DECEPTION DETECTION

NLP PROJECT

PRESENTED BY: GROUP 33

GROUP MEMBERS: SHREYANSH SRIVASTAV(2022485), SATWIK GARG(2022461), ASHUTOSH DWIVEDI(2022116)

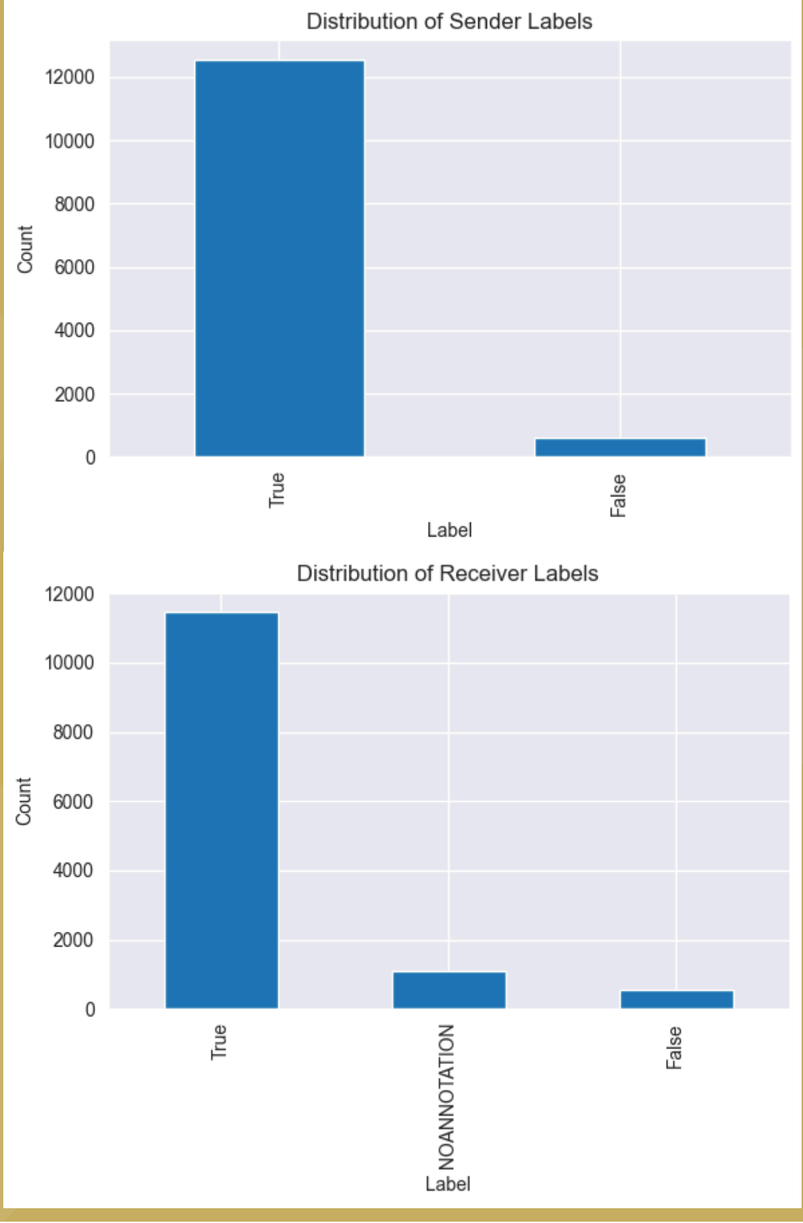
Introduction: Tackling Online Deception

- Online interactions demand trust, but Deception is a significant risk for relationships and negotiations.
- *Examples:* Misleading game allies (Diplomacy), phishing scams, social media misinformation.
- Automated detection is challenging: Lies often subtle, context-driven; true intent data scarce.
- Diplomacy game offers a rich context for studying strategic deception in long-term interactions.
- Peskov et al. (2020) created the valuable Diplomacy dataset with unique *sender intent* & *receiver perception* labels.
- Building on this, we test Graph-based Relational Analysis for Player Honesty In Tactical Exchanges GRAPHITE to potentially enhance lie detection using this data.

Related Work: Learning from Past Work

- **Foundation: Peskov et al. (2020) - Diplomacy Dataset**
  - **Established the core dataset with sender/receiver labels for strategic deception.**
  - **Used baselines (LogReg, LSTM, Power features); showed task difficulty (low Lie F1).**
- **Advanced Technique: Wongkamjan et al. (2025) - CTRL-D**
  - **Focused on detecting deceptive negotiation offers in Diplomacy.**
  - **Used Counterfactual Reinforcement Learning (RL) & CICERO's value function.**
  - **Calculated "Bait," "Switch," "Edge" metrics for a BERT classifier.**
  - **Reported high precision but lower recall; emphasized "strategic friction."**

Dataset Description



Train set : 189 conversations : 13132 messages

Validation set : 21 conversations : 1416 messages

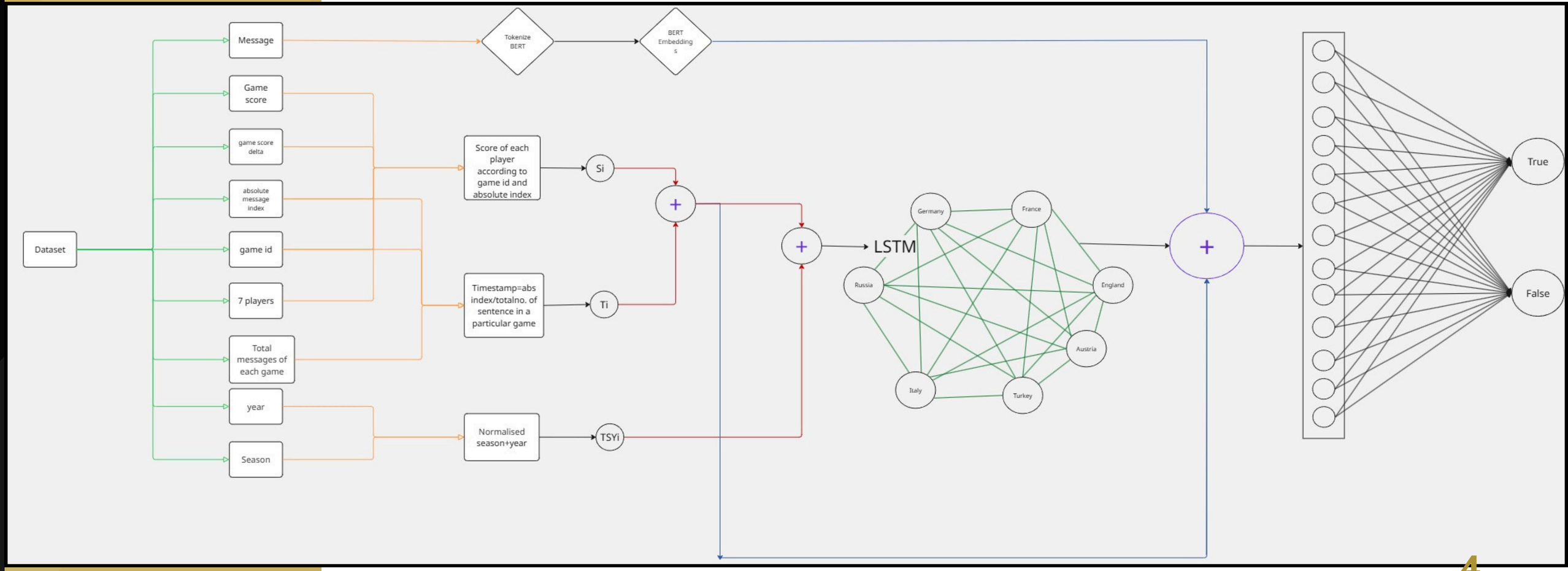
Test set : 42 conversations : 2741 messages

This dataset consists of the following features-

- speakers
- receivers
- messages
- sender_labels
- receiver_labels
- game_score
- score_delta
- absolute_message_index
- relative_message_index
- season
- years
- game_id

GRAPHITE

Graph-based Relational Analysis for Player Honesty In Tactical Exchanges



GRAPHITE

Graph-based Relational Analysis for Player Honesty In Tactical Exchanges

- There are **7 players** in each game:
- England, France, Germany, Italy, Austria, Russia, and Turkey.
- We have constructed a **fully connected graph** with 7 nodes where each node is a **player.** So there will be **7C2 = 21** edges . All the edges between them represent the relationship score between them which is calculated by a **LSTM model.** The inputs of the LSTM model are –
  - Score(Si) of each player at that point in time calculated according to game id and absolute index
  - Timestamp(Ti)=abs index/total no. of sentence in a particular game
  - Normalized Season+year(TSYi)
- We have also taken the **message** as input and applied **BERT tokenization** and **embedding(Ei)** to it.

Now the relationship score between two conversational(sender and receiver) players(from LSTM),embedded text(Ei),Si and Ti are concatenated to give a fully connected layer with 778 neurons and the output is calculated from that.

RESULTS

Model	Macro F1	Lie F1	Accuracy
Context LSTM + Power	0.533	0.13	
Sender (LLM)	0.4529	0.1397	0.6359
Receiver (LLM)	0.4230	0.1151	0.5774
Sender (GRAPHITE)	0.5781	0.2636	0.8125
Receiver (GRAPHITE)	0.5242	0.1086	0.8873

Model	Macro F1	Lie F1
Harbingers	0.528	0.246
Harbingers+Power	0.529	0.237
LSTM	0.538	0.137
Context LSTM	0.558	0.192
Context LSTM+ BERT	0.527	0.135
Context LSTM + Power	0.572	0.27
Context LSTM+Power+BERT	0.561	0.209
Bag of words	0.543	0.191
Bag of words+Power	0.549	0.202

Table 1: Baseline(sender) on test set.

Model	Macro F1	Lie F1
Harbingers	0.459	0.147
Harbingers+Power	0.451	0.155
LSTM	0.538	0.136
Context LSTM	0.543	0.15
Context LSTM+ BERT	0.533	0.151
Context LSTM + Power	0.533	0.13
Context LSTM+Power+BERT	0.536	0.124
Bag of words	0.515	0.137
Bag of words+Power	0.516	0.139

Table 2: Baseline(receiver) on test set.

CONCLUSION

- GRAPHITE outperforms humans in explicit deception detection, achieving a Lie F1 score of **0.2636** on the sender task compared to the human score of **0.225**, demonstrating its ability to identify subtle cues beyond human intuition.
- Despite extreme class imbalance, the model effectively detects rare deceptive instances, showcasing the strength of the class weighting strategy and the fusion of multi-modal features including BERT embeddings and game state data.
- Graph-based relationship modeling significantly boosts performance, highlighting that understanding the dynamic context between players is essential for accurate deception detection—not just the message content.

THANK YOU!