Deception Detection

GROUP - 38
Members : Muthuraj Vairamuthu, Syam Sai Santosh Bandi, Pratham Sibal

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Introduction

Deception detection plays a vital role in domains like politics, negotiation, and online safety.

We explore deception in Diplomacy, a strategic game where players communicate, form alliances, and often lie to win.

Our project builds on the work of Peskov et al. (2020), who labeled messages in Diplomacy as truthful or deceptive.

We develop models that combine linguistic patterns and game-specific metadata to identify deceptive messages with high accuracy.

The aim: Go beyond traditional models and enhance lie detection using modern NLP and custom transformer-based architectures.

Related Work

Early research (e.g., DePaulo et al., 2003) found that linguistic cues such as hedging, emotional tone, and uncertainty signal deception.

With advances in NLP, models evolved from basic pattern matching to context-aware neural architectures.

Peskov et al. (2020) introduced the Diplomacy dataset, labeling messages as truth or lie based on player intent.

- Their best model, Context LSTM + Power, used both message content and in-game power metrics.
- Achieved Macro F1: 57.2%, Lie F1: 27.2% — approaching human-level performance (Macro F1: 58.1%).

It Takes Two to Lie: One to Lie, and One to Listen

Denis Peskov, Benny Cheng
Ahmed Elgohary, Joe Barrow
Computer Science, University of Maryland
{dpeskov, bcheng96, elgohary, jdbarrow}@umd.edu

Cristian Danescu-Niculescu-Mizil
Information Science
Cornell University
cristian@cs.cornell.edu

Jordan Boyd-Graber
iSchool, Language Science, UMIACS, LSC
University of Maryland
jbg@umiacs.umd.edu

Abstract

Trust is implicit in many online text conversations—striking up new friendships, or asking for tech support. But trust can be betrayed through deception. We study the language and dynamics of deception in the negotiation-based game Diplomacy, where seven players compete for world domination by forging and breaking alliances with each other. Our study with players from the Diplomacy community gathers 17,289 messages annotated by the sender for their *intended* truthfulness and by the receiver for their *perceived* truthfulness. Unlike existing datasets, this captures deception in long-lasting relationships, where the interlocutors strategically combine truth with lies to advance objectives. A model that uses power dynamics and conversational contexts can predict when a lie occurs nearly as well as human players.

1 Introduction

Message	Sender's intention	Receiver's percep.
If I were lying to you, I'd smile and say "that sounds great." I'm honest with you because I sincerely thought of us as partners.	Lie	Truth
You agreed to warn me of unexpected moves, then didn't . . . You've revealed things to England without my permission, and then made up a story about it after the fact!	Truth	Truth
. . . I have a reputation in this hobby for being sincere. Not being duplicitous. It has always served me well. . . . If you don't want to work with me, then I can understand that . . .	Lie	Truth
(Germany attacks Italy)		
Well this game just got less fun	Truth	Truth
For you, maybe	Truth	Truth

Table 1: An annotated conversation between Italy (white) and Germany (gray) at a moment when their relationship breaks down. Each message is annotated by

Methodology:
Overall Approach

1. RoBERTa + Metadata Cross-Attention Architecture:

- A multimodal transformer-based framework that fuses semantic language representations with structured contextual metadata to enhance deception detection in strategic discourse. The core architecture is built upon RoBERTa-base, a large-scale pretrained encoder optimized for deep syntactic and semantic abstraction.

- Cross-attention mechanism bridges the high-dimensional token embeddings from RoBERTa with dense metadata vectors representing game-theoretic constructs like power disparity, temporal phase, and player identities.

- This architecture captures interaction-aware deception patterns, where language is interpreted differently based on situational context, such as a promise from a dominant player during a critical turn.

RoBERTa-Based Model – Novelty :

This model replaces shallow LSTM-based encoders with RoBERTa's deep pre-trained contextual understanding and introduces a novel cross-attention mechanism that dynamically fuses game metadata with textual embeddings.

Methodology:
Overall Approach

2. Custom Transformer with Harbinger Token Masking:

- A compact transformer-based architecture trained from scratch to identify deception by highlighting lexical uncertainty cues. It avoids pretrained bias and metadata dependence, instead learning from handcrafted linguistic signals associated with hedging and evasion.
- Harbinger tokens—curated markers like *"maybe"*, *"perhaps"*, and *"honestly"*—are encoded as a binary mask and concatenated with token embeddings, enabling the model to distinguish potentially deceptive phrasing patterns.
- This model captures deception-relevant textual subtleties through direct sequence modeling and short-term context (up to 2 prior messages), offering interpretability and strong generalization without external feature reliance.

Harbinger Transformer – Novelty

Built from scratch, this lightweight transformer uniquely integrates a harbinger mask to emphasize linguistic cues commonly tied to deception, such as hedging or cognitive distancing. Without relying on metadata or pretraining, it offers interpretable modeling aligned with conversational flow and deception signaling.

ROBERTA

HARBINGER TRANSFORMER

General Settings:

- Seed: 42
- Transformer Base Model: roberta-base
- Max Sequence Length: 256
- Batch Size: 16
- Number of Epochs: 15
- Dropout Rate: 0.3
- Gradient Clipping: Max Grad Norm = 1.0

Optimization:

- Learning Rate: 3e-5
- Warmup Ratio: 0.1
- Weight Decay: 0.01

Metadata Encoder:

- Input Dimension: 79
- Hidden Layer Size: 512

Cross-Attention Settings:

- Attention Heads: 12

Loss Function:

- Type: Asymmetric Focal Loss
- Focal Loss Parameters:
  - γpos=5.0
  - γneg=2.0
  - α=0.88
- Positive Class Weight: 5.0

Hyperparameters:

- Batch size: 32
- Learning rate: 1×10^-4
- Number of epochs: 15
- Embedding dimension: 127
- Attention heads: 4
- Transformer layers: 2
- Hidden dimension: 256
- Dropout: 0.3
- Maximum sequence length: 300

Training Strategy:

- Used BCEWithLogitsLoss with class weight pos_weight≈21.2 to address severe class imbalance
- Precision-recall curve used for threshold optimization post-training
- Best model checkpoint selected based on highest validation F1 score.

Evaluation Metrics:

- Macro F1 Score
- Lie Class F1 Score
- Accuracy

Results

Metric	Value
Loss	0.0655
Accuracy	0.8508
F1 Score	0.2631
Macro F1 Score	0.5900
Lie F1 Score	0.2631
Truth F1 Score	0.9170

Table 1: Test Results for RoBERTa Model

RoBERTa-Based Model: Breaking the Benchmark

The RoBERTa model leveraged deep contextual embeddings and metadata-aware cross-attention to surpass the paper's Context LSTM+Power model.

It achieved a Macro F1 Score of 0.5900, exceeding the paper's 0.572, and a Truth F1 Score of 0.9170, with Lie F1 Score nearly matching the paper's best.

- Improves upon the original benchmark by learning subtle deception cues using cross-modal alignment of text and context.
- Outperforms human baseline (Macro F1 0.581) in deception detection.

Class	Precision	Recall	F1-Score	Support
0.0	0.93	0.88	0.90	2501
1.0	0.17	0.27	0.21	240
Overall Metrics:				
Accuracy		0.8238		
F1 Score		0.2121		
AUC		0.6112		
Macro Avg	0.55	0.57	0.56	2741
Weighted Avg	0.86	0.82	0.84	2741

Table 2: Test Set Performance for Transformer Model

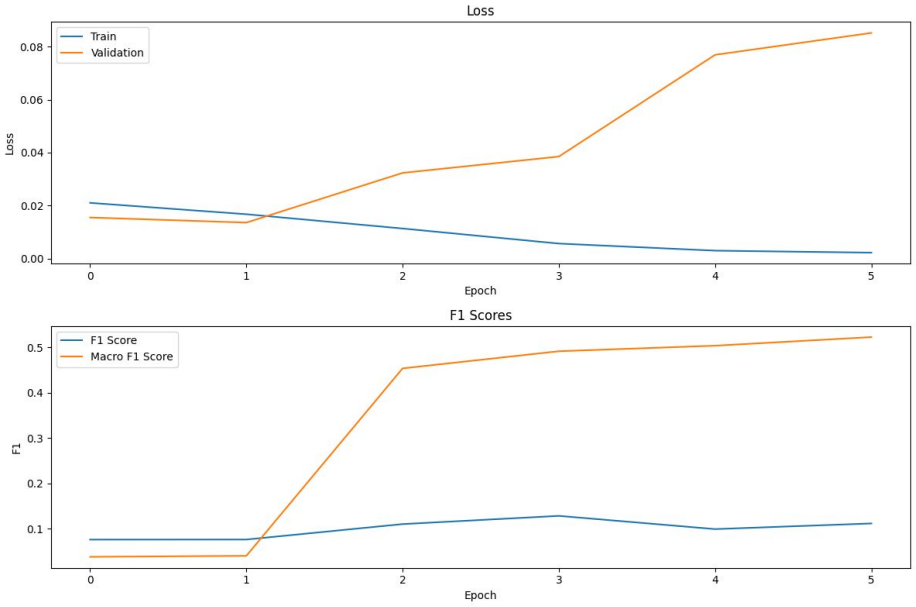
Transformer-Based Model with Harbinger Tokens: Lightweight yet Powerful

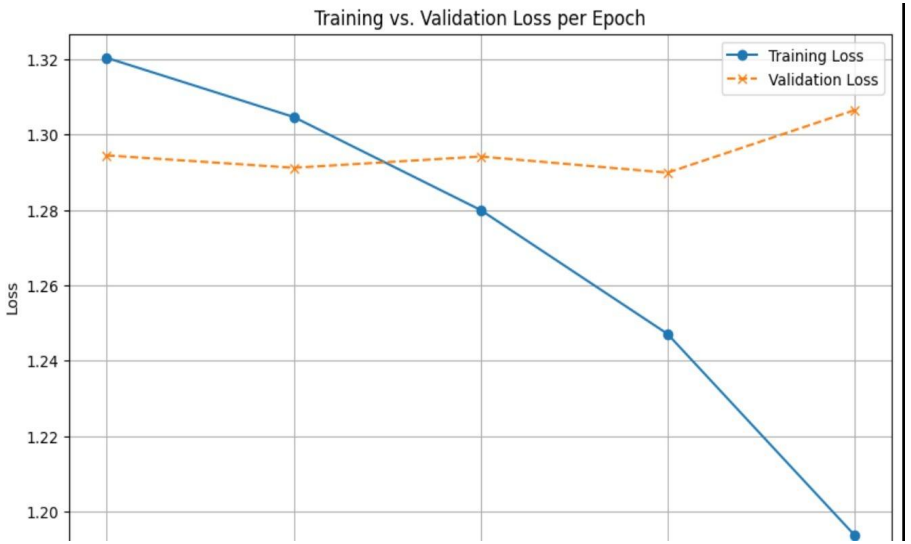
A handcrafted, interpretable transformer architecture focusing on deception-indicative linguistic signals such as hedging and uncertainty cues.

Despite its simplicity, it achieved Macro F1 of 0.56 and Test Accuracy of 0.8238, outperforming traditional baselines like Bag of Words (BoW).

- Performs competitively while being highly efficient and easily deployable.
- Excels in capturing uncertainty without metadata dependency.

Results





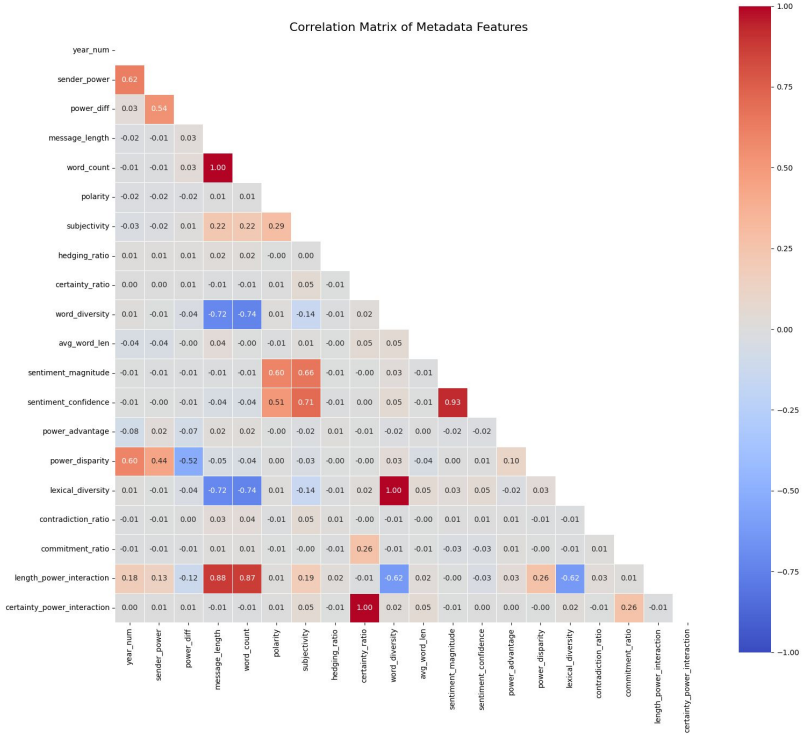
The RoBERTa model showed smooth convergence with minimal overfitting—training loss consistently fell while validation loss rose only slightly in later epochs.

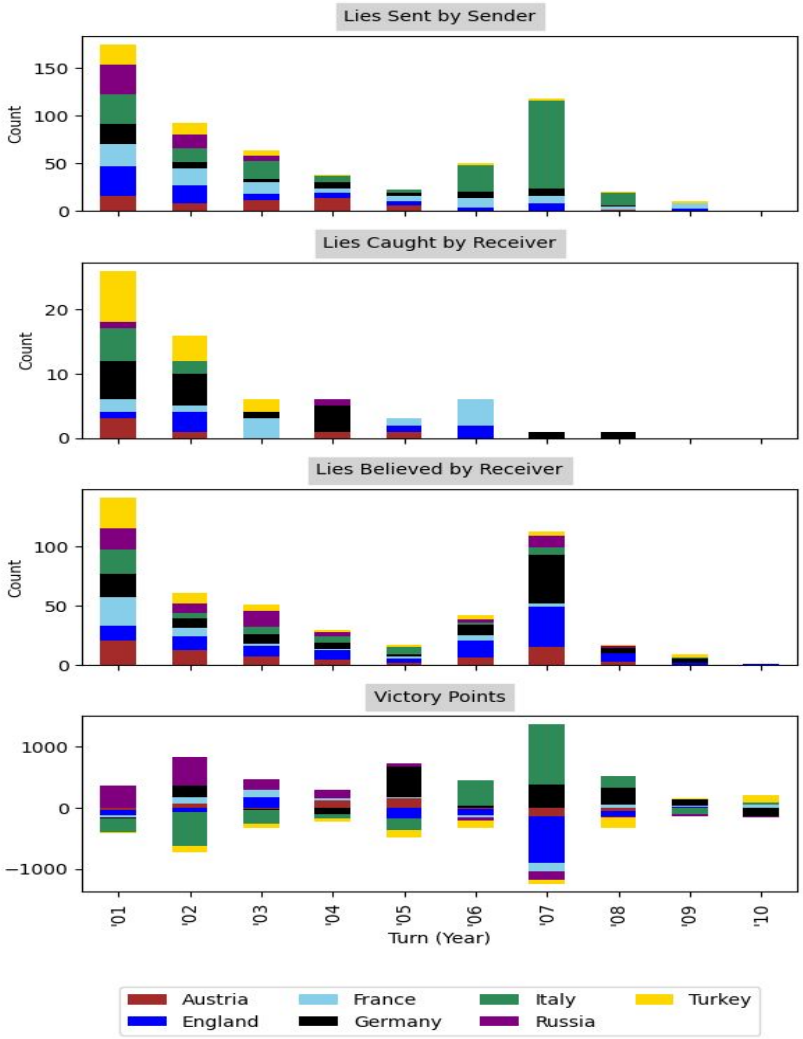
The Transformer model displayed stable training with a strong drop in loss from 1.32 to 1.20, with validation loss holding steady for most epochs.

Both models demonstrate well-regularized training and effective convergence.

Graphs confirm model reliability and proper hyperparameter tuning.

Further EDA and Correlation





Discussion and Conclusion

- Both models surpassed the paper's results, with RoBERTa excelling in Macro F1 and Transformer in accuracy.
- We aim to improve generalization in RoBERTa with early stopping and enhance lie detection in the Transformer via class rebalancing techniques.
- Future steps include ensemble modeling and hyperparameter tuning to push performance further.

Roberta Model.pt

Github link