57th | ACL 2019 Florence
ANNUAL MEETING July 28th - August 2nd
of the Association for Computational Linguistics

Towards Multimodal Sarcasm Detection (An *Obviously* Perfect Paper)

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, Soujanya Poria
{sacastro,vrncapr,mihalcea}@umich.edu, {hazarika,rogerz}@comp.nus.edu.sg, sporia@sutd.edu.sg

What is MUStARD?

GOAL

Build a **dataset** for multimodal **sarcasm detection** (**video**+**audio**+**text**)
Evaluate simple **baselines**

DATASET

- **690** one-utterance videos (avg. duration 5s).
- Balanced, labeled as (non-) sarcastic.
- **Transcripts** and **preceding context video** (avg. duration 14s) are also included.

Example sarcastic utterance from MUStARD

Motivation

NEED

Sarcasm expressed through verbal and non-verbal cues, such as
Change of tone

1)

Chandler : Yes and we are *very* excited about it.

2) **Sheldon** :
Its just a *privilege* to watch your mind at work.

- **Text** : suggests a compliment.
- **Audio** : neutral tone.
- **Video** : straight face.

Incongruencies across modalities

Over-emphasis on words

Multi-modality in sarcasm is largely **unexplored**.

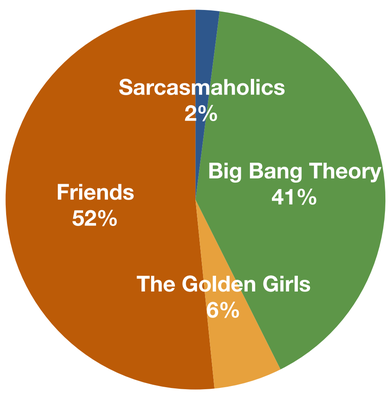
EXAMPLE

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

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

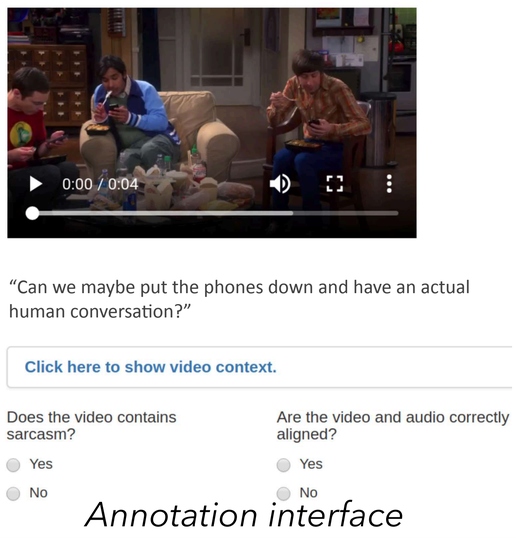
Annotation

PROCEDURE



Two judges annotated 6,421 videos coming from ***The Big Bang Theory*** and 624 videos coming from ***Friends***, ***The Golden Girls***, and ***Sarcasmaholics Anonymous***.

- Kappa scores of 0.23 and 0.58, respectively.
- A **third** judge broke the ties.
- Low-quality and least agreed instances filtered out, providing **balanced** dataset of 690 instances.



Annotation interface

Experiments

FEATURES

Text: [CLS] token representation from the last 4 layers, from BERT (case sensitive)
Video: avg. ResNet-152 *pool5* layer.
Audio: MFCC, melspectrogram, spectral centroid and their associated temporal derivatives.

RESULTS

- Video performs best among unimodal variants.
- Bi-modal stronger with inclusion of text.

Algorithm	Modality	Precision	Recall	F-Score
Majority	-	25.0	50.0	33.3
Random	-	49.5	49.5	49.8
SVM	T	65.1	64.6	64.6
	A	65.9	64.6	64.6
	V	68.1	67.4	67.4
	T+A	66.6	66.2	66.2
	T+V	**72.0**	**71.6**	**71.6**
	A+V	66.2	65.7	65.7
	T+A+V	71.9	71.4	71.5
$\Delta_{multi-unimodal}$ Error rate reduction		↑ 3.9% ↑ 12.2%	↑ 4.2% ↑ 12.9%	↑ 4.2% ↑ 12.9%

Speaker dependent

- Speaker independent setup is more-challenging.
- Video features might contain speaker bias.

Algorithm	Modality	Precision	Recall	F-Score
Majority	-	32.8	57.3	41.7
Random	-	51.1	50.2	50.4
SVM	T	60.9	59.6	59.8
	A	**65.1**	62.6	62.7
	V	54.9	53.4	53.6
	T+A	64.7	**62.9**	**63.1**
	T+V	62.2	61.5	61.7
	A+V	64.1	61.8	61.9
	T+A+V	64.3	62.6	62.8
$\Delta_{multi-unimodal}$ Error rate reduction		↓ 0.4% ↓ 1.1%	↑ 0.3% ↑ 0.8%	↑ 0.4% ↑ 1.1%

Speaker independent

ANALYSIS

Speaker	Utterance
Sheldon	**Darn. If you weren't busy, I'd ask you to join us.**
Chandler	**I'm sorry, we don't have your sheep.**
Chandler	**I am sorry, it was a one time thing. I was very drunk and it was someone else's subconscious.**

Correct sarcastic prediction by T+V based model but not the T only model.

Conclusion

- We provide a dataset for Multimodal Sarcasm with *video* + *audio* + *text* consisting of 690 videos.
- We evaluate several baselines.

FUTURE

- ➢ Multimodal fusion.
- ➢ Multiparty modeling.
- ➢ Other neural baselines.
- ➢ Speaker localization.
- ➢ Sarcasm in conversational context.

Dataset and software:
https://github.com/soujanyaporia/MUStARD

UNIVERSITY OF MICHIGAN
NUS National University of Singapore
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN