Multimodal Sarcasm

Santiago Castro*, Devamanyu Hazarika*, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, Soujanya Poria

(* equal contribution)

Goal

Build a **dataset** for multimodal **Sarcasm Detection** (**video**+**audio**+**text**) and provide **baselines**.

Motivation

—Really?

Is it sarcastic? We don't know.
Previous work have focused on text only. We propose to use **videos** (along with audio and transcription text).

—Really?
Audio: neutral tone
Video: neutral face
(sarcastic)

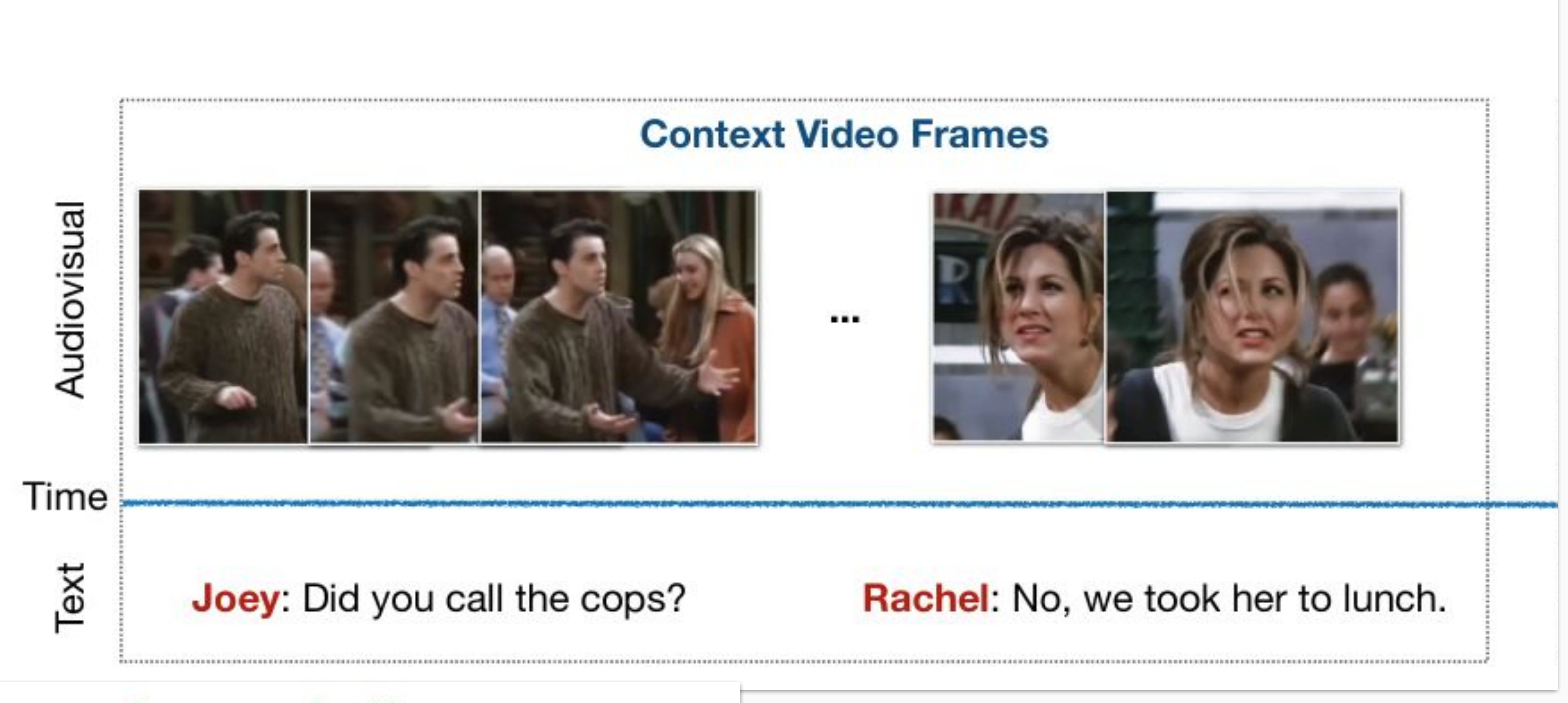
—Really?
Audio: rising tone
Video: smile
(non-sarcastic)

Data Collection

1. **2 people** annotated 6421 videos coming from **The Big Bang Theory** episodes and 624 videos coming from **Friends**, **The Golden Girls**, and **Sarcasmaholics Anonymous**.
2. Kappa scores of 0.2326 and 0.5877, respectively.
3. A **third person** broke ties.
4. We filtered out the bad quality ones and least agreed, and crafted a **balanced** dataset.

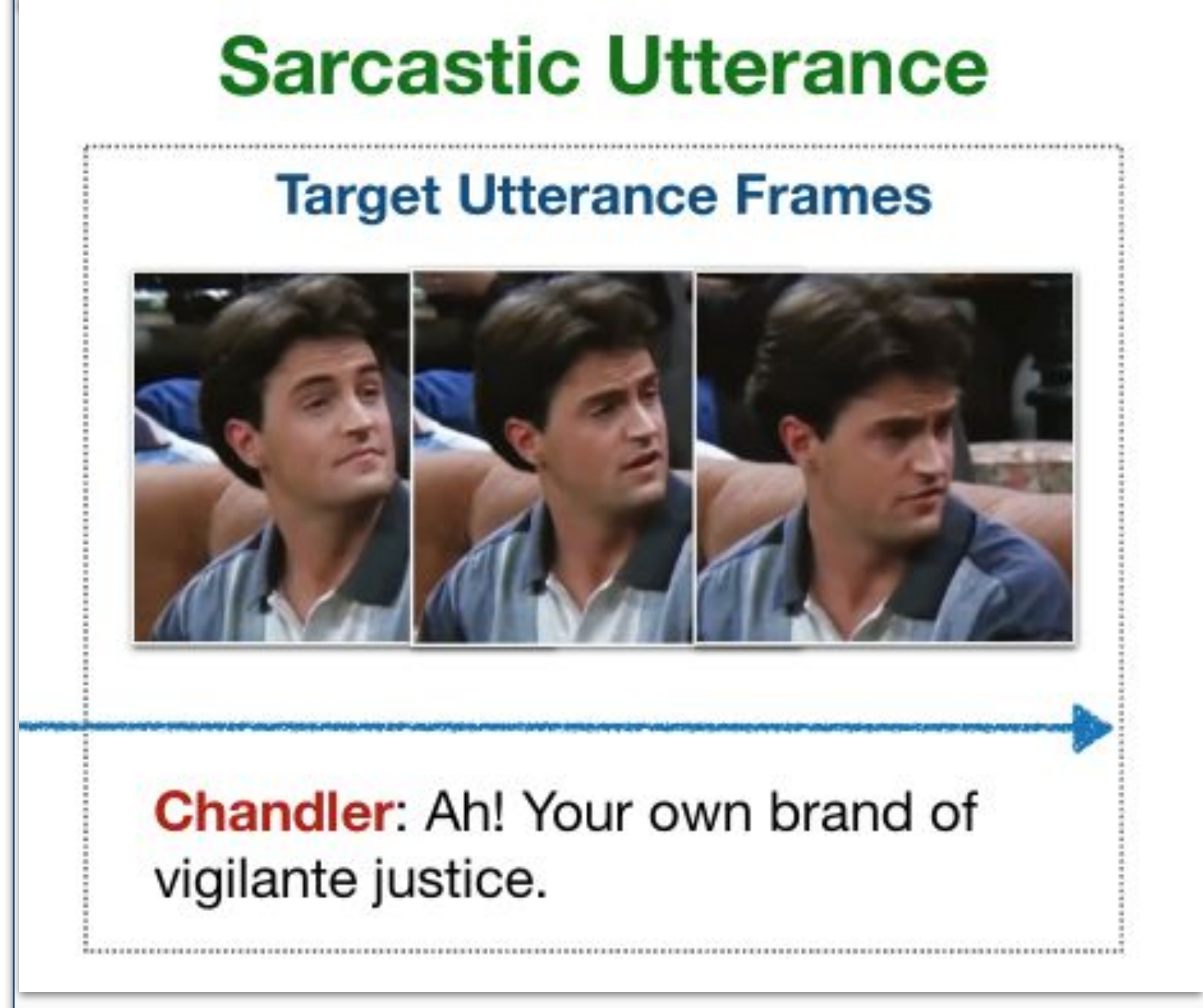
Dataset

- **690** one-utterance videos, avg. duration 5s.
- Balanced, labeled as **sarcastic or non-sarcastic**.
- Come with **transcripts** and preceding **context** video, avg. duration of 14s.



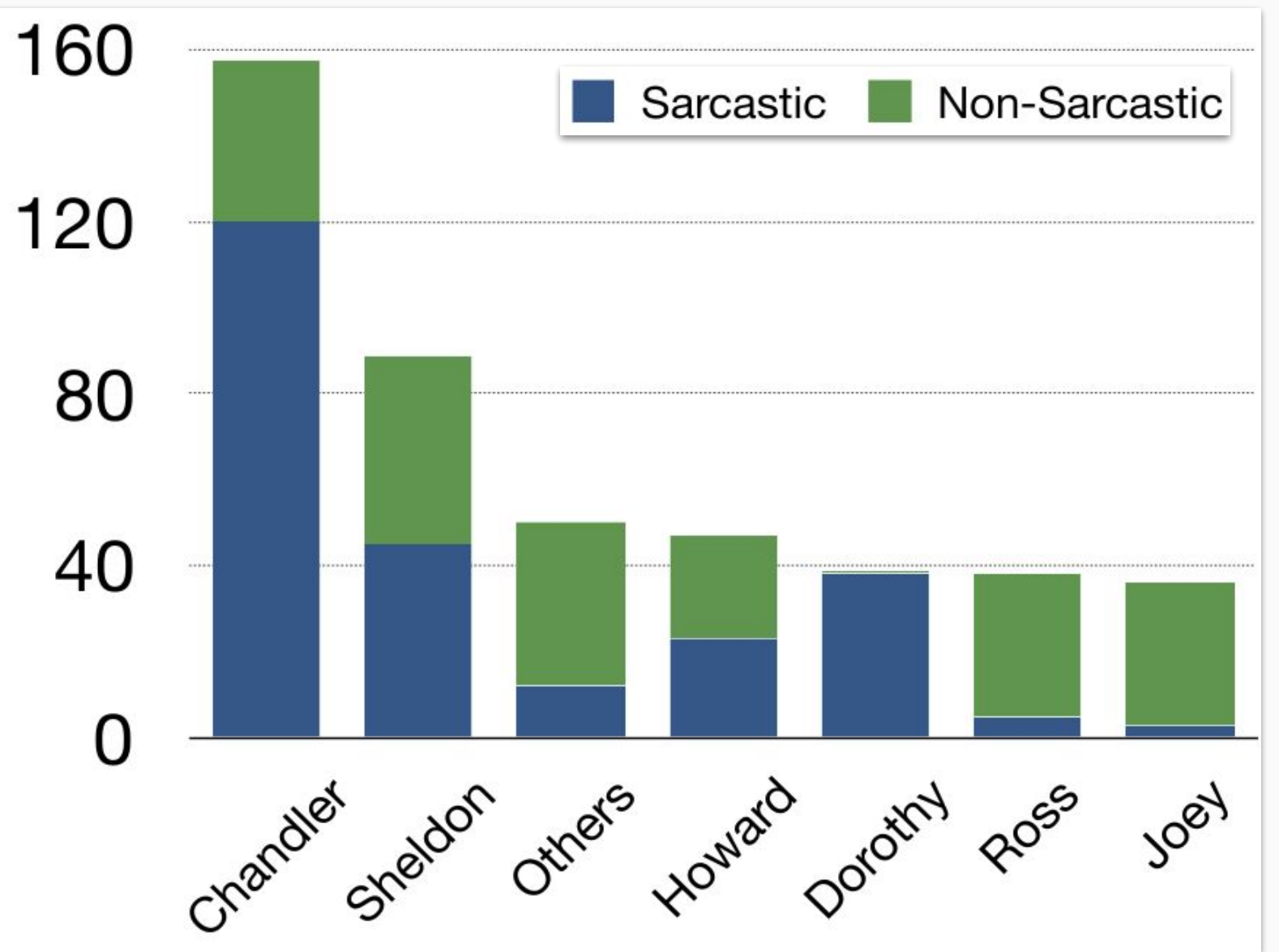
Joey: Did you call the cops?

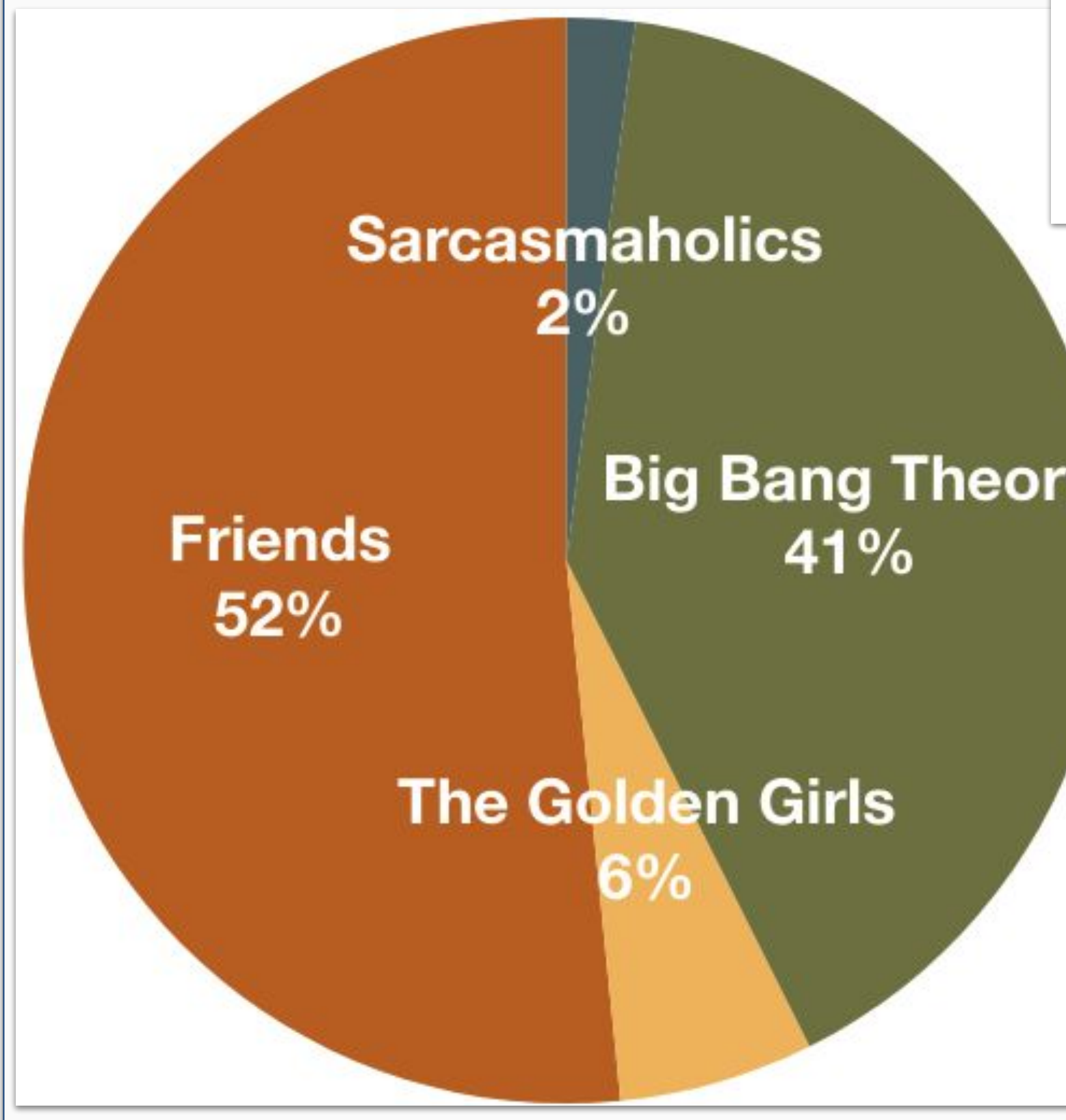
Rachel: No, we took her to lunch.

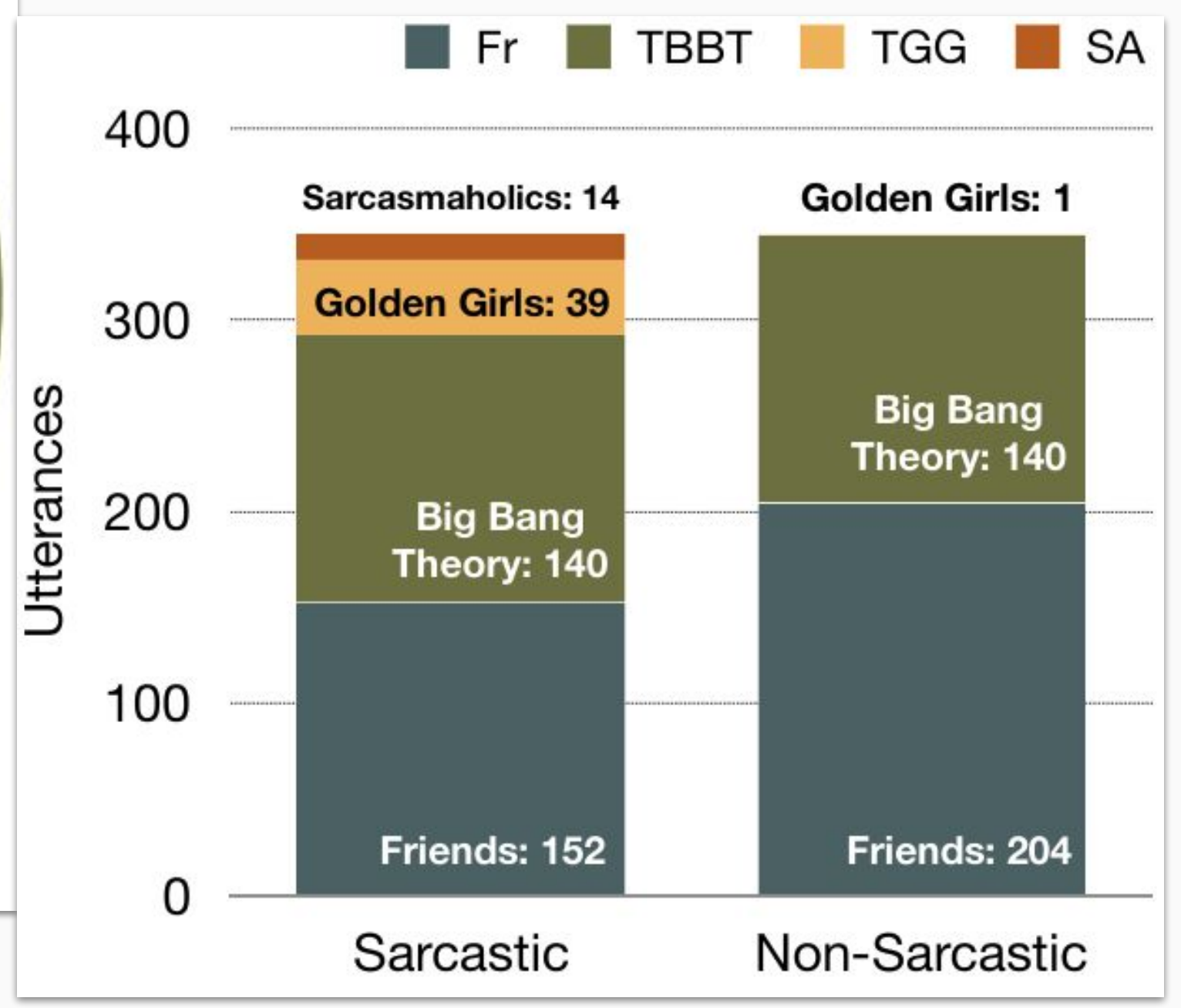


Chandler: Ah! Your own brand of vigilante justice.

Stats







Baselines

Algorithm	Modality	Precision	Recall	F-Score
Majority	-	25.0	50.0	33.3
Random	-	49.5	49.5	49.8
SVM	T	60.5	59.8	58.9
	A	66.4	66.2	66.1
	V	68.6	68.5	68.4
	T+A	66.7	66.5	66.4
	A+V	67.1	66.9	66.8
	T+V	**71.8**	**71.7**	**71.7**
	T+A+V	67.1	66.9	66.8
$\Delta_{multi-unimodal}$ Error rate reduction		↑ 3.2% ↓ 10.2%	↑ 3.2% ↓ 10.2%	↑ 3.3% ↓ 10.4%

Text: [CLS] token repr. from the last 4 layers, from BERT-base cased.

Video: avg. ResNet-152 *pool5* layer.

Audio: MFCC, melspectrogram, spectral centroid and their associated temporal derivatives.

Conclusion

- We provide a dataset for Sarcasm study with video+audio+text with 690 videos.
- We provide some strong baselines.

Future Work

- Better fusion of the modalities.
- More advanced Neural methods.
- Main speaker localization.