DECEPTION DETECTION

Manya Agrawal
2022281
manya22281@iiitd.ac.in
IIIT-DELHI

Mohmmad Ayaan
2022302
ayaan22302@iiitd.ac.in
IIIT-DELHI

Shubham K. Dwivedi
2022494
shubham22494@iiitd.ac.in
IIIT-DELHI

Abstract

Our proposed machine learning system analyzes deception in QANTA Diplomacy dataset which contains 17,289 annotated Messages from the Diplomacy game. The classification of truthful versus deceptive messages by our models depends on linguistic evidence from messages and their metadata alongside game dynamic information. The best model shows this particular achievement while data analysis shows deceptive messages typically use ambiguous language along with power disparities between participants. This research makes advancement in deception analysis for strategic settings while showing the difficulties involved in modeling subjective content. Research needs to incorporate elements from player game data along with in-game conditions to achieve better performance accuracy.

1 Problem Definition

In the game of so-called Diplomacy, the players engage in a strategic communication which might be to form alliances, negotiate plans, and sometimes deceive others to gain a competitive edge. This project aims to predict whether a message exchanged between players is deceptive (containing false information) or truthful. The model will leverage the convos happening in the game and metadata, with accuracy to measure the performance.

2 Introduction

Diplomacy and security functions together with online communication require deception detection to establish trust which determines their outcomes. Players in the board game Diplomacy create alliances and then betray one another to control pre-World War I Europe territories thus making the game an excellent platform to study deception in strategic text-based interactions. The long-term relationships between players within the Diplomacy game reconstruct the natural progression of international negotiations through multiple rounds of truth and deceitful statements. Players promise peace at first to build trust but initiate attacks afterward which our project models as one of its key objectives.

3 Objective

Our goal is to build a prediction system that determines deceptive or truthful status in messages using textual content along with metadata points such as sender-receiver pairs and game scoring data. Our research focuses on resolving the labeling ambiguity which occurs because sender intentions and receiver interpretations influence the provided deception annotations.

4 Related Work

Psychology and computational linguistics have established extensive studies about detecting deception. Researchers during the earliest stages of deception detection worked with separate communication channels by utilizing Ekman's facial expression methods (Ekman1992) while identifying vague pronouns as defined by Newman et al. (Newman et al.2003). Modern research utilizes LIAR datasets to detect fake news according to Wang (Wang2017).

Researchers can study deception in text-based negotiations using the Diplomacy dataset which Peskov et al. (Peskov et al.2020) introduced as an innovative tool. The researchers produced performance estimates reaching 60% accuracy using logistic regression and LSTMs but these results were limited by human annotation problems. The research conducted by Wu et al. [2019] examined courtroom deceptions through text-video integration but failed to produce results that extended beyond their study context.

The current research utilizes transformer mod-

els that run on BERT combined with score difference metadata to extend existing investigations from Peskov et al. despite little investigation in prior work. The analysis of both textual content and metadata seeks to resolve scalability issues that occur when detecting deception in real-world text-only situations.

5 Dataset

The Diplomacy dataset comprises 17,289 messages spanning 12 games. Each message is annotated from two perspectives: the sender's intended truthfulness and the receiver's perceived truthfulness.

6 Methodology

6.1 Baseline Model-1

Context LSTM model is the baseline model to estimate the actual and suspected lies by capturing the sequential dependencies in conversations. With pre-trained GloVe embeddings, a message-level BiLSTM, and a conversation-level LSTM, which well simulates the deception patterns for receiver-based and sender-based labeling. Hyperparameters must be optimized, with methods such as dropout regularization and class imbalance handling to guarantee dependability. It is therefore a reliable basis for consideration of models with attention mechanisms or transformers.

Table 1: Performance Metrics for Baseline Model-1 for Suspected-Lie

Metric	Training	Validation	Test
True Precision	0.9741	0.9661	0.9349
False Precision	0.0928	0.0401	0.0856
True Recall	0.6988	0.5542	0.9260
False Recall	0.6237	0.4894	0.0970
True F1	0.8138	0.7044	0.9304
False F1	0.1615	0.0741	0.0909
Micro Precision	0.6952	0.5518	0.8707
Micro Recall	0.6952	0.5518	0.8707
Micro F1	0.6952	0.5518	0.8707
Macro Precision	0.5334	0.5031	0.5102
Macro Recall	0.6612	0.5218	0.5115
Macro F1	0.4876	0.3892	0.5107
Loss	0.4765	0.5332	0.9492

Results:

6.2 Baseline Model-2

Context LSTM + BERT optimizes the baseline solution by enabling word embedding interpretation from pre-trained BERT encoder which results in better semantic detection during conversational analysis. The system uses BERT encoding to start before the hierarchical LSTM framework with message-level BiLSTM and conversation-level LSTM components to analyze message internal structures and message-to-message relationships. By merging BERT's deep semantic processing with LSTMs' sequential information modeling capability the model gains better ability to find deceptive patterns both in sender and receiver labeling contexts. The robustness and stability of the system result from fine-tuning the hyperparameters along with using dropout regularization and class imbalance handling components. Transformers serve as a strong potential improvement after the model showed substantial F1-score enhancements above the baseline during evaluation.

Table 2: Performance Metrics for Baseline Model-2 for Suspected-Lie

Metric	Training	Validation	Test
True Precision	0.9697	0.9807	0.9387
False Precision	0.0734	0.0400	0.1073
True Recall	0.6220	0.1642	0.8883
False Recall	0.6060	0.9149	0.1879
True F1	0.7579	0.2814	0.9128
False F1	0.1309	0.0766	0.1366
Micro Precision	0.6213	0.1917	0.8416
Micro Recall	0.6213	0.1917	0.8416
Micro F1	0.6213	0.1917	0.8416
Macro Precision	0.5215	0.5103	0.5230
Macro Recall	0.6140	0.5396	0.5381
Macro F1	0.4444	0.1790	0.5247
Loss	0.6260	0.5755	1.0008

Results:

6.3 Try Model-1(RoBERTa)

- **Data Preprocessing:** The input dataset consists of JSONL files, where each line corresponds to a game conversation involving multiple participants. Each message is annotated with two binary labels: *actual-lie* and *suspected-lie*. For each message, we extract context messages, speaker and receiver identities, and optional handcrafted features.

- **Tokenization:** The RoBERTa tokenizer splits each message and its context into tokens, with a maximum length of 128 tokens. The model

optionally incorporates up to two previous messages as context. Padding and truncation are applied when needed.

- **Model Architecture:** The system builds on HuggingFace's `roberta-base` encoder. Each message is contextualized with its surrounding discourse. An `nn.Embedding` layer converts speaker and receiver IDs into learnable embeddings. These are concatenated with the message embedding and optional handcrafted feature vectors to form the final input. The model includes two parallel binary classification heads for predicting actual and suspected lies.

- **Training:** To address class imbalance, a separate Focal Loss is used for each label. The optimizer is AdamW with a learning rate of 1e-5 and weight decay of 0.01. A linear learning rate scheduler with warmup stabilizes training. We apply gradient clipping and perform periodic validation. The model checkpoints the best weights based on the average macro F1 score across both tasks.

- **Results:**

Table 3: Evaluation Metrics

Metric	Actual Lie	Suspected Lie
Accuracy	0.8395	0.8735
Precision	0.8025	0.8091
Recall	0.8595	0.8424
Macro F1	0.8176	0.8236
Micro F1	0.8395	0.8735
Class F1 (Truthful)	0.8808	0.9172
Class F1 (Deceptive)	0.7543	0.7299

Metric	Combined
Accuracy	0.8564
Macro F1	0.8206

6.4 Try Model-2

- **Results:**

Table 4: Model 2: Baseline Model (Suspected Lie)

Metric	Model 1 (Epoch: 12)	Model 2 (Epoch: 7)
Test Micro Prec.	0.8446	0.8416
Test Micro Rec.	0.8446	0.8416
Test Micro F1	**0.8446**	0.8416
Test Macro Prec.	0.5172	**0.5230**
Test Macro Rec.	0.5174	**0.5381**
Test Macro F1	0.5173	**0.5247**
Test True F1	**0.9148**	0.9128
Test False F1	0.1198	**0.1366**
Test Loss	**0.9418**	1.0008
Val. Micro Prec.	**0.8877**	0.8675
Val. Micro Rec.	**0.8877**	0.8675
Val. Micro F1	**0.8877**	0.8675
Val. Macro Prec.	0.5111	**0.5153**
Val. Macro Rec.	0.5221	**0.5423**
Val. Macro F1	0.5106	**0.5121**
Val. True F1	**0.9402**	0.9285
Val. False F1	0.0809	**0.0957**
Val. Loss	**0.4950**	0.5462
Train Micro F1	0.6170	**0.6213**
Train Macro F1	0.4275	**0.4444**
Train True F1	**0.7568**	0.7579
Train False F1	0.0982	**0.1309**
Train Loss	**0.5306**	0.6260
Best Epoch	12	**7**
Train Duration	13m 4s	**12m 51s**

Table 5: Model 2: Baseline Model (Actual Lie)

Metric	Model 1	Model 2
Test Micro Prec.	**0.8446**	0.8369
Test Micro Rec.	**0.8446**	0.8369
Test Micro F1	**0.8446**	0.8369
Test Macro Prec.	0.5172	**0.5289**
Test Macro Rec.	0.5174	**0.5340**
Test Macro F1	0.5173	**0.5308**
Test True Prec.	0.9155	**0.9185**
Test True Rec.	**0.9140**	0.9012
Test True F1	**0.9148**	0.9098
Test False Prec.	0.1189	**0.1394**
Test False Rec.	0.1208	**0.1667**
Test False F1	0.1198	**0.1518**
Test Loss	0.9418	**0.9417**

7 Other Approaches Tried

We also experimented with existing deception detection architectures such as **FakeBERT** and **DaBERTa**, which have shown promising results in general fake news and sentiment analysis tasks. However, when applied to the Diplomacy dataset, these models underperformed significantly in comparison to our baseline models. The f-scores and validation performance were notably low, likely due to domain-specific nuances and limited context understanding. As a result, we shifted our focus towards fine-tuning **RoBERTa**-based architectures, which provided better generalization and contextual representation for deception classification in our setting.

8 Discussion/Analysis

The question examines the way linguistic markers and contextual data signals deception events in strategic environments. Experimental results display that deceptive speech contains hedging statements ("maybe") and flattery techniques that match psychological analysis described by Ekman (Ekman1992). The players who score higher deceive with greater confidence because their strong position minimizes their risk. The pattern of behavior matches the predictions from game theory analysis of information inequality. The training process becomes complicated when senders and receivers provide different annotations such as "Deceived" and "Caught." The system identifies straightforward messages correctly with higher accuracy compared to messages that remain unclear. Because these models process text only they are unable to detect important nonverbal cues from messages. The subjective nature of these annotations decreases agreement between different annotators which negatively affects Macro F1 scores.

8.1 Observations

The attention mechanism of BERT helps identify deceptive keywords such as "trust" and "promise" which suggests it might be interpretable. The small number of deceptive messages in the dataset distorts prediction results towards truthfulness. The game-based design of the dataset cannot extend its results to different domains although its long-term exchange patterns remain distinct.

Conclusion and Future Work:

Our findings highlight the interplay of linguistic ambiguity, power dynamics, and subjective annotations in strategic deception. The research provides innovative approaches to both social interaction analysis with NLP and trust modeling understanding.

- To improve context, including multimodal data (such as player history and message timing).
- Explore unsupervised methods to handle noisy annotations.
- The scientific method can be used to analyze similar data sources including online forums and diplomatic cables.
- The development of transparent models should enable traceability of deceit signs for real-world utilization (for instance, negotiation instruction).

Any interesting Question?

Does player aptitude correlate to their use of misinformation? During game play in Diplomacy expert users combine factual and deceitful statements to establish connectivity while enhancing their position. The functionality of your model should enable classification of players according to their deception activities including both their frequency of lies and their rate of achieving detection avoidance (undetected lies). Your current Diplomacy-Dataset allows tracking of speakers between games so preprocessing the dataset for this purpose seems feasible.

Acknowledgments

We sincerely thank our instructors and peers for their insightful guidance and constructive feedback, which greatly contributed to the development and refinement of this research.

References

[Peskov et al.2020] Peskov, D., et al. (2020). It Takes Two to Lie: One to Lie, and One to Listen. In *Proceedings of ACL 2020*.

[Ekman1992] Ekman, P. (1992). *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. Norton.

[Newman et al.2003] Newman, M. L., et al. (2003). Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*.

[Guo et al.2023] Guo, X., et al. (2023). Audio-Visual Deception Detection: DOLOS Dataset. *arXiv:2303.12745*.

[Wang2017] Wang, W. Y. (2017). LIAR: A Benchmark Dataset for Fake News Detection. In *Proceedings of ACL 2017*.