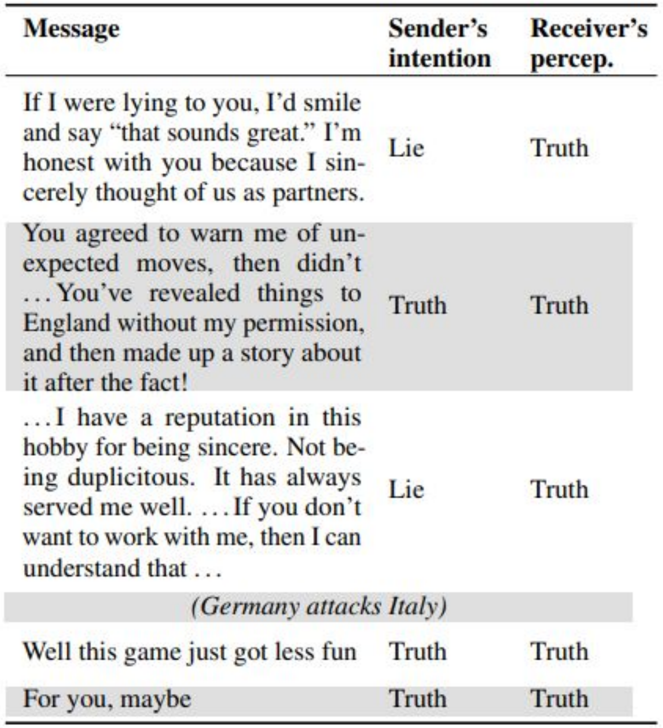
Deception Detection: Lie Detection in Diplomacy via Machine Learning

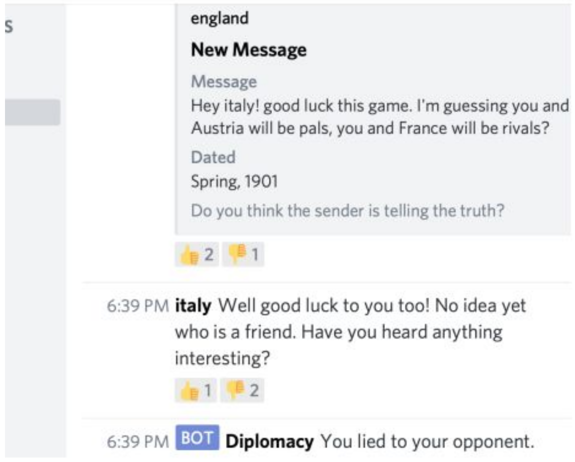
Group 29
Palak Bhardwaj (2022344)
Grishma Bellani (2022189)
Piyush Narula (2022354)

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

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



england

New Message

Message

Hey italy! good luck this game. I'm guessing you and Austria will be pals, you and France will be rivals?

Dated

Spring, 1901

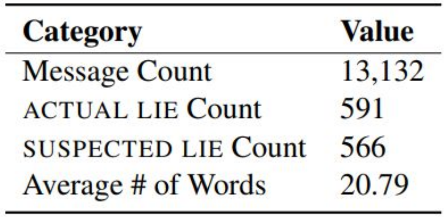
Do you think the sender is telling the truth?

👍 2 👎 1

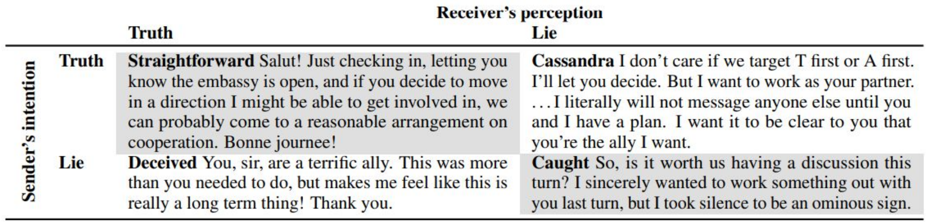
6:39 PM **italy** Well good luck to you too! No idea yet who is a friend. Have you heard anything interesting?

👍 1 👎 2

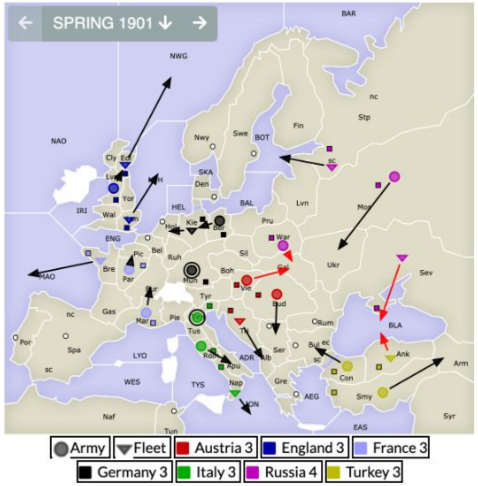
6:39 PM BOT **Diplomacy** You lied to your opponent.

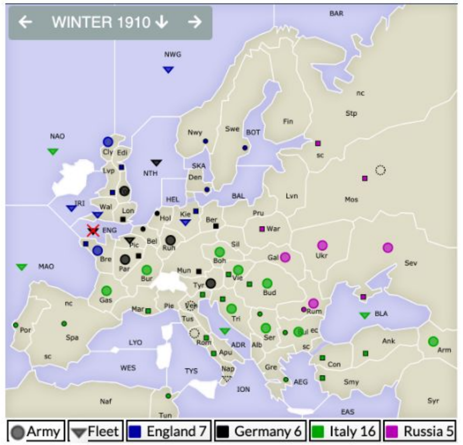


Category	Value
Message Count	13,132
ACTUAL LIE Count	591
SUSPECTED LIE Count	566
Average # of Words	20.79



		Receiver's perception	
		Truth	Lie
Sender's intention	Truth	**Straightforward** Salut! Just checking in, letting you know the embassy is open, and if you decide to move in a direction I might be able to get involved in, we can probably come to a reasonable arrangement on cooperation. Bonne journee!	**Cassandra** I don't care if we target T first or A first. I'll let you decide. But I want to work as your partner. ... I literally will not message anyone else until you and I have a plan. I want it to be clear to you that you're the ally I want.
	Lie	**Deceived** You, sir, are a terrific ally. This was more than you needed to do, but makes me feel like this is really a long term thing! Thank you.	**Caught** So, is it worth us having a discussion this turn? I sincerely wanted to work something out with you last turn, but I took silence to be an ominous sign.



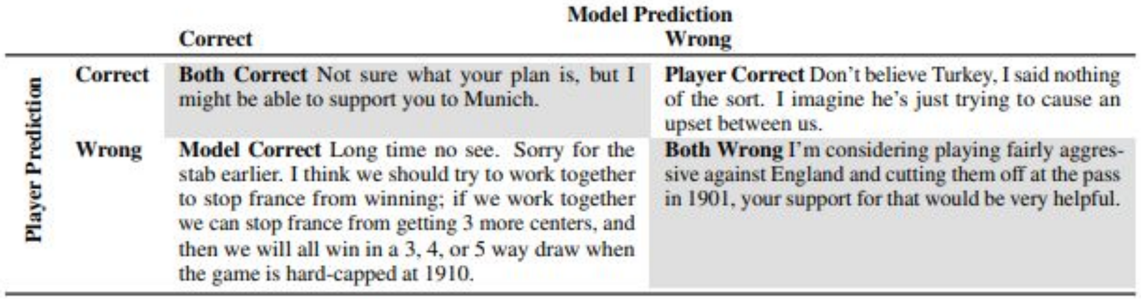


(Peskov et al., 2020)

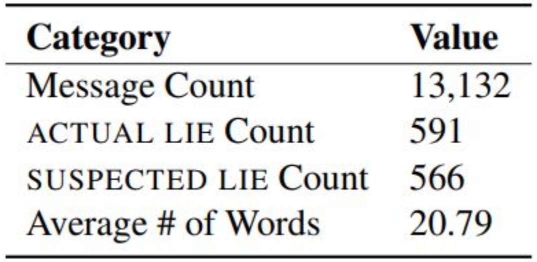
The research investigates online relationship deception patterns by utilizing information from "It Takes Two to Lie: One to Lie, and One to Listen" (Peskov et al., 2020). It presents a novel approach to deception detection in diplomatic communication using deep learning models.

About Dataset & Experimental Setup

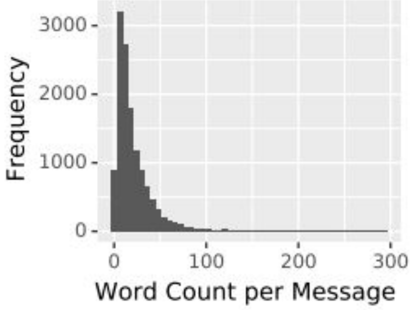
- Total Messages: 17,289 messages from 12 Diplomacy games
- Labeling: ~5% messages marked as fraudulent
- Each message includes:
  - Sender & Receiver information
  - Truth Status: "true", "false", or "None"
- The model uses Sentence-BERT embeddings (dim 384), linguistic (dim 6), and other embeddings (country: 16, season: 8, year: 8), with hidden dim 64 and dropout 0.4.
- Training involved grid search over learning rates [5e-5, 1e-4, 2e-4], batch sizes [4, 8], for 15 epochs with early stopping and positive class weighted by (num-truths/num-lies)*1.5.



		Model Prediction	
		Correct	Wrong
Player Prediction	Correct	**Both Correct** Not sure what your plan is, but I might be able to support you to Munich.	**Player Correct** Don't believe Turkey, I said nothing of the sort. I imagine he's just trying to cause an upset between us.
	Wrong	**Model Correct** Long time no see. Sorry for the stab earlier. I think we should try to work together to stop france from winning; if we work together we can stop france from getting 3 more centers, and then we will all win in a 3, 4, or 5 way draw when the game is hard-capped at 1910.	**Both Wrong** I'm considering playing fairly aggressive against England and cutting them off at the pass in 1901, your support for that would be very helpful.



Category	Value
Message Count	13,132
ACTUAL LIE Count	591
SUSPECTED LIE Count	566
Average # of Words	20.79



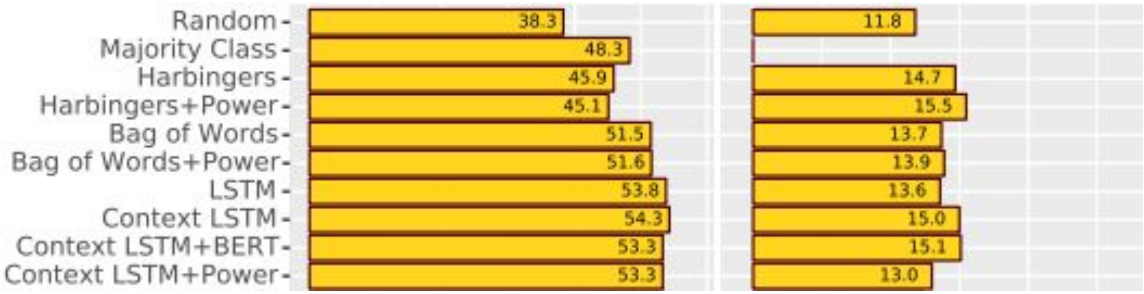
- Test Dataset - 2 games out of 12 games
- Validation Data- 1 game out of 12 games
- Training Data - 9 games out of 12 games

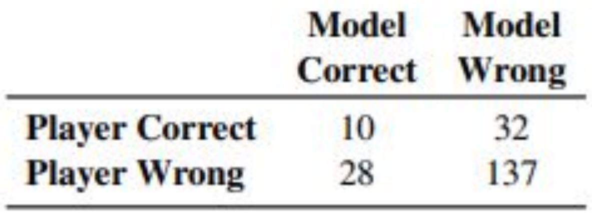
Baseline Models

- **Context LSTM + Power:**
  - Achieved the best performance with a Macro F1 score of 55.13 for detecting actual lies

- **Context LSTM :**
  - Performed slightly worse than the primary baseline with a Macro F1 score of 53.7 and the Lie F1 score was 12.58.

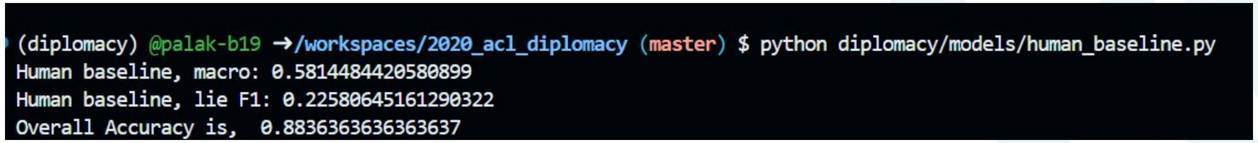
- **Other Models Tested:**
  - Bagofwords.py
  - Harbingers.py
  - Randomandmajoritybaselines.py





	Model Correct	Model Wrong
Player Correct	10	32
Player Wrong	28	137

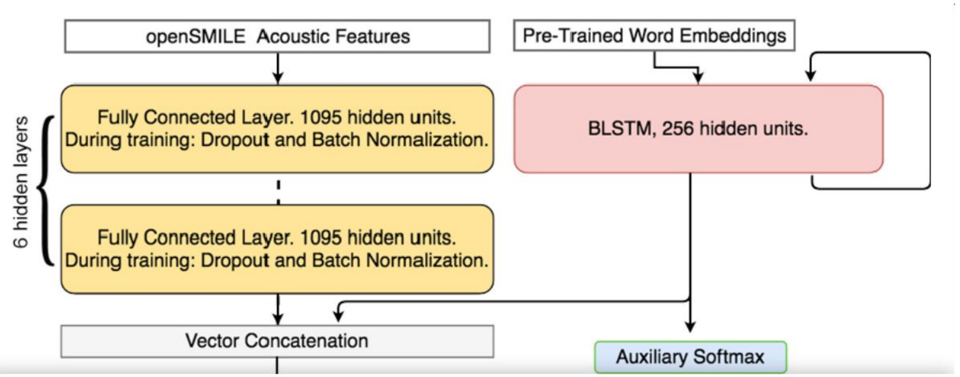
Humanbaseline.py



```
(diplomacy) @palak-b19 →/workspaces/2020_acl_diplomacy (master) $ python diplomacy/models/human_baseline.py
Human baseline, macro: 0.5814484420580899
Human baseline, lie F1: 0.22580645161290322
Overall Accuracy is,  0.8836363636363637
```

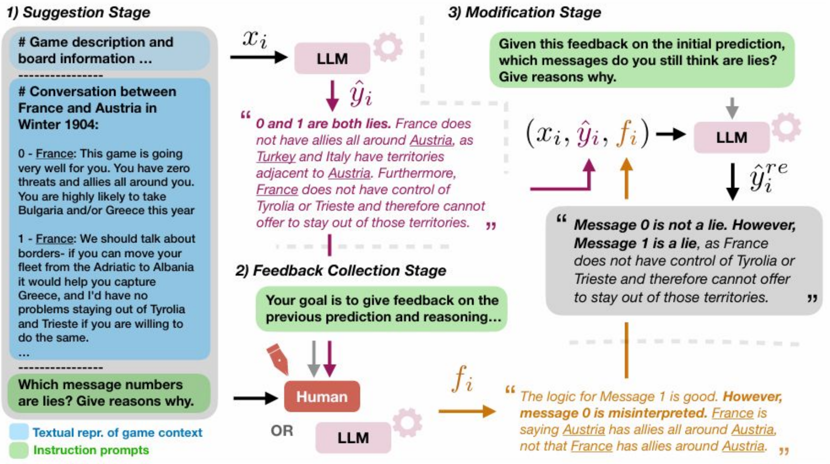
Related Work

Ref 1: Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection


openSMILE Acoustic Features
Pre-Trained Word Embeddings
Fully Connected Layer. 1095 hidden units. During training: Dropout and Batch Normalization.
BLSTM, 256 hidden units.
6 hidden layers
Fully Connected Layer. 1095 hidden units. During training: Dropout and Batch Normalization.
Vector Concatenation
Auxiliary Softmax

1. A deep learning model for deception detection using both acoustic and lexical features, including sentiment indicators and syntactic patterns.
2. Used the Columbia X-Cultural Deception Corpus. Sentence-BERT is used to generate semantic embeddings with word count.

Ref 2: LLMs are Superior Feedback Providers:



Game description and board information ...

Conversation between France and Austria in Winter 1904:

0 - France: This game is going very well for you. You have zero threats and allies all around you. You are highly likely to take Bulgaria and/or Greece this year

1 - France: We should talk about borders- if you can move your fleet from the Adriatic to Albania it would help you capture Greece, and I'd have no problems staying out of Tyrolia and Trieste if you are willing to do the same.

...

Which message numbers are lies? Give reasons why.

“ ***0 and 1 are both lies.*** *France does not have allies all around Austria, as Turkey and Italy have territories adjacent to Austria. Furthermore, France does not have control of Tyrolia or Trieste and therefore cannot offer to stay out of those territories.* ”

Your goal is to give feedback on the previous prediction and reasoning...

“ *The logic for Message 1 is good.* ***However, message 0 is misinterpreted.*** *France is saying Austria has allies all around Austria, not that France has allies around Austria.* ”

Given this feedback on the initial prediction, which messages do you still think are lies? Give reasons why.

“ ***Message 0 is not a lie. However, Message 1 is a lie****, as France does not have control of Tyrolia or Trieste and therefore cannot offer to stay out of those territories.* ”

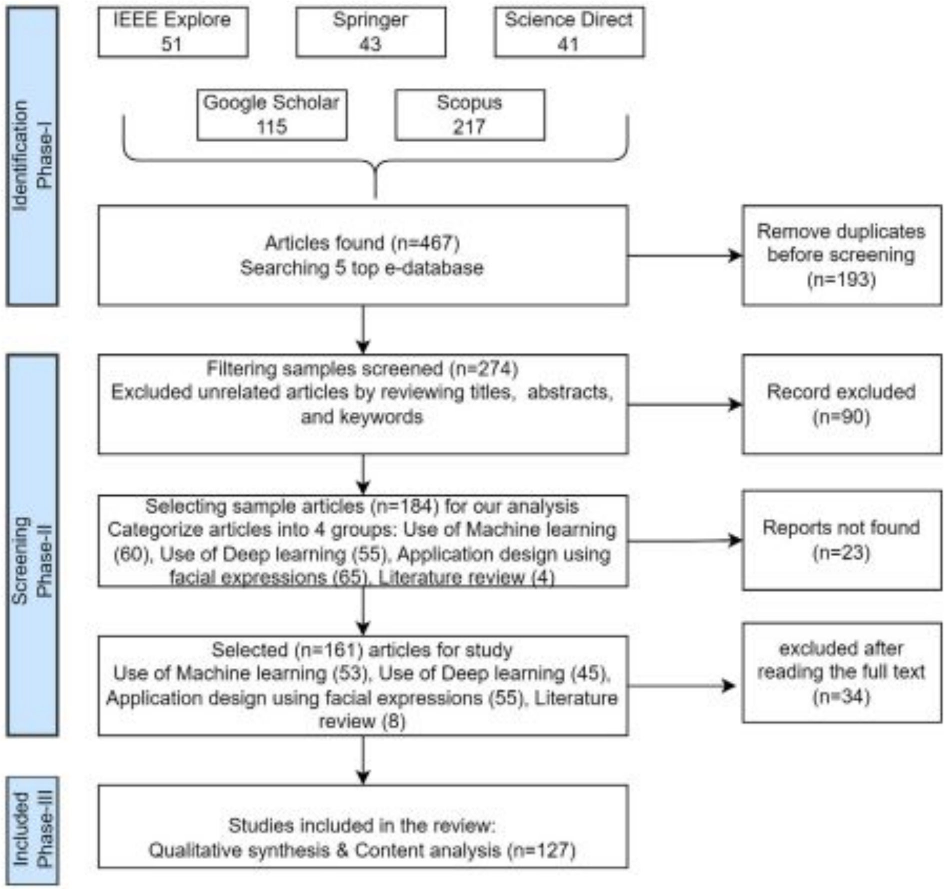
Output - Code Output

1.COT Reasoning implemented after inspiration from the paper attached.
2.The implemented code generates an output and a rationale, the rationale is the then sent again for a better output generation.
3. Llama3 was used to replicate paper results.

Related Work

Ref 3: Deception Detection Using Machine Learning and Deep Learning Techniques

1. Prome et al. emphasize using LSTMs and linguistic features (e.g., sentiment, pronouns) for effective deception detection in text.
2. The model employs a bi-directional LSTM with Sentence-BERT embeddings and LIWC-inspired features like VADER sentiment and modifier counts.
3. A weighted loss function is used to address class imbalance, dynamically adjusting the positive weight to prioritize deceptive messages.



Identification Phase-I

IEEE Explore 51

Springer 43

Science Direct 41

Google Scholar 115

Scopus 217

Articles found (n=467)
Searching 5 top e-database

Remove duplicates before screening (n=193)

Screening Phase-II

Filtering samples screened (n=274)
Excluded unrelated articles by reviewing titles, abstracts, and keywords

Record excluded (n=90)

Selecting sample articles (n=184) for our analysis
Categorize articles into 4 groups: Use of Machine learning (60), Use of Deep learning (55), Application design using facial expressions (65), Literature review (4)

Reports not found (n=23)

Selected (n=161) articles for study
Use of Machine learning (53), Use of Deep learning (45), Application design using facial expressions (55), Literature review (8)

excluded after reading the full text (n=34)

Included Phase-III

Studies included in the review:
Qualitative synthesis & Content analysis (n=127)

Fig. 1. Proposed review methodology for sample collection and analysis.

Methodology

Self attention - Bidirectional LSTM + Oversampling

- Applied on the fused input features (Hb) after concatenating sentence embeddings, linguistic features, country, season, year, scores, etc.
- Handles variable message lengths using packed sequences.
- Learns contextual dependencies in both directions, improving deception pattern recognition across the dialogue.
- Output is normalized and passed to attention for better interpretability
- Used to generate 384-dimensional embeddings for each message using paraphrase-MiniLM-L6-v2.
- “[EMPTY]” messages are encoded as zero vectors.
- Captures semantic meaning of messages, enabling better linguistic and contextual understanding.
- These embeddings form the core part of the model’s input (Eb), later fused with other features

LLM feedback loop

- Sets up the Groq API and initializes the OpenAI client for making requests
- Uses a language model to predict whether game statements are lies or truths.
- Implements checkpointing to save and resume processing without starting over.
- Processes dialogues in batches with error handling and prediction logging.

Discussion and Results

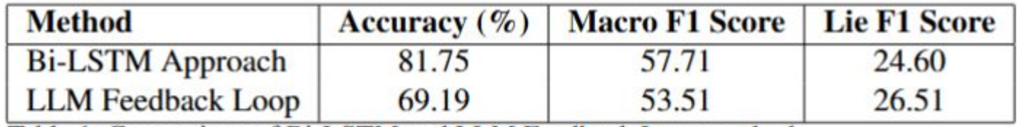
Discussion

- The system uses a neural network with self-attention for deception detection in strategic communication texts.
- Sentence embeddings are generated using a pre-trained model, combined with six linguistic features (e.g., word count, self-references).
- Contextual metadata includes game scores, seasons, and nations, processed using a BiLSTM with self-attention.
- Class imbalance is addressed using a weighted loss function and oversampling, with evaluation based on macro-F1 and false class (lie) detection.

Results

Macro_F1	57.22
Lie_F1	24.77
Accuracy	81.75%

Through semantic embedding techniques which use Sentence-BERT **(paraphrase-MiniLM-L6-v2**) the program extracts profound semantic message content surpassing basic word frequency patterns.This pre-trained language model encodes contextual meaning and nuances in communication.



Method	Accuracy (%)	Macro F1 Score	Lie F1 Score
Bi-LSTM Approach	81.75	57.71	24.60
LLM Feedback Loop	69.19	53.51	26.51

Table 1. Comparison of Bi-LSTM and LLM Feedback Loop methods

Conclusion and Future Work

Conclusion

- Enhanced Deception Detector uses Sentence-BERT, spaCy, VADER, and metadata for deception detection.
- It adopts a multimodal approach, combining semantic, linguistic, and contextual game data for a comprehensive analysis.
- The model effectively handles data imbalance using weighted loss functions and oversampling techniques.
- Performance is further enhanced through parameter optimization methods such as grid search and early stopping.

Future Work

- Use CNNs or RNNs for automated linguistic feature extraction, enhancing generalizability.
- Apply hierarchical modeling across message, turn, and game levels to capture multi-scale context.
- Implement SHAP or LIME for interpretability of key features influencing predictions.
- Analyze model decisions to support trust and debugging.