Deception Detection using Real-life Trial Data

Verónica Pérez-Rosas
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
vrncapr@umich.edu

Mohamed Abouelenien
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
zmohamed@umich.edu

Rada Mihalcea
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
mihalcea@umich.edu

Mihai Burzo
Mechanical Engineering
University of Michigan, Flint
Flint, MI 48502, USA
mburzo@umich.edu

Deception Detection using Real-life Trial Data

Verónica Pérez-Rosas
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
vrncapr@umich.edu

Mohamed Abouelenien
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
zmohamed@umich.edu

Rada Mihalcea
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
mihalcea@umich.edu

Mihai Burzo
Mechanical Engineering
University of Michigan, Flint
Flint, MI 48502, USA
mburzo@umich.edu

Consider the Hypothetical Scenario:

- You are on trial
- For (allegedly) murdering your nephew
- You know you're innocent
- But there are "witnesses" against you...

It can happen

“Of that, I’m innocent...”

-Tyrion Lannister

But there are 2 options

1.- Trial by Combat

“I’m Oberyn Martell.”

-Oberyn Martell

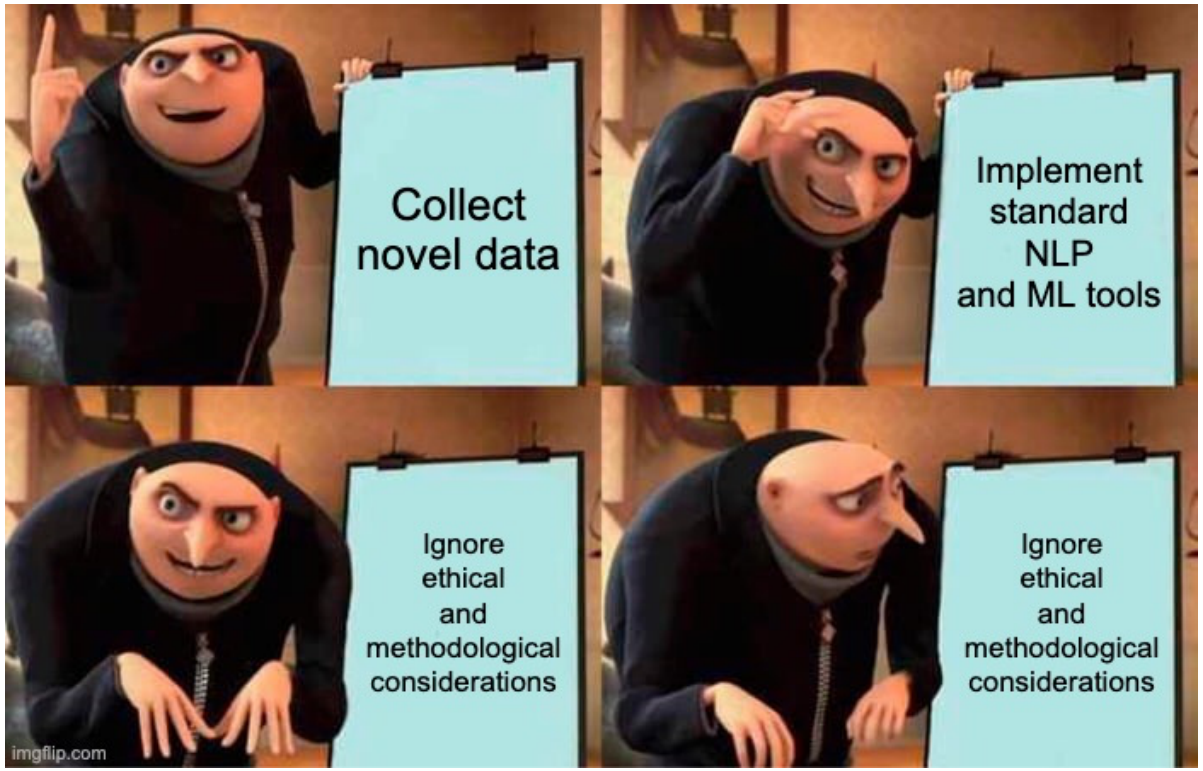
2.- Trial by Deception Detection using Real-life Trial Data

Which would you choose?

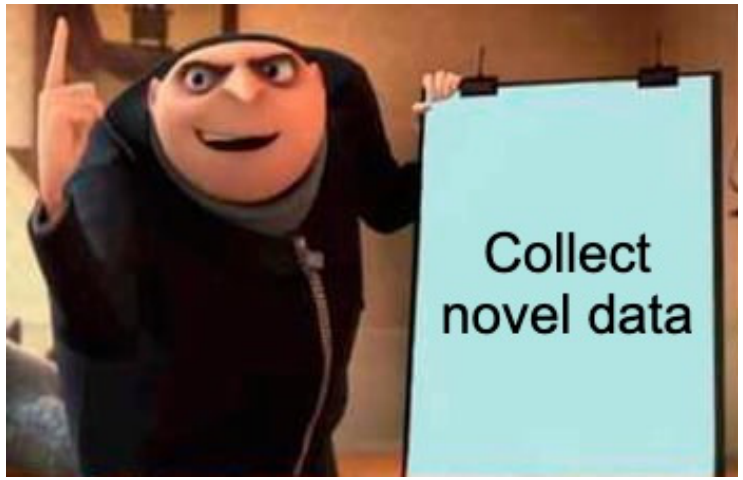
Pros: Handsome
Pros: Hispanic (?)
Cons: Overconfident

Pros: Data Science
Cons: University of Michigan
Cons: Overconfident

Lets talk about option 2...

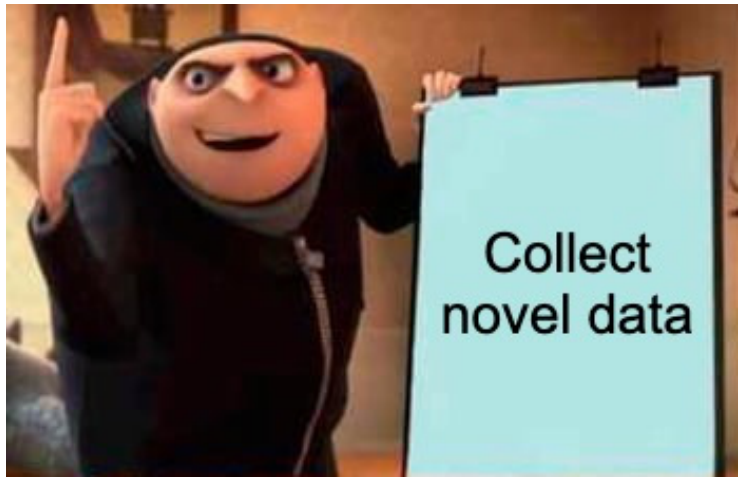


- Public recorded trials.
- Defendant or witness clearly identified physically and verbally.
- “Three different trial outcomes that helped us to correctly label a certain trial video clip as deceptive or truthful: guilty verdict, non-guilty verdict, and exoneration.”



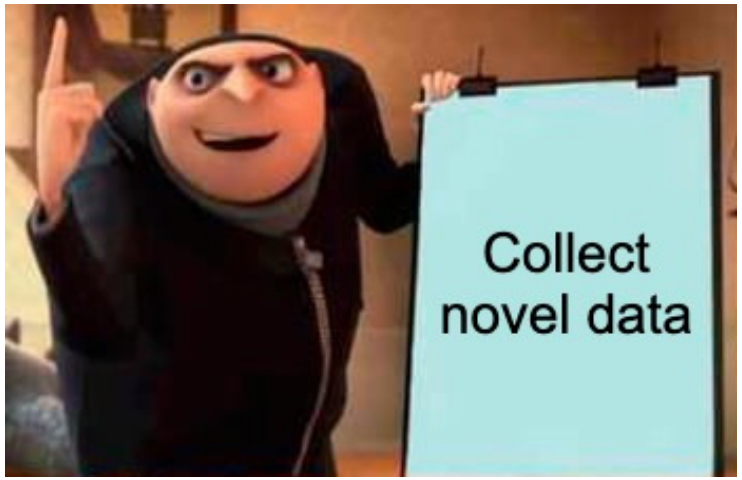
- 121 videos
  - 61 deceptive
    - 28 seconds per video
    - 35 male
    - ages 16 to 60

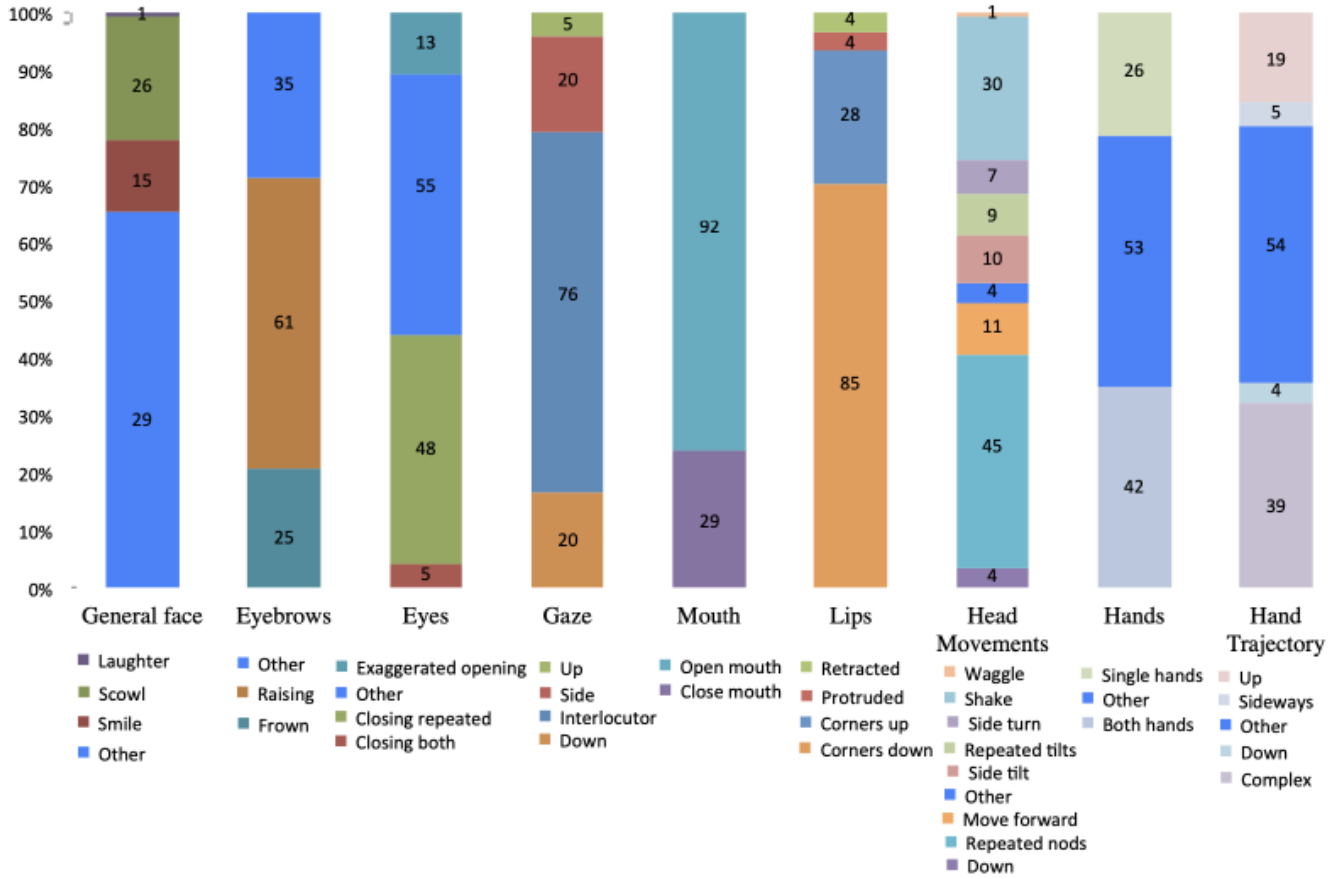
- Video clips are transcribed via crowd sourcing using Amazon Mechanical Turk.
- Obtained transcriptions were manually verified to avoid spam and ensure their quality.



- 8,055 words
  - 66 words per transcript

- The multimodal annotation was performed by two annotators who had to choose one label that dominated for each of the nine gestures





- Unigrams and bigrams from the bag-of-words of the transcripts with frequencies greater than 10.
- Binary feature for each of the 40 available gesture labels.
- “In our current work we are not focusing on the insights that can be gained from linguistic analyses”

- Decision Trees
- Random Forests

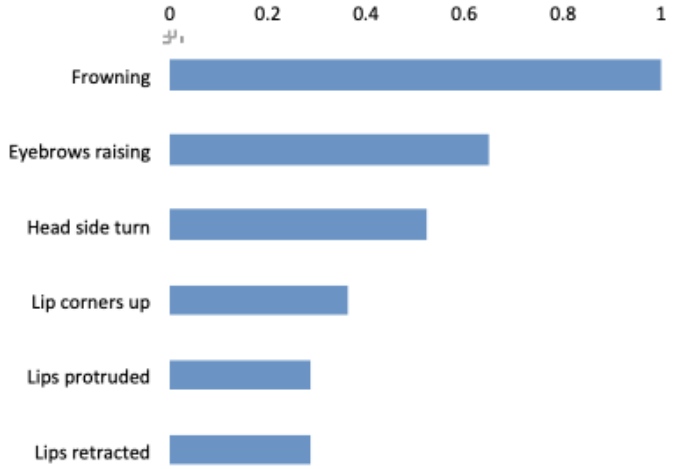
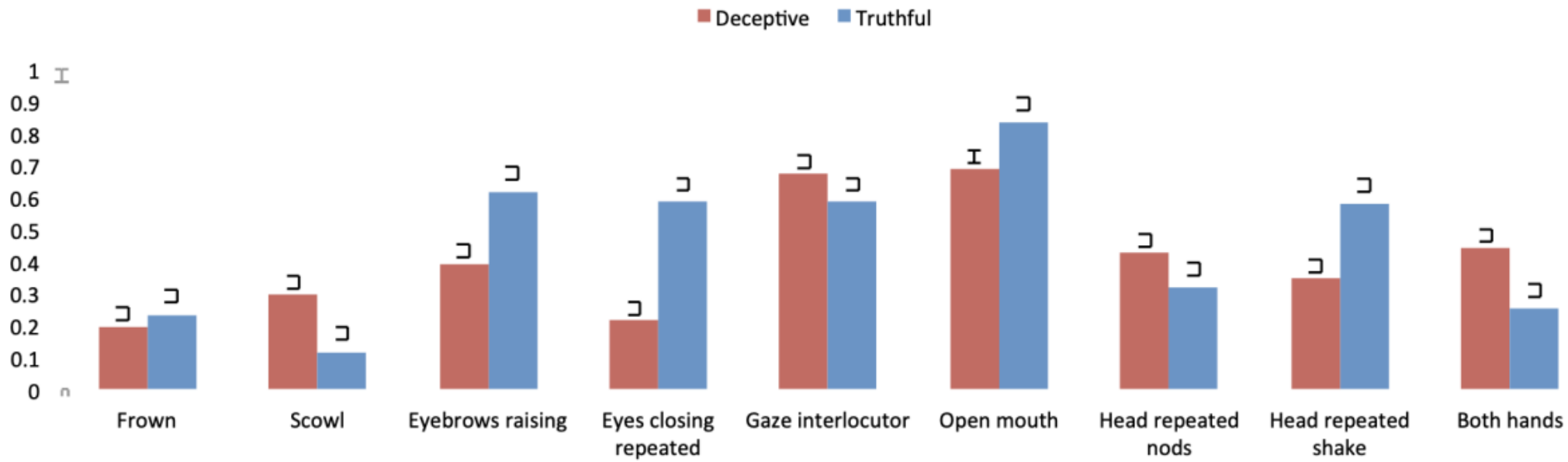


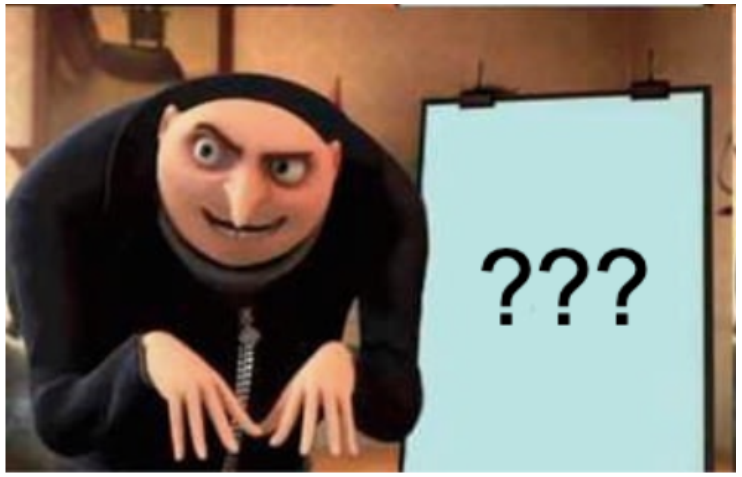
Figure 4: Weights of top non-verbal features

Feature Set	DT
All	**75.20%**
– Hand gestures	71.90%
– Facial displays	59.50%
– Bigrams	66.94%
– Unigrams	61.98%

Table 4: Feature ablation study.



- Evaluate human ability to identify deceit on trial.
- Each annotated a total of 121*4 = 484 instances WITHOUT pay.
- “Annotators were not offered a monetary reward and we considered their judgments to be honest as they participated voluntarily in this experiment.”

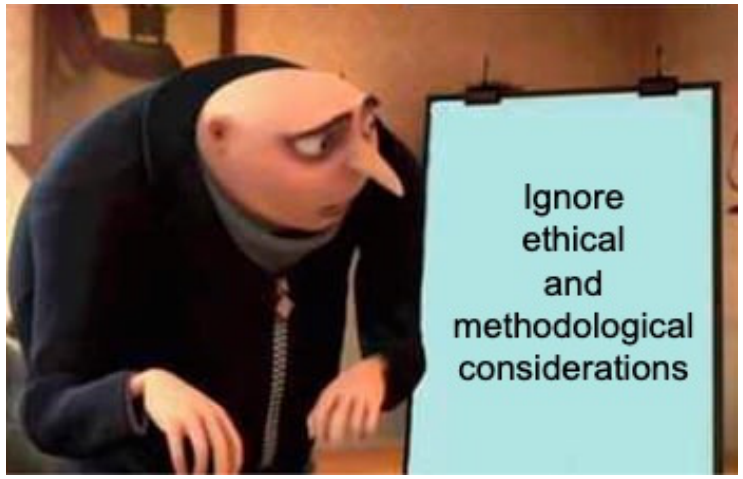


	Text	Audio	Silent video	Full video
A1	54.55%	51.24%	45.30%	56.20%
A2	47.93%	55.37%	46.28%	53.72%
A3	50.41%	59.50%	47.93%	59.50%
Sys	60.33%	NA	68.59%	75.20%

Table 6: Performance of three annotators (A1, A2, A3) and the developed automatic system (Sys) on the real-deception dataset over four modalities.

Why so many red flags?

- “Three different trial outcomes that helped us to correctly label a certain trial video clip as deceptive or truthful: guilty verdict, non-guilty verdict, and exoneration.”



“Moreover, the average number of exonerations per year increased from 3.03 in 1973-1999 to 4.29 between 2000 and 2013.”

Deception Detection using Real-life Trial Data

Verónica Pérez-Rosas
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
vrncapr@umich.edu

Mohamed Abouelenien
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
zmohamed@umich.edu

Rada Mihalcea
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
mihalcea@umich.edu

Mihai Burzo
Mechanical Engineering
University of Michigan, Flint
Flint, MI 48502, USA
mburzo@umich.edu

- 121 transcripts with ~66 words each

Small & biased data

Is it possible that people behave differently if they know their words and appearances will be recorded?”

Meng-Jin Lin, cineast*

*Could be.

- 121 transcripts with ~66 words each

Small & biased data

"I do think one potential bias is that all of the data comes from publicly-released trials. However, many trials, including juvenile and highly-sensitive trials, are not released."

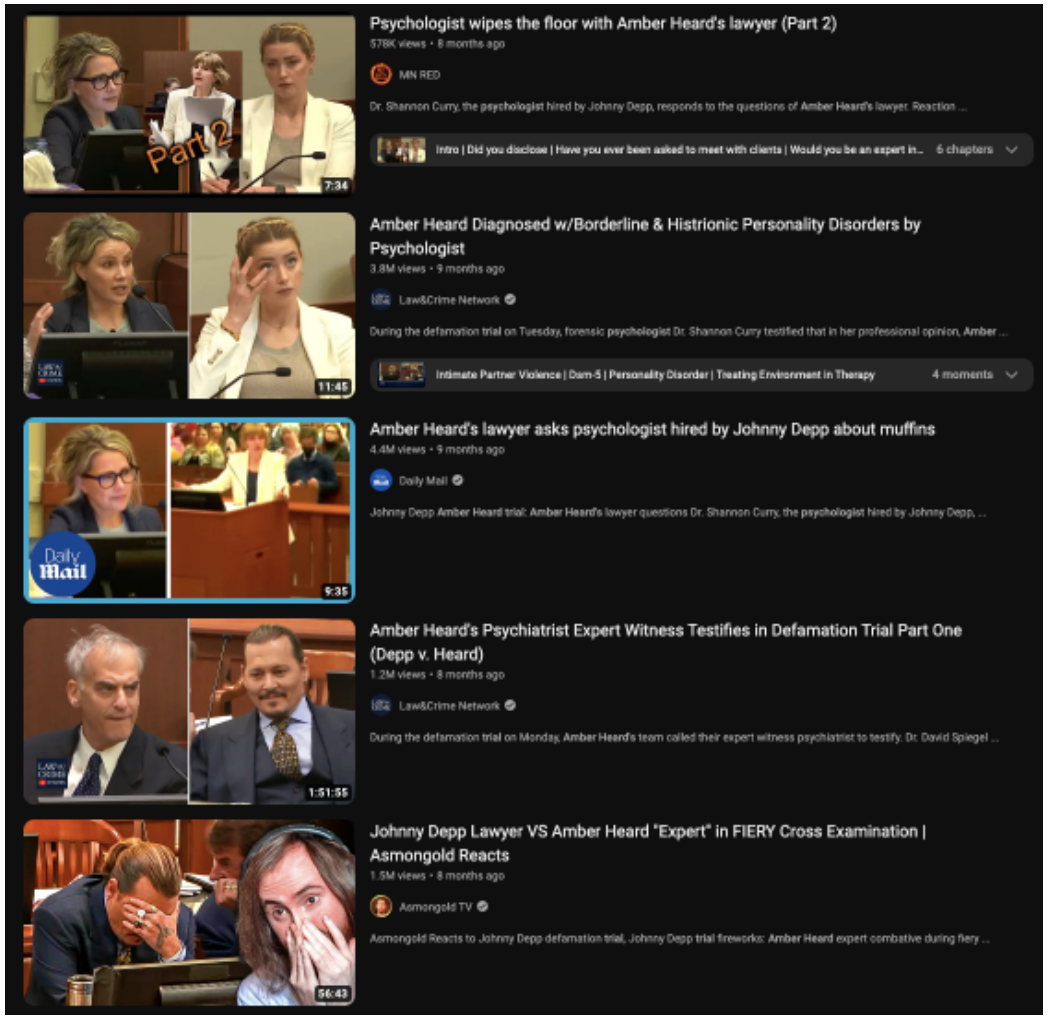
Adhitya Venkatraman, CSI Miami enjoyer*

*No shame.

Are we missing some sources of biases in the data?

Discuss...

What about professional liars?



- The multimodal annotation was performed by two annotators who had to choose one label that dominated for each of the nine gestures

- For the first 56 videos the average agreement was 75.16%.
- After discussions, the rest were labeled by only one annotator.

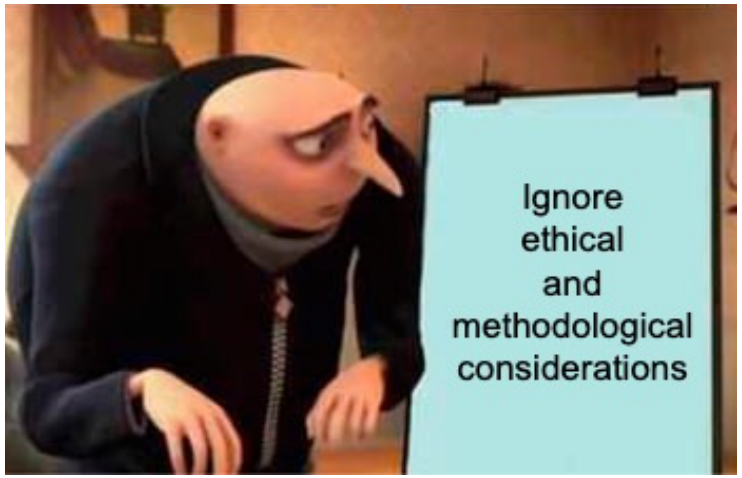
- The multimodal annotation was performed by two annotators who had to had to choose one label that dominated for each of the nine gestures

"Raising eyebrows and frowning are objectively not mutually exclusive categories."

Mike Hardy, Anatomy expert*

*No innuendo.

- The multimodal annotation was performed by two annotators who had to choose one label that **dominated** for each of the nine gestures



*"Facial expressions also play a critical role in the identification of deception. **Ekman** defined micro-expressions as **relatively short** involuntary expressions, which can be indicative of deceptive behavior."*

Not Paul Ekman.

- “In our current work we are not focusing on the insights that can be gained from linguistic analyses”

Why?
What could we learn from this?

Discuss...

- “Annotators were not offered a monetary reward and we considered their judgments to be honest as they participated voluntarily in this experiment.”

“These annotators then each gave ~4 hours of unpaid work to the researchers.”

Devon Rojas, Labor economist*

*Should be.

- Evaluate human ability to identify deceit on trial.

	Text	Audio	Silent video	Full video
A1	54.55%	51.24%	45.30%	56.20%
A2	47.93%	55.37%	46.28%	53.72%
A3	50.41%	59.50%	47.93%	59.50%
Sys	60.33%	NA	68.59%	75.20%

Table 6: Performance of three annotators (A1, A2, A3) and the developed automatic system (Sys) on the real-deception dataset over four modalities.

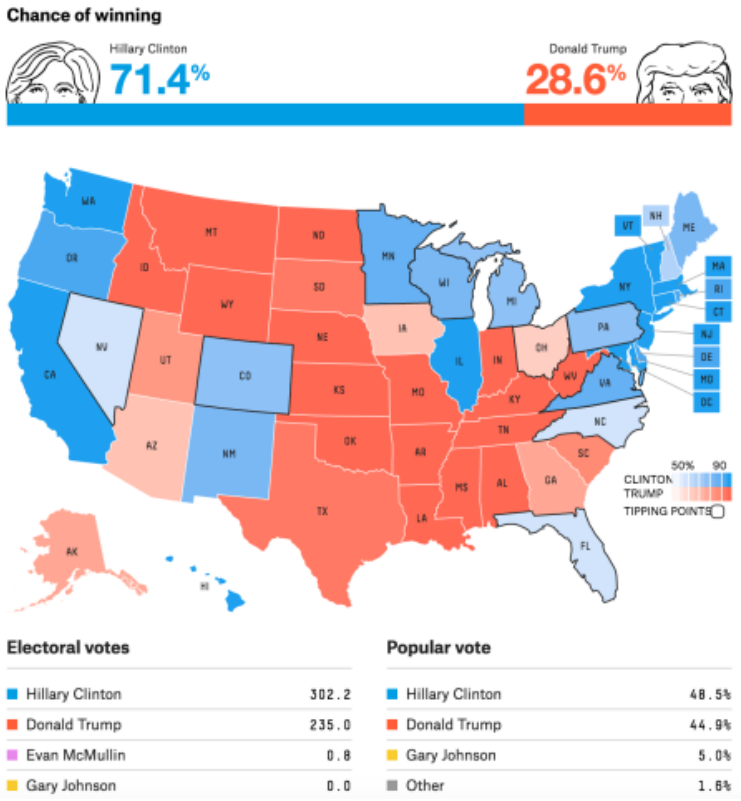
“That’s not fair.”

Me.

Implications

Feature Set	DT
All	**75.20%**
– Hand gestures	71.90%
– Facial displays	59.50%
– Bigrams	66.94%
– Unigrams	61.98%

Table 4: Feature ablation study.



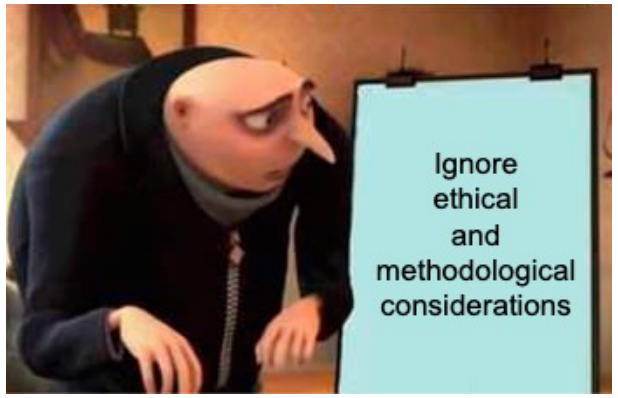
“Is the problem of detecting deceit inherently ethically concerning? Who does it benefit? Who does it harm?”

What protections could be put in place in the future to prevent these potential harms?

Kaitlyn Zhou, best TA who always gives A’s*

Discuss...

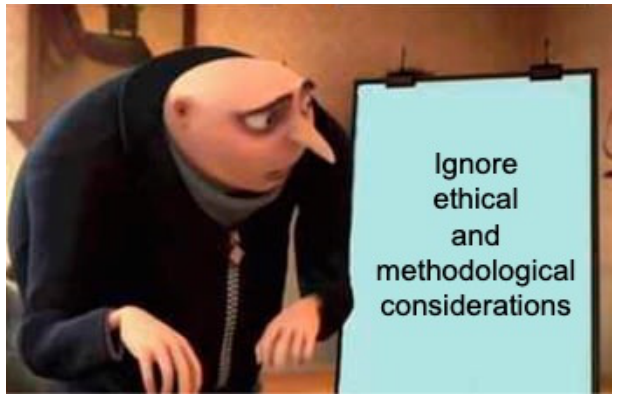
*Hopefully



“In CS 184 we read a lot about the COMPAS algorithm which produces risk scores that should be used to determine an individual’s eligibility for various welfare programs but was often used instead to extend sentences. [...] the algorithm was found to assign scores that perpetuated existing harm towards marginalized communities.”

Clare Chua, CS 184 propagandist*

*Blatantly.



“It is known that facial recognition is often more accurate when detecting lighter skinned people, so I am concerned that the facial expression detection will be inaccurate in the same way. […] I am also worried about the analysis of body language, since **neurodivergent** people may have different body language that could be seen as deceptive.

Sophia Angelica Ramsey*

*I’ve got nothig here.

“The motivations behind undertaking the project, the authors highlight the perils of ‘falsely accusing the innocents and freeing the guilty.’ However, the paper does not actively discuss nor showcase the classification algorithm’s results on that front. […] we have no information on how the algorithm fares with falsely exonerating the guilty and falsely accusing the innocent.”

Allex Desronvil, poet*

“Their model predicts who will be found guilty, as opposed to who is actually guilty.”

Nourya Cohen, lawyer*

*But no from Michigan.

“How should we generally think about policies that seek to standardize benchmarking for AI in government/criminal applications?”

Joel A Johnson, self-published sci-fi author*

At what accuracy% would you be confortable implementing the algorithm?

Discuss...

*It’s easier these days.

Which would you choose?

Data Science Person