A Survey of Multimodal Sarcasm Detection

Shafkat Farabi[1], Tharindu Ranasinghe[2], Diptesh Kanojia[3], Yu Kong[4], Marcos Zampieri[1]
[1]George Mason University, USA,
[2]Lancaster University, UK
[3]University of Surrey, UK,
[4]Michigan State University, USA
mfarabi@gmu.edu

Abstract

Sarcasm is a rhetorical device that is used to convey the opposite of the literal meaning of an utterance. Sarcasm is widely used on social media and other forms of computer-mediated communication motivating the use of computational models to identify it automatically. While the clear majority of approaches to sarcasm detection have been carried out on text only, sarcasm detection often requires additional information present in tonality, facial expression, and contextual images. This has led to the introduction of multimodal models, opening the possibility to detect sarcasm in multiple modalities such as audio, images, text, and video. In this paper, we present the first comprehensive survey on multimodal sarcasm detection - henceforth MSD - to date. We survey papers published between 2018 and 2023 on the topic, and discuss the models and datasets used for this task. We also present future research directions in MSD.

1 Introduction

Sarcasm is a sophisticated linguistic phenomenon wherein individuals articulate thoughts using words that convey the opposite of their intended meaning [Tiwari *et al.*, 2023]. The Cambridge English Dictionary defines sarcasm as *"The use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way"*. Sarcasm is prevalent in user generated content across social media platforms such as Twitter (now known as X), Facebook, and Reddit, as well as in popular culture, including sitcoms and movies. Many use sarcasm to convey contempt, anger, humor, or derogatory sentiments [Maynard and Greenwood, 2014].

Sarcasm is often expressed using incongruity between spoken words and the intended sentiment. It frequently relies on usage of hyperbole and reference to contextual world knowledge [Chaudhari and Chandankhede, 2017]. These strategies make automatically identifying sarcasm a challenging yet interesting task for many applications. For example, the figurative nature of sarcasm makes it an often-quoted challenge for sentiment analysis [Joshi *et al.*, 2017]. Detection of elements of sarcasm helps to resolve seemingly contradictory sentiments like *"The restaurant was so clean that I could barely avoid stepping into the puddle!"* [Badlani *et al.*, 2019]. Maynard and Greenwood [2014] show that correctly detecting sarcasm can significantly improve sentiment analysis systems. Similarly, sarcasm plays a crucial role in offensive speech and humor identification [Frenda, 2018]. Finally, applications that model mental health on social media can also benefit from sarcasm identification. Rothermich *et al.* [2021] show a correlation between sarcasm use and mental conditions such as anxiety and depression.

The importance of sarcasm identification for better understanding communication cannot be overstated. We find growing interest in the problem within the AI, Computer Vision, Speech Processing, and NLP communities that motivates us to present this survey. To the best of our knowledge, this is the first comprehensive survey on MSD filling an important gap in the literature. We survey over 60 papers that present datasets and computational approaches to detect sarcasm and we describe them in detail in this survey. The remainder of this paper is organized as follow: Section 2 discusses text-based sarcasm detection as compared to multimodal approaches. Visuo-Textual detection of sarcasm is discussed in Section 3 while Section 4 discusses Audio-Visual & Textual detection. Section 5 concludes this survey and presents avenues for future work.

2 Textual *vs.* Multimodal Sarcasm Detection

A clear majority of the previous works in automatic sarcasm detection have focused on text classification. Often portrayed as a supervised machine learning problem, several datasets have been introduced for text-based sarcasm detection. The biggest and most popular such dataset, `SARC` [Khodak *et al.*, 2018], contains 533 million sarcastic and 1.3 million non-sarcastic data collected from Reddit. `iSarcasm` is another such popular dataset [Oprea and Magdy, 2020] containing 777 sarcastic and 3707 non-sarcastic sentences, all collected from Twitter. Numerous studies have been conducted on these datasets over the years, encompassing both conventional machine learning models in the early stages and more recent deep learning approaches such as transformers [Hazarika *et al.*, 2018; Liu *et al.*, 2022b]. These datasets and methods on text based sarcasm detection are widely discussed in surveys such as [Chaudhari and Chandankhede, 2017; Joshi

et al., 2017; Verma *et al.*, 2021; Salini and HariKiran, 2023; Alqahtani *et al.*, 2023].

There is yet another form of sarcasm prevalent on social media where users opt for accompanying text with images to express sarcasm. In these cases, the text conveys a meaning that contrasts the content of the image. Figure 1 presents an example where the text and the image taken individually are not sarcastic but paired together; they express sarcastic intent.

(a) Thanks again for the full fries!

(b) Another perfect pizza from <user>!

Figure 1: Sarcasm using an image accompanying some text.

Additionally, humans can utilize their facial expressions and voice tone to supplement what they are saying in order to express sarcasm. In this case, video, audio, and text are all necessary to express sarcasm. Figure 2 shows such a scenario. In this case, the contextual conversation leading up to the sarcastic remark is necessary to appreciate the irony.

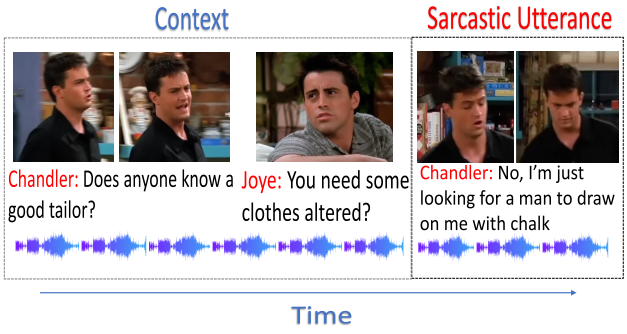


Figure 2: Sarcasm conveyed through text (dialogue), audio (tone), and video (facial expression). The context is important in these cases.

Cai *et al.* [2019] pointed out that identifying sarcasm solely based on text is sometimes impossible. In fact, necessary cues to understand sarcasm are often present in the facial cues of the speaker and/or media accompanying the text. Hence, automated models tasked with detecting sarcasm need to be able to take in visual (and sometimes auditory) information to complement the textual data. Following this, there has been an increased effort in the research community to design automatic systems to detect multimodal sarcasm.

Numerous datasets have been curated for MSD with data collected from social media and TV shows such as sitcoms. Curating large human annotated datasets is, however, time consuming and extensive. Accurately annotating sarcastic utterances is a particularly challenging task due to its intrinsic subjectivity. Sarcasm is expressed using underlying incongruity, but this incongruity can be explicitly obvious, or implicitly presented without any negative sentiment phrases; and the degree of incongruity can vary [Mishra *et al.*, 2016]. Furthermore, the annotator's judgement of sarcasm is also known to be effected by their cultural upbringing [Joshi *et al.*, 2016].

A multitude of deep learning frameworks have been proposed that can learn from these datasets. Initial works focused on using separate encoders such as ResNets and BERT to encode the data and then proposed novel techniques to fuse these higher-level features. Later works build on top of these approaches by introducing more complex fusion techniques. More recent studies have moved towards an approach of tuning multimodal encoders such as CLIP, VilBERT, and VisualBERT for this specific task.

Methodology The previous surveys on sarcasm detection focus solely on text [Joshi *et al.*, 2017]. In this paper, we fill this important research gap by summarizing the datasets and state-of-the-art methods on MSD in two categories: **(1) Visuo-Textual and (2) Audio-Visual and Textual Datasets**. In order to find research papers on sarcasm detection, we searched for relevant papers on *Google Scholar* and scientific databases such as ACM Digital Library, IEEE Explore, ACL Anthology, Springer, and CVF Open Access. We use keywords such as *multimodal*, *sarcasm detection*, *sarcasm detection from images*, *social media sarcasm detection*, and others. Most of the papers we report findings from were published in reputed venues such as AAAI, ACL, CVPR, EMNLP, NAACL, and others. We survey papers that curate datasets for MSD or propose a method that can perform substantially well on existing benchmark datasets. In Figure 3, we present a comparison of the number of studies published in text only sarcasm vs. multi-modal sarcasm detection, between the years 2018-2023.

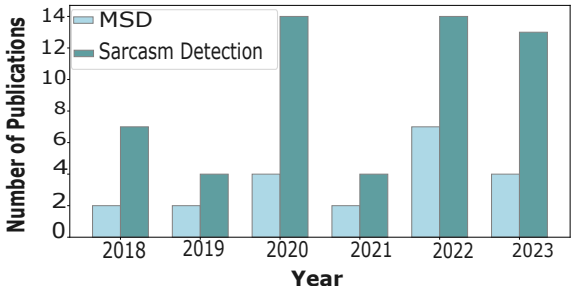


Figure 3: Number of papers on multimodal sarcasm detection and text only sarcasm detection.

Dataset Name	Source	Sarcastic Samples	Non-sarcastic Samples	Additional Remarks
MSDD [Cai *et al.*, 2019]	Twitter	10560	14075	Tweets containing a picture. Annotated using hashtags in the tweet
MSDD 2.0 [Qin *et al.*, 2023]	Twitter	11651	12980	Enhanced version of MSDD, spurious cues removed manually corrected annotations
[Schifanella *et al.*, 2016]	Instagram, Twitter, Tumbler	22,025	20,025	Labelled using hashtags, a subset human annotated
Silver-Standard Dataset [Sangwan *et al.*, 2020]	Instagram	10,000	10,000	Labelled using hashtags
Gold-Standard Dataset [Sangwan *et al.*, 2020]	Instagram	1600	10,000	Human-annotations for sarcasm samples
[Das and Clark, 2018a]	Facebook	20,120	21,230	Not all samples are multimodal. 98.26% samples have accompanying images
MORE [Desai *et al.*, 2022]	[Schifanella *et al.*, 2016] & [Sangwan *et al.*, 2020]	3510	-	Contains natural language explanation of the sarcasm with non-sarcastic form as well

Table 1: Summary of datasets for **Visuo-Textual** sarcasm detection.

We observe that papers on MSD account for significant number of publications on sarcasm detection in this period. Hence, a summary of the literature on MSD is essential to aid future work. We expect this survey to help both researchers already working on MSD as well as researchers new to the task. Furthermore, we believe this is a very timely survey given that the recent introduction of Large Language Models (LLMs) is likely to spark more interest in research on vision and language processing applications.

3 Visuo-Textual Sarcasm Detection

The use of images accompanied by text (visuo-textual) are a common way to express sarcasm on social media. Sarcasm expressed through such medium relies heavily on incongruity between the image and text modality. In the following sections, we describe the datasets collected and the approaches to visuo-textual sarcasm detection.

3.1 Datasets

We summarized all datasets for visuo-textual sarcasm detection in Table 1. Schifanella *et al.* [2016] presented one of the first datasets comprised of text and image pairs collected from user posts on Instagram, Twitter, and Tumblr. The authors collect 10,000 sarcastic and 10,000 non-sarcastic posts from Instagram and Tumbler each, and 2,005 sarcastic and 2,005 non-sarcastic posts from Twitter. The authors also provide human annotations for 1,000 sarcastic images from Instagram and 1,000 sarcastic images from Tumblr. Cai *et al.* [2019] presented a similar dataset collected from Twitter. Later works refer to this dataset as the MMSD dataset. We present instances from the MMSD dataset in Figure 4. The dataset contains 19,818 training, 2,410 validation, and 2,409 test examples. Of these, 10,560 samples are sarcastic, and 14,075 non-sarcastic. Tweets containing hashtags similar to #sarcasm are labelled as sarcastic examples and non-sarcastic otherwise.

Later, Qin *et al.* [2023] enhanced the MMSD dataset by removing spurious cues. They point out that many positive samples in MMSD dataset contain hashtags and emojis that might serve as an easy giveaway to the sarcastic nature of the data. They further point out that some negative samples in MMSD are mis-annotated. They manually re-annotate the negative samples and remove the spurious cues from the positive samples. This version of the dataset is named MMSD2.0 and contains 9,572 sarcastic and 10,240 non-sarcastic train samples. The number of sarcastic samples in the validation and test sets changed to 1,042 and 1,037, respectively. The non-sarcastic samples in the validation and test sets changed to 1,368 and 1,372, respectively.

(a) What a joy to wake up the morning after thanksgiving dinner at the hatch flat !!

(b) the nice thing is that after i get my driveway shoveled, i can start shoveling the road .

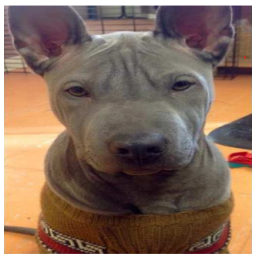
Figure 4: Examples from the MMSD [Cai *et al.*, 2019] Dataset.

Sangwan *et al.* [2020] release a similar dataset composed of image and text pairs collected from Instagram. They release two versions of the dataset of different sizes. The bigger version, named 'Silver-Standard Dataset', consists of 20,000 Instagram posts, evenly distributed among positive and negative samples. Similar to MMSD, they determine positive/negative samples based on presence or lack thereof of hashtags in the post (#sarcasm, #sarcastic, etc.). They also released a smaller version of the dataset, named 'Gold-Standard Dataset', containing only 1,600 human annotated samples. We present examples from this dataset in Figure 5.

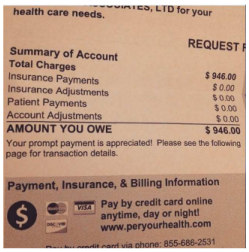
Das and Clark [2018a] introduced a partially multimodal sarcasm detection dataset collected from Facebook. This

Method	MMSD				MMSD 2.0				Silver	Gold
	acc.	pre.	rec.	f1.	acc.	pre.	rec.	f1.	acc.	acc.
[Cai *et al.*, 2019]	83.44	76.57	84.15	80.18	70.57	64.84	69.05	66.88	-	-
VisualBERT [Li *et al.*, 2019]	83.51	76.66	82.94	79.68	-	-	-	-	-	-
LXMERT [Tan and Bansal, 2019]	83.93	77.83	82.59	80.14	-	-	-	-	-	-
[Xu *et al.*, 2020]	84.02	77.97	83.42	80.60	-	-	-	-	-	-
ViLBERT [Lu *et al.*, 2019]	84.68	77.52	86.37	81.71	-	-	-	-	-	-
[Pan *et al.*, 2020]	86.05	80.87	85.08	82.92	80.03	76.28	77.82	77.04	-	-
[Liang *et al.*, 2021]	86.10	85.39	85.80	85.60	-	-	-	-	-	-
[Liang *et al.*, 2022]	87.55	87.02	86.97	87.00	79.83	75.82	78.01	76.90	-	-
[Liu *et al.*, 2022a]	87.36	81.84	86.48	84.09	76.50	73.48	71.07	72.25		
[Qin *et al.*, 2023]	88.33	82.66	88.65	85.55	85.64	80.33	88.24	84.10	-	-
[Wang *et al.*, 2020]	88.51	82.95	89.39	86.05	-	-	-	-	-	-
[Pramanick *et al.*, 2022]	90.82	-	-	88.20	-	-	-	-	-	-
[Tian *et al.*, 2023]	93.49	-	-	93.21	-	-	-	-	-	-
[Ding *et al.*, 2022]	93.85	93.57	94.89	94.06	-	-	-	-	-	-
[Sangwan *et al.*, 2020]	-	-	-	-	-	-	-	-	84.22	71.5
GPT4 [Lin *et al.*, 2024]	75.88	-	-	75.08	-	-	-	-	-	-
InstructBLIP [Yang *et al.*, 2023]	73.10	-	-	-	-	-	-	-	-	-

Table 2: A summary of approaches and their performance on **Visuo-Textual** datasets.



(a) Someone is excited for sweater season.



(b) Life is just going great this year so far for me.

Figure 5: Example samples from the 'Silver-Standard Dataset'.

dataset contains 20,120 sarcastic and 21,230 non-sarcastic samples, 98.26% of which include both an image and text. Desai *et al.* [2022] investigated a related but novel research field of sarcasm explanation generation from multimodal input. In simple terms, given a sarcastic image-text pair, the goal of this task is to generate a natural language explanation for the pair to be considered sarcastic. To this end, they proposed the MORE dataset containing 3,510 sarcastic utterances, including texts and images. Each utterance is accompanied by a natural language explanation of sarcasm and a non-sarcastic version.

3.2 Methods

The datasets described in the last section have been explored using deep learning approaches. Here we broadly classify these approaches into three classes: (1) traditional deep learning models that use separate encoders for image and text, (2) multimodal transformers, and (3) LLM-based approaches with prompt engineering. We describe the approaches next and summarize their performance on two benchmark datasets in Table 2.

Traditional deep learning models Schifanella *et al.* [2016] were the first to propose that visual contextual features are necessary to decode sarcastic intent from text. Their proposed approach concatenates extracted visual and textual features. Although, the authors demonstrated that visual modality helps in detecting sarcasm in social media, they did not investigate the nature of this relationship, nor did they attempt to engineer specific methods based on the nature of this relation. Das and Clark [2018a] utilized CNN from [Das and Clark, 2018b] to extract image features, and many low level features from post description and user reactions. Their approach can adapt to detecting sarcasm in both text only and image-text multi-modal scenarios in a low-data environment. A key limitation is the dependency on audience reaction. Furthermore, the CNN used for extracting features was noted to incorrectly associate location of the image with sarcastic intent [Das and Clark, 2018b]. Cai *et al.* [2019] extracted image attributes and introduced it as a third modality in an attempt to boost performance. The authors identified that multimodal feature fusion by means of simple concatenation is insufficient, and improved this aspect by introducing hierarchical fusion. They performed early fusion by initializing text modality Bi-LSTM with features from the visual modality, and performed representation fusion and modality fusion following the works of Gu *et al.* [2018]. Although the authors attempted to design a sophisticated inter-modal feature fusion technique, they did not try to analyze and take advantage of how information from these modalities interplay with each other.

Pan *et al.* [2020] noted that inter and intra modal incongruity play an important role in sarcasm identification. They took advantage of this by proposing a BERT and ResNet based model that can concentrate on both inter-modal and intra-modal incongruity. They accomplished this by introducing attention both in intra-modality and inter-modality

fashion. The incongruity between modalities can be disordered and unstructured, which the authors of Pan *et al.* [2020] aimed to teach their model solely using data. In order to make this learning process more explicit, Xu *et al.* [2020] proposed the D&R (Decomposition and Reconstruction) network. They projected the image and text representations in a common subspace, and unique sub-spaces orthogonal to the common space. They later fused features from the unique sub-spaces in an attempt to focus the model more on contrasting elements in modalities. They also extracted adjective-noun pairs (ANP) from the images, and applied ANP aware cross modal attention to make the model more aware of semantic associations between cross modal contexts. To make the learning of inter modal incongruity more structured, Liang *et al.* [2021] proposed constructing modality specific and cross-modal dependency graphs from the features extracted through BERT and ViT [Dosovitskiy *et al.*, 2020]. They processed the information stored within these graphs by using interactive graph convolutional networks (GCN). In a later work, Liang *et al.* [2022] improved further by engineering a method that only focuses on relevant patches of the image that relate to sarcastic cues in the text, achieved by refining the construction of a cross modal graph, by focusing on the objects in the image. More specifically, they followed Anderson *et al.* [2018] to extract image-attribute pairs from the image. Next, the cross modal graph was generated by using the similarity between image attributes and text words. One drawback of their approach is the dependency on external knowledge to determine word similarity which is a crucial part of their construction of cross modal graph.

The methods discussed thus far focus on using different encoders for different modalities and focus on effectively fusing multi-modal representations in a manner that caters to the incongruity within and between the modalities. Teaching deep networks to find such associations between high level features of different modalities is difficult in a low data environment, even with clever techniques.

Multimodal Transformers that can encode text and image to a common feature space are becoming popular. Wang *et al.* [2020] benchmarked a few of these methods (VisualBERT, LXMERT, ViLBERT) for MSD. While a common feature space for multi-modal encodings may help, comparatively smaller pre-training available for these multimodal encoders result to substandard performance compared to uni-modal encoders like BERT and ResNet. However, to show a common feature space for encoding does help, they introduced a trainable bridge between a text-only and image-only encoder, to align them. Along with a 2D intra-attention module for feature fusion, achieving good performance. Qin *et al.* [2023] presented `multi-view CLIP`, which further solidifies the idea that a common vision-language feature space facilitates better performance on MSD. They use CLIP [Radford *et al.*, 2021], a popular multi-modal feature extractor, along with clever engineering with transformers for feature fusion.

Prompting and LLMs Ding *et al.* [2022] explored prompt-tuning for multimodal sarcasm detection. They modeled sarcasm detection as a masked language prediction task and integrated it with a ViT for image encoding and an inter-modality attention transformer to predict the sarcasm level of the text. They used a dot product based similarity assessment similar to the approach of Radford *et al.* [2021] for providing supervision for training the model.

Lin *et al.* [2024] studied the ability of multimodal LLMs to identify social abuse in social media memes in a nuanced fashion. They released a multi-task benchmark (GOAT Benchmark) containing different task categories, one of which is multimodal visuo-textual sarcasm and consists of data sampled from MMSD dataset. They benchmarked several popular LLMs like GPT4 and LLaVa-1.5 on these tasks with template prompting. Yang *et al.* [2023] released another multimodal benchmark for LLMs titled MM-BigBench containing sarcasm detection as a task and samples data from MMSD. They benchmarked various LLMs like GPT4, LLaMa, OpenFlamingo, Blip, and InstructBLIP etc. Despite being initial works touching multimodal sarcasm detection with LLMs, sarcasm was not the primary focus here. Furthermore, they did not experiment with prompt tuning the LLMs to improve detection performance, focusing on LLM capabilities in a zero shot scenario.

4 Audio-Visual & Textual Sarcasm Detection

In this modality, sarcasm is detected via audio and video recordings of dialogue accompanied by text captions. This type of sarcasm is very prevalent in sitcoms, TV shows, and stand-up comedy. While sarcasm can also exist in short video social media platforms such as TikTok and YouTube, all the datasets and methods developed to tackle this modality have focused on sitcoms and TV shows.

4.1 Datasets

A summary of datasets developed for audio-visual and textual sarcasm is presented in Table 3. The most widely-used dataset is `MUStARD` [Castro *et al.*, 2019] which contains sarcastic clips from popular sitcom TV shows, namely *Friends*, *The Golden Girls*, *Sarcasmaholics Anonymous*, and *The Big Bang Theory*. For non-sarcastic utterances, the authors reuse data from a multimodal emotion recognition dataset called `MELD` [Poria *et al.*, 2019]. The dataset is balanced, manually annotated and contains 690 samples. Each utterance contains a video clip, audio, and captions in text with necessary contextual conversation leading to the utterance. The context includes audio, video, captions and speaker identifiers, as often these clips contain a conversation between multiple parties.

Numerous research studies have extended the `MUStARD` dataset in several directions. Chauhan *et al.* [2020] introduced sentiment and emotion labels in the `MUStARD` dataset, thereby building `SE-MUStARD`, showing that these labels improve sarcasm detection. Ray *et al.* [2022] further enhanced the `SE-MUStARD` dataset to almost double its size and introduce emotion, valence, arousal, and sarcasm-type labels. They also corrected nearly 399 annotation errors in Chauhan *et al.* [2020]'s emotion labels. They published this corrected and enhanced dataset as `MUStARD++`. `MUStARD++` is enhanced with 264 new videos from 'The Big Bang Theory', and 'The Silicon valley'. The total number of sarcastic utterances in this dataset is 601. In order to keep it balanced, additional non-sarcastic examples were also included.

Dataset Name	Source	Sarcastic Samples	Non-sarcastic Samples	Additional Remarks
MUStARD [Castro *et al.*, 2019]	TV Shows (YouTube)	345	345	Includes clips from sitcoms, with contextual data, speaker information. Annotated manually. Non-sarcastic samples collected from MELD [Poria *et al.*, 2019]
SE-MUStARD [Chauhan *et al.*, 2020]	MUStARD	345	345	Adds sentiment and emotion labels to MUStARD Annotated manually.
MUStARD++ [Ray *et al.*, 2022]	TV Shows (YouTube)	601	601	Enhanced MUStARD with additional videos and labels Provides corrections for some labels in MUStARD.
MUStARD++ Balanced [Bhosale *et al.*, 2023]	MUStARD++ & House MD	691	674	Extended to balance the sarcasm types
SEEmoji MUStARD [Chauhan *et al.*, 2022]	MUStARD++	691	601	Augmented with emojis, sentiment, and emotions
Spanish Multimodal Sarcasm [Alnajjar and Hämäläinen, 2021]	Archer, South Park	90	869	Voice and text are in Spanish, manually annotated

Table 3: Summary of datasets for **Audio-Visual & Textual** sarcasm detection.

Bhosale *et al.* [2023] noticed that the newly introduced 'sarcasm types' category in MUStARD++ is imbalanced. In order to address this issue, they augmented the dataset with 90 sarcastic and 74 non-sarcastic samples taken from the TV series 'House MD'. They manually annotated the 'sarcasm types' label of the newly introduced data and named this extended dataset MUStARD++ Balanced. As a by-product of this effort, they increased the diversity by adding new data. In more recent work, Chauhan *et al.* [2022] published another version of MUStARD called SEEmoji MUStARD. The authors noted that emotions and sentiments are sometimes implicit and difficult to decipher from text. But emojis can play an important role by alluding to the implicit emotions embedded in the text by the speaker. Hence, they appended appropriate emojis from a pool of 25 most popular ones used in social media, along with the sentiment (positive/negative/neutral) and emotion labels of the emojis to each sample of this dataset.

Finally, Alnajjar and Hämäläinen [2021] presented a dataset in Spanish comprised of clips taken from *Archer* and *South park* TV shows. The dataset contains 960 utterances, of which only 90 are sarcastic, and 869 are non-sarcastic. The dataset does not contain any train-test splits. The authors include two different dialects of Spanish, and manually annotate samples.

4.2 Methods

Designing methods to detect sarcasm from audio-visual and textual datasets is more difficult than visuo-textual datasets. These datasets are comprised of video and audio, along with transcript of the audio. Often it contains multiple persons having a conversation. The cues for sarcasm such as intra-modal incongruity can be manifested in a nuanced manner through facial expression, voice tone, and hand gestures [Castro *et al.*, 2019].

A summary of all methods and their performance applied to audio-visual and textual sarcasm detection is presented in Table 4. The methods experimenting with these datasets are deep learning based, and these can be broadly classified into three classes: (1) Traditional Deep Learning Approaches fusing the multimodal features by concatenating them, (2) Approaches using Multimodal Attention for feature fusion, and (3) Approaches using Multi-Task learning where sentiment classification is an auxiliary task.

Traditional deep learning approaches Castro *et al.* [2019] was the first study that demonstrates audio and video can help boost performance on MSD, as can the relevant context. Being an initial work, their proposed framework is rather straight forward, using BERT Librosa, and ResNet-152 for feature extraction, followed by feature concatenation and prediction using an SVM. Alnajjar and Hämäläinen [2021] took on a similar approach of training an SVM to predict sarcasm from concatenated modality specific features. However, they are the only work dealing with Audio-visual and textual detection of sarcasm in a non-English language (Spanish). Their study re-affirms the importance of multiple modalities for detecting sarcasm, as suggested by Castro *et al.* [2019]. However, a benchmarking of the current state-of-the-art MUsTARD dataset frameworks on their Spanish Multimodal Sarcasm dataset is absent from this work.

Furthermore, both suffer from limitations due to lack of investigations in complex multimodal fusion techniques, without the advantage of multiparty conversation relationships, and the utilization of SVM over neural networks.

Multimodal Attention Wu *et al.* [2021] proposed a multimodal fusion technique that can identify and use information pertaining to inter-modal incongruities. They proposed IWAN model with a focus on such incongruities in the form of positive spoken words paired with negative tone/facial expression, achieved through an attention-based word level scoring mechanism using features from BERT, ResNet and OpenSmile [Eyben *et al.*, 2010]. Notably, this technique improved sarcasm detection but they modeled word-tone level incongruity, and left exploration of contextual incongruities for future. Aggarwal *et al.* [2023] filled this gap by proposing the use of multi-headed bimodal attention, targeting incorporation of multimodal incongruities in a global scenario. More

Method	Dataset	Accuracy	Precision	Recall	F1 Score
[Castro *et al.*, 2019]	MUStARD	-	72.6	71.6	71.6
[Chauhan *et al.*, 2020]	Se-MUStARD	-	73.4	72.8	72.6
IWAN[Wu *et al.*, 2021]	MUStARD	-	75.2	75.2	75.1
[Ray *et al.*, 2022]	MUStARD	-	74.2	74.2	74.2
[Aggarwal *et al.*, 2023]	MUStARD	79.32	78.1	77.42	77.6
MuLOT [Pramanick *et al.*, 2022]	MUStARD	78.57	-	-	-
[Ray *et al.*, 2022]	MUStARD ++	-	70.3	70.3	70.3
[Tiwari *et al.*, 2023]	MUStARD ++	-	73.2	73.2	73.3
[Bhosale *et al.*, 2023]	MUStARD ++	-	73.5	72.8	73.1
[Bhosale *et al.*, 2023]	MUStARD ++ Balanced	-	73.8	73.5	73.6
[Chauhan *et al.*, 2022]	SEEmoji-MUStARD	-	77.9	76.9	76.7
[Alnajjar and Hämäläinen, 2021]	Spanish Multimodal Sarcasm	93.1	-	-	-

Table 4: A summary of approaches and their performance on **Audio-Visual and Textual** datasets.

complex methods of modeling cross-modal incongruity were hindered by the size of the MUStARD dataset. To circumvent this limitation, Pramanick *et al.* [2022] proposed the MuLOT framework, where cross-modal incongruity is learned using optimal-transport, while self-attention is introduced to tackle lack of intra-modal incongruity.

Sarcasm can also be identified by readers' gaze pattern. Tiwari *et al.* [2023] studied this phenomena by incorporating gaze features for multi-modal sarcasm detection. They collected gaze information for a subset of MUStARD++ dataset and designed a framework to predict gaze information from textual utterances, demonstrating gaze features with text, video, and audio, improve task performance on MUStARD++ dataset.

Not unlike research in visuo-textual sarcasm detection, the trend is now shifting towards using multimodal transformers. The reason for this preference is the fact that multimodal transformers are more capable of identifying both intra and inter modal dependencies from data. Bhosale *et al.* [2023] employed a ViFi-CLIP [Rasheed *et al.*, 2023], a video-text encoder, to encode the video frames as well as the text in a common representation space. They also used a Wav2vec 2.0 [Baevski *et al.*, 2020], a self supervised transformer based speech encoder, fine tuned on speech emotion recognition to encode the audio.

Multi-Tasking with Auxiliary Sentiment Classification: Chauhan *et al.* [2020] explored the role of speaker sentiment in sarcasm identification. They augmented the MUStARD dataset with emotion and sentiment labels, used attention for aggregating the features and trained their model in a multi-task learning approach where sentiment classification is the auxiliary task. The complex role of sentiment and emotion in the context of sarcasm detection was also explored by Ray *et al.* [2022]. They introduced the MUStARD++ dataset, and utilized a collaborative gating strategy for multimodal feature fusion with an extensive ablation study on the effect of speaker information and the modalities. In a later work, Chauhan *et al.* [2022] explored this further by attaching emojis that often have sentiments contrasting that of the sentence. They proposed an emoji-aware-multi-modal-multitask deep learning framework using emotion and sentiment classification as an auxiliary task and evaluate on SEEmoji MUStARD. These studies demonstrate that auxiliary information pertaining to the speaker emotion and sentiment help in detecting irony and sarcasm.

5 Conclusion And Future Directions

This paper presented the first comprehensive survey of MSD. We presented popular datasets as well as computational approaches used for this task. As the interest on MSD continues to grow, we see the following directions for future research.

Multilingual datasets As evidenced in this survey, the bulk of work on MSD is on English data, leaving a critical gap within applications developed for other languages. A notable exception is the work by Alnajjar and Hämäläinen [2021] on Spanish. We hope this survey encourages the creation of larger and more comprehensive multilingual data to aid research on MSD on languages other than English.

Perspectivism Identifying sarcasm is a highly subjective task for humans. Different people see sarcasm differently, and this is reflected in dataset annotation. Ground truth labels in annotated MSD datasets are based on annotations by a single person or on the aggregation of multiple annotations. We believe that developing MSD models that consider perspectives from multiple annotators, as in [Weerasooriya *et al.*, 2023] is a more realistic way of representing the problem and it should be explored in the future.

Inter-task dependencies Sarcasm is related to other forms of non-literal language such as humor, and also offensive language and hate speech. The recent HahaCkathon shared task at SemEval [Meaney *et al.*, 2021], for example, introduced a dataset annotated with respect to humor and offense. This opens the possibility of exploring inter-task dependencies, and to use multi-task learning where MSD can also be modeled jointly with other related tasks.

LLMs The recent introduction of a new generation of LLMs is a promising direction for research in MSD. We believe that models that are able to model image and text (e.g., GPT-4) should be further explored for MSD as they have proven to achieve state-of-the-art performance on multiple vision and language tasks.

References

[Aggarwal *et al.*, 2023] Sajal Aggarwal, Ananya Pandey, and Dinesh Kumar Vishwakarma. Multimodal sarcasm recognition by fusing textual, visual and acoustic content via multi-headed attention for video dataset. In *WCONF*, 2023.

[Alnajjar and Hämäläinen, 2021] Khalid Alnajjar and Mika Hämäläinen. ¡Qué maravilla! multimodal sarcasm detection in Spanish: a dataset and a baseline. In *NAACL*, 2021.

[Alqahtani *et al.*, 2023] Amal Alqahtani, Lubna Alhenaki, and Abeer Alsheddi. Text-based sarcasm detection on social networks: A systematic review. *IJACSA*, 14(3), 2023.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[Badlani *et al.*, 2019] Rohan Badlani, Nishit Asnani, and Manan Rai. An ensemble of humour, sarcasm, and hate speech for sentiment classification in online reviews. In *W-NUT 2019*, November 2019.

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*, 2020.

[Bhosale *et al.*, 2023] Swapnil Bhosale, Abhra Chaudhuri, Alex Lee Robert Williams, Divyank Tiwari, Anjan Dutta, Xiatian Zhu, Pushpak Bhattacharyya, and Diptesh Kanojia. Sarcasm in sight and sound: Benchmarking and expansion to improve multimodal sarcasm detection. *arXiv preprint arXiv:2310.01430*, 2023.

[Cai *et al.*, 2019] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *ACL*, 2019.

[Castro *et al.*, 2019] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *ACL*, 2019.

[Chaudhari and Chandankhede, 2017] Pranali Chaudhari and Chaitali Chandankhede. Literature survey of sarcasm detection. In *WiSPNET*, 2017.

[Chauhan *et al.*, 2020] Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *ACL*, 2020.

[Chauhan *et al.*, 2022] Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. An emoji-aware multitask framework for multimodal sarcasm detection. *Knowledge-Based Systems*, 257:109924, 2022.

[Das and Clark, 2018a] Dipto Das and Anthony J Clark. Sarcasm detection on facebook: A supervised learning approach. In *ICMI*, 2018.

[Das and Clark, 2018b] Dipto Das and Anthony J. Clark. Sarcasm detection on flickr using a cnn. In *ICCBD*, 2018.

[Desai *et al.*, 2022] Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *AAAI*, 2022.

[Ding *et al.*, 2022] Daijun Ding, Hu Huang, Bowen Zhang, Cheng Peng, Yangyang Li, Xianghua Fu, and Liwen Jing. Multi-modal sarcasm detection with prompt-tuning. In *ACAIT*, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Eyben *et al.*, 2010] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *ACM MM*, 2010.

[Frenda, 2018] Simona Frenda. The role of sarcasm in hate speech. a multilingual perspective. In *SEPLN*, 2018.

[Gu *et al.*, 2018] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. Hybrid attention based multimodal network for spoken language classification. In *ACL*, 2018.

[Hazarika *et al.*, 2018] Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. CASCADE: Contextual sarcasm detection in online discussion forums. In *COLING*, 2018.

[Joshi *et al.*, 2016] Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. How do cultural differences impact the quality of sarcasm annotation?: A case study of Indian annotators and American text. In *SIGHUM*, 2016.

[Joshi *et al.*, 2017] Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22, 2017.

[Khodak *et al.*, 2018] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. A large self-annotated corpus for sarcasm. In *LREC*, 2018.

[Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[Liang *et al.*, 2021] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *ACM MM*, 2021.

[Liang *et al.*, 2022] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *ACL*, 2022.

[Lin *et al.*, 2024] Hongzhan Lin, Ziyang Luo, bo Wang, Ruichao Yang, and Jing Ma. Goat-bench: Safety insights

to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*, 2024.

[Liu *et al.*, 2022a] Hui Liu, Wenya Wang, and Haoliang Li. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *EMNLP*, 2022.

[Liu *et al.*, 2022b] Yiyi Liu, Yequan Wang, Aixin Sun, Xuying Meng, Jing Li, and Jiafeng Guo. A dual-channel framework for sarcasm recognition by detecting sentiment conflict. In *NAACL*, 2022.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for v-l tasks. In *NeurIPS*, 2019.

[Maynard and Greenwood, 2014] Diana Maynard and Mark Greenwood. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC*, 2014.

[Meaney *et al.*, 2021] JA Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *SemEval*, 2021.

[Mishra *et al.*, 2016] Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI*, 2016.

[Oprea and Magdy, 2020] Silviu Oprea and Walid Magdy. iSarcasm: A dataset of intended sarcasm. In *ACL*, 2020.

[Pan *et al.*, 2020] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *EMNLP*, 2020.

[Poria *et al.*, 2019] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, 2019.

[Pramanick *et al.*, 2022] Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *WACV*, 2022.

[Qin *et al.*, 2023] Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In *ACL*, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Rasheed *et al.*, 2023] Hanoona Rasheed, Muhammad Uzair khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Finetuned clip models are efficient video learners. In *CVPR*, 2023.

[Ray *et al.*, 2022] Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. A multimodal corpus for emotion recognition in sarcasm. In *LREC*, 2022.

[Rothermich *et al.*, 2021] Kathrin Rothermich, Ayotola Ogunlana, and Natalia Jaworska. Change in humor and sarcasm use based on anxiety and depression symptom severity during the covid-19 pandemic. *Journal of psychiatric research*, 140:95–100, 2021.

[Salini and HariKiran, 2023] Yalamanchili Salini and J. HariKiran. Sarcasm detection: A systematic review of methods and approaches. In *ICSMDI*, 2023.

[Sangwan *et al.*, 2020] Suyash Sangwan, Md Shad Akhtar, Pranati Behera, and Asif Ekbal. I didn't mean what i wrote! exploring multimodality for sarcasm detection. In *IJCNN*, 2020.

[Schifanella *et al.*, 2016] Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. Detecting sarcasm in multimodal social platforms. In *ACM MM*, 2016.

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, 2019.

[Tian *et al.*, 2023] Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. Dynamic routing transformer network for multimodal sarcasm detection. In *ACL*, 2023.

[Tiwari *et al.*, 2023] Divyank Tiwari, Diptesh Kanojia, Anupama Ray, Apoorva Nunna, and Pushpak Bhattacharyya. Predict and use: Harnessing predicted gaze to improve multimodal sarcasm detection. In *EMNLP*, 2023.

[Verma *et al.*, 2021] Palak Verma, Neha Shukla, and AP Shukla. Techniques of sarcasm detection: A review. In *ICACITE*, 2021.

[Wang *et al.*, 2020] Xinyu Wang, Xiaowen Sun, Tan Yang, and Hongbo Wang. Building a bridge: A method for image-text sarcasm detection without pretraining on image-text data. In *NLPBT*, 2020.

[Weerasooriya *et al.*, 2023] Tharindu Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher Homan, and Ashiqur KhudaBukhsh. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *EMNLP*, 2023.

[Wu *et al.*, 2021] Yang Wu, Yanyan Zhao, Xin Lu, Bing Qin, Yin Wu, Jian Sheng, and Jinlong Li. Modeling incongruity between modalities for multimodal sarcasm detection. *IEEE MultiMedia*, 28(2):86–95, 2021.

[Xu *et al.*, 2020] Nan Xu, Zhixiong Zeng, and Wenji Mao. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *ACL*, 2020.

[Yang *et al.*, 2023] Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. Mm-bigbench: Evaluating multimodal models on multimodal content comprehension tasks. *arXiv preprint arXiv:2310.09036*, 2023.