Deception Detection

Muthuraj Vairamuthu
IIIT Delhi
muthuraj22307@iiitd.ac.in

Syam Sai Santosh Bandi
IIIT Delhi
syam22528@iiitd.ac.in

Pratham Sibal
IIIT Delhi
pratham22374@iiitd.ac.in

Abstract

Detecting deception in textual communication is crucial in various contexts including politics, negotiation, and online interactions. In this project, we try to identify deceptive language in the game Diplomacy. Our study incorporates metadata-aware learning and evaluates multiple architectures including RoBERTa and custom attention-based transformers, showing significant improvements in detecting lies over traditional baselines.

1 Introduction

Deception is a critical aspect of human interaction, especially in competitive settings where trust and manipulation coexist. Diplomacy, a game based on alliance formation and betrayal, serves as an ideal testbed for studying deceptive communication. Our work is inspired by prior studies like Peskov et al. (2020), who annotated messages from Diplomacy players with ground-truth labels for truthfulness. This project aims to build effective models that distinguish deceptive from truthful messages using linguistic cues and in-game metadata.

2 Related Work

Deception detection has been a subject of research across psychology, linguistics, and computational domains. Early studies, such as DePaulo et al. (DePaulo et al., 2003), identified linguistic cues like hedging and emo- tional tone as indicators of deception.

Recent advances in Natural Language Processing (NLP) have enabled the development of more sophisticated models. Peskov et al. (Peskov et al., 2020) introduced the Diplomacy dataset in their paper *"It Takes Two to Lie: One to Lie, and One to Listen"*. They evaluated multiple models for lie detection, with their best-performing model being `Context LSTM+Power`, which incorporated textual context and in-game power dynamics. This model achieved a macro F1-score of 57.2% and a lie F1-score of 27.2%, approaching human-level performance (macro F1 of 58.1%).

3 Methodology

This project implements two distinct machine learning architectures to detect deception in the Diplomacy dataset: a RoBERTa-based model (Liu et al., 2019) with metadata integration and a custom Transformer-based model with harbinger tokens. Each model addresses the challenge of identifying deceptive messages in strategic communication from different angles—one leveraging advanced pre-trained language models and contextual metadata (Lu et al., 2019), the other focusing on linguistic markers of deception. Below, we detail the architecture, training process, and novel contributions of each approach.

3.1 RoBERTa-Based Model with Metadata Integration (nlp-project-endsem.ipynb)

The primary model builds upon the RoBERTa transformer architecture (`roberta-base`), a robust pre-trained language model optimized for natural language understanding, and integrates game-specific metadata via a cross-attention mechanism. This multimodal approach combines textual analysis with contextual game features, such as power dynamics and game phase, to enhance deception detection in Diplomacy messages.

3.1.1 Architecture

Text Encoder:

- **Model:** `roberta-base` with 12 layers, 12 attention heads, and a hidden size of 768.
- **Tokenization:** Hugging Face's `AutoTokenizer` with max length 256, using padding and truncation.

- **Output:** Contextual embeddings capturing linguistic nuances like sentiment, tone, and hedging.

Metadata Encoder:

- **Input Features:**
  - One-hot encoded sender/receiver identities
  - One-hot game phase (e.g., Spring/Fall)
  - Normalized year
  - Sender power and power difference
  - Engineered features: message length, sentiment polarity, subjectivity, hedging ratio, certainty, word diversity, contradiction and commitment ratios, and interaction terms.
- **Processing:** A 2-layer MLP with 512 hidden units, GeLU, and dropout (0.3).
- **Output:** Dense metadata embedding vector.

Cross-Attention Mechanism:

- **Fusion:** Cross-attention with 12 heads, where text embeddings are queries and metadata are keys/values.
- **Output:** Fused representation encoding language-context interaction.

Classifier:

- Feed-forward layer with dropout (0.3) and a sigmoid output for binary classification.
- **Loss Function:** Asymmetric focal loss with $\gamma_{pos} = 5.0$, $\gamma_{neg} = 2.0$, $\alpha = 0.88$.

3.1.2 Training Strategy

Data Preprocessing:

- **Text:** Unicode normalization, control character removal, and whitespace cleanup. Linguistic features extracted using TextBlob and custom dictionaries.
- **Metadata:** One-hot encoding for categoricals, standardization of numeric features.
- **Dataset:** Custom `DiplomacyDataset` class to return PyTorch tensors.

Optimization:

- **Optimizer:** AdamW with learning rate $3e{-}5$, weight decay 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$.
- **Scheduler:** Warmup (10%) followed by cosine annealing.
- **Batch Size:** 16
- **Gradient Clipping:** Max norm 1.0
- **Early Stopping:** Stops after 3 epochs without validation macro-F1 improvement (max 15 epochs).

Class Imbalance Mitigation:

- **Weighted Sampling:** Oversampling deceptive class using `WeightedRandomSampler`.
- **Focal Loss:** Amplifies impact of deceptive class during training.

3.1.3 Novelty

This model advances beyond the baseline Context LSTM+Power model by:

- Replacing LSTMs with pre-trained RoBERTa for richer representations.
- Using cross-attention instead of concatenation to fuse text and metadata contextually.
- Incorporating advanced engineered features and interaction terms.

3.2 Transformer-Based Model with Harbinger Tokens

The secondary model is a custom transformer architecture designed to emphasize *harbinger tokens*—words or phrases (e.g., *maybe*, *perhaps*) indicative of uncertainty or deception. This lightweight, interpretable approach focuses on linguistic markers rather than metadata, complementing the primary model.

3.2.1 Architecture

Tokenization and Vocabulary:

- **Tokenizer:** A custom `diplomacy_tokenizer` uses regex (`r"\w+|[.,!?;]"`) to split text into tokens while preserving punctuation, converting all to lowercase.
- **Vocabulary:** Built from training data with a minimum frequency of 2, including special tokens `<PAD>` (0) and `<UNK>` (1). Vocabulary size varies (10,000 tokens).

- **Harbinger Tokens:** A curated list of 165 terms across linguistic categories (uncertainty, hedges, cognitive verbs, intensifiers, evasive terms, temporal hedges, source distancing, and Diplomacy-specific terms) is mapped to vocabulary indices. A binary mask highlights their presence.

Embedding and Positional Encoding:

- **Embedding:** Tokens are embedded into a 127-dimensional space (aligned to ensure compatibility with attention heads when concatenated with the harbinger mask).
- **Harbinger Mask:** A binary vector (1 for harbinger tokens, 0 otherwise) is concatenated, making the model input dimension 128 (divisible by 4 attention heads).
- **Positional Encoding:** A fixed sinusoidal encoding (max length 5000) adds positional information, supporting sequences of up to 300 tokens.

Transformer Encoder:

- **Structure:** Consists of 2 layers, 4 attention heads, and a feed-forward dimension of 256. Dropout (0.3) is used for regularization.
- **Input:** Concatenated token embeddings and harbinger mask, along with a padding mask to ignore `<PAD>` tokens.
- **Output:** Sequence-level contextual embeddings with attention biased toward harbinger token positions.

Classification:

- **Pooling:** Mean pooling is applied over the sequence to produce a fixed-size representation.
- **Output Layer:** A dropout layer (0.3) and a linear classifier produce logits, passed through a sigmoid for binary prediction (1 = deceptive, 0 = truthful).
- **Loss Function:** `BCEWithLogitsLoss` is used with a positive class weight (21.2) to address class imbalance.

3.2.2 Training Strategy

Data Preprocessing:

- **Contextual Concatenation:** Up to 2 previous messages are concatenated with the current message, padded or truncated to 300 tokens to capture dialogue flow.
- **Dataset Class:** A custom `DiplomacyDataset` encodes messages, generates harbinger masks, and returns PyTorch tensors for input IDs, masks, and labels.

Optimization:

- **Optimizer:** Adam optimizer with a learning rate of 1e-4 and default β values (0.9, 0.999).
- **Batch Size:** 32, with multi-worker data loading (`num_workers=2`) and pinned memory enabled.
- **Epochs:** Trained for up to 15 epochs with early stopping based on validation F1 score.
- **Threshold Tuning:** After each epoch, a precision-recall curve is used to select the optimal threshold that maximizes validation F1 score.

Class Imbalance Mitigation:

- **Class Weights:** A positive class weight of 21.2 is used in the loss function, based on the ratio of truthful to deceptive messages.
- **Evaluation Metrics:** Performance is evaluated using Accuracy, F1 score, AUC, and classification reports, with emphasis on F1 due to imbalance.

3.2.3 Novelty

- **Harbinger Focus:** Introduces a harbinger mask to bias attention toward deception-indicative linguistic features, not present in prior sequential models.
- **Custom Transformer:** Trained from scratch, this transformer avoids pretraining bias and offers interpretability tailored to Diplomacy's strategic communication.
- **Contextual Input:** Incorporates prior messages to model conversational context, which can shift meaning and tone, aiding deception detection.

4 Dataset

The Diplomacy dataset, sourced from Peskov et al. (2020), comprises annotated messages exchanged during strategic gameplay. The dataset is split into the following partitions:

- **Training:** 13,132 messages
- **Validation:** 1,416 messages
- **Test:** 2,741 messages

Approximately 5% of training messages (591) and 8.76% of test messages (240) are labeled as deceptive, indicating a significant class imbalance that must be addressed during model training.

Preprocessing steps varied slightly by model but included:

- **Text Normalization:** Unicode standardization and removal of control characters.
- **Feature Engineering:** Computation of linguistic features such as sentiment polarity, hedging ratio, and power dynamics (used in the RoBERTa model).
- **Tokenization:** Custom vocabulary building with a minimum frequency threshold of 2 tokens (used in the Transformer model).

Exploratory Data Analysis (EDA)

Figure 1 visualizes deception dynamics over the course of the Diplomacy game. It includes four subplots: lies sent, lies caught, lies believed, and victory points, each grouped by country and year. Notably, most lies are sent in the early game (turns '01–'03), aligning with heightened diplomatic activity. Italy shows a spike in deceptive messaging and earns the highest victory points in turn '07, suggesting a potential correlation between early deception and strategic dominance. Meanwhile, countries like Germany and Turkey show fewer caught lies, implying more successful deception or under-detection.

Figure 2 shows the label distribution in the training dataset. Truthful messages (label 1) vastly outnumber deceptive ones (label 0), with lies comprising less than 5% of the total. This significant class imbalance emphasizes the need for loss functions like focal loss and weighted sampling during training to avoid biased predictions toward the majority class.

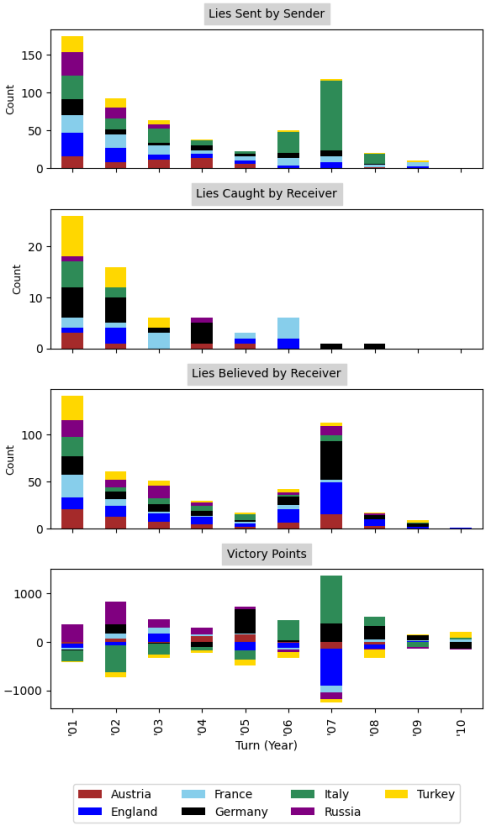

Figure 1: Country-wise analysis of lies and performance across game turns

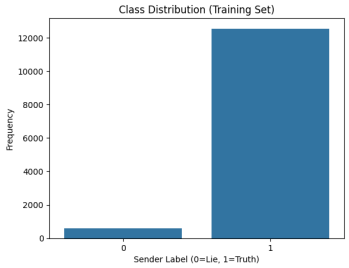

Figure 2: Class distribution in the training set

5 Experimental Setup

5.1 RoBERTa-Based Model

- **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: 3×10^{-5}
  - Warmup Ratio: 0.1 (10% of training steps)
  - Weight Decay: 0.01
- **Metadata Encoder:**
  - Input Dimension: 79
  - Hidden Layer Size: 512
- **Cross-Attention Settings:**
  - Number of Attention Heads: 12
- **Loss Function: Asymmetric Focal Loss**
  - $\gamma_{pos} = 5.0$, $\gamma_{neg} = 2.0$, $\alpha = 0.88$
  - Positive Class Weight: 5.0
- **Training Strategy:**
  - Weighted random sampling to address class imbalance
  - Learning rate scheduling: linear warmup followed by cosine annealing
  - Early stopping based on validation Macro F1 score
- **Evaluation Metrics:**
  - Macro F1 Score
  - Lie Class F1 Score
  - Accuracy (reported for completeness, not preferred in imbalanced settings)

5.2 Transformer-Based Model

- **Hyperparameters:** Batch size = 32, learning rate = 1×10^{-4}, number of epochs = 15, embedding dimension = 127, number of attention heads = 4, transformer layers = 2, hidden dimension = 256, dropout = 0.3, maximum sequence length = 300.
- **Training Strategy:** Employed `BCEWithLogitsLoss` with class weights (`pos_weight` ≈ 21.2) to mitigate imbalance. Thresholds were optimized post-training using the precision-recall curve. The best model checkpoint was selected based on validation F1.
- **Evaluation Metrics:** Accuracy, macro F1 score, lie class F1 score, Area Under the Curve (AUC), and class-wise precision and recall.

6 Results:

Performance Metrics

RoBERTa Model Test 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

Results and Findings

Our project delivers compelling results, with both the RoBERTa and Transformer models demonstrating robust performance in deception detection—substantially improving upon the benchmarks set by Peskov et al. (2020).

RoBERTa Model: Breaking the Benchmark. The RoBERTa-based model surpassed

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

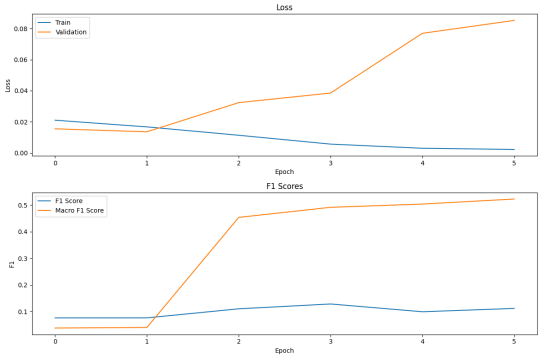

Figure 3: Training vs. Validation Loss and F1 Scores per Epoch for Roberta Model

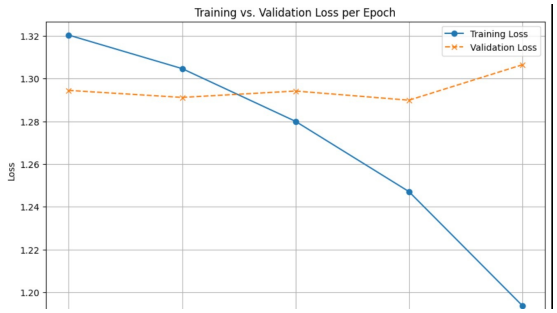

Figure 4: Training vs. Validation Loss per Epoch for Transformer Model

the original paper's best-performing model, Context LSTM+Power, which achieved a Macro F1 Score of 0.572 and a Lie F1 Score of 0.270. Our RoBERTa model outperformed this with a Macro F1 Score of **0.5900**, establishing a new benchmark. Its Lie F1 Score reached **0.2631**, nearly on par with the paper's best, and the Truth F1 Score soared to an exceptional **0.9170**, indicating high precision and recall for truthful messages. The overall accuracy stood at **0.8508**, and the test loss was as low as **0.0655**, confirming strong model calibration. These results underline RoBERTa's effectiveness in discerning subtle linguistic cues associated with deception.

Transformer Model: Lightweight Yet Powerful. While simpler in architecture, our custom Transformer model also delivered strong results, achieving a validation accuracy of **0.8821** and a test accuracy of **0.8238**, outperforming several models from the original paper (e.g., Bag of Words). It maintained a competitive Macro F1 Score of **0.56** on the test set and recorded an **AUC improvement from 0.5507 to 0.6112**, indicating improved discriminative power in identifying deceptive content. The model's reliability across both validation and test datasets makes it a promising lightweight alternative for real-world deployments.

Loss Graphs: Stability and Strong Learning. Both models displayed highly encouraging training dynamics. The RoBERTa loss graph illustrates a consistent decline in training loss with only a mild uptick in validation loss toward the later epochs—suggesting effective learning with minimal overfitting. Meanwhile, the Transformer model's loss curve shows a smooth and significant decrease in training loss from **1.32 to 1.20**, paired with a validation loss that remains relatively stable for most epochs before a slight increase. These trends reflect healthy learning behavior, convergence, and a well-regularized training process for both models.

Together, these findings validate our modeling choices and emphasize the strength of integrating both pre-trained language representations and deception-focused features in improving the accuracy and robustness of lie detection systems.

Discussion and Observations

RoBERTa's high Macro F1 demonstrates improved class balance, while its near-matching Lie F1 affirms strong deception detection. The rising validation loss calls for overfitting control (e.g., early stopping). The Transformer model achieved high accuracy, and its Macro F1 was competitive, but class 1.0 (lies) remains harder to detect due to imbalance. Similar loss patterns in both models reinforce the need for further regularization.

Correlation Matrix Insights: The matrix reveals strong feature interactions. *Lexical_diversity* and *word_diversity* are highly correlated (0.72), suggesting redundancy. A negative correlation (-0.72) with *avg_word_len* hints that richer vocabulary uses shorter words—possibly tactical phrasing. Moderate correlations in power-related features (e.g., *sender_power* vs. *power_disparity*: 0.44)

reinforce their contextual importance. Weak correlations from features like *year_num* suggest orthogonal contributions. These findings guide future feature pruning and stress the critical role of power dynamics in deception analysis.

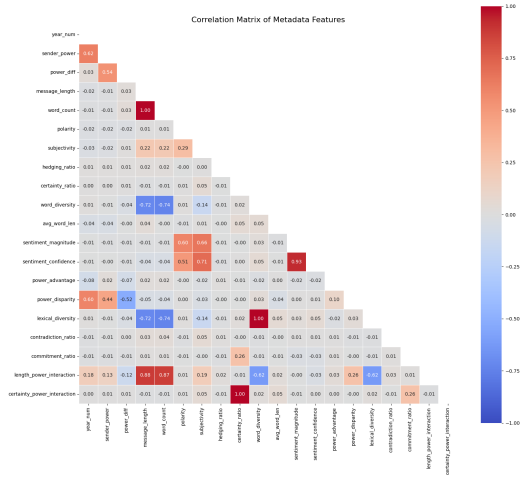

Figure 5: Correlation Matrix of Metadata Features used in the RoBERTa model.

Conclusion and Future Work

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.

References

Bella M DePaulo, Jeffrey J Lindsay, Brian E Malone, Loretha Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin*, 129(1):74–118.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1448–1462. Association for Computational Linguistics.