Bias in Multimodal Emotion Recognition Models

Student Name: Julian Wyatt
Supervisor Name: Noura Al-Moubayed

Abstract —

Context/Background - Multimodal Emotion Recognition is the process of extracting emotion and opinion by leveraging more than one modality such as facial imaging, textual and vocal information. Bias is inherent in this system as it will perform better for some protected group of individuals such as race or gender, which raises ethical concerns where sub-groups of people are disadvantaged. This could also amplify cultural and social bias in society.

Aims - This paper seeks to evaluate the bias in a multimodal emotion recognition system, while experimenting with potential mitigation strategies to alleviate the overall bias.

Method - We utilise an early fusion method for merging textual data (from bi-LSTM latent space), visual data (from 3D-convolutional projected latent space), and finally vocal data (from a projected latent space), to build our multimodal emotion recognition system. From here, we experiment with different combinations of debiasing techniques for each modality and compare our metrics with recent literature.

Keywords — Multimodal Emotion Recognition, Debias, Explainable AI, Computer Vision, NLP

I INTRODUCTION

A Emotion and Emotion Recognition

Emotion is a person's subconscious cognitive mood and state of mind, which we naturally portray through the pitches and tones in our voice, our facial expressions and what we say. Within recent years, the development of AI allows us to recognise, model and predict emotion in a range of scenarios, thus generating machine based emotional intelligence. This gives rise to a list of potential uses, from e-health systems for detecting spontaneous pain based facial expressions (Littlewort et al. 2009) to conversational chat bots leveraging mental health analysis (Antony et al. 2021). Within the growing technology world, especially during the COVID-19 pandemic, systems that are accurately analyse emotion are becoming increasingly important for automating tasks and AI based socialising.

B Multimodal Emotion Recognition

Advanced emotion recognition systems leverage all available data by using multiple modalities. This means that if we wish to analyse a video review to assess the reviewers emotion and opinion of an entity, we can utilise their facial expressions, the tones and pitches in their voice, and finally the textual information. Not only does this improve the accuracy of the prediction of their emotion, but should also make the overall predictions more explainable. For example Liu et al. (2020) proposed a dynamic attention-based recurrent neural network to make visual and textual recommendations more explainable, highlighting the potential for explanations in multimodal systems. Popular multimodal emotion recognition models in literature, aim to predict which emotion the audio, visual and textual content relates to (Mittal et al. 2020).

C Bias in Data Science

Bias is very prevalent in the majority of datasets. Commonly, this is a product from class imbalance in datasets. For example we could have an imaging dataset like MNIST (dataset for handwritten digits). If we took a subset of that data consisting of 1 image containing a written 1, and 999 images containing a written 7, then we tried to predict whether a given image is a 1 or a 7, the model will predict 7 nearly every time. Clearly, if we extrapolate this idea to a population demographic dataset, we will experience the same issues. Buolamwini & Gebru (2018) highlight the extent of the demographic imbalance in multiple datasets. The ratio of lighter skinned people to darker skinned people equated to 79.6% for the IJB-A dataset and 86.2% for the Adience dataset. Therefore the models which use these datasets could be argued to be unfair and biased towards white males as models will learn the inherent probability distribution between demographics.

Across the steps of the data science process, bias can emerge in a wide range of cases. Mehrabi et al. (2019), outline 23 sources of bias, which range from historical bias (Suresh & Guttag 2019) (*the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection*), evaluation bias (Suresh & Guttag 2019) (*happens during model evaluation*) and algorithmic bias (Baeza-Yates 2018) (*when the bias is not present in the input data and is added purely by the algorithm*). Furthermore, the same paper goes on to provide different definitions of fairness with AI. Most notably, the equality of odds definition: *We say that a predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A* (such as gender or race) *and outcome Y, if $\hat{Y}$ and A are independent conditional on Y.* Mathematically, this is shown as (Hardt et al. 2016):

$$Pr\{\hat{Y}=1|A=0,Y=y\} = Pr\{\hat{Y}=1|A=1,Y=y\}, \quad y \in \{0,1\} \tag{1}$$

Where $\hat{Y}, A, y$ is the predictor, the protected attributed and the class respectively. In other words, the probability of true positives to false positives should be the same for protected and unprotected groups.

Determining this bias can be very difficult in many models, giving rise to the explainability of machine learning models. It can generally be hypothesised that the more explainable a model is, the lower it's performance is at doing a task (Gunning & Aha 2019). This is especially apparent in models such as decision trees, which use branching paths to determine some classification problem. Thus making them excellent at explaining the model, but limits their overall potential by fitting them to human-understandable rules. On the other end of the spectrum, deep learning models can have fabulous performance due to their stochastic gradient descent to attempt to find a global minimum for a function. However, their inputs and outputs are heavily mathematical and lots of steps have to be taken to start making them explainable (Zhang et al. 2020). By making these deep learning models more explainable, we can in turn determine where the bias is being introduced within the model and experiment with mitigation strategies.

D Objectives

This paper seeks to address the research question: *How biased are the machine learning models used for multimodal emotion recognition and what affect do debiasing techniques have on these models?* To advance the current literature, the following objectives should be completed:

- **Multimodal Emotion Recognition** - Reproduce architecture from modern literature which predicts an emotion class from a video-based dataset using textual, spoken, and visual information.
- **Measure Performance and Bias** - Measure the quality of the architecture and the effect of the bias in the system and compare this to the state of the art.
- **Baseline debiasing techniques** - Perform baseline debiasing techniques to each modality in feature space and remeasure the metrics.
- **Propose debiasing technique** - Using a cycle consistency adversarial architecture, synthesise a single modality and predict a stereotype, such that over time the stereotype becomes indistinguishable.

II RELATED WORK

The related work is split across the following 2 sections, initially covering the background problem, and also covering the bias and mitigation strategies across different modalities.

A Multimodal Emotion Recognition

As defined above, multimodal emotion recognition is the process of leveraging multiple modalities such as text, speech and images to extract emotion, sentiment and opinion. For textual emotion recognition, there are many methods in recent literature which perform this task. The main method recently, is to use a bi-directional long short term (bi-LSTM) architecture (Xu et al. 2019). This architecture receives the textual data via a numerical encoding through an embedding network such as GloVe (Pennington et al. 2014). From the rise of transformer based networks, the GloVe representation has been traded for BERT (Li et al. 2019), with experimentation across a combination of embeddings (Naseem et al. 2020).

Speech emotion recognition (SER) is still a growing field, where new models vary slightly from the previous, seeking for real-time analysis. Originally, Yu et al. (2013) proposed learning the emotion classes via a 7-layer dense neural network (the paper predates the standardisation of using convolutions) after translating the speech data into image data using a short time fourier transform. Whereas, more recently Lech et al. (2020) incorporated the modern standard, and experiment with preprocessing, and it's effect on performance to reach real time analysis. This was incorporated by reducing the bandwidth and companding (compressing the data, then expanding the data) the sound data. Furthermore, some architectures focus on the preprocessing by calculating the important features prior to using deep learning methods, such as the openSMILE software (Poria et al. 2015).

Visual emotion recognition is well researched, invoking the communication between computer vision and deep learning disciplines; therefore, there are many proposed approaches in literature. Recently, the OpenFace toolkit (Baltrusaitis et al. 2018), and the method from (Bulat & Tzimiropoulos 2017) are used as preprocessing techniques to locate and shrink noisy images to only include faces, where (Bulat & Tzimiropoulos 2017) tends to perform better for faces which are more obscured. Following the preprocessing, the extracted faces from these techniques are passed into a deep learning model. For example Samadiani et al. (2021), use a novel LSTM implementation, along with a convolutional network, whose outputs are joined via multimodal techniques.

Multimodal emotion recognition models form the bridge between the previous 3 topics. There are several methods for merging the different modalities, which cover two bands: early fusion and late fusion. Early fusion seeks to fuse modalities in feature space, whereas late fusion fuses modalities in semantic space (Snoek et al. 2005). For example Tsai et al. (2019) used a cross-modal transformer architecture to learn the important features between modalities on unaligned multimodal data. Whereas (Yu et al. 2021), a late fusion method, projects the semantic uni-modal representations into a lower dimension using linear network layers and their unique Unimodal Label Generation Modules. Additionally, there are models such as (Zadeh et al. 2017) which find the cartesian product of each of the modalities in feature space, accurately combining them. Very recent approaches such as (Yakaew et al. 2021), use bimodality (ignoring the textual dimension), with a custom architecture to determine whether the mouth is open or closed, and only combine the modes when the actor is talking. Finally, Atmaja & Akagi (2020), combine both early and late fusion techniques via support vector regression to improve accuracy.

B Debiasing Techniques

Adversarial Debias (Zhang et al. 2018). Developing the generative adversarial network architecture (Goodfellow et al. 2014), we use a predictor network of varying complexity (acting as the generator), and an adversary network (acting as the discriminator). The predictor is some gradient-based machine learning network which solves a regression or classification problem; while the adversary network seeks to predict the protected attribute, from Eq. 1. The predictor network's weights are then updated by

$$\nabla_W \mathcal{L}_P - \text{proj}_{\nabla_W \mathcal{L}_A} \nabla_W \mathcal{L}_P - \alpha \nabla_W \mathcal{L}_A \tag{2}$$

After training, the predictor architecture should converge on the equality of odds definition (Eq. 1).

Textual Embedding Debias. (Kaneko & Bollegala 2019) propose a method for reducing the bias within pre-trained embeddings such as GloVe. Using an autoencoder, they devise a loss function (Eq. 3), such that they remove the bias from the gendered words, maintain the information from gender-neutral words and minimise the reconstruction loss to keep the meaning in the embedding.

$$\mathcal{L} = \lambda_f \mathcal{L}_f + \lambda_m \mathcal{L}_m + \lambda_g \mathcal{L}_g + \lambda_r \mathcal{L}_r \tag{3}$$

Furthermore, double-hard debias (Wang et al. 2020), implement another post-processing technique. Generally, this method involves applying dimensionality reduction to biased words, then using hard debias to neutralise the gender subspace in these biased words.

Convolutional Network Debias. (Yucer et al. 2020) propose a different adversarial approach for improving racial bias within images. They use a cycle consistency loss function (Zhu et al. 2017) for augmenting images from one racial class to another, which produced an increase in accuracy within the minority classes. David et al. (2020) sought to avoid the adversarial approaches, as they are notorious for having a sharp manifold, by learning concepts within the image at each convolutional layer (these include abstract things like colour or texture, or specific objects such as an oven or a dog). Then, adding an orthogonalisation regularisation term (Eq. 4) to the loss function, which learns to remove the projection bias in the class embeddings.

$$\mathcal{L}_{debias}(\beta') = \sum_{c} \Big(\frac{\beta_c'^{\top} v}{\|\beta_c'\|_2 \|v\|_2} \Big)^2 \tag{4}$$

Debiasing Vocal Data (Gorrostieta et al. 2019). Depending on the nature of the speech analysis, slightly different debiasing methods can be used, for example the adversarial technique from (Zhang et al. 2018) could be repeated when using a convolutional network. Alternatively, building from that technique, we use a method for debiasing generic deep learning architectures. The loss function for the network (Eq. 5) adds a weighted equality loss, which is itself calculated in Eq. 6. This seeks to minimise the recall predictions for male vs female protected classes. The function $r_{a,g}()$ computes the recall for activation class a (0 or 1) and gender class (f for female and m for male).

$$\mathcal{L}_N = \mathcal{L}_P + \alpha \mathcal{L}_{eq} \tag{5}$$

$$\mathcal{L}_{eq} = |r_{0,m}(\hat{y}_{bin}, y_{bin}) - r_{0,f}(\hat{y}_{bin}, y_{bin})| + |r_{1,m}(\hat{y}_{bin}, y_{bin}) - r_{1,f}(\hat{y}_{bin}, y_{bin})| \tag{6}$$

III SOLUTION

A Architecture

As the focus for this paper is to highlight the biases in these systems, we will keep the architecture similar to (Zadeh et al. 2017). Thus, the underlying early fusion will form a 3 dimensional manifold, by finding the cartesian product of the features discussed below.

The textual features will incorporate a pre-trained GloVe embedding network (Pennington et al. 2014), which is passed through a bi-LSTM (Xu et al. 2019). The vocal features will be extracted from the openSMILE software (Poria et al. 2015), using the large emotion feature configuration set. This contains 6,552 low-level description features including intensity, loudness, pitch, means and standard deviation. To make this latent representation more meaningful, we will use the autoencoder dimensionality reduction method. Finally, the visual features will be extracted using the openFACE open source facial recognition library (Baltrusaitis et al. 2018), which is then passed through a 3D-convolutional network to encode the temporal properties extracted from openFACE. Thus, we obtain a latent space representation of the facial data.

B Debiasing Approaches

To examine how each method affects the overall bias and fairness, we will investigate every combination of the following debiasing techniques, using male and female subgroups as our protected classes. Therefore, this enables the meaningful evaluation of each debiasing technique and how they can collaborate in this multimodal context.

For the textual features, we will utilise the GloVe embedding debias method from (Kaneko & Bollegala 2019). This builds a latent representation of the dictionary used such that we protect the masculine and feminine properties in gendered words, the gender neutrality in gender-neutral words and finally, remove the bias in stereotypical words. More specifically, we use an autoencoder with the loss function in Eq. 3, to minimise 4 different criteria: retaining gender-related words (masculine and feminine), the stereotypical and gender neutral words should be orthogonal to the gendered words, and finally the reconstruction loss.

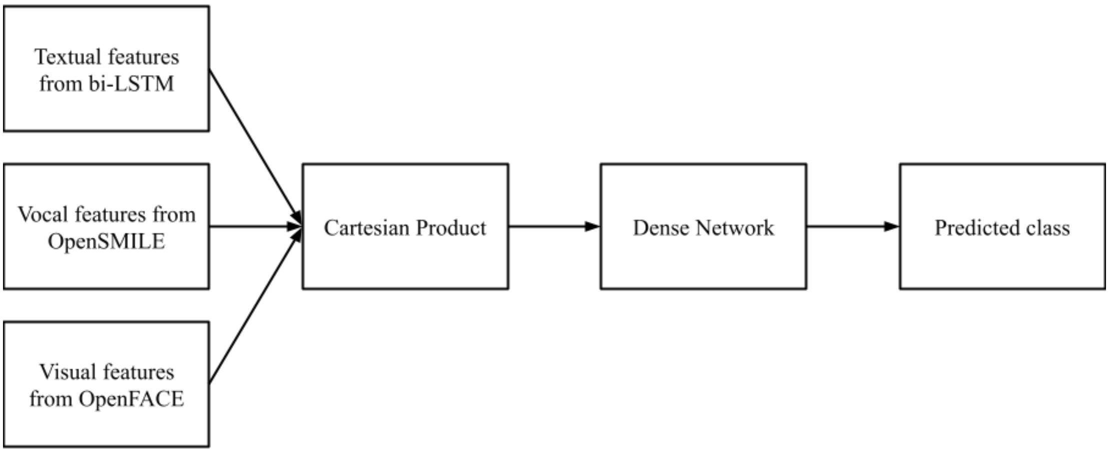
Textual features from bi-LSTM
Vocal features from OpenSMILE
Visual features from OpenFACE
Cartesian Product
Dense Network
Predicted class

Figure 1: Multimodal Emotion Recognition Architecture

As the convolutional debiasing techniques focus on a different problem set, we will use the adversarial debiasing technique (Zhang et al. 2018) for the visual network. Thus, updating the weights of the convolutional network adversarially (Eq. 2), to minimise the equality of odds definition.

The vocal features will be debiased like (Gorrostieta et al. 2019), which means that we add an equality equation to the loss function (Eq. 5 and 6). Therefore, through training, the protected male and female classes will become indistinguishable with respect to the weights in the neural network. Finally, these deep learning based debiasing techniques will include a low weighted mean square error loss equation to help maintain the overall accuracy when debiasing.

C cycleBias

In addition to the above debiasing techniques, we propose to explore a technique combining (Zhang et al. 2018) and (Yucer et al. 2020). We synthesise the vocal data using a generative adversarial network (GAN), such that the adversary seeks to predict the protected class, and the generator converts the vocal data into it's latent representation. From here, we wish to reverse this network's function using the cycle consistency loss (Zhu et al. 2017). Overall, this process would allow us to visualise the effect that the adversarial technique causes on the latent representation of the model. The overall loss equation is as seen in Eq. 7, where the loss for both GANs is seen in Eq. 8. This is a minmax problem where the generator seeks to minimise its loss while the adversary seeks to maximise its loss. A generic architecture for this model can be seen in Fig. 2. This shows how the cycle consistency loss is produced. Furthermore, after convergence of the adversarial model, the model should reach the equality of odds definition outlined earlier by removing the inherent bias in the vocal latent space.

$$\mathcal{L} = \mathcal{L}_{GAN_A} + \mathcal{L}_{GAN_B} + \lambda\mathcal{L}_{cyc} \tag{7}$$

$$\mathcal{L}_{GAN_A}(G, D_Y, X, Y) = \mathbf{E}_{y\ p_{data}(y)}[logD_y(y)] + \mathbf{E}_{x\ p_{data}(x)}[log(1 - D_y(G(x))] \tag{8}$$

D Datasets

In order to perform our evaluation and analysis of the architecture, we will use the OMGEmotion dataset (Barros et al. 2018). This provides utterances from a selection of 1 minute (average)

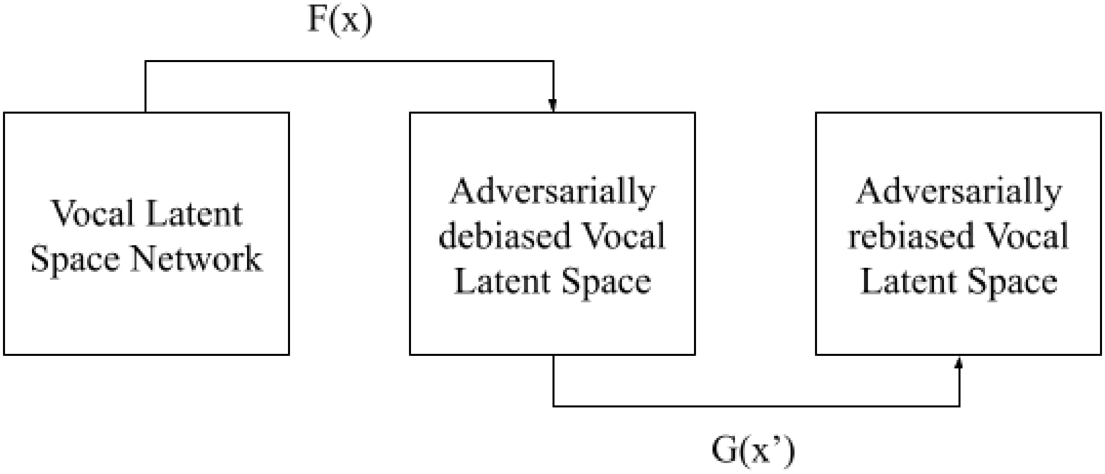
F(x)
Vocal Latent Space Network
Adversarially debiased Vocal Latent Space
Adversarially rebiased Vocal Latent Space
G(x')

Figure 2: Cycle Consistency Debiasing Architecture

YouTube videos, that represents the diversity across protected groups, along with the transcribed textual data from those utterances. These have been transcribed using Google's Speech Recognition system which may be biased. However, we are assuming the transcriptions are accurate due to the popularity and scale of Google's model.

There are 7 potential emotions to classify each utterance, and these include: anger, disgust, fear, happiness, sadness, surprise and neutral. Additionally, we will estimate the valence and arousal values for each utterance, thus modelling the emotion recognition as a 2-D regression problem, allowing us to measure the effect of debiasing techniques depending on classification vs regression problems. Furthermore, a fundamental issue with emotion recognition datasets is the subjectivity inherent in labeling the datasets. The OMGEmotion dataset raises this issue and collates the data through collaboration of many annotators, deciding on ”gold standard” values using averages from a frequency distribution taken from an average of 5 annotators.

Moreover, we will use the IEMOCAP dataset (Busso et al. 2008), which differs from OMGEmotion as the dataset was recorded from actors, acting out emotions through a script. Therefore, this allows us to gain a better idea how the model performs in terms of accuracy, however there is not the same diversity of footage as IEMOCAP includes only 10 actors.

E Evaluation

The architecture in Fig. 1, allows for debiasing the feature networks. Therefore, we will assess the severity of the bias in this baseline system. Next, we will apply the mentioned debiasing techniques to determine a combination of those methods that help to alleviate the underlying bias. From here, we can evaluate our original and debiased models to the state of the art models, thus assessing how the debiasing methods affect the overall performance and how the methods compliment each other.

To evaluate the model, we will use 4 primary metrics: accuracy, F1-score, CCC and TPR_{diff}. Accuracy is the ratio of correct predictions against total predictions. By increasing this we reach a ”better” model. F1-score gives a better general performance score from calculating the harmonic mean of the precision and recall values (Eq. 9). In Eq. 10, TP is the number of true positive predictions, FP is the number of false positives and FN is the number of false negatives.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{9}$$

$$precision = \frac{TP}{TP + FP}\ , \quad recall = \frac{TP}{TP + FN} \tag{10}$$

Furthermore, we use the Concordance Correlation Coefficient (CCC), and TPR metrics from (Gorrostieta et al. 2019). The CCC is a regression metric calculated using Eq. 11; where μ and σ are the mean and variances for the prediction and the gold standard and finally ρ is the Pearson's Correlation Coefficient between the prediction and the gold standard. TPR_{diff} (True positive rate), is an "equality of odds" metric thus quantifying the bias in the model, and is calculated by finding the true positive rate (TPR) for low activation instances for the male and female subgroups and again for high activation instances. We then find the difference of those two activation instances, and sum them.

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{11}$$

When measuring the CCC for arousal and valence on the OMGEmotion dataset, Zheng et al. (2018) produce a best score of 0.397 for arousal and 0.515 for valence, and a best average of 0.454. These use different combinations of textual, audio and visual modalities with differing techniques. One of the top recent results for IEMOCAP is from Mittal et al. (2020), by incorporating a novel data-driven multiplicative fusion technique. They reported a 0.824% F1 score and a 82.7% mean accuracy.

IV VALIDITY

Our problem: *How biased are the machine learning models used for multimodal emotion recognition and what affect do debiasing techniques have on these models?* is paramount for the development of machine learning models in the future. The bias in these "closed-box" (hitherto known as "black-box") systems will affect millions of people on a daily basis, therefore it is important to understand what makes these models biased and how we can mitigate that bias. For example, insurance or credit card application systems may use a machine learning component to assess the suitability of someone applying. This will result in a potential for the system to become biased towards a protected class of people, thus introducing discrimination. It is therefore clear to understand the importance of assessing bias within machine learning. While it is harder to see the importance of bias in multimodal emotion recognition systems in general, it could still apply as a potential for CCTV based crime prediction by assessing emotion.

To cover the entirety of the problem, the related research outlines the current standard across all aspects of multimodal emotion recognition systems. We outline how each modality is processed and therefore how exactly we can combine the modalities. The research for this field is still growing with a few notable papers mentioned which are only a few weeks old. We also outline several processes of debiasing each modality separately, with the potential for some methods being used across modality. Consequently, this allows us to explore the effect that the modern debiasing techniques have on an intricate architecture such as multimodal emotion recognition.

The process of solving the research problem, outlined in the methodology, provides a stable and reliable solution to the overarching problem. To begin, we will use state-of-the-art metrics for assessing the severity of the bias in the form of TPR_{diff}, F1-score and CCC. Furthermore, through the evaluation of different combinations of debiasing techniques, we can learn the effect they have on the architecture. Moreover, when visualising the latent representation of each stage of the cycleBias model, we can explore how that adversarial technique influences the bias and performance within the vocal latent representation manifold.

V RESULTS

VI EVALUATION

VII CONCLUSIONS

References

Antony, C., Pariyath, B., Safar, S., Sahil, A. & Nair, A. R. (2021), 'Emotion recognition-based mental healthcare chat-bots: A survey', *Available at SSRN 3768304* .

Atmaja, B. T. & Akagi, M. (2020), Multitask learning and multistage fusion for dimensional audiovisual emotion recognition, *in* 'ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', pp. 4482–4486.

Baeza-Yates, R. (2018), 'Bias on the web', *Communications of the ACM* **61**(6), 54–61.

Baltrusaitis, T., Zadeh, A., Lim, Y. C. & Morency, L. (2018), Openface 2.0: Facial behavior analysis toolkit, *in* '2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)', pp. 59–66.

Barros, P., Churamani, N., Lakomkin, E., Sequeira, H., Sutherland, A. & Wermter, S. (2018), The OMG-Emotion Behavior Dataset, *in* '2018 International Joint Conference on Neural Networks (IJCNN)', IEEE, pp. 1408–1414.

Bulat, A. & Tzimiropoulos, G. (2017), How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks), *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1021–1030.

Buolamwini, J. & Gebru, T. (2018), Gender shades: Intersectional accuracy disparities in commercial gender classification, *in* S. A. Friedler & C. Wilson, eds, 'Proceedings of the 1st Conference on Fairness, Accountability and Transparency', Vol. 81 of *Proceedings of Machine Learning Research*, PMLR, New York, NY, USA, pp. 77–91.
URL: *http://proceedings.mlr.press/v81/buolamwini18a.html*

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S. & Narayanan, S. S. (2008), 'Iemocap: Interactive emotional dyadic motion capture database', *Language resources and evaluation* **42**(4), 335–359.

David, K. E., Liu, Q. & Fong, R. (2020), 'Debiasing convolutional neural networks via meta orthogonalization', *arXiv preprint arXiv:2011.07453* .

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014), 'Generative adversarial networks', *arXiv preprint arXiv:1406.2661* .

Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R. & Kane, J. (2019), Gender de-biasing in speech emotion recognition., *in* 'INTERSPEECH', pp. 2823–2827.

Gunning, D. & Aha, D. (2019), 'Darpa's explainable artificial intelligence (xai) program', *AI Magazine* **40**(2), 44–58.

Hardt, M., Price, E., Price, E. & Srebro, N. (2016), Equality of opportunity in supervised learning, *in* D. Lee, M. Sugiyama, U. Luxburg, I. Guyon & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 29, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf*

Kaneko, M. & Bollegala, D. (2019), 'Gender-preserving debiasing for pre-trained word embeddings', *arXiv preprint arXiv:1906.00742* .

Lech, M., Stolar, M., Best, C. & Bolia, R. (2020), 'Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding', *Frontiers in Computer Science* **2**, 14.
URL: *https://www.frontiersin.org/article/10.3389/fcomp.2020.00014*

Li, X., Bing, L., Zhang, W. & Lam, W. (2019), 'Exploiting bert for end-to-end aspect-based sentiment analysis', *arXiv preprint arXiv:1910.00883* .

Littlewort, G. C., Bartlett, M. S. & Lee, K. (2009), 'Automatic coding of facial expressions displayed during posed and genuine pain', *Image and Vision Computing* **27**(12), 1797–1803.

Liu, P., Zhang, L. & Gulla, J. A. (2020), 'Dynamic attention-based explainable recommendation with textual and visual fusion', *Information Processing & Management* **57**(6), 102099.
URL: *https://www.sciencedirect.com/science/article/pii/S0306457319301761*

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2019), 'A survey on bias and fairness in machine learning', *arXiv preprint arXiv:1908.09635* .

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. & Manocha, D. (2020), M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 1359–1367.

Naseem, U., Razzak, I., Musial, K. & Imran, M. (2020), 'Transformer based deep intelligent contextual embedding for twitter sentiment analysis', *Future Generation Computer Systems* **113**, 58–69.

Pennington, J., Socher, R. & Manning, C. D. (2014), Glove: Global vectors for word representation, *in* 'Empirical Methods in Natural Language Processing (EMNLP)', pp. 1532–1543.
URL: *http://www.aclweb.org/anthology/D14-1162*

Poria, S., Cambria, E. & Gelbukh, A. (2015), Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis, *in* 'Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Lisbon, Portugal, pp. 2539–2544.
URL: *https://www.aclweb.org/anthology/D15-1303*

Samadiani, N., Huang, G., Hu, Y. & Li, X. (2021), 'Happy emotion recognition from unconstrained videos using 3d hybrid deep features', *IEEE Access* **9**, 35524–35538.

Snoek, C. G., Worring, M. & Smeulders, A. W. (2005), Early versus late fusion in semantic video analysis, *in* 'Proceedings of the 13th annual ACM international conference on Multimedia', pp. 399–402.

Suresh, H. & Guttag, J. V. (2019), 'A framework for understanding unintended consequences of machine learning', *arXiv preprint arXiv:1901.10002* .

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P. & Salakhutdinov, R. (2019), Multimodal transformer for unaligned multimodal language sequences, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 6558–6569.
URL: *https://www.aclweb.org/anthology/P19-1656*

Wang, T., Lin, X. V., Rajani, N. F., McCann, B., Ordonez, V. & Xiong, C. (2020), 'Double-hard debias: Tailoring word embeddings for gender bias mitigation', *arXiv preprint arXiv:2005.00965* .

Xu, G., Meng, Y., Qiu, X., Yu, Z. & Wu, X. (2019), 'Sentiment analysis of comment texts based on bilstm', *IEEE Access* **7**, 51522–51532.

Yakaew, A., Dailey, M. N. & Racharak, T. (2021), 'Multimodal sentiment analysis on video streams using lightweight deep neural networks'.

Yu, D., Seltzer, M. L., Li, J., Huang, J.-T. & Seide, F. (2013), 'Feature learning in deep neural networks-studies on speech recognition tasks', *arXiv preprint arXiv:1301.3605* .

Yu, W., Xu, H., Yuan, Z. & Wu, J. (2021), 'Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis', *arXiv preprint arXiv:2102.04830* .

Yucer, S., Akçay, S., Al-Moubayed, N. & Breckon, T. P. (2020), Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops', pp. 18–19.

Zadeh, A., Chen, M., Poria, S., Cambria, E. & Morency, L.-P. (2017), 'Tensor fusion network for multimodal sentiment analysis', *arXiv preprint arXiv:1707.07250* .

Zhang, B. H., Lemoine, B. & Mitchell, M. (2018), Mitigating unwanted biases with adversarial learning, *in* 'Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society', pp. 335–340.

Zhang, Y., Tiňo, P., Leonardis, A. & Tang, K. (2020), 'A survey on neural network interpretability', *arXiv preprint arXiv:2012.14261* .

Zheng, Z., Cao, C., Chen, X. & Xu, G. (2018), 'Multimodal emotion recognition for one-minute-gradual emotion challenge', *arXiv preprint arXiv:1805.01060* .

Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. (2017), Unpaired image-to-image translation using cycle-consistent adversarial networks, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 2223–2232.