Deception Detection in Online Mafia Game Interactions

Lisa Fu
Department of Computer Science, Stanford University

Introduction

We all lie, for a variety of reasons. Sometimes, they can be little white lies, perhaps to flatter or protect. In more devastating cases, they can lead to innocent on death row, or result in evasion of criminal cases. This project aims to answer the following:

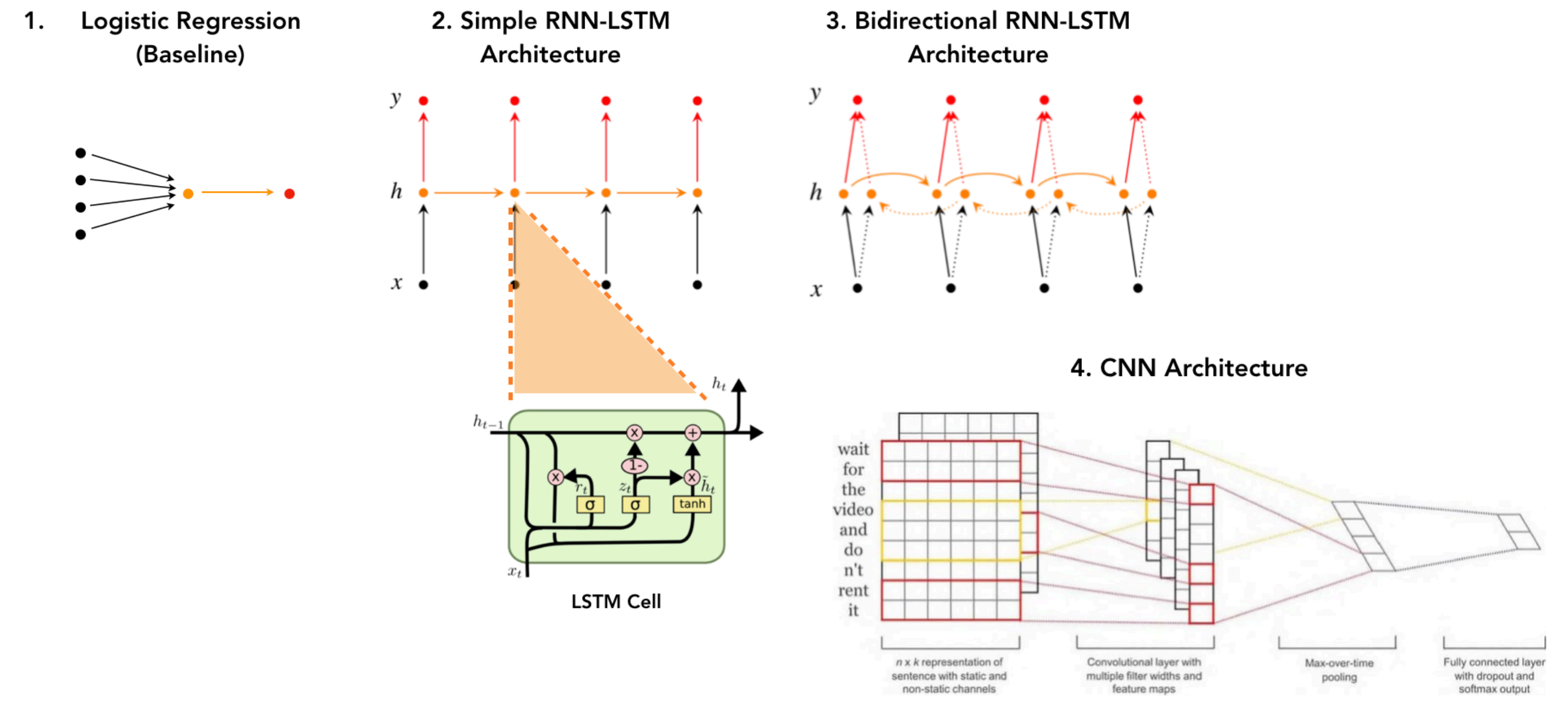
- What are the linguistic subtleties and differences that may reflect deceptive vs. truthful language?
- How can deep learning and NLP techniques be applied to differentiate these cues?

"Mafia" as Proxy for Deception

- Party game simulating conflict between uninformed majority ("**Innocents**") and informed minority ("**Mafia**")
- Two phases: Mafia secretly "murder" an innocent during "night", Innocents vote to eliminate suspect during "day"
- Game ends when all Mafia eliminated or when # of Mafia members > # of Innocents

Methodology

Our primary dataset is the Mafiascum Dataset, a corpus of text scraped from 700+ games of Mafia on an internet forum (Mafiascum)[1]. After sanitizing, we were able to work with 3500 labeled text documents, with each document containing text for all of single player's interactions throughout a single game of Mafia.

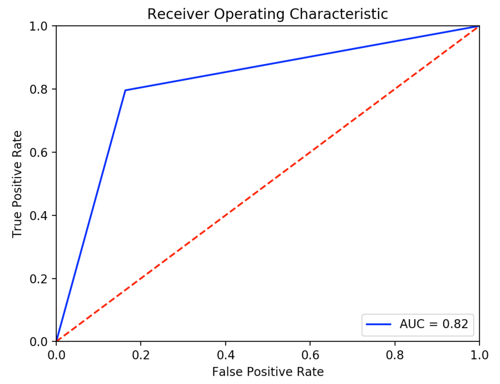

1. Logistic Regression (Baseline)
2. Simple RNN-LSTM Architecture
3. Bidirectional RNN-LSTM Architecture
4. CNN Architecture

LSTM Cell

y
h
x
h_{t-1}
h_t
x_t
r_t
z_t
$\tilde{h}_t$
σ
σ
tanh
1-
wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

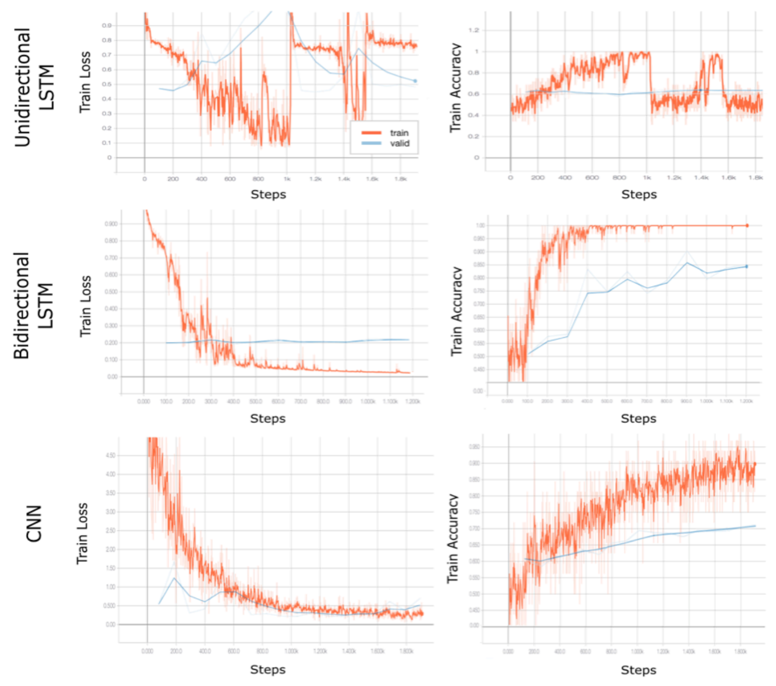
Results

Model	AUROC	Accuracy	Precision	Recall	F_1
Logistic Regression	0.816	0.816	0.815	0.820	0.815
Unidirectional LSTM	0.880	0.888	0.874	0.880	0.875
Bidirectional LSTM	0.945	0.947	0.948	0.945	0.946
CNN	0.948	0.949	0.948	0.948	0.948

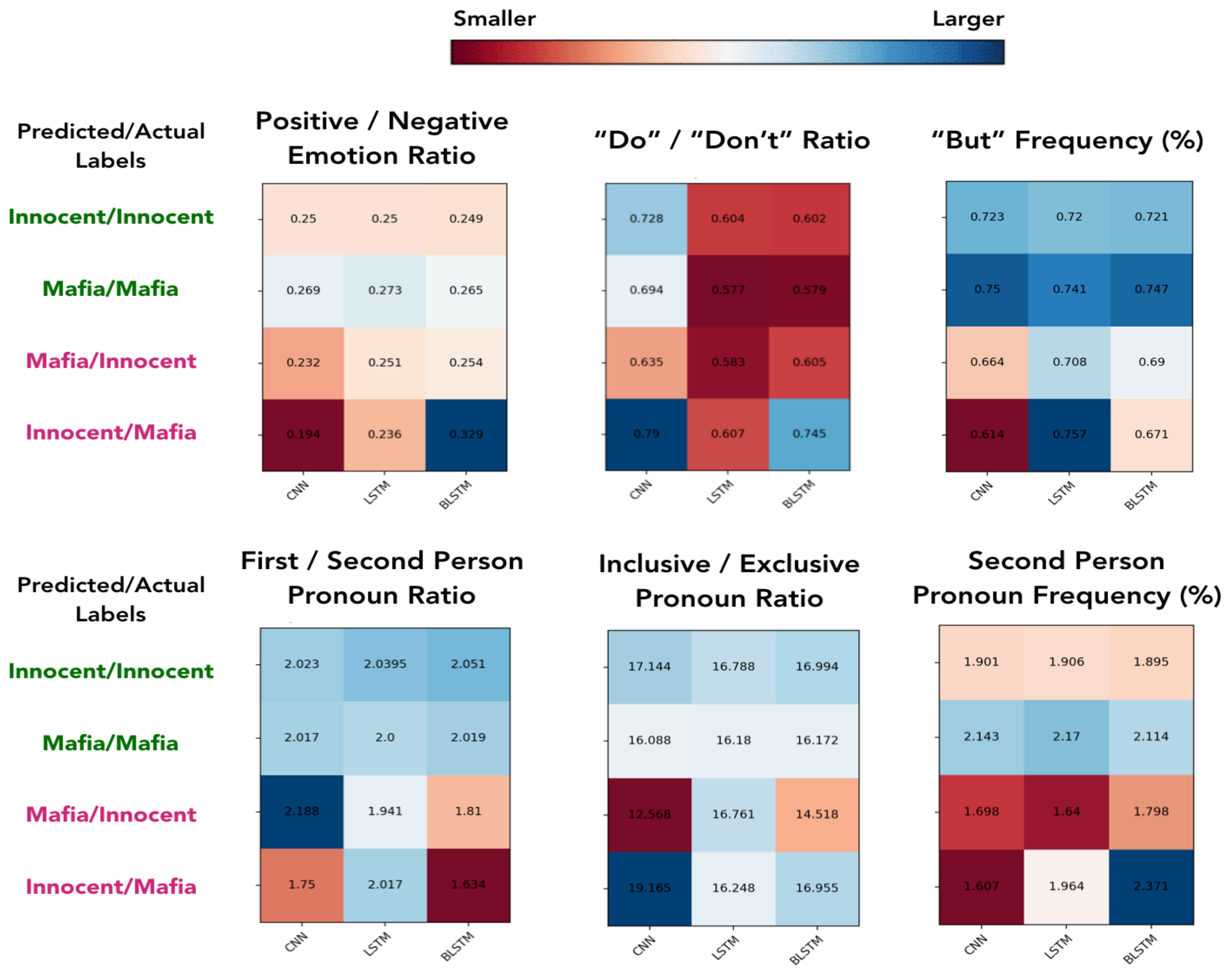


Baseline: Bag of Words + Logistic Regression (AUROC = 0.82)

Loss/Accuracy Curves for RNN-LSTM, Bidirectional RNN-LSTM, and CNN architecture models (AUROC = 0.880, 0.945, 0.948, respectively)



Ratio/Frequency of Handpicked Linguistic Signals
(chosen based on qualitative analysis and deception research)



Positive / Negative Emotion Ratio

Predicted/Actual Labels	CNN	LSTM	BLSTM
Innocent/Innocent	0.25	0.25	0.249
Mafia/Mafia	0.269	0.273	0.265
Mafia/Innocent	0.232	0.251	0.254
Innocent/Mafia	0.194	0.236	0.329

"Do" / "Don't" Ratio

Predicted/Actual Labels	CNN	LSTM	BLSTM
Innocent/Innocent	0.728	0.604	0.602
Mafia/Mafia	0.694	0.577	0.579
Mafia/Innocent	0.635	0.583	0.605
Innocent/Mafia	0.79	0.607	0.745

"But" Frequency (%)

Predicted/Actual Labels	CNN	LSTM	BLSTM
Innocent/Innocent	0.723	0.72	0.721
Mafia/Mafia	0.75	0.741	0.747
Mafia/Innocent	0.664	0.708	0.69
Innocent/Mafia	0.614	0.757	0.671

First / Second Person Pronoun Ratio

Predicted/Actual Labels	CNN	LSTM	BLSTM
Innocent/Innocent	2.023	2.0395	2.051
Mafia/Mafia	2.017	2.0	2.019
Mafia/Innocent	2.188	1.941	1.81
Innocent/Mafia	1.75	2.017	1.634

Inclusive / Exclusive Pronoun Ratio

Predicted/Actual Labels	CNN	LSTM	BLSTM
Innocent/Innocent	17.144	16.788	16.994
Mafia/Mafia	16.088	16.18	16.172
Mafia/Innocent	12.568	16.761	14.518
Innocent/Mafia	19.165	16.248	16.955

Second Person Pronoun Frequency (%)

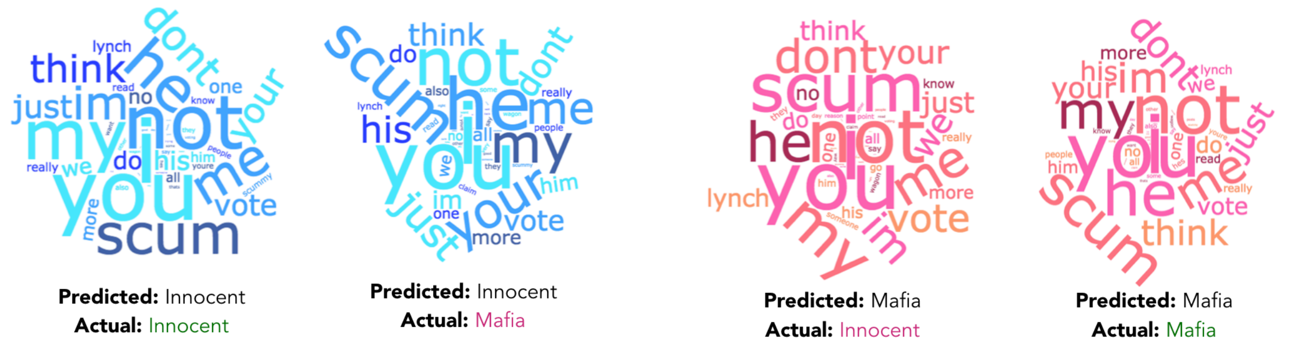
Predicted/Actual Labels	CNN	LSTM	BLSTM
Innocent/Innocent	1.901	1.906	1.895
Mafia/Mafia	2.143	2.17	2.114
Mafia/Innocent	1.698	1.64	1.798
Innocent/Mafia	1.607	1.964	2.371

Analysis

Linguistic Cues

- Other person oriented pronouns more correlated with deception (in agreement with [1][2]) – disassociate from lie
- Negative emotion more correlated with deception (agreement with [2])
- Higher frequency of "but" in deceitful language (agreement with [1])
- "Do" more frequently in truthful language (vs. "Don't" in deceitful language)

Word Cloud of Term Frequencies



Predicted: Innocent **Actual:** Innocent

Predicted: Innocent **Actual:** Mafia

Predicted: Mafia **Actual:** Innocent

Predicted: Mafia **Actual:** Mafia

Conclusion and Future Improvements

Model Evaluation

- High test accuracy suggests overfitting of models
- Need more finetuning of hyperparameters
- Experiment with feature vector representation
- Performance of Bi-LSTM greater than Uni-LSTM, this is expected due to greater contextual association in bidirectionality

Exploring other proxies for deception

- Child-child conflict resolution dialogue (blame vs. truthtelling)
- Include multiple modalities for deception (audio, behavioral, textual)

Acknowledgements

Thank you to the CS224N teaching team for your support (and Annie Hu for being my project mentor), the engaging class material/guest speakers, and for the opportunity to conduct hands-on experiments in a custom area of interest!

References

[1] de Ruiter, Bob Kachergis, George. (2018). The Mafiascum Dataset: A Large Text Corpus for Deception Detection.
[2] Dou, J., Liu, M., Muneer, H. Schlussel, A. (2015). What Words Do We Use to Lie?: Word Choice in Deceptive Messages. CHI.