

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

Santiago Castro<sup>†\*</sup>, Devamanyu Hazarika<sup>Φ\*</sup>, Verónica Pérez-Rosas<sup>†</sup>,  
Roger Zimmermann<sup>Φ</sup>, Rada Mihalcea<sup>†</sup>, Soujanya Poria<sup>ℓ</sup>

<sup>†</sup>Computer Science & Engineering, University of Michigan, USA

<sup>Φ</sup>School of Computing, National University of Singapore, Singapore

<sup>ℓ</sup>Information Systems Technology and Design, SUTD, Singapore

{sacastr, vrncapr, mihalcea}@umich.edu,

{hazarika, rogerz}@comp.nus.edu.sg, sporia@sutd.edu.sg

## Abstract

Sarcasm is often expressed through several verbal and non-verbal cues, e.g., a change of tone, overemphasis in a word, a drawn-out syllable, or a straight looking face. Most of the recent work in sarcasm detection has been carried out on textual data. In this paper, we argue that incorporating multimodal cues can improve the automatic classification of sarcasm. As a first step towards enabling the development of multimodal approaches for sarcasm detection, we propose a new sarcasm dataset, Multimodal Sarcasm Detection Dataset (MUSTARD<sup>1</sup>), compiled from popular TV shows. MUSTARD consists of audiovisual utterances annotated with sarcasm labels. Each utterance is accompanied by its context of historical utterances in the dialogue, which provides additional information on the scenario where the utterance occurs. Our initial results show that the use of multimodal information can reduce the relative error rate of sarcasm detection by up to 12.9% in F-score when compared to the use of individual modalities. The full dataset is publicly available for use at <https://github.com/soujanyaporia/MUSTARD>.

## 1 Introduction

Sarcasm plays an important role in daily conversations by allowing individuals to express their intent to mock or display contempt. It is achieved by using irony that reflects a negative connotation. For example, in the utterance: *Maybe it's a good thing we came here. It's like a lesson in what not to do*, the sarcasm is explicit as the speaker expresses learning of a lesson in a positive light when in reality, she means it in a negative way. However, there are also scenarios where sarcasm lacks explicit linguistic markers, thus requiring additional cues that

can reveal the speaker's intentions. For instance, sarcasm can be expressed using a combination of verbal and non-verbal cues, such as a change of tone, overemphasis in a word, a drawn-out syllable, or a straight looking face. Moreover, sarcasm detection involves finding linguistic or contextual incongruity, which in turn requires further information, either from multiple modalities (Schifanella et al., 2016; Mishra et al., 2016a) or from the context history in a dialogue.

This paper explores the role of multimodality and conversational context in sarcasm detection and introduces a new resource to further enable research in this area. More specifically, our paper makes the following contributions: (1) We curate a new dataset, MUSTARD, for multimodal sarcasm research with high-quality annotations, including both multimodal and conversational context features; (2) We exemplify various scenarios where incongruity in sarcasm is evident across different modalities, thus stressing the role of multimodal approaches to solve this problem; (3) We introduce several baselines and show that multimodal models are significantly more effective when compared to their unimodal variants; and (4) We also provide preceding turns in the dialogue which act as context information. Consequently, we surmise that this property of MUSTARD leads to a new sub-task for future work: *sarcasm detection in conversational context*.

The rest of the paper is organized as follows. Section 2 summarizes previous work on sarcasm detection using both unimodal and multimodal sources. Section 3 describes the dataset collection, the annotation process, and the types of sarcastic situations covered by our dataset. Section 4 explains how we extract features for the different modalities. Section 5 shows the experimental work around the new dataset while Section 6 analyzes it. Finally, Section 7 offers conclusions and discusses open

\*Equal contribution.

<sup>1</sup>MUSTARD is an abbreviation for MULTImodal SARCasm Dataset. Similar to how “mustard” adds spice to our food and meals, we believe sarcasm adds spice to our interactions and lives.

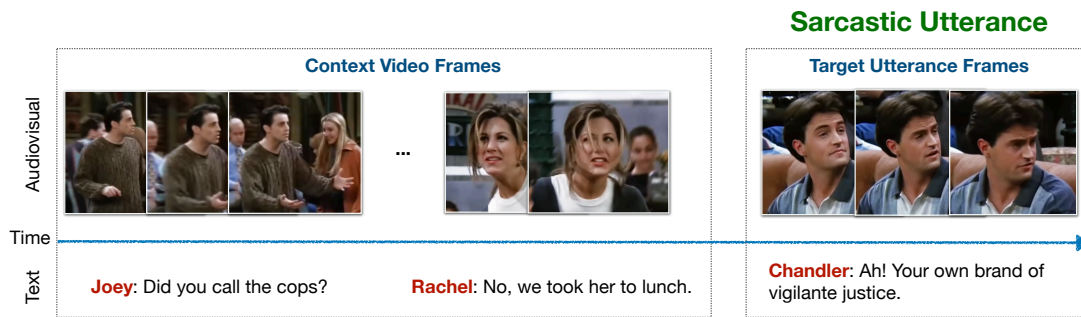


Figure 1: Sample sarcastic utterance in the dataset along with its context and transcript.

problems related to this resource.

## 2 Related Work

Automated sarcasm detection has gained increased interest in recent years. It is a widely studied linguistic device whose significance is seen in sentiment analysis and human-machine interaction research. Various research projects have approached this problem through different modalities, such as text, speech, and visual data streams.

**Sarcasm in Text:** Traditional approaches for detecting sarcasm in text have considered rule-based techniques (Veale and Hao, 2010), lexical and pragmatic features (Carvalho et al., 2009), stylistic features (Davidov et al., 2010), situational disparity (Riloff et al., 2013), incongruity (Joshi et al., 2015), or user-provided annotations such as hashtags (Liebrecht et al., 2013).

Resources in this domain are collected using Twitter as a primary data source and are annotated using two main strategies: manual annotation (Riloff et al., 2013; Joshi et al., 2016a) and distant supervision through hashtags (Davidov et al., 2010; Abercrombie and Hovy, 2016). Other research leverages context to acquire shared knowledge between the speaker and the audience (Wallace et al., 2014; Bamman and Smith, 2015). A variety of contextual features have been explored, including speaker’s background and behavior in online platforms (Rajadesingan et al., 2015), embeddings of expressed sentiment and speaker’s personality traits (Poria et al., 2016), learning of user-specific representations (Wallace et al., 2016; Kolchinski and Potts, 2018), user-community features (Wallace et al., 2015), as well as stylistic and discourse features (Hazarika et al., 2018). In our dataset, we capitalize on the conversational format and provide context by including preceding utterances along with speaker identities. To the best of our knowledge, there is no prior work which deals

with the task of sarcasm detection in conversation.

**Sarcasm in Speech:** Sarcasm detection in speech has mainly focused on the identification of prosodic cues in the form of acoustic patterns that are related to sarcastic behavior. Studied features include mean amplitude, amplitude range, speech rate, harmonics-to-noise ratio, and others (Cheang and Pell, 2008). Rockwell (2000) presented one of the initial approaches to this problem that studied the vocal tonalities of sarcastic speech. They found slower speaking rates and greater intensity as probable markers for sarcasm. Later, Tepperman et al. (2006) studied prosodic and spectral features of sound — both in and out of context — to determine sarcasm. In general, prosodic features such as intonation and stress are considered important indicators of sarcasm (Bryant, 2010; Woodland and Voyer, 2011). We take motivation from this previous research and include similar speech parameters as features in our dataset and baseline experiments.

**Multimodal Sarcasm:** Contextual information for sarcasm in text can be included from other modalities. These modalities help in providing additional cues in the form of both common or contrasting patterns. Prior work mainly considers multimodal learning for the readers’ ability to perceive sarcasm. Such research couples textual features with cognitive features such as the gaze-behavior of readers (Mishra et al., 2016a,b, 2017) or electro/magneto-encephalographic (EEG/MEG) signals (Filik et al., 2014; Thompson et al., 2016). In contrast, there is limited work exploring multimodal avenues to understand sarcasm conveyed by the opinion holder. Attardo et al. (2003) presented one of the preliminary explorations on this topic where different phonological and visual markers for sarcasm were studied. However, this work did not analyze the interplay of the modalities. More recently, Schifanella et al. (2016) presented a multimodal approach for this task by considering vi-

sual content accompanying text in online sarcastic posts. They extracted semantic visual features from images using pre-trained networks and fused them with textual features. In our work, we extend these notions and propose to analyze video-based sarcasm in dialogues. To the best of our knowledge, ours is the first work to propose a resource on video-level sarcasm. Joshi et al. (2016b) proposed a dataset similar to us, i.e., based on the TV show *Friends*. However, their corpus only includes the textual modality and is thus not multimodal in nature. Furthermore, we also analyze multiple challenges in sarcasm that call for multimodal learning and provide an evaluation setup for future works to test upon.

### 3 Dataset

To enable the exploration of multimodal sarcasm detection, we introduce a new dataset (MUSARD) consisting of short videos manually annotated for their sarcasm property.

#### 3.1 Data Collection

To collect potentially sarcastic examples, we conduct web searches on differences sources, mainly YouTube. We use keywords such as *Friends sarcasm*, *Chandler sarcasm*, *Sarcasm 101*, *Sarcasm in TV shows*. Using this strategy, we obtain videos from three main TV shows: *Friends*, *The Golden Girls*, and *Sarcasmaholics Anonymous*. Note that during this initial search, we focus exclusively on sarcastic content. To obtain non-sarcastic videos, we select a subset of 400 videos from MELD, a multimodal emotion recognition dataset derived from the *Friends* TV series, originally collected by Poria et al. (2018). In addition, we collect videos from *The Big Bang Theory*, a TV show whose characters are often perceived as sarcastic. We obtain videos from seasons 1–8 and segment episodes using laughter cues from its audience. Specifically, we use open-source software for laughter detection (Ryokai et al., 2018) to obtain initial segmentation boundaries and fine-tune them using the subtitles’ timestamps.

The collected set consists of 6,421 videos. Note that although some of the videos in our initial pool include information about their sarcastic nature, the majority of our videos are not labeled. Thus, we conduct a manual annotation as described next.

#### 3.2 Annotation Process

We built a web-based annotation interface that shows each video along with its transcript and re-

Video 1



“Can we maybe put the phones down and have an actual human conversation?”

[Click here to show video context.](#)

Does the video contains sarcasm?

- Yes
- No

Are the video and audio correctly aligned?

- Yes
- No

Figure 2: Graphical user interface used by the annotators to label the videos in our dataset.

quests annotations for sarcasm. We also ask the annotators to flag misaligned videos, i.e., cases where the audio or video is not properly synchronized. The interface allows the annotators to watch a context video consisting of the previous video utterances, whenever they deem it necessary. Given the large number of videos to be annotated, we request annotations in batches of four videos at a time. Our web interface is shown in Fig. 2.

We conduct the annotation in two steps. First, we annotate the videos from *The Big Bang Theory*, as it contains the largest set of videos. Second, we annotate the remaining videos, belonging to the other sources. The annotation is conducted by two graduate students who have first been provided with easy examples of explicit sarcastic content, to illustrate sarcasm in videos. Each annotator labeled the full set of videos independently.

For the first step, after annotating the first part – consisting of 5,884 utterances from *The Big Bang Theory* – we noticed that the majority of them were labeled as non-sarcastic (98% were considered as non-sarcastic by both). In addition, our initial inter-annotation agreement was low (Kappa score is 0.1463). We thus decided to stop the annotation process and reconcile the annotation differences before proceeding further. The annotators discussed their disagreements for a subset of 20 videos, and then re-annotated the videos. This time, we obtained an improved inter-annotator agreement of 0.2326. The annotation disagreements were reconciled by a third annotator by identifying the disagreement cases, watching the videos again and

deciding which is the correct label for each one.

Next, we annotate the second part, consisting of 624 videos drawn from Friends, The Golden Girls, and Sarcasmaholics Anonymous. As before, the two annotators label each video independently. The inter-annotator agreement was calculated with a Kappa score of 0.5877. Again, the differences were reconciled by a third annotator.

The resulting set of annotations consists of 345 videos labeled as sarcastic and 6,020 videos labeled as non-sarcastic for a total of 6,365 videos.

### 3.3 Transcriptions

Since we collect videos from several sources, some of them had subtitles or transcripts readily available. This is particularly the case for videos from Big Bang Theory and MELD. We use the MELD transcriptions directly. For Big Bang Theory, we extracted the transcript by applying manual substring matching on the episode subtitles. The remaining videos are manually transcribed.

### 3.4 Sarcasm Dataset: MUSTARD

To enable our experiments, which focus explicitly on the multimodal aspects of sarcasm, we decided to work with a balanced sample of sarcastic and non-sarcastic videos. We thus obtain a balanced sample from the set of 6,365 annotated videos. We start by selecting all videos marked as sarcastic from the full set, and then we randomly obtain an equally sized non-sarcastic sample from the non-sarcastic subset by prioritizing the ones annotated by a larger number of annotators. Our dataset thus comprises 690 videos with an even number of sarcastic and non-sarcastic labels. Source, character, and label-ratio statistics are shown in Figs. 3 and 4.

In the remainder of this paper, we use the term utterance while referring to the videos in our dataset. We extend the definition of an utterance<sup>2</sup> to include consecutive multi-sentence dialogues of the same speaker to prioritize completeness of information. As a result, 61.3% of the utterances from the dataset are single sentences, while the remaining utterances consist of two or more sentences. Each utterance in our dataset is coupled with its context utterances, which are preceding turns by the speakers participating in the dialogue. Some of the context videos contain multi-party dialogue between speakers participating in the scene. The number of turns in the context is manually set to include a coherent background of the target utterance.

<sup>2</sup>An utterance is usually defined as a unit of speech bounded by breaths or pauses.

Statistics	Utterance	Context
Unique words	1991	3205
Avg. utterance length (tokens)	14	10
Max. utterance length (tokens)	73	71
Avg. duration (seconds)	5.22	13.95

Table 1: Dataset statistics by utterance and context.

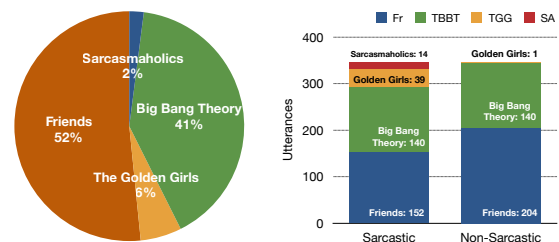
Table 1 shows general statistics for the utterances in our dataset.

Each utterance and its context consists of three modalities: video, audio, and transcription (text). Also, all the utterances are accompanied by their speaker identifiers. Fig. 1 illustrates a sarcastic utterance along with its associated context in the dataset. Fig. 4b provides the list of major characters present in the dataset. Fig. 4a details the distribution of labels per character. Some of the characters, such as *Chandler* and *Sheldon*, occupy major portions of the dataset. This is expected since they play comic roles in the shows. To avoid speaker bias of such popular characters, we also include non-sarcastic samples for these characters. In contrast, the dataset intentionally includes minor roles such as *Dorothy* from The Golden Girls, who is entirely sarcastic throughout the corpus. This allows the study of speaker bias for sarcasm detection.

### 3.5 Qualitative Aspects

Sarcasm detection in text often requires additional information that can be leveraged from associated modalities. Below, we analyze some cases that require multimodal reasoning. We exemplify using instances from our proposed dataset to further support our claim of sarcasm being often expressed in a multimodal way.

**Role of Multimodality:** Fig. 5 presents two cases where sarcasm is expressed through the incongruity between modalities. In the first case, the language modality indicates fear or anger, whereas the facial modality lacks any visible sign of anxiety



(a) Source across the dataset. (b) Labels across source shows.

Figure 3: Ratio of the TV shows composing the dataset.







Figure 7: Context importance in sarcasm detection.

lowed to extract each of them is described below:

**Text Features:** We represent the textual utterances in the dataset using BERT (Devlin et al., 2018), which provides a sentence representation  $\mathbf{u}_t \in \mathbb{R}^{d_t}$  for every utterance  $u$ . In particular, we average the last four transformer layers of the first token (`[CLS]`) in the utterance – using the BERT-Base model – to get a unique utterance representation of size  $d_t = 768$ . We also considered averaging Common Crawl pre-trained 300 dimensional GloVe word vectors (Pennington et al., 2014) for each token; however, it resulted in lower performance as compared to BERT-based features.

**Speech Features:** To leverage information from the audio modality, we obtain low-level features from the audio data stream for each utterance in the dataset. Through these features, we intend to provide information related to pitch, intonation, and other tonal-specific details of the speaker (Tepperman et al., 2006). We utilize the popular speech-processing library Librosa (McFee et al., 2018) and perform the processing pipeline described next. First, we load the audio sample for an utterance as a time series signal with a sampling rate of 22050 Hz. Then we remove background noise from the signal by applying a heuristic vocal-extraction method.<sup>3</sup> Finally, we segment the audio signal into  $d_w$  non-overlapping windows to extract local features that include MFCC, melspectrogram, spectral centroid and their associated temporal derivatives (delta). Segmentation is done to achieve a fixed length representation of the audio sources which are otherwise variable in length across the dataset. All the extracted features are concatenated together to compose a  $d_a = 283$  dimensional joint representation  $\{\mathbf{u}_i^a\}_{i=1}^{d_w}$  for each window. The final audio representation of each utterance is obtained by calculating the mean across the window segments, i.e.

<sup>3</sup>[http://librosa.github.io/librosa/auto\\_examples/plot\\_vocal\\_separation.html#sphx-glr-auto-examples-plot-vocal-separation-py](http://librosa.github.io/librosa/auto_examples/plot_vocal_separation.html#sphx-glr-auto-examples-plot-vocal-separation-py)

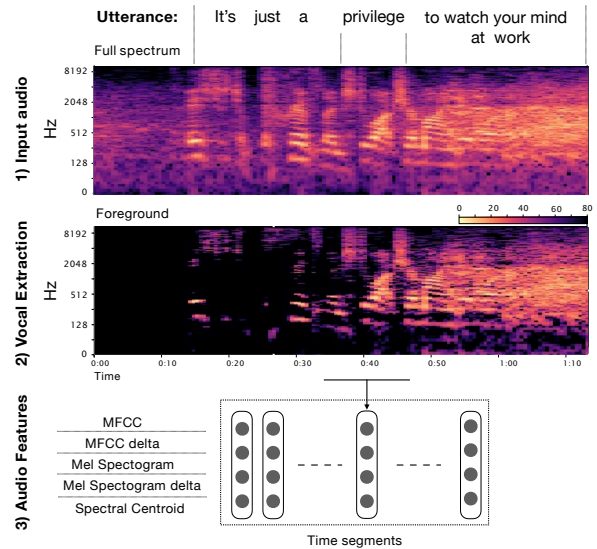


Figure 8: Feature extraction for audio modality.

$$\mathbf{u}_a = \frac{1}{d_w} (\sum_i \mathbf{u}_i^a) \in \mathbb{R}^{d_a}.$$

**Video Features:** We extract visual features for each of the  $f$  frames in the utterance video using a `pool5` layer of an ImageNet (Deng et al., 2009) pretrained ResNet-152 (He et al., 2016) image classification model. We first preprocess every frame by resizing, center-cropping and normalizing it. To obtain a visual representation of each utterance, we compute the mean of the obtained  $d_v = 2048$  dimensional feature vector  $\mathbf{u}_i^v$  for every frame:  $\mathbf{u}_v = \frac{1}{f} (\sum_i \mathbf{u}_i^v) \in \mathbb{R}^{d_v}$ . While we could use more advanced visual encoding techniques (e.g., recurrent neural network encoding techniques), we decide to use the same averaging strategy as with the other modalities.

## 5 Experiments

To explore the role of multimodality in sarcasm detection, we conduct multiple experiments evaluating each modality separately and also combinations of modalities provided in the dataset. Additionally, we investigate the role of context and speaker information for improving predictions.

### 5.1 Experimental Setup

We perform two main sets of evaluations. The first set involves conducting five-fold cross-validation experiments where the folds are randomly created in a stratified manner. This is done to ensure label balance across folds. In each of the  $K$  iterations, the  $k^{th}$  fold acts as a testing set while the remaining are used for training. Validation folds can be obtained from a part of the training folds. As the folds

are created in a randomized manner, there is overlap between speakers across training and testing sets, thus resulting in a *speaker-dependent* setup. The second set of evaluations restrict the inclusion of utterances from the same speaker to be either in the training or testing sets. Utterances from The Big Bang Theory, The Golden Girls and Sarcasmaholics Anonymous are made part of the training set while Friends is used as a testing set.<sup>4</sup> We call this the *speaker-independent* setup. Motivation for such a setup is discussed in Section 6.

During our experiments, we use precision, recall, and F-score as the main evaluation metrics, weighted across both sarcastic and non-sarcastic classes. The weights are obtained based on the class ratios. For speaker-dependent scenario, we report results by averaging across the five-fold cross-validation results.

## 5.2 Baselines

The experiments are conducted using three main baseline methods:

**Majority:** This baseline assigns all the instances to the majority class, i.e., non-sarcastic.

**Random:** This baseline makes random/chance predictions sampled uniformly across the test set.

**SVM:** We use Support Vector Machines (SVM) as the primary baseline for our experiments. SVMs are strong predictors for small-sized datasets and at times outperform neural counterparts (Byvatov et al., 2003). We use the SVM classifiers from scikit-learn (Pedregosa et al., 2011) with an RBF kernel and a scaled gamma. The penalty term  $C$  is kept as a hyper-parameter which we tune based on each experiment (we choose between 1, 10, 30, 500, and 1000). For the speaker dependent setup we scale the features by subtracting the mean and dividing them by the standard deviation. Multiple modalities are combined using early fusion, where the features drawn from the different modalities are concatenated together.

## 6 Multimodal Sarcasm Classification

Table 2 presents the classification results for sarcasm prediction in the speaker-dependent setup. The lowest performance is obtained with the Majority baseline which achieves 33.3% weighted F-score (66.7% F-score for non-sarcastic class and

<sup>4</sup>Split details are released along with the dataset for consistent comparison by future works.

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	<b>72.0</b>	<b>71.6</b>	<b>71.6</b>
	A+V	66.2	65.7	65.7
T+A+V	71.9	71.4	71.5	
$\Delta_{multi-unimodal}$		$\uparrow 3.9\%$	$\uparrow 4.2\%$	$\uparrow 4.2\%$
Error rate reduction		$\uparrow 12.2\%$	$\uparrow 12.9\%$	$\uparrow 12.9\%$

Table 2: Speaker-dependent setup. All results are averaged across five folds where each fold present weighted F-score across both sarcastic and non-sarcastic classes.

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	<b>65.1</b>	62.6	62.7
	V	54.9	53.4	53.6
	T+A	64.7	<b>62.9</b>	<b>63.1</b>
	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	
$\Delta_{multi-unimodal}$		$\downarrow 0.4\%$	$\uparrow 0.3\%$	$\uparrow 0.4\%$
Error rate reduction		$\downarrow 1.1\%$	$\uparrow 0.8\%$	$\uparrow 1.1\%$

Table 3: Multimodal sarcasm classification. Evaluated using an speaker-independent setup. Note: T=text, A=audio, V=video.

0% for sarcastic). The pre-trained features for the visual modality provide the best performance among the unimodal variants. The addition of textual features through concatenation improves the unimodal baseline and achieves the best performance. The tri-modal variant is unable to achieve the best score due to a slightly sub-optimal performance from the audio modality. Overall, the combination of visual and textual signals significantly improves over the unimodal variants, with a relative error rate reduction of up to 12.9%.

We manually investigate the utterances where the bimodal textual and visual model predicts sarcasm correctly while the unimodal textual model fails. In most of these samples, the textual component does not reveal any explicit sarcasm (see Fig. 9). As a result, the utterances require additional cues, which it avails from the multimodal signals.

The speaker-independent setup is more challenging as compared to the speaker-dependent scenario, as it prevents the model from registering speaker-specific patterns. The presence of new speakers in the testing set requires a higher degree of generalization from the model. Our setup also segregates at the source level, thus the testing involves an en-

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.

Figure 9: Sample sarcastic utterances correctly predicted by T+V but not only T model in the speaker-dependent setup. The utterances either do not have explicit sarcastic markers or need commonsense-reasoning to detect ironies, such as *being drunk in someone else's subconscious*.

tirely new environment concerning all the modalities. We believe that the speaker-independent setup is a strong test-bed for multimodal sarcasm research. The increased difficulty of this task is also noticed in the model training, which now requires a smaller error margin (or higher  $C$  value) of the SVM's decision function to provide good test performance.

Table 3 presents the performance of our baselines in the speaker-independent setup. In this case, the multimodal variants do not greatly outperform the unimodal counterparts. Unlike Table 2, the audio channel plays a more important role, and it is slightly improved by adding text. By inspecting the correctly predicted sarcastic examples by text plus audio but not by text, we observe a tendency of higher mean pitch (mean fundamental frequency) with respect to those incorrectly predicted, as Attardo et al. (2003) suggested. Failure cases seem to contain particular patterns of high pitch, also studied by Attardo et al. (2003), but in average they seem to have normal pitch. In this sense, future work can focus on analyzing the temporal localities of the audio channel.

In this setup, video features do not seem to work well. We hypothesize that, because the visual features are about object features (not specific to sarcasm) and the model is shallow, these features may make the model capture character biases which make them unsuitable for the speaker-independent setup. This is also suggested by the statistics in Fig. 10 which we describe in the next section. By looking at the incorrect predictions by the best model, we infer that models should better capture the mismatches between the main speaker facial expressions and the emotions of what is being said.

**The Role of Context and Speaker Information:** We investigate whether additional informa-

Setup	Features	Precision	Recall	F-Score
Speaker Dependent	T	65.1	64.6	64.6
	+ context	65.5	65.1	65.0
	+ speaker	<b>67.7</b>	<b>67.2</b>	<b>67.3</b>
	Best (T + V)	72.0	71.6	<b>71.8</b>
	+ context + speaker	71.9	71.4	71.5
Speaker Independent	T	<b>60.9</b>	59.6	59.8
	+ context	57.9	54.5	54.1
	+ speaker	60.7	<b>60.7</b>	<b>60.7</b>
	Best (T + A)	64.7	<b>62.9</b>	<b>63.1</b>
	+ context + speaker	<b>65.2</b>	<b>62.9</b>	63.0
		64.7	<b>62.9</b>	<b>63.1</b>

Table 4: Role of context and utterance's speaker. Note: T=text, A=audio, V=video.

tion, such as an utterance's context (i.e., the preceding utterances, cf. Section 3.5) and the speaker identification, are helpful for the predictions. Context features are generated by averaging the representations of the utterances (as per Section 4) present in the context. For the speakers, we use a one-hot encoding vector with size equal to the total unique speakers in a training fold.

Table 4 shows the results for both evaluation settings for the textual baseline and the best multimodal variant. For the context features, we see a slight improvement in the best variant of the speaker independent setup (text plus audio); however, in other models, there is no improvement. A possible reason could be the loss of temporal information when pooling across the conversation.

For the speaker features, we see an improvement in the speaker-dependent setup for the textual modality. Due to the speaker overlap across splits, the model can leverage speaker regularities for sarcastic tendencies. However, we do not observe the same trend for the best multimodal variant (text + video) where the score barely improves. To understand this result, we visualize the correct predictions made by this model. The results, as seen in Fig. 10, show a correlation between the class distributions among the overall ground truth and the correctly predicted instances per speaker. As this model does not use speaker information, this correlation indicates that the multimodal variant is able to learn speaker-specific information transitively through the input features, rendering additional speaker input redundant. Lastly, in the speaker independent setup, the speaker information does not lead to improvement. This is also expected as there is no speaker overlap between the splits.



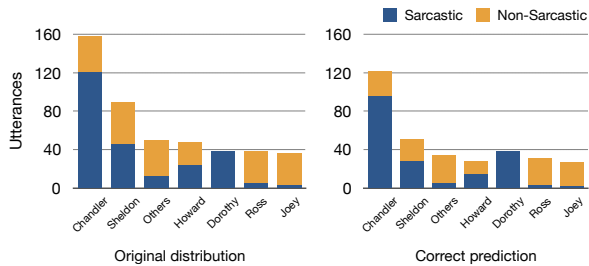


Figure 10: Correlation in speaker-specific sarcastic tendencies of the top-7 speakers. Predictions are obtained from the best performing model from Table 2. Speaker identifier features are not used.

## 7 Conclusion and Future Work

In this paper, we provided a systematic introduction to multimodal learning for sarcasm detection. To enable research on this topic, we introduced a novel dataset, MUSTARD, consisting of sarcastic and non-sarcastic videos drawn from different sources. By showing multiple examples from our curated dataset, we demonstrate the need for multimodal learning for sarcasm detection. Consequently, we developed models that leverage three different modalities, including text, speech, and visual signals. We also experimented with the integration of context and speaker information as additional input for our models.

The results of the baseline experiments supported the hypothesis that multimodality is important for sarcasm detection. In multiple evaluations, the multimodal variants were shown to significantly outperform their unimodal counterparts, with relative error rate reductions of up to 12.9%.

Moreover, while conducting this research, we identified several challenges that we believe are important to address in future research work on multimodal sarcasm detection.

**Multimodal fusion:** So far, we have only explored early fusion for multimodal classification. Future work could investigate advanced spatiotemporal fusion strategies (e.g., Tensor-Fusion (Zadeh et al., 2017), CCA (Hotelling, 1936)) to better encode the correspondence between modalities. Another direction could be to create fusion strategies that can better model incongruity among modalities to identify sarcasm.

**Multiparty conversation:** The dialogues represented in our dataset are often multi-party conversations. Advanced techniques to learn multimodal relationships could incorporate better relationship modeling (Majumder et al., 2018), and exploit models that provide gesture, facial and pose information

about the people in the scene (Cao et al., 2018).

**Neural baselines:** As we strove to create a high-quality dataset with rich annotations, we had to trade-off corpus size. Moreover, the occurrence of sarcastic utterances itself is scanty. To focus on effects induced by multimodal experiments, we chose a balanced version of the dataset with a limited size. This, however, arises the problem of over-fitting in complex neural models. As a consequence, in our initial experiments, we noticed that SVM classifiers perform better than their neural counterparts, such as CNNs. Future work should try to overcome this issue with solutions involving pre-training, transfer learning, domain adaption, or low-parameter models.

**Sarcasm detection in conversational context:** Our proposed MUSTARD is inherently a dialogue level dataset where we aim to classify the last utterance in the dialogue. In a dialogue, to classify an utterance at time  $t$ , the preceding utterances at time  $< t$  can be considered as its context. In this work, although we utilize conversational context, we ignore modeling various key conversation specific factors such as interlocutors’ goals, intents, dependency, etc. (Poria et al., 2019). Considering these factors can improve context modeling necessary for *sarcasm detection in conversational context*. Future work should try to leverage these factors to improve the baseline scores reported in this paper.

**Main speaker localization:** We currently extract visual features ubiquitously for each frame. As gesture and facial expressions are important features for sarcasm analysis, we believe the capability for models to identify the speakers in the multiparty videos is likely to be beneficial for the task.

Finally, we believe the resource introduced in this paper has the potential to enable novel research in multimodal sarcasm detection.

## Acknowledgements

We are grateful to Gautam Naik for his help in curating part of the dataset from online resources. This research was partially supported by the Singapore MOE Academic Research Fund (grant #T1 251RES1820), by the Michigan Institute for Data Science, by the National Science Foundation (grant #1815291), by the John Templeton Foundation (grant #61156), and by DARPA (grant #HR001117S0026-AIDA-FP-045).

## References

- Gavin Abercrombie and Dirk Hovy. 2016. Putting sarcasm detection into context: The effects of class imbalance and manual labelling on supervised machine classification of twitter conversations. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 107–113.
- Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor*, 16(2):243–260.
- David Bamman and Noah A Smith. 2015. Contextualized sarcasm detection on twitter. *ICWSM*, 2:15.
- Gregory A Bryant. 2010. Prosodic contrasts in ironic speech. *Discourse Processes*, 47(7):545–566.
- Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gisbert Schneider. 2003. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences*, 43(6):1882–1889.
- Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.
- Paula Carvalho, Luís Sarmento, Mário J Silva, and Eugénio De Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it’s so easy;- . In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.
- Henry S Cheang and Marc D Pell. 2008. The sound of sarcasm. *Speech communication*, 50(5):366–381.
- Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Ruth Filik, Hartmut Leuthold, Katie Wallington, and Jemma Page. 2014. Testing theories of irony processing using eye-tracking and erps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3):811.
- Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. Cascade: Contextual sarcasm detection in online discussion forums. *Proceedings of the 27th International Conference on Computational Linguistics*, page 1837–1848.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.
- Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016a. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99.
- Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 757–762.
- Aditya Joshi, Vaibhav Tripathi, Pushpak Bhattacharyya, and Mark J Carman. 2016b. Harnessing sequence labeling for sarcasm detection in dialogue from tv seriesfriends’. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 146–155.
- Y Alex Kolchinski and Christopher Potts. 2018. Representing social media users for sarcasm detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1115–1121.
- CC Liebrecht, FA Kunneman, and APJ van Den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37. New Brunswick, NJ: ACL.
- Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2018. Dialoguernn: An attentive rnn for emotion detection in conversations. *arXiv preprint arXiv:1811.00405*.
- Brian McFee, Matt McVicar, Stefan Balke, Carl Thomé, Vincent Lostanlen, Colin Raffel, Dana Lee, Oriol Nieto, Eric Battenberg, Dan Ellis, Ryuichi Yamamoto, Josh Moore, WZY, Rachel Bitner, Keunwoo Choi, Pius Friesch, Fabian-Robert

- Stöter, Matt Vollrath, Siddhartha Kumar, nehz, Simon Waloschek, Seth, Rimvydas Naktinis, Douglas Repetto, Curtis "Fjord" Hawthorne, CJ Carr, João Felipe Santos, JackieWu, Erik, and Adrian Holovaty. 2018. [librosa/librosa: 0.6.2](#).
- Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017. Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 377–387.
- Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016a. Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI*, pages 3747–3753.
- Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016b. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1095–1104.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vij. 2016. A deeper look into sarcastic tweets using deep convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1601–1612.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.
- Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *arXiv preprint arXiv:1905.02947*.
- Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 97–106. ACM.
- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714.
- Patricia Rockwell. 2000. Lower, slower, louder: Vocal cues of sarcasm. *Journal of Psycholinguistic Research*, 29(5):483–495.
- Kimiko Ryokai, Elena Durán López, Noura Howell, Jon Gillick, and David Bamman. 2018. Capturing, representing, and interacting with laughter. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 358. ACM.
- R Schifanella, P de Juan, J Tetreault, L Cao, et al. 2016. Detecting sarcasm in multimodal social platforms. In *ACM Multimedia*, pages 1136–1145. ACM.
- Joseph Tepperman, David Traum, and Shrikanth Narayanan. 2006. "yeah right": Sarcasm recognition for spoken dialogue systems. In *Ninth International Conference on Spoken Language Processing*.
- Dominic Thompson, Ian G Mackenzie, Hartmut Leuthold, and Ruth Filik. 2016. Emotional responses to irony and emoticons in written language: evidence from eda and facial emg. *Psychophysiology*, 53(7):1054–1062.
- Tony Veale and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *ECAI*, volume 215, pages 765–770.
- Byron C Wallace, Eugene Charniak, et al. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1035–1044.
- Byron C Wallace, Laura Kertz, Eugene Charniak, et al. 2014. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 512–516.
- Silvio Amir Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *CoNLL 2016*, page 167.
- Jennifer Woodland and Daniel Voyer. 2011. Context and intonation in the perception of sarcasm. *Metaphor and Symbol*, 26(3):227–239.
- Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.