A Multimodal Corpus for Emotion Recognition in Sarcasm

Anupama Ray, Shubham Mishra, Apoorva Nunna, Pushpak Bhattacharyya
IBM Research India, Department of Computer Science and Engineering, IIT Bombay

INTRODUCTION

- Sarcasm is a sophisticated linguistic articulation where the explicit or surface meaning of what is said is often incongruous with the underlying intended meaning
- **Goal**: Given a sarcastic utterance, we aim to find the intended/implicit emotion behind the sarcastic utterance
- **Motivation**: Sarcasm is a complex linguistic artifact, often a result of some emotion-induced. Not just detecting sarcasm, but also understanding the intended emotion in the presence of sarcasm would improve the quality of chatbots, online review analysis etc.
- **Contribution**: We release an extended data resource, where we have doubled an existing multimodal dataset called MUStARD. We identified and corrected labeling errors in MUStARD and added labels for emotion, valence, arousal, and sarcasm-type. Performed exhaustive experimentation to benchmark multimodal fusion models for emotion detection in sarcasm

BACKGROUND CONCEPTS

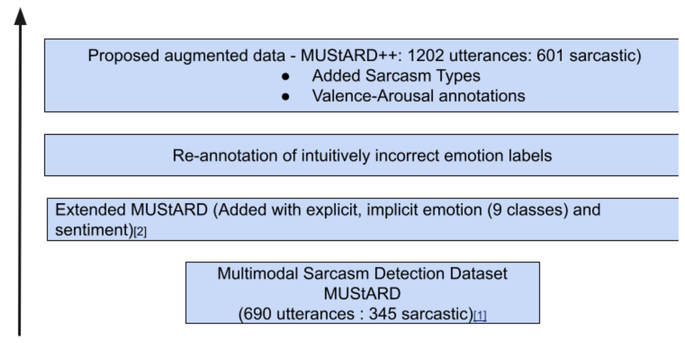
- Types of Sarcasm: There are 4 major types of sarcasm
- **Propositional**: Knowing the context is necessary to understand this sarcasm
- **Illocutionary**: The cues for sarcasm arise from non-textual modalities
- **Embedded**: Incongruity is presented directly through phrases and words
- **Like-Prefixed**: Like-phrase is used to express denial/incongruity
- Valence: Pleasantness associated with an input
- Arousal: Perceived intensity of an emotion

ANNOTATION DETAILS

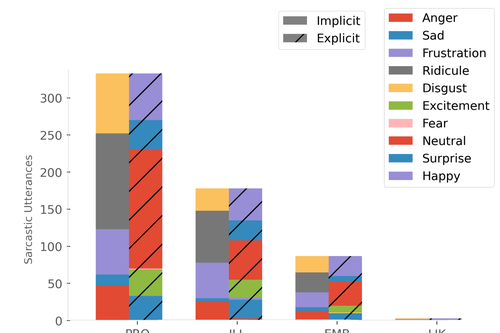
- Number of annotators = 7 (4 Male + 3 Female) from diverse backgrounds
- Kappa scores for Inter-Annotator Agreement
  - IAA for emotion annotation = 0.595
  - IAA for valence annotation = 0.638
  - IAA for arousal annotation = 0.689
- The manual annotation task pointed out the need for an emotion class that is not covered by basic emotions thereby leading to introduction of label 'Ridicule'

DATASET

Our corpus, MUStARD++ contains 1202 utterances all of them available both in text mode and have their corresponding video snippets. We provide the context, speaker information, and the source TV show for each instance. Each utterance is labeled with the presence of sarcasm, the sarcasm type, the implicit and explicit emotions, valence, and arousal ratings (from 1-9)



Evolution of MUStARD++



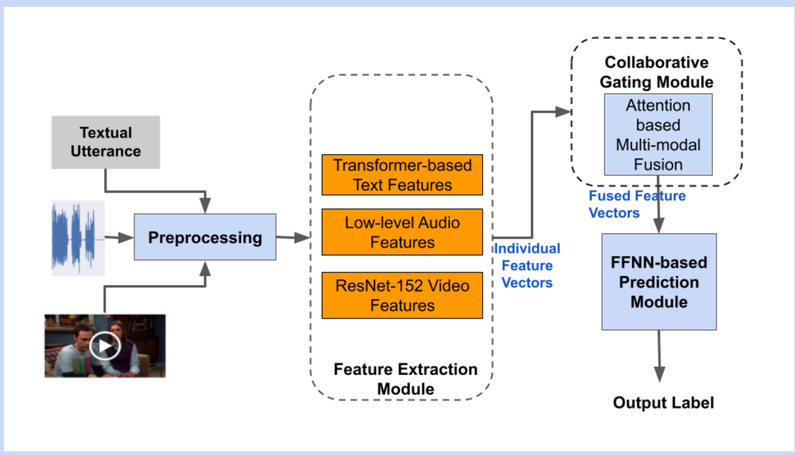
Distribution of Emotion over Sarcasm Type in Proposed Dataset

EXAMPLE

Example from MUStARD++ (context in yellow, utterance in red)

Explicit Emotion = Surprise
Implicit Emotion = Ridicule
Sarcasm Type = Embedded
Valence = 4, Arousal = 8

METHODOLOGY



Best Feature Extraction Models: BART features for text, MFCC, spectrogram, prosodic features for audio and ResNET-152 features for video

Multi-modal fusion: Collaborative Gating (with pair-wise attention followed by aggregation)[3]

RESULTS

Methods	*Speaker Independent*			*Speaker Dependent*		
	P	R	F1	P	R	F1
(Castro et al., 2019)	64.7	62.9	63.1	72.1	71.7	71.8
(Chauhan et al., 2020)	69.53	66.0	65.9	73.40	72.75	72.57
Proposed MUStARD*	**72.1**	**72**	**72**	**74.2**	**74.2**	**74.2**
%ΔMUStARD	↑ 3.69%	↑ 9.09%	↑ 9.25%	↑ 1.08%	↑ 1.99%	↑ 2.24%
Proposed MUStARD++	70.2	70.2	70.2	70.3	70.3	70.3

Sarcasm detection results (weighted average) on MUStARD and MUStARD++. Proposed MUStARD refers to the best model on MUStARD. Proposed MUStARD++ refers to the result of our best model for sarcasm detection on MUStARD++

	Speaker Independent						Speaker Dependent					
	w/o Context			w Context			w/o Context			w Context		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
T	67.9	67.7	67.7	69.3	69.2	69.2	69.4	69.3	69.3	70.2	70	70
A	63.9	63.5	63.6	64.3	64.1	64.1	65.3	65.2	65.2	65.0	64.9	64.9
V	59.5	59.4	59.4	60.3	60.0	60.0	61.8	61.7	61.7	61.6	61.4	61.5
T+A	68.8	68.6	68.7	**70.2**	**70.2**	**70.2**	**69.8**	**69.5**	**69.5**	69.2	69.1	69.1
A+V	65.7	65.4	65.5	67.5	67.3	67.4	64.9	64.5	64.5	64.2	64.0	64.0
V+T	68.2	68.1	68.1	67.9	67.6	67.6	69.1	69.0	69.0	69.4	69.1	69.1
T+A+V	**69.5**	**69.4**	**69.4**	69.6	69.5	69.6	69.6	69.3	69.3	**70.6**	**70.3**	**70.3**

Sarcasm detection results for MUStARD++, Weighted Average

	Speaker Independent						Speaker Dependent					
	w/o Context			w Context			w/o Context			w Context		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
T	33±0.9	33.6±1	33.3±0.9	32.3±0.7	32.7±0.6	32.5±0.6	29.9±0.9	30.3±0.8	30.1±0.8	30.2±0.9	30.8±0.9	30.5±0.9
A	26.7±1.1	27.1±1.4	26.8±1.1	24.9±1.0	26.3±1.4	25.5±1.2	24.3±0.8	24.7±0.6	24.5±0.7	26.7±1.2	26.9±1.2	26.8±1.2
V	28.8±0.9	29.4±1.3	29±1.1	28.5±1.2	29.2±1.4	28.8±1.3	30.3±1.4	31.4±1.2	30.6±1.4	28.7±0.8	30.08±1.1	29.1±1.0
T+A	31.5±1.7	31.6±1.8	31.6±1.7	32.1±0.7	32.04±0.6	32.03±0.6	29.1±1.6	29.2±1.5	29.1±1.5	31.2±2	31.8±1.8	31.4±1.8
A+V	25.9±1.9	26.3±2	26.1±1.9	28.2±1.1	28.3±1.2	28.2±1.1	29.7±0.6	30.6±0.9	30.1±0.7	25.2±1.0	25.2±0.9	25.2±0.9
V+T	31.9±0.8	32.5±0.7	32.2±0.7	32.7±1.1	33.3±1.0	33.0±1.1	31.1±0.8	31.2±0.7	31.1±0.7	31.8±0.	31.9±0.5	31.8±0.6
T+A+V	31.2±1	31.6±0.1	31.4±1	28.9±1.3	29.0±1.4	28.9±1.3	30.9±0.3	30.5±0.4	30.7±0.3	31.6±1.5	31.3±1.3	31.5±1.4

Mean, std-dev of 5 runs for Implicit Emotion Classification (Multiclass) on MUStARD++

	Speaker Independent						Speaker Dependent					
	w/o Context			w Context			w/o Context			w Context		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
T	38.5±0.8	39.2±0.9	38.8±0.8	38.2±1.2	38.8±1.3	38.5±1.3	38.7±0.5	39.3±0.5	39±0.5	38.9±0.4	39.7±0.6	39.3±0.5
A	26.4±0.9	28±1.4	27.1±1	27.1±1.1	28.1±1.3	27.6±1.2	28.1±1.1	29.3±1.0	28.6±1.1	28.4±0.7	31.6±1.3	29.6±0.8
V	25.1±0.6	25.9±0.6	25.5±0.6	24.4±0.7	24.9±0.9	24.6±0.8	25.7±1.4	36.1±0.8	27.7±0.8	27.0±0.8	29.6±1.6	28.0±1
T+A	38.9±1.1	39.5±1.3	39.2±1.2	39.2±0.6	39.5±0.6	39.3±0.6	39.1±0.8	39.7±0.7	39.4±0.7	39.1±0.5	39.5±0.7	39.3±0.6
A+V	26.4±1.4	26.5±1.5	26.4±1.4	26.2±1.22	26.3±1.6	26.2±1.5	27.6±1	28.2±1.2	27.9±1.1	27.8±0.5	28±0.4	27.9±0.4
V+T	38.6±0.7	39.2±0.8	38.8±0.8	40.5±0.7	41.2±0.7	40.8±0.7	39.8±0.1	40±0.2	39.8±0.2	39.9±0.4	40.3±0.6	40±0.5
T+A+V	37.8±0.1	38.3±0.8	38.0±0.9	39.5±0.8	39.6±0.9	39.5±0.9	40±0.6	39.8±0.5	39.9±0.9	39.7±1.3	39.4±1.2	39.5±1.2

Mean, std-dev of 5 runs for Explicit Emotion Classification (Multiclass) on MUStARD++

OBSERVATIONS

- In emotion detection, all modality combinations that include text performed better since text modality considers the actual spoken content unlike other modalities which only focus on audio and visual features
- In both sarcasm and emotion classification, context information improved performance

CONCLUSIONS & FUTURE WORK

- This paper presents a multimodal sarcasm dataset that can be used in the area of sarcasm detection and emotion recognition. We double the multimodal sarcasm dataset MUStARD[1], while adding fine-grained information like valence-arousal ratings and sarcasm type
- **Future Work**:
  - Sarcasm type information can choose the right modality combination for a given utterance for sarcasm detection and emotion recognition.
  - Using arousal and valence to investigate its effect on emotion classification and sarcasm detection
  - Use emotion labels to improve sarcasm detection

REFERENCES

1. Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection (an _obviously_ perfect paper). arXiv preprint arXiv:1906.01815.
2. Chauhan, D. S., Dhanush, S. R., Ekbal, A., & Bhattacharyya, P. (2020, July). Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4351-4360).
3. Liu, Y., Albanie, S., Nagrani, A., & Zisserman, A. (2019). Use what you have: Video retrieval using representations from collaborative experts. arXiv preprint arXiv:1907.13487.

Repository Link - https://github.com/apoorva-nunna/MUStARD_Plus_Plus

Contact Information: apoorvanunna@cse.iitb.ac.in, pb@cse.iitb.ac.in