Cognition Aware Multi-modal Sarcasm Detection

Divyank Pratap Tiwari
Department of Computer Science and Engineering
IIT Bombay
divyanktiwari.96@gmail.com

Diptesh Kanojia
Cse, Surrey
dipteshk@gmail.com

Pushpak Bhattacharyya
Department of Computer Science and Engineering
IIT Bombay
pushpakbh@gmail.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 the code, gaze data, and our best models for further research.

1 Problem Definition

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. Our primary hypothesis 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).

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

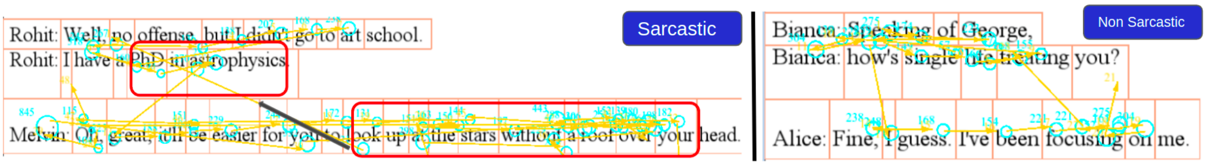

Figure 1: Sample images from a Gaze data collection setup which shows saccadic movements (yellow lines) and fixations (blue circles) for 1) a sarcastic (left image) and 2) a non-sarcastic dialogue (right image).

scanpath is a line graph that contains fixations as nodes and saccades as edges.

2 Motivation

2.1 Sarcasm: A Challenging Problem

We discussed in the previous section that sarcasm detection is a challenging problem to solve. It is because the sentences in such cases convey a different or opposite meaning for a sentence by using words of opposite meaning. The words cannot be useful mostly in this case to get to the deep meaning of the sentence. This is the reason methods like eye tracking which capture the way of human thinking using the eye movements are useful in such case. The frequency of the eye moving back and forth is often corresponding to the level of complexity in the sentence.

2.2 Gaze: A Useful Resource

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.
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.

2.3 Motivation for Multimodality

The task of sentiment analysis has been in field of research for a long time now and for a lot of years it was being performed only on the text data as the input . Due to recent boom of the internet and digitization in the world, loads and loads of data is getting uploaded every day on the internet. Social media websites like twitter allow users to post text along with images and videos to express their thoughts on the site. This has created an opportunity for collection of huge amounts of Sentiment data which is multi-modal in nature. Text is a very important modality when it comes to understanding of sentiments and opinion s involved in the data, but it it is insufficient in many cases. The visual modality can be very useful in providing information about facial gestures. The audio and visual modalities when combined with the text can provide much better information about the opinion present in the input In case sarcastic data, role of other modalities becomes very crucial, because with only a sarcastic text it is tough to identify the sentiment involved, only when the image and audio features are considered, we can do sentiment analysis for sarcastic data effectively. When it comes to sarcastic data, multimodality becomes very important in order to predict the correct sentiment or emotions for the data.
For Example:

if we have an image of a person speaking *"oh wow well done"* with a sarcastic expression.
Than if we only consider the textual modality, the sentiment which will be predicted is positive but, only when we explore the features of the image we would get to know that the correct sentiment is negative.

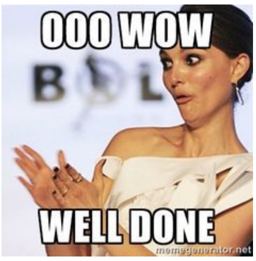

Figure 2: Sarcasm by visual cues [1]

2.4 Applications

- **Chat bots:** Intellect of chat bots can be enhanced if they are able to detect sarcasm in a customers query. The chat bots would be able to give more meaningful and relevant replies to queries having sarcasm.
  For Example:
  Customer: Thank you so much for your great service, thanks for dropping me in London and my luggage in surrey.
  Reply from bot: Thanks for the appreciation.
  Here the bot was not able to detect the disgust present in the customers comments and could not help too.

- **Online Reviews :** Many times Customers who are not happy with some product or a service, write negative reviews in sarcastic form. The system if not able to detect the deep meaning of the reviews, would classify such reviews as positive reviews and thus there will be hindrance in betterment of their services.

3 Literature Survey

3.1 Sarcasm Definition

The use of sarcasm to convey disgust is frequently mentioned. The fundamental character of it is, it could be difficult to determine the speaker's aim when they are "speaking one thing but meaning the other" or when there is incongruity.
Based on (Joshi et al., 2016), sarcasm is considered as a 6-tuple representation:
(u, p, p', S, H, C)
u = Utterance
p = Literal Proposition
p' = Intended Proposition
H = Hearer/Listener
C = Context
S = Speaker

Although it is well known that sarcasm is typically used to convey a negative emotion, it is important to analyse the reasons behind this choice of expression.

3.2 Sarcasm types

In the paper (Joshi et al., 2016), four different types of sarcasm are mainly mentioned.

- **Propositional:** Statements which require context involved to be known in order to under stand the sarcasm, otherwise they look as simple prepositions.
  Example: Yeah, right! that looks exactly like ganesh.
  Such statements could be understood only if ganesh's personality is known.

- **Embedded**: These statements include incongruity built within the words and phrases themselves.
  Example: Yes, I relish the thought of a stranger covering my body with oil and rubbing it.

- **Illocutionary:** This type of sarcasm requires other modalities apart from just text, like video and audio in order to be interpreted.
  Example: Oh wow, well done. (Big eyes and clap), this statement would only be understood when visual features are seen.

- **Like-Prefixed:** In these cases, a Like expression exists that presents an implicit denial of the claim stated in the statement.
  Example: Like you give any importance to me!

3.3 Gaze and NLP

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.

3.4 Approaches to Sarcasm Detection

Transformers (Vaswani et al., 2017) architecture-based approaches have increased in prevalence within NLP and also within sarcasm detection literature. This is most notably due to their ability to pick up semantic and syntactic relationships within text. Various rule-based and machine learning based approaches to sarcasm detection have been discussed in (Joshi et al., 2017); they also present a linguistic perspective to sarcasm detection. On the dataset released with the SemEval 2018 Shared Task 3 (Van Hee et al., 2018), (Potamias et al., 2020) offered an RCNN-RoBERTa methodology, where a RoBERTa transformer was used with BiLSTM to enhance F1-scores from cutting-edge neural network classifiers for the task of sarcasm detection. This paper also reports that the RCVV-RoBERTa approach achieved an F1-score of 90.0 on the Riloff dataset (Riloff et al., 2013). Several methods for sarcasm detection are discussed by (Shangipour ataei et al., 2020). in their article from 2020. A BERT (Devlin et al., 2019) model without concatenated layers, BERT encodings with a Logistic Regression model, and other language models like IAN (Ma et al., 2017) that are trained and assessed on a Twitter-based sarcastic dataset are among them. With an F1-score of 73.4 in those evaluations, the BERT language model without any additional layers performs the dataset's best. (Ray et al) proposes a Multimodal approach to sarcasm detection, involving various transformer and neural network-based architectures to extract features from the audio, video and text modalities, they achieved a macro-F1 score of 70.2% on the MUStARD++ dataset, a sarcasm annotated dataset, with utterances from famous sit-coms. Some existing literature investigates methods for performing sarcasm detection in Arabic (Abu Farha and Magdy, 2021), where an extensive set of experiments are performed on different transformer architectures, that include mBERT, XLM-RoBERTa (Conneau et al., 2020) and language-specific models like MARBERT (Abdul-Mageed et al., 2021). In a low-resource environment, the most effective model in this study achieves an F1-score of 58.4. A weighted average Ensemble of a CNN, LSTM, and Gated Recurrent Unit (GRU) based architectures is trained with GloVe (Pennington et al., 2014) word embeddings to identify sarcasm, as demonstrated in (Goel et al., 2022). The Ensemble outperformed comparative studies by up to 8% on SARC (Khodak et al., 2018), a Reddit comments dataset. (Bouazizi and Otsuki Ohtsuki, 2016) used a pattern-based approach to the task. This study emphasizes the role of four sets of features obtained based on different sarcasm types, the study also analyses the contribution of these features towards the classification task. This pattern-based study achieved 83.1% accuracy and 91.1% precision on the task of sarcasm detection. After transformers came into the picture, the popularity of the machine learning approaches has been declining. Some studies include (Reyes and Rosso, 2011) and (Barbieri et al., 2014) which used a Naive Bayes and Decision Tree model, respectively, in order to identify sarcasm where both achieve the best F1 scores over 70 on their chosen datasets.

3.5 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.

3.6 Test of Significance

In case of human involvement in a project for annotation, it becomes very important to prove the significance of the results generated using those human annotations. This is because, to rely on results produced by some experiments on a dataset one needs to completely trust the authenticity of the annotations i.e. the annotations which were performed were not done casually and have some meaning to it.
Paired Students T-test is one such way of testing significance where the means of two samples are compared and p-value is produced. Hypothesis Tests use samples to infer the properties of an entire population.
There are two kinds of Hypothesis as mentioned below

- **NULL Hypothesis:** The group means are equal(samples represent same population)
- **Alternative Hypothesis:** The groups have unequal means

In case of two sample independent T test: p value is a probability that represents how similar or different the two samples are from each other. We also define a significance level, mostly 0.05. If the p value $<$ Significance level, than the two samples are significantly different.

4 Sarcasm Datasets

Some of the important datasets with sarcasm data include MUStARD, MUStARD++, ZuCo, MaSaC.

4.1 MUStARD++

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, and 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 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).

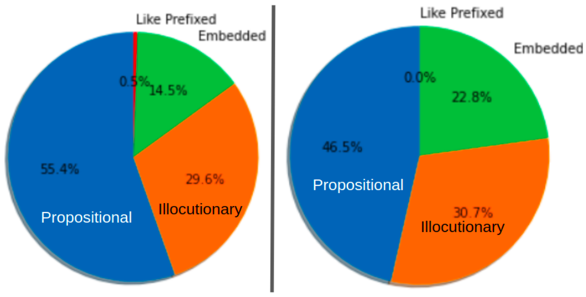

Figure 3: Sarcasm-type distribution from **D1** (left) and **D2** (right) datasets.

4.2 MaSaC

(Bedi et al., 2021) worked on the problem of **Multi-modal Sarcasm detection and Humour classification**, again in the language of Hindi, but interestingly in Code-Mixed situation along with English. They publicly release their code-mixed dataset for research on github[2]. It contains both English words in Hindi and Hindi words in English as shown below in 4

Maya:blackberry के फोन्स सिग्नल्स नगरवाद क़रीब रखो तो सिग्नल से ब्रेन डैमेज हो सकता

Maya:abhee kisee का esemes aaega lekin मुख्य riyaakaaree तो नहीं

Figure 4: Examples from MaSaC

Similar to M2H2, MaSaC also contains humour labels for the data, however it also contains sarcasm labels for each utterance.
Some other details are as follows

- **Utterances:** 15K utterances
- **Episodes:** 50 episodes (400 scenes)
- **Language:** Hindi+English code-mixed
- **Source:** Sarabhai vs. Sarabhai

[2]https://github.com/LCS2-IIITD/MSH-COMICS.git

4.3 Image+Text Sarcasm Data

While looking for multimodal datasets another commonly faced situation is that, relatively more datasets labelled as multimodal consider images to be their visual modality instead of video as desired by us. Nonetheless, the following dataset created by (Sangwan et al., 2020) contains instances each of which have a text along with an image associated and was built for the task of sarcasm detection. For testing their proposed approach for sarcasm detection they compiled two Instagram based datasets.

- **Silver dataset**
  - Posts: 10K sarcastic and 10K non-sarcastic
  - Annotation method: Hashtag based
- **Gold dataset**
  - Posts: 1600 sarcastic
  - Annotation method: Manual

Since the data is based on Instagram posts, quite often the images themselves contain some text within which could also be a carrier of incongruity and hence the authors take advantage of the transcript extracted from the image and take advantage of that too by treating it as a third modality. The following figures 5 and 6 show the kind of data this dataset holds

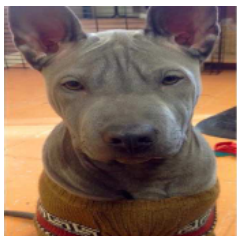

Figure 5: Example 1: Text incongruous with image *(Caption:Someone is excited for sweater season)*

Figure 6: Example 2: Sarcasm within the transcript in the image

4.3.1 Other Multimodal Gaze Datasets

Zuco and Zuco2.0 are two datasets that can be useful resources for gaze and EEG features data. This has 130 Gb of data annotated with gaze features and EEG features.

5 Feature Extraction Techniques: Uni-modal Features

There are techniques to find good quality embedding of feature representations for the separate uni-modal data i.e. for visual, audio, and text data, it is very important to capture useful information in the uni-modal representations of the data so that when these representations are fused, quality information from all modalities are captured. The techniques which are popular for extracting these uni-modal features are mentioned in the following sections.

5.1 Visual Feature Extraction

5.1.1 Facet Library

It is open-source research from google and it is very useful to get an understanding of the data and its structure that is being used. Two tools facets overview and facets dive can be used for this purpose.

5.1.2 Resnet-152 encoder

These are very deep-layered networks with convolution layers, but as in case of standard neural networks with convolution layers, the problem of vanishing/exploding gradients arises. To overcome this skip connections were introduced in the the resnet architecture and the resnet architecture had 152 layers in it. These capture visual features effectively when pre tarined on some large image dataset.

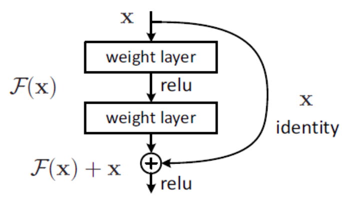

Figure 7: Skip connection in RESNET [3]

5.2 Audio Feature Extraction

Audio features can play an important role in sentiment and emotion analysis, the pitch, tone, and speed of the speech are useful in determining the sentiment of the speaker, for example, if the sentiment involved is anger then most likely the audio will be loud and the tone will not be soft. Some of the Popular tools and techniques used to extract audio features are:

5.2.1 openSMILE

openSMILE (open-source Speech and Music Interpretation by Large-space Extraction) is an open-source software used for audio feature extraction and is also important for the task of classification of music labels.

5.2.2 COVAREP

A COLLABORATIVE VOICE ANALYSIS REPOSITORY which has made access to most of the audio processing and extraction-related algorithms easier. It has made the research more reproducible as reproducing algorithms from original papers was a tougher task. This covarep can be used for extraction of audio features from signal for the task of sentiment and emotion analysis.

5.3 Textual Feature Extraction

Lot of research has been done to generate good quality embedding's for the text data which capture lots of information in the text for example the Distributional similarity between words or phrases of the text, also semantic information needs to be captured from the text.
Some of the Popular word embedding techniques used are:

5.3.1 Glove Pre-trained embedding

GloVe (Pennington et al., 2014) is an algorithm for extracting vector representations for words in a given document. It makes use of the global word-word co-occurrence matrix of a dataset in order to generate the word vectors for the words which is why it captures global semantics or context for the words.

5.3.2 ELMO embedding

Embeddings from Language Models (ELMo) is a technique used for extracting vector representations of words from sentences and these vector representations provide information about the word sense too. The difference between Elmo and the above embedding i.e. glove is that there can be different representations of the same word , when the word is being used in different context's.

It uses 2 layered Bi-Lstm's for the generating word embedding's.

6 Multimodal Fusion

Now after we have the uni-modal representations of all the modalities separately, a task of fusion needs to be performed which is basically combining all the unimodal vectors and generating a single vector for the complete data across all the modalities involved. Some of the techniques for multimodal fusion are mentioned below:

6.1 Gating

(Liu et al., 2020) presents a very detailed work which aims at ensuring good quality representation when videos are involved with other modalities like text, and audio etc., Their main aim is to focus on building a compact representation that finds application in a number of video understanding tasks, such as video retrieval, clustering and summarization. To this extent they propose a multimodal fusion framework, called, 'Collaborative Gating' that ensures that video and text that correspond to each other stay similar in representation, as compared to when they are unassociated. They treat the video, audio, and embedded text as three different modalities. Since this methodology internally utilizes attention, we take inspiration from this work to perform multimodal fusion in our project.

6.2 Concatenation

This is the most basic way of fusing uni-modal representations, a concatenation operation is performed among all the vector representations to generate a single fused multimodal vector representation.

6.3 Dynamic Fusion Graph

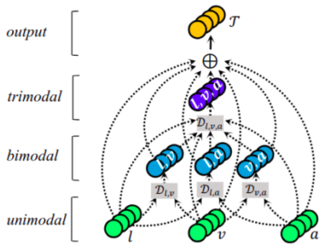

Figure 8: Dynamic Fusion Graph [4]

This Dynamic Fusion Graph explicitly models the n-modal interactions in a hierarchical manner as well as it has the capability of altering its network/structure based on the importance of n-modal dynamics. Dense feed forward neural networks are used to generate representations for all possible combinations of modalities and finally in the last layer concatenation of all these representation is performed to generate the final representation having features from all three modalities.

7 Gaze Annotation

We instructed 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.

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

7.1 Annotation & Feature Validity

We compute **inter-annotator agreement** using a pair-wise Fleiss' kappa (Scott, 1955), which *resulted 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 1), respectively. In Table 2, 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.

8 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 feature further in a multi-task setting of sarcasm detect 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 jokes are the key reasons for the inter-annotator agreement value being low. 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.

	Average Fixation Duration			IA Regression Path Duration		
	μ_Pos ± σ_Pos	μ_Neg ± σ_Neg	p	μ_Pos ± σ_Pos	μ_Neg ± 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 2: 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.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. Arbert marbert: Deep bidirectional transformers for arabic.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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.

Francesco Barbieri, Horacio Saggion, and Francesco Ronzano. 2014. Modelling sarcasm in Twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–58, Baltimore, Maryland. Association for Computational Linguistics.

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Multi-modal sarcasm detection and humor classification in code-mixed conversations. *IEEE Transactions on Affective Computing*, pages 1–1.

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.

Mondher Bouazizi and Tomoaki Otsuki Ohtsuki. 2016. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

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

Priya Goel, Rachna Jain, Anand Nayyar, Shruti Singhal, and Muskan Srivastava. 2022. Sarcasm detection using deep learning and ensemble learning. *Multimedia Tools and Applications*, 81(30):43229–43252.

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.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

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.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1006–1011, Austin, Texas. 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.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm.

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.

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. 2020. Use what you have: Video retrieval using representations from collaborative experts.

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.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4068–4074.

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).

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).

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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.

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.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

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.

Antonio Reyes and Paolo Rosso. 2011. Mining subjective knowledge from customer reviews: A specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 118–124, Portland, Oregon. Association for Computational Linguistics.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

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.

Suyash Sangwan, Md Shad Akhtar, Pranati Behera, and Asif Ekbal. 2020. I didn't mean what i wrote! exploring multimodality for sarcasm detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

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.

Taha Shangipour ataei, Soroush Javdan, and Behrouz Minaei-Bidgoli. 2020. Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 67–71, Online. Association for Computational Linguistics.

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*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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*.