Predict and Use: Harnessing Predicted Gaze to Improve Multimodal Sarcasm Detection

Divyank Pratap Tiwari[1], Diptesh Kanojia[2], Anupama Ray[3],
Apoorva Nunna[1], Pushpak Bhattacharyya[1]
[1]Computing for Indian Language Technology, IIT Bombay, India.
[2]Surrey Institute for People-Centred AI, University of Surrey, United Kingdom.
[3]IBM Research India
[1]{213050029, 203050028, pb}@iitb.ac.in
[2]d.kanojia@surrey.ac.uk, [3]anupamar@in.ibm.com

Abstract

Sarcasm is a complex linguistic construct with incongruity at its very core. Detecting sarcasm depends on the actual content spoken and tonality, facial expressions, the context of an utterance, and personal traits like language proficiency and cognitive capabilities. In this paper, we propose the utilization of synthetic gaze data to improve the task performance for *multimodal sarcasm detection* in a conversational setting. We enrich an existing multimodal conversational dataset, *i.e.,* MUStARD++ with gaze features. With the help of human participants, we collect gaze features for $<$ 20% of data instances, and we investigate various methods for gaze feature prediction for the rest of the dataset. We perform extrinsic and intrinsic evaluations to assess the quality of the predicted gaze features. We observe a performance gain of up to 6.6% points by adding a new modality, *i.e.,* collected gaze features. When both collected and predicted data are used, we observe a performance gain of 2.3% points on the complete dataset. Interestingly, with *only* predicted gaze features, too, we observe a gain in performance (1.9% points). We retain and use the feature prediction model, which maximally correlates with collected gaze features. Our model trained on combining collected and synthetic gaze data achieves SoTA performance on the MUStARD++ dataset. To the best of our knowledge, ours is the first predict-and-use model for sarcasm detection. We publicly release[1] the code, gaze data, and our best models for further research.

1 Introduction

Sarcasm originates from the Greek word *sarkasmós* adapted from *sarkázein*, which means a sneering or cutting remark. Sarcasm depends on "bitter, caustic, and other ironic expressions that are usually directed against an individual." (Gibbs, 1986). It is a complex linguistic phenomenon that gets expressed with words that mean the opposite of what the speaker intends to say; *e.g., I love being ignored* expresses the bitterness of the speaker. The roots of sarcasm lie in *incongruity* (Joshi et al., 2015), which makes computational sarcasm detection a challenging problem; and the NLP community has attempted to tackle this problem using innovative approaches. Sarcasm detection in the text has largely been attempted by focusing on lexical indicators (Bamman and Smith, 2021), sentiment incongruity (Joshi et al., 2015), *etc.,* in both rule-based and learning-based systems (Abulaish and Kamal, 2018). However, sarcasm is also expressed through tonal changes and/or facial expressions. Hence researchers have started investigating modalities other than text, *viz.,* audio and video, to help detect sarcasm (Castro et al., 2019a; Cai et al., 2019; Gupta et al., 2021; Chauhan et al., 2022; Ray et al., 2022). Mishra et al. (2017a) observed that gaze features are helpful in detecting sarcasm within short sentences without context, which is our inspiration. In a conversational setting, *sarcasm often results from an earlier utterance*, which is the problem we focus on in this work. To the best of our knowledge, ours is the first attempt at multimodal detection of sarcasm using gaze behaviour in a conversational setting.

1.1 Gaze Terminology

A **fixation** is a relatively longer stay of gaze on an object (word), and **saccades** refer to quick shifting of gaze between two positions of rest (Mishra et al., 2017b). An Interest Area (IA) is a part of the screen that is of interest to us. In these areas, the text is displayed and *each word is a separate and unique IA*. Forward and backward saccades are called **progressions** and **regressions**, respectively, while a **scanpath** is a line graph that contains fixations as nodes and saccades as edges.

We use the MUStARD++ dataset (Ray et al.,

[1] https://www.cfilt.iitb.ac.in/emnlp23sarcgaze

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15933–15948
December 6-10, 2023 ©2023 Association for Computational Linguistics

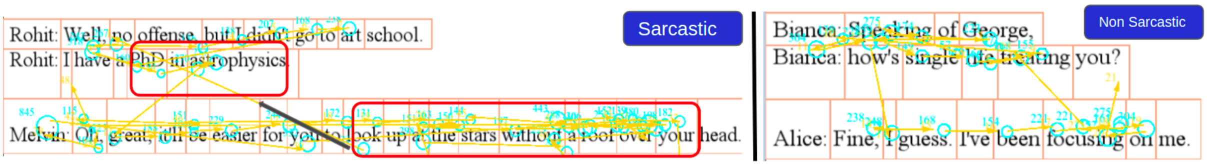

Rohit: Well, no offense, but I didn't go to art school.
Rohit: I have a PhD in astrophysics
Melvin: Oh, great, it'll be easier for you to look up at the stars without a roof over your head.
Sarcastic
Bianca: Speaking of George,
Bianca: how's single life treating you?
Alice: Fine, I guess. I've been focusing on me.
Non Sarcastic

Figure 1: Sample images from a Gaze data collection setup which shows saccadic movements (yellow lines) and fixations (blue circles) for 1) a sarcastic (left image) and 2) a non-sarcastic dialogue (right image).

2022) which is a multimodal conversational dataset with videos annotated for sarcasm, sarcasm type and emotions. This data has several video frames as visuals linked with the utterance that is marked sarcastic or non-sarcastic Our primary hypothesis in this work is that there are distinctive eye movement patterns when a human reader is processing sarcasm due to the presence of incongruous words within the utterance or previously spoken sentences (Mishra et al., 2016b). Unlike previous studies, we perform the task of sarcasm detection in a conversational setting, exploiting multimodality and gaze features. Figure 1 illustrates gaze fixations (blue circles w/ bigger circles for longer duration) and progressions-regressions for a sarcastic, and a non-sarcastic utterance.

Gaze features, however, are costly in terms of resources- subjects, data, time, and money. One major contributions of our work is predicting gaze features and harnessing predicted features for sarcasm detection. We thus venture into generating and using synthetic data for sarcasm detection. Overall, our Contributions are:

- A novel method for generating synthetic data from collected gaze features.
- Enriching the MUStARD++ dataset with eye-tracking/gaze features for 1155 samples collected from 5 human participants. This will be useful for research in eye-tracking-based sarcasm and similar language phenomena detection.
- Comparing various gaze feature prediction techniques and utilizing gaze data, both collected and synthetic, to achieve SoTA performance (2.3% point gain) for multimodal sarcasm detection.

1.2 Motivation

From Figure 1, it can be observed that the non-sarcastic utterance has a significantly lower regressive eye movement (yellow lines) as compared to the sarcastic utterance. The number of fixations is also lower in number. In the sarcastic utterance, we see a lot of regression on the part of the text containing “look up at the stars without a roof over your”, we also observe regressive movement towards the previous utterance in the context- towards“PhD in astrophysics”. Such indicators can also be used to explain the origin of sarcasm from a conversational context. However, we observe that the non-sarcastic example (right) also has a few regressive paths leading to previous utterances, which will happen for any reader, given they would like to understand the context in the dialogue fully. We believe capturing these regressions and progressions present in gaze data can help detect sarcasm and generate similar gaze data for new samples, as fixations, movements, and regressions can be learned from them. We also believe the creation of quality synthetic eye-tracking data will be useful in reducing dependency on highly time-consuming human eye-tracking annotations.

2 Related Work

Existing studies demonstrate how cognitive features have been used to improve performance for various NLP tasks. User understandability of sarcasm can be evaluated with the help of gaze behaviour (Mishra et al., 2016a), where incongruity in the text induces gaze behaviour characterized by longer fixation durations, repeated regressions, and also scan path complexity (Mishra et al., 2017b). Previously, sarcasm detection based on only textual input has shown minor improvements with the help of gaze-based features (Mishra et al., 2016b, 2017a). Gaze behaviour has also been used to identify a reader’s native language (Berzak et al., 2017), as well as to detect grammatical errors in compressed sentences (Klerke et al., 2015a, 2016). Klerke et al. (2015b) also show that gaze behaviour can be used to evaluate the output of Machine Translation systems better than automated metrics. Similarly, gaze-based features have also been shown to help the task of cognate and false friends' detection (Kanojia et al., 2021). Gaze behaviour has also been used to evaluate how a reader would rate the quality of a piece of text (Mathias et al., 2018). Similarly, Mathias et al. (2020b) also perform the task of essay grading in a zero-shot setting using only gaze-based features and show the efficacy of gaze-based features for performing NLP tasks (Mathias et al., 2020a). However, existing research does not discuss the correlation of multimodal features (like visual and audio) with gaze-based features, and does not investigate these features for multimodal sarcasm detection in a conversational setting. In the subsection below, we discuss the literature on multimodal studies in NLP.

Lack of data has been a common problem in cases of both sarcasm as well as cognitive NLP. Numerous efforts have been made in building gaze feature predictors in order to reduce dependency on gold gaze data by producing high quality synthetic gaze data. Study in Takmaz (2022) utilizes "adapter" in a language model to match the results of a fully fine tuned language model for predicting eye tracking features with a highly efficient network in terms of the number of parameters. Ding et al. (2022) propose a Bi-LSTM-based network that, with the help of a few psycho-linguistic features, predicts eye tracking features. The paper states that the readability of a text reflected in the linguistic features is important to predict eye movement patterns (Scarborough et al., 2009). The creation of synthetic gaze data has also been performed in multilingual settings. In Srivastava (2022), a model trained on a completely different set of languages predicts gaze data for a completely new language.

2.1 Multimodal NLP

Existing literature on multimodal sentiment classification refers to the MOUD (Pérez-Rosas et al., 2013) and MOSI (Zadeh et al., 2016) datasets and the IEMOCAP dataset (Busso et al., 2008) for the task of multimodal emotion recognition. Poria et al. (2017) propose the use of a bidirectional contextual long short-term memory (bc-LSTM) architecture for both tasks and show improvements over baseline on all three datasets. However, Majumder et al. (2018) later propose context modelling with a hierarchical fusion of multimodal features and achieve improved performance in a monologue setting. In the conversation setting, Hazarika et al. (2018) propose using a Conversational Memory Network (CMN) to leverage contextual information from the conversation history and achieve improved performance. Novel multimodal neural architectures (Wang et al., 2019; Pham et al., 2019) and multimodal fusion approach (Liang et al., 2018; Tsai et al., 2018) have propelled the deployment of computational models. Efficient multimodal fusion approaches have also been discussed in (Sahay et al., 2020; Tsai et al., 2019; Liu et al., 2018).

For multimodal sarcasm detection, a recent survey discusses the datasets and approaches in detail (Bhat and Chauhan, 2022). The MUStARD dataset (Castro et al., 2019b) provides clips compiled from popular TV shows, including Friends, The Golden Girls, The Big Bang Theory, and Sarcasmaholics Anonymous, annotated with sarcasm labels. Ray et al. (2022) extend upon this dataset by adding emotion labels and additional clips while also benchmarking for the multimodal sarcasm detection task. They call this extended dataset *MUStARD++* and utilise feature fusion and a feed-forward network to predict the sarcasm label. The authors show an F1-score of 70.2% points using audio, text and video modalities.

Our work utilises a similar approach with the additional gaze modality and also reproduces the baseline experiments. With this work, we aim to underpin how gaze-based features perform in a multimodal setting and if they correlate well with feature sets other than textual (visual and audio). We also investigate predicting gaze-based features to save annotation time/cost for multimodal studies.

3 Dataset and Gaze Annotation

MUStARD++ is a multimodal dataset that consists of textual utterances with context, audio, and video from a corresponding clip. This data has been acquired from publicly available sources for five television shows: Friends, The Big Bang Theory (seasons 1–8), The Golden Girls, Burnistoun, and The Silicon Valley. Each dialogue is presented as a combination of the main 'utterance' and the 'context' in which it was uttered. It contains a total of 1,202 instances, out of which 601 are sarcastic, and 601 are non-sarcastic. Along with sarcasm annotation, the dataset also provides additional information like an emotion class, valence, arousal, and sarcasm type. We chose this dataset for our experiments and performed gaze annotation on 231 samples, where 129 are sarcastic, and 102 are non-sarcastic. To avoid any skew, the sarcastic instances are chosen to encompass all four types of sarcasm

	Average Fixation Duration			IA Regression Path Duration		
	μ_Pos $\pm$ σ_Pos	μ_Neg $\pm$ σ_Neg	p	μ_Pos $\pm$ σ_Pos	μ_Neg $\pm$ σ_Neg	p
P1	208.0 ± 15.1	217.8 ± 13.7	0.0011	657.3 ± 305.3	495.4 ± 190.7	0.0140
P2	209.6 ± 16.3	224.6 ± 27.5	0.0147	572.5 ± 232.2	466.2 ± 221.0	0.0274
P3	241.6 ± 14.0	253.6 ± 21.1	0.0124	638.2 ± 130.8	502.0 ± 102.1	0.0001
P4	252.1 ± 10.4	241.2 ± 11.9	0.0001	727.4 ± 269.2	568.5 ± 160.1	0.0030
P5	212.6 ± 17.9	226.7 ± 16.2	0.0084	952.9 ± 280.3	696.3 ± 218.5	0.0002

Table 1: Two-sampled T-test statistics for average fixation duration and interest area regression path duration for Positive labels (Sarcastic) and Negative labels (Non-sarcastic) for participants P1-P5.

with a distribution similar to the one in the source data from MUStARD++. The selected instances include dialogues with short contexts (in the range of 2-5 speaker turns) as well as long contexts (6-13 speaker turns).

3.1 Datasets

For our multimodal sarcasm detection experiments, we now have three variants as datasets. The first variant is the complete dataset from MUStARD++, *i.e.,* **D1**. Since we are only able to acquire gaze data over 231 samples out of 1202 as discussed above, **D2 is the other variant**, which consists of a total of 1,155 data instances (231 samples x 5 participants). Please note that textual, audio and video features for the 231 samples remain the same while *gaze features vary for each participant in this D2 variant*. For the portion of samples we do not get manually collected gaze data, we choose to predict the gaze tracking features as described below (Section 4.1), and **call it D3**. This variant, *i.e.,* **D3** consists of 971 samples in total. We show the train/test split statistics in Table 2. We also try to maintain a balance between various types of sarcasm in these instances, the distribution details of which are provided in Figure 2; and we provide the details of the gaze annotation process below.

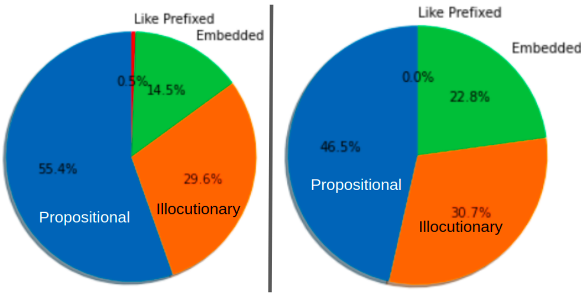


Figure 2: Sarcasm-type distribution from **D1** (left) and **D2** (right) datasets.

3.2 Gaze Annotation

We instruct five annotators to read the 'textual utterances with its context' on the screen and ask them to provide annotations for the *implied binary sentiment* in the dialogue, *i.e.,* positive or negative. These samples are shuffled, and the experiment builder software is allowed to choose a random instance from the 231 samples to be presented next on the screen. We do not instruct the annotators to look for sarcasm to avoid the Priming Effect, *i.e.,* if sarcasm is expected beforehand, it becomes easier to process. It may have resulted in unattentive participation by annotators (Sáchez-Casas et al., 1992). It ensures the ecological validity of our experiment as 1) the participant has no clue which utterance to expect, and no special attention is paid to either class from the instances, and 2) it also ensures attentive participation. Our annotators are graduate students between the ages of 22-27 with good proficiency in the English language. Annotator selection was made after ensuring they had English as the medium of instruction through undergraduate and their ongoing post-graduate degree program. We ensure that they consent to record their eye movement pattern to be used for this research.

We provide two unrecorded samples at the start of the experiment to acquaint them with the annotation process. While annotating for sentiment over 231 samples, we provide our annotators with a short break after every 30 samples to ensure minimal annotator fatigue, and re-calibrate for their eye movements after each break. The head movement was minimised using a chin-rest during the annotation process. The gaze tracking device used is an SR-Research Eyelink-1000 (monocular remote mode with a sampling rate of 500Hz) that captures the eye movement of the reader/annotator.

3.3 Annotation & Feature Validity

We compute **inter-annotator agreement** using a pair-wise Fleiss' kappa (Scott, 1955), which *re-*

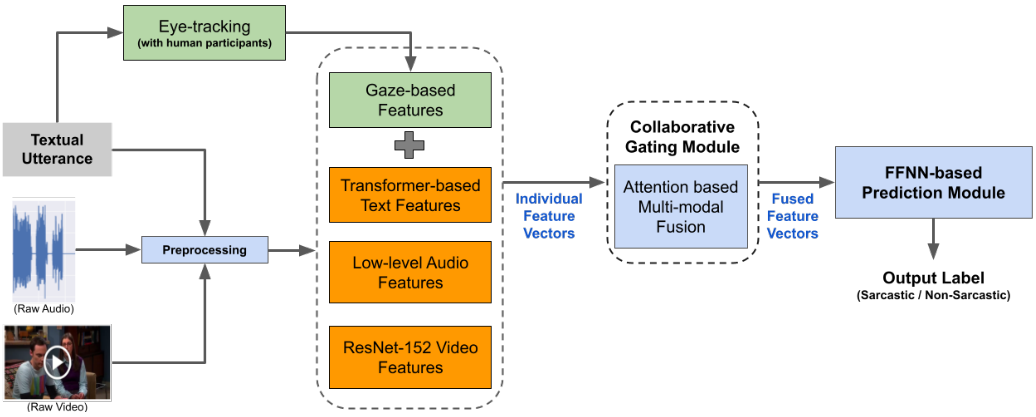

Eye-tracking
(with human participants)
Textual Utterance
(Raw Audio)
(Raw Video)
Preprocessing
Gaze-based Features
Transformer-based Text Features
Low-level Audio Features
ResNet-152 Video Features
Individual Feature Vectors
Collaborative Gating Module
Attention based Multi-modal Fusion
Fused Feature Vectors
FFNN-based Prediction Module
Output Label
(Sarcastic / Non-Sarcastic)

Figure 3: The **architecture diagram** for multimodal sarcasm detection setup shows how *gaze-based features are introduced* in the pipeline with other features (*text/audio/video*) while a *collaborative gating mechanism fuses these features* for predicting labels via a feed-forward neural network.

sulted in a statistically significant (p<0.05) moderate agreement (0.41) among our annotators. To validate features for our experiment, we chose a standard gaze-based feature and a saccadic regression-based feature, *i.e.,* average fixation duration and interest area regression path duration (Table 8), respectively. In Table 1, we show the analysis from a two-sampled t-test over feature data from each participant. We observe that for each participant (P1-P5), the difference between sarcastic and non-sarcastic instances is statistically significant, which further motivates us to use these features for sarcasm detection/classification.

Variant	Train	Test	Total
D1	947	255	1202
D2	172	59	231
D3	776	195	971

Table 2: Train/test split statistics for dataset variants.

4 Our Approach

An architecture diagram for our setup is shown in Figure 3. We reproduce the multimodal sarcasm detection experiments as in Ray et al. (2022) as the baseline, with the addition of *gaze features as the fourth modality* in addition to text, speech and visual modalities. For **textual features**, we utilize the pre-trained BART language model (Lewis et al., 2019) and obtain *embeddings for both textual utterance and the context* from the dialogue. BART provides a feature vector representation $\mathbf{x}_t \in \mathbb{R}^{d_t}$ for every instance $\mathbf{x}$. We encode the text using the BART Large model with $d_t = \mathit{1024}$ and use the mean of the last four transformer layer representations to get a unique embedding representation for both the utterance and the context.

For **audio features**, like in Ray et al. (2022), we sampled the audio signal at 22.5KHz as a pre-processing step. Since the audio has background noise and canned laughter (we deal with sitcoms), we used the vocal-separation method[2] to process it. We extract three low-level features: *Mel Frequency Cepstral Coefficients (MFCC), Mel spectrogram (using Librosa library (McFee et al., 2022))*, and *prosodic features using OpenSMILE*[3]. We split the audio signal into equal segments of 1-second duration each to maintain consistent feature representation in all instances. Since the audio signal length varies with utterances, this segmentation helps in keeping the vector size constant across the dataset. For each segment, we extract MFCC, Mel spectrogram and prosodic features of size d_m, d_s, d_p respectively. Then we take the average across segments to get the final feature vector. Here $d_m = \mathit{128}$, $d_s = \mathit{128}$, $d_p = \mathit{35}$, so our audio feature vector is of size $d_a = d_m + d_s + d_p = \mathit{291}$.

For **visual features from the videos**, we use a pool-5 layer from pre-trained ResNet-152 (He et al., 2016) image classification model. To improve the video representation and reduce noise, we extract the keyframes to be passed to ResNet-152. The computer vision community widely uses key frame extraction, which is defined as the frames that form the most appropriate summary of a given video (Jadon and Jasim, 2019). We use an open-source tool, Katna[4], to perform key-frame extraction. For the final feature vectors, we average the

[2] https://librosa.org/doc/main/auto_examples/plot_vocal_separation.html
[3] https://audeering.github.io/opensmile/
[4] https://katna.readthedocs.io/en/latest/

vectors of each key frame of an instance (context and utterance) extracted from ResNet-152. The size of the final video feature representation is $d_v = 2048$.

For **gaze-based features**, we obtain a total of 31 features from the SR Research Experiment DataViewer software[5], but do not use them all for our experiments. We employ the KBest feature selection method from the *scikit-learn* library (Pedregosa et al., 2011) to optimize the features. Given the sarcasm label along with gaze-based features, the method resulted in the selection of **a total of 25 correlated gaze-based features**. Due to space constraints, we provide this list in Appendix C with a feature description in Table 8.

4.1 Gaze Feature Prediction

We provide details of our gaze feature prediction model here. To remind, we predict 25 gaze-based features, but, to report the correlations between the actually collected gaze features and the predicted gaze features, we choose the three most important features*, i.e.,* the *average fixation duration*, the *regression path duration*, and the *regression count*. In a sarcastic utterance, if the human eye rests on the text for a longer duration, it signifies the presence of some incongruity, which may be because of sarcasm; this phenomenon is captured in the feature named average fixation duration. Similarly, suppose the eye regresses back to the context again and again. In that case, this depicts that there is some difference between the surface meaning and the deep meaning of the utterance that is creating complexity in understanding. This we capture in the regression features. Since collecting gaze data requires human effort and is costly in terms of time and money, *we try to predict these gaze-based features for our D3 variant of the MUStARD++ dataset* (see 3.1 on datasets for an explanation of D3), containing 971 samples. We used SVM (Cortes and Vapnik, 1995) as well as feed-forward NN (FFNN) (Bebis and Georgiopoulos, 1994), convolutional NN (CNN) (O'Shea and Nash, 2015), RoBERTa (Liu et al., 2019b), and adapters-infused RoBERTa (Pfeiffer et al., 2020) to compare the quality of the gaze features predicted using these techniques. We evaluate the quality of the gaze features by calculating two different correlation metrics, *i.e.,* Pearson correlation (Freedman et al., 2007) and Spearman correlation (Spearman, 1904) between the predicted gaze and actual collected gaze feature values for a specific set of samples. These correlation coefficients are given in Table 9, 10 and 11 in **Appendix D**. Our best-performing model for the gaze features prediction task is a feed-forward neural network-based architecture. The BART encoding for the text utterances (discussed in section 4) is used as input to the FFNN in a matrix form. The architecture constitutes 3 hidden layers, an input layer having 1024 nodes (equal to the size of text embedding). The 4 layers are followed by fully connected layers with a single output node, the gaze feature values being the ground truth labels for the prediction task.

4.2 Experimental Setup

To perform feature fusion, we use a collaborative gating mechanism (Liu et al., 2019a) on all the features described above. We first compute projection $\Psi^{(i)}(V)$ where $i \in \{t, tc, a, ac, v, vc, g\}$ and t, a, v, c, g are *text, audio, video, corresponding context, and gaze*. This mechanism implements two tasks, 1) finding the attention vector prediction over provided input vectors and 2) performing expert response modulation using the computed attention vector prediction. For response modulation of each modality projection, we perform

$$\Psi^{(i)}(V) = \Psi^{(i)}(V) \circ \sigma(T^{(i)}(V)) \qquad (1)$$

where σ is an element-wise *sigmoid* activation and $\circ$ is the element-wise multiplication, *i.e.,* Hadamard product (Horn, 1990). These modulated projections are then concatenated and passed to fully connected linear layers (ReLU) followed by a *softmax* layer to predict target class probability distribution. We use the standard cross-entropy loss for sarcasm detection/classification. For training, we perform hyper-parameter search with dropout in range of [0.1, 0.2, 0.3, 0.4, 0.5, 0.6], learning rate in [0.001, 0.0001], batch size [256, 384], shared embedding size [2048, 1024] and the projection embedding size [1024, 256].

For benchmarking our results, we train over three iterations with different (randomly chosen) seed values of 42, 200, and 1005. The seed values ensure that the dataset samples are randomly shuffled. Our experiments were performed using a single nVidia RTX A6000, where each iteration takes approximately 2 hours. We report the mean ± standard deviation on the test splits in terms of

[5] https://www.sr-research.com/experiment-builder/

macro precision, recall and F1-scores. We note that the quality of predicted gaze features plays a key role in the performance of our multimodal sarcasm detection task, as described in the section below.

5 Results and Discussion

Modality	macro-P	macro-R	macro-F1
Vid + Aud + Text + Gaze	0.816 ± 0.008	0.814 ± 0.009	0.815 ± 0.009
Vid + Text + Gaze	0.832 ± 0.013	0.832 ± 0.012	0.832 ± 0.013
Vid + Aud + Gaze	0.792 ± 0.011	0.784 ± 0.010	0.785 ± 0.011
Aud + Text + Gaze	0.802 ± 0.008	0.801 ± 0.007	0.802 ± 0.008
Vid + Gaze	0.864 ± 0.014	0.862 ± 0.014	0.862 ± 0.014
Aud + Gaze	0.774 ± 0.002	0.768 ± 0.004	0.769 ± 0.004
Text + Gaze	0.830 ± 0.013	0.829 ± 0.012	0.829 ± 0.012
Vid + Aud + Text	0.75 ± 0.008	0.749 ± 0.009	0.749 ± 0.009
Vid + Text	0.731 ± 0.013	0.725 ± 0.012	0.725 ± 0.013
Vid + Aud	0.688 ± 0.011	0.683 ± 0.010	0.684 ± 0.011
Aud + Text	0.732 ± 0.008	0.732 ± 0.007	0.732 ± 0.008
Vid	0.654 ± 0.014	0.655 ± 0.014	0.655 ± 0.014
Aud	0.588 ± 0.002	0.585 ± 0.004	0.586 ± 0.004
Text	0.718 ± 0.013	0.718 ± 0.012	0.718 ± 0.012

Table 3: Results obtained on the **D2 dataset** via experiments with **Video (Vid), Audio (Aud), Textual (Text), and Gaze-based (Gaze)** features where macro-P/R/F1 are macro Precision, Recall and F1-score, respectively. These results are compared with results obtained on same dataset without using gaze features for the sarcasm detection task.

We first describe the results on the D2 variant in Table 3. This table reports the results while ablating on various feature combinations. A clear performance improvement can be observed when the gaze modality is added. Compared to the baseline, *i.e.,* all modalities except gaze, there is a significant gain of 6.6% points when all features, including gaze, are used for sarcasm detection. We do reproduce the baseline experiments on the D1 variant and show the results in Table 4, where we also observe a slight gain (0.8% points) on the best baseline score provided by MUStARD++ authors when they used the standard video, audio and textual modalities. We compare the results on D1 from baseline experiments with combinations of video, audio, and textual modalities with our experiments on D1 with the same feature combinations.

We observe from Table 4, that adding predicted gaze-based features lowers the performance benchmark significantly. However, a combination of PredGaze and Gaze added to each combination still outperforms all the baseline combinations with up to 6.8% point gain by using only video, PredGaze, and Gaze (highlighted with underlined in the table). On using all the available features, we still outperform the baseline scores by 2.3% points (highlighted using bold). We also observe that by only using predicted and collected gaze-based (PredGaze+Gaze) features, we are able to obtain an F1-score of 0.679, which is encouraging. It shows that our model can predict sarcasm with some certainty even without using standard modalities like video, audio, or text, only on the basis of eye movement patterns. This, however, encouraged us to probe the efficacy of the predicted gaze features.

Reproduced **Baseline Experiments on D1** *without gaze*			
Modality	**macro-P**	**macro-R**	**macro-F1**
Vid + Aud + Text (*baseline*)	0.710 ± 0.006	0.710 ± 0.006	*0.710 ± 0.006*
Vid + Text	0.693 ± 0.002	0.693 ± 0.003	0.693 ± 0.003
Vid + Aud	0.688 ± 0.029	0.688 ± 0.028	0.688 ± 0.028
Aud + Text	0.674 ± 0.031	0.674 ± 0.031	0.674 ± 0.031
Vid	0.591 ± 0.004	0.591 ± 0.004	0.591 ± 0.004
Aud	0.645 ± 0.005	0.645 ± 0.005	0.645 ± 0.005
Text	0.690 ± 0.003	0.690 ± 0.003	0.690 ± 0.003
Our Experiments on D1 *with collected & predicted gaze*			
Vid + Aud + Text + PredGaze + Gaze	0.732 ± 0.001	0.732 ± 0.001	**0.733 ± 0.001**
Vid + Text + PredGaze + Gaze	0.718 ± 0.001	0.718 ± 0.001	0.718 ± 0.001
Vid + Aud + PredGaze + Gaze	0.712 ± 0.006	0.709 ± 0.006	0.711 ± 0.006
Aud + Text + PredGaze + Gaze	0.724 ± 0.001	0.723 ± 0.001	0.724 ± 0.001
Vid + PredGaze + Gaze	0.662 ± 0.006	0.660 ± 0.005	0.659 ± 0.004
Aud + PredGaze + Gaze	0.695 ± 0.004	0.695 ± 0.004	0.694 ± 0.004
Text + PredGaze + Gaze	0.721 ± 0.003	0.722 ± 0.003	0.722 ± 0.003
PredGaze + Gaze	0.680 ± 0.003	0.679 ± 0.004	0.679 ± 0.004

Table 4: Results obtained on the **D1 dataset** via baseline experiments with **Video (Vid), Audio (Aud), and Textual (Text), Collected Gaze-based (Gaze) and Predicted Gaze-based (PredGaze)** features where macro-P/R/F1 are macro Precision, Recall and F1-score, respectively. The best F1-score achieved on D1 is highlighted in **bold**.

Our Experiments on the D3 variant, where only predicted gaze-based features are available, also show an improvement of at most 1.9% points when video, audio and text are the modalities used along with predicted gaze features. We compare the results of sarcasm detection on this D3 dataset with and without gaze features in Table 6, accuracies without the use of gaze features being the baseline for this experiment. When used along with other modalities, *i.e.,* video, text and audio, the model was able to beat the corresponding baseline by 1.9% points which is encouraging as this improvement came from complete synthetic data. We report the accuracies from the best performing model among the models we experimented with.

5.1 Discussion

Upon looking at the results from our task while ablating for feature combinations, we also perform

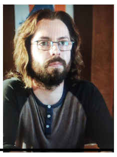
Final Utterance	**Ground Truth**	**Prediction (w/o Gaze)**	**Prediction (with Gaze)**	**Sarcasm Type**	**Video Frame**
BERNADETTE: And I love that I work and do all the cleaning, and you're okay with that	Sarcastic	Non-Sarcastic	Sarcastic	Propositional	
CHANDLER Oh, uh, no thanks. I just had an M&M.	Sarcastic	Non-Sarcastic	Sarcastic	Illocutionary	
Moderator: Sarcastic? Us? Nooo.	Sarcastic	Non-Sarcastic	Non-Sarcastic	Illocutionary	
Gilfoyle: But the fact that you're so sorry makes it all better.	Sarcastic	Non-Sarcastic	Non-Sarcastic	Embedded	

Table 5: **Qualitative analysis**, which presents instances from the dataset and their sarcasm label predictions with and without the presence of gaze in the input.

Modality	**macro-P**	**macro-R**	**macro-F1**
Vid + Aud + Text + PredGaze	**0.699** ± **0.007**	**0.699**± **0.007**	**0.700** ± **0.007**
Vid + Text + PredGaze	0.655 ± 0.001	0.655 ± 0.001	0.655 ± 0.001
Vid + Aud + PredGaze	0.658 ± 0.010	0.658 ± 0.009	0.658 ± 0.009
Aud + Text + PredGaze	0.689 ± 0.004	0.688 ± 0.004	0.689 ± 0.004
Vid + PredGaze	0.611 ± 0.007	0.612 ± 0.007	0.612 ± 0.008
Aud + PredGaze	0.659 ± 0.001	0.659 ± 0.001	0.659 ± 0.001
Text + PredGaze	0.696 ± 0.004	0.697 ± 0.004	0.697 ± 0.004
PredGaze	0.583 ± 0.004	0.583 ± 0.004	0.583 ± 0.004
Vid + Aud + Text	0.681 ± 0.007	0.683 ± 0.007	0.683 ± 0.007
Vid + Text	0.642 ± 0.001	0.641 ± 0.001	0.641 ± 0.001
Vid + Aud	0.648 ± 0.010	0.648 ± 0.009	0.648 ± 0.009
Aud + Text	0.678 ± 0.004	0.678 ± 0.004	0.678 ± 0.004
Vid	0.584 ± 0.007	0.582 ± 0.007	0.581 ± 0.008
Aud	0.656 ± 0.001	0.655 ± 0.001	0.655 ± 0.001
Text	0.66 ± 0.004	0.658 ± 0.004	0.658 ± 0.004

Table 6: Results obtained on the **D3 dataset** via experiments with **Video (Vid), Audio (Aud), Textual (Text), and Predicted Gaze-based (PredGaze)** features where macro-P/R/F1 are macro Precision, Recall, and F1-score. These results are compared with results obtained on the same dataset without using gaze features for the sarcasm detection task.

a qualitative analysis of data samples which encourages the use of gaze as a modality. We randomly choose two samples which are shown at the top in Table 5. The first two samples (rows one and two) show the final textual utterance in the first column, which are examples of 'Propositional' and 'Illocutionary' sarcasm. Starting with a positive and ending with an implied negative sentiment has incongruity, which should be detected with the help of computational models. However, the label for this sample was *incorrectly predicted by all feature combinations*, and *only after introducing gaze-based features*, the model correctly predicts the label. We also show a representative image from the video clip during the final utterance. The image is only a single frame, but our analysis of the clip shows that the character maintains a similar expression throughout the utterance except for the final few frames when she utters the part, 'with that'. We believe that selecting key frames from the video using an automated method may not be very effective in such a case; therefore, even the visual features are unable to help.

However, we have other examples which are still hard to label accurately. The last two rows in Table 5 show two examples where the first is a sarcastic utterance of the illocutionary type. Ironically, the word 'Sarcastic' is present within the utterance itself. Given the full context of this utterance and, on observing the clip, audio-based features should have helped in this scenario. The moderator utters a very grumpy 'Nooo.' and high-pitched 'Sarcastic? Us?'. We believe that the low-level audio features are not able to capture these tonal changes, but the failure of other feature sets also begets further investigation. Also, the video clip shows the scene to be focused on the whole group, and there is a lesser chance of detecting key frames for visual features to capture subtle changes. Due to space constraints, we provide full context and utterance for all these samples in another table present in the Appendix B, *i.e.,* Table 7.

Similarly, in the last example, despite sufficient conversational context (can be seen from Table 7), neither of the computational models is able to capture the sarcasm present in the final utterance. There is sufficient incongruity present in this sample instance of the embedded type of sarcasm. However, the final utterance, “But the fact that you’re so sorry makes it all better” can come across as an emotional and caring statement.

6 Conclusion and Future Work

This paper discussed the use of gaze-based features for the task of sarcasm detection in a multimodal and conversational setting. We propose the use of textual, audio, and video in combination with the gaze modality by showing a substantial improvement in performance with the addition of collected gaze-based features. We collect gaze data over a small number of samples and predict these features for a larger portion of the data, both of which we will release with the code and the best models from our experiments. With predicted gaze-based features, however, we observe a small improvement in the task performance in this case. To the best of our knowledge, our results indicate that adding collected gaze-based features certainly improves task performance in every feature combination, proving the efficacy of gaze-based features. Our qualitative analysis also suggests that better audio and visual features should help improve task performance.

In future, we would like to improve the quality of predicted gaze-based features further in a multi-task setting of sarcasm detection and gaze prediction.

Limitations

Our work has certain limitations, as gaze data collection is challenging. Multimodal datasets are also scarce, and it’s challenging to benchmark the performance of this approach over multiple datasets. We release the complete gaze data with annotator-provided sentiment labels, but our inter-annotator agreement is only moderate. The subjectivity of sarcasm and cultural contexts present in humour, are the key reasons for the inter-annotator agreement value being lower than expected. The understanding of sarcasm varies from person to person depending upon the age, culture, context, familiarity with the characteristics present in the utterance, *etc.* This makes sarcasm a very hard and cognitively loaded phenomenon for even linguists to annotate. Collection of eye-tracking/gaze data is a tedious and costly process; it requires hours of human participation without any loss of concentration of the annotator. Transformers-based models, in the case of video, audio, as well as text, require large amounts of data to be able to generalise and perform well. Thus, dataset contribution becomes essential to push boundaries and enable more research in the field.

Ethics Statement

MUStARD++ used in our experiments is ethically verified in the previous works that used the dataset (Ray et al., 2022; Castro et al., 2019b). We took consent from all 5 annotators for the gaze annotations, which involved tracking the participant’s eye while they read the text displayed on a screen. We also pay the annotators for their time and efforts in the annotation.

References

Muhammad Abulaish and Ashraf Kamal. 2018. Self-deprecating sarcasm detection: An amalgamation of rule-based and machine learning approach. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 574–579.

David Bamman and Noah Smith. 2021. Contextualized sarcasm detection on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):574–577.

George Bebis and Michael Georgiopoulos. 1994. Feed-forward neural networks. *Ieee Potentials*, 13(4):27–31.

Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. 2017. Predicting native language from gaze. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 541–551.

Aruna Bhat and Aditya Chauhan. 2022. Multimodal sarcasm detection: A survey. In *2022 IEEE Delhi Section Conference (DELCON)*, pages 1–7.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019a. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019b. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

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

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Xiao Ding, Bowen Chen, Li Du, Bing Qin, and Ting Liu. 2022. CogBERT: Cognition-guided pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3210–3225, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Raymond W Gibbs. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology: General*, 115(1):3.

Sundesh Gupta, Aditya Shah, Miten Shah, Laribok Syiemlieh, and Chandresh Maurya. 2021. Filming multimodal sarcasm detection with attention. In *International Conference on Neural Information Processing*, pages 178–186. Springer.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Roger A Horn. 1990. The hadamard product. In *Proc. Symp. Appl. Math*, volume 40, pages 87–169.

Shruti Jadon and Mahmood Jasim. 2019. Video summarization using keyframe extraction and video skimming. *arXiv preprint arXiv:1910.04792*.

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)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Diptesh Kanojia, Prashant Sharma, Sayali Ghodekar, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2021. Cognition-aware cognate detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3281–3292, Online. Association for Computational Linguistics.

Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard. 2015a. Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 97–105.

Sigrid Klerke, Sheila Castilho, Maria Barrett, and Anders Søgaard. 2015b. Reading metrics for estimating task efficiency with mt output. In *Proceedings of the sixth workshop on cognitive aspects of computational language learning*, pages 6–13.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1528–1533.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal language analysis with recurrent multistage fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 150–161.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019a. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia. Association for Computational Linguistics.

Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, 161:124–133.

Sandeep Mathias, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharya. 2020a. A survey on using gaze behaviour for natural language processing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4907–4913. International Joint Conferences on Artificial Intelligence Organization. Survey track.

Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Eyes are the windows to the soul: Predicting the rating of text quality using gaze behaviour. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2352–2362.

Sandeep Mathias, Rudra Murthy, Diptesh Kanojia, and Pushpak Bhattacharyya. 2020b. Cognitively aided zero-shot automatic essay grading. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 175–180, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Brian McFee, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Dan Ellis, Jack Mason, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, viktorandreevichmorozov, Keunwoo Choi, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Adam Weiss, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Matt Vollrath, Taewoon Kim, and Thassilo. 2022. librosa/librosa: 0.9.1.

Abhijit Mishra, Kuntal Dey, and Pushpak Bhattacharyya. 2017a. 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)*, pages 377–387.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016a. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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)*, pages 1095–1104, Berlin, Germany. Association for Computational Linguistics.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2017b. Scanpath complexity: Modeling reading effort using gaze information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Keiron O'Shea and Ryan Nash. 2015. An introduction to convolutional neural networks.

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.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. 2022. A multimodal corpus for emotion recognition in sarcasm. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6992–7003, Marseille, France. European Language Resources Association.

Rosa M Sáchez-Casas, José E García-Albea, and Christopher W Davis. 1992. Bilingual lexical processing: Exploring the cognate/non-cognate distinction. *European Journal of Cognitive Psychology*, 4(4):293–310.

Saurav Sahay, Eda Okur, Shachi H Kumar, and Lama Nachman. 2020. Low rank fusion based transformers for multimodal sequences. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 29–34, Seattle, USA. Association for Computational Linguistics.

Hollis S Scarborough, Susan Neuman, and David Dickinson. 2009. Connecting early language and literacy to later reading (dis) abilities: Evidence, theory, and practice. *Approaching difficulties in literacy development: Assessment, pedagogy and programmes*, 10:23–38.

William A Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

C Spearman. 1904. nthe proof and measurement of association between two things, oamerican j.

Harshvardhan Srivastava. 2022. Poirot at CMCL 2022 shared task: Zero shot crosslingual eye-tracking data prediction using multilingual transformer models. In *CMCL Shared Task on Multilingual and crosslingual prediction of human reading behavior*.

Ece Takmaz. 2022. Team DMG at CMCL 2022 shared task: Transformer adapters for the multi- and cross-lingual prediction of human reading behavior. In *CMCL Shared Task on Multilingual and crosslingual prediction of human reading behavior*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7216–7223.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

A Appendix

The appendix section contains five tables referred to in the paper, **on the three pages below**.

B Appendix: 1

Refer to Table 7 for the examples used in Discussion section and related error analysis.

C Appendix: 2

Refer to Table 8. The table contains all the 25 gaze features and their definitions used in our experiment.

D Appendix: 3

The three tables, *i.e,* Table 9, 10 and 11 compare the quality of predicted gaze feature values with the original collected gaze feature values for three different features. We choose these three features to include both regressions as well as fixation features in the comparison study. The models that are compared include a feed-forward neural network model, a support vector machine, and a transformer model: RoBERTa-base. The transformer model was also used with a pretrained adapter infused with it (Takmaz, 2022). We observe that the feed-forward NN produced the best correlation values and was used finally to predict all 25 gaze feature values for the 971 samples of the MUStARD++ dataset.

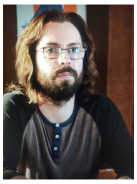
Final Utterance	**Ground Truth**	**Prediction (w/o Gaze)**	**Prediction (with Gaze)**	**Sarcasm Type**	**Video Frame**
Howard: And I love that you're strong and independent. BERNADETTE: And yet, I still love when you hold a door for me. Howard: "I love that I'm kind of a slob around here, and... you're okay with that." BERNADETTE: Uh-huh. BERNADETTE: And I love that I work and do all the cleaning, and you're okay with that.	Sarcastic	Non-Sarcastic	Sarcastic	Propositional	
PERSON: Hey, you guys the water is great. You should really go in! CHANDLER Oh, uh, no thanks. I just had an M&M.	Sarcastic	Non-Sarcastic	Sarcastic	Illocutionary	
Scott: No seriously. Moderator: Sarcastic? Us? Nooo.	Sarcastic	Non-Sarcastic	Non-Sarcastic	Illocutionary	
Monika: Things are just really different at Raviga, and I had nothing to do with the decision. Guys, I'm so, so sorry. Dinesh: Wait, so the company that offered us the most until Richard talked them into offering us the least is now offering us nothing? Gilfoyle: But the fact that you're so sorry makes it all better.	Sarcastic	Non-Sarcastic	Non-Sarcastic	Embedded	

Table 7: Error Analysis table, presents instances from the dataset and their sarcasm label predictions with and without the presence of gaze in the input.

Gaze Feature	Feature Description
Avg. Blink Duration	Mean of all blink duration's in a Dialogue/ trial.
Avg. Fixation Duration	Average duration(in milliseconds) of all selected fixations in a trial.
Total Regression Duration	Total time of eye regression in a trial.
Run Count	Total runs/count of fixations in a trial.
First Fixation Duration	Time for which the eye fixated first time in a trial.
Total Duration	Total Duration for a trial.
Fixation count	Total number of fixations in a trial.
Max. Fixation Duration time	Maximum time for which eye fixated in a trial.
Min. Fixation Duration Time	Minimum time for which eye fixated in a trial.
Interest Area Count	Number of Interest Areas in a trial.
IP Duration	Duration of Interest Period in milliseconds.
Out Regression Count	Total number of Regression in a trial.
Regression In count	Number of times regression happened to a lower id interest area.
Fixation Duration Median	Meadian of fixation durations in a trial.
Max Pupil Size	Largest size of the pupil in the trial recording.
Mean Pupil Size	Mean of the pupil sizes in a trial recording.
Min. Pupil Size	Smallest pupil size in trial recording.
Min Pupil Size x	X position of the pupil at the time when pupil size is minimum.
Interest Area Run count	Mean of number of times the interest area was entered and left.
Saccade count	Total number of saccades in a trial.
Sample count	Total number of samples in the trial.
Fixation Duration SD	Standard deviation of all fixation durations.
Saccade Amplitude SD	Standard deviation of all saccade amplitudes.
Visited IA count	Total number of times the interest area was visited.
RT	Reaction time associated with the trial.

Table 8: Gaze features and their description, these are the final set of gaze features that were used in the sarcasm detection experiment.

Total Regression Duration		
Models	**Pearson corr**	**Spearmann corr**
FFNN	**-0.32**	**0.50**
CNN	-0.23	.47
SVM	-0.185	-0.44
Roberta-base	-0.062	0.13
Roberta-base + pretrained adapter	-0.062	0.13

Table 9: Pearson correlation and Spearman correlation coefficient values to compare the quality of predicted gaze feature: total regression path duration, using different models

Out Regression Count		
Models	**Pearson corr**	**Spearmann corr**
FFNN	**0.41**	**0.213**
CNN	0.341	0.184
SVM	0.2515	0.12676
Roberta-base	0.078	0.11
Roberta-base + pretrained adapter	0.078	0.11

Table 10: Pearson correlation and Spearman correlation coefficient values to compare the quality of predicted gaze feature: total regression count, using different models

Average Fixation Duration		
Models	**Pearson corr**	**Spearmann corr**
FFNN	**-0.204**	**-0.32**
CNN	-0.186	-0.28
SVM	-0.08002	-0.07692
Roberta-base	0.063	-0.041
Roberta-base + pretrained adapter	0.063	-0.041

Table 11: Pearson correlation and Spearman correlation coefficient values to compare quality of predicted gaze feature: average fixation duration, using different models