results.