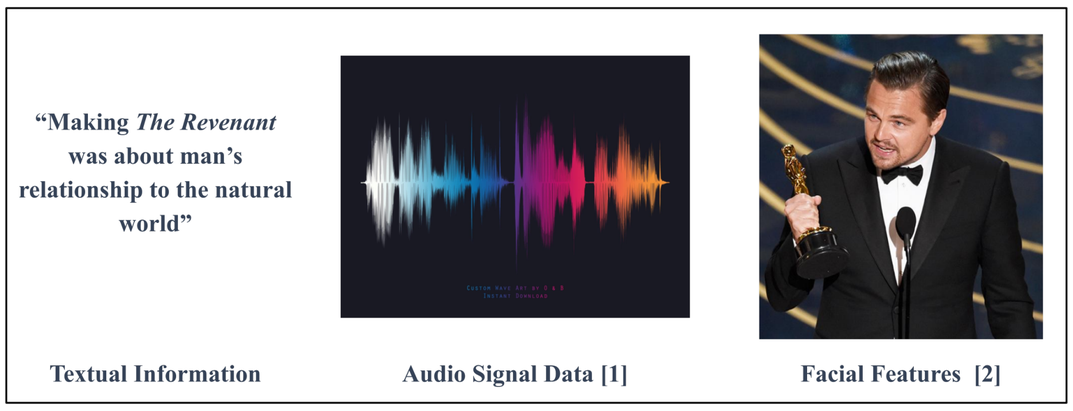
Bias in Multimodal Emotion Recognition Models

Student: Julian Wyatt - mbtj48

Sponsor: Noura Al-Moubayed Affiliations: Durham University

I - ABSTRACT & INTRODUCTION

Context/Background - Emotionally intelligent systems are becoming increasingly prevalent and systemic within the developing technology world. By leveraging this developing intelligence, we can build a multimodal based emotional model to learn emotion, sentiment and opinion from a person. Multimodal Emotion Recognition (MER) is the process of extracting emotion and opinion by leveraging more than one modality such as facial imaging, textual and vocal information. We can choose to predict valence and arousal (2D regression problem), or predict emotions as classes.



Unfortunately, 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, potentially amplifying cultural and social bias in society. This bias arises due to a variety of issues from *historical bias* to *algorithmic bias*. We can mathematically describe bias using the equality of odds definition (Hardt et al. 2016):

$$Pr\{\hat{Y} = 1 \mid A = 0, Y = y\} = Pr\{\hat{Y} = 1 \mid A = 1, Y = y\}, \quad y \in \{0,1\}.$$

Where $\hat{Y}$, A, y is the predictor, the protected attributed, and the class respectively. In other words, this means that the probability of true positives to false positives should be the same for protected and unprotected groups. Through the investigation of where the bias emerges, we can begin to explore a diverse range of mitigation strategies to be analysed and evaluated.

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 MER system. From here, we experiment with different combinations of debiasing techniques for each modality and compare our metrics with recent literature.

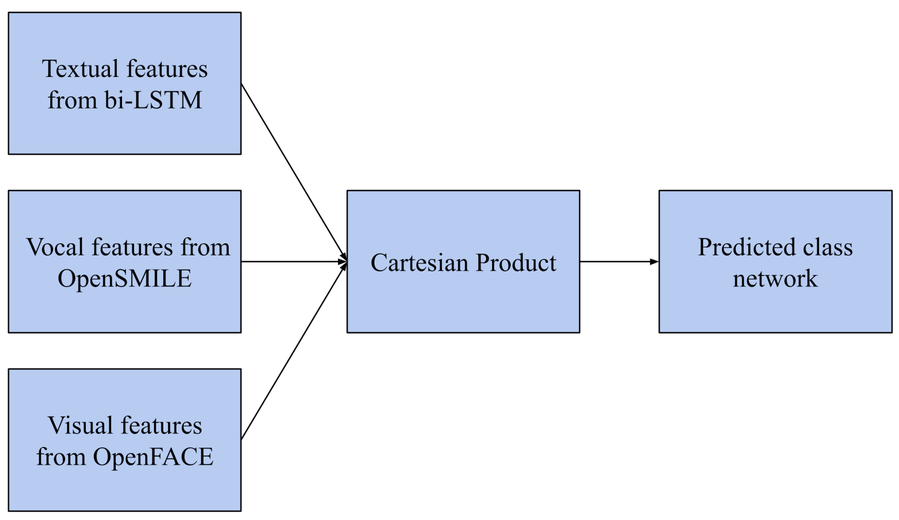
II - RESEARCH QUESTION & OBJECTIVES

How biased are the machine learning models used for multimodal emotion recognition and what affect do debiasing techniques have on these models?

- Reproduce architecture from modern literature which predicts an emotion class from a video-based dataset using textual, vocal and facial imaging information.
- Measure the quality of the architecture and the effect of the bias in the system and compare this to the state of the art.
- Perform baseline debasing techniques of gendered stereotypes for each modality in feature space and remeasure the metrics.
- Using a cycle consistency adversarial architecture, synthesise a single modality such that over time the stereotype becomes indistinguishable (Zhang et al. 2018), and then restore the original information by rebiasing with another adversarial network.

III - METHODOLOGY

Architecture - We develop a generic MER architecture to focus on debiasing the feature networks, which we combine via the cartesian product (Zadeh et al. 2017). The textual information is learned by using a pre-trained GloVe embedding network, passed through a bi-LSTM. Our visual features are gathered via the openFACE open source facial recognition library (Baltrusaitis et al. 2018), and inputted into a 3D convolution network to encode the temporal properties. Finally, we extract 6,552 low level descriptions of the audio from openSMILE (Poria et al. 2015), then inputted into an autoencoder to make the latent representation more meaningful.



Debiasing Approaches - By investigating a combination of the following techniques we can evaluate the collaboration of the methods. For textual features, we use the GloVe embedding debias method (Kaneko & Bollegala 2019) by building a latent representation of the dictionary such that we protect the meaning of masculine and feminine words, and remove the bias from stereotypical words. Therefore, this provides us with a loss function with 4 criteria, including the feminine words, masculine words, orthogonal gender neutral words and the reconstruction loss:

$$\mathscr{L} = \lambda_{\mathrm{f}}\mathscr{L}_{\mathrm{f}} + \lambda_{\mathrm{m}}\mathscr{L}_{\mathrm{m}} + \lambda_{\mathrm{g}}\mathscr{L}_{\mathrm{g}} + \lambda_{\mathrm{r}}\mathscr{L}_{\mathrm{r}}$$

For the convolutional information, we use the adversarial debiasing technique (Zhang et al. 2018). The adversary predicts the protected attribute, while the predictor network works as normal, but its weights are updated by:

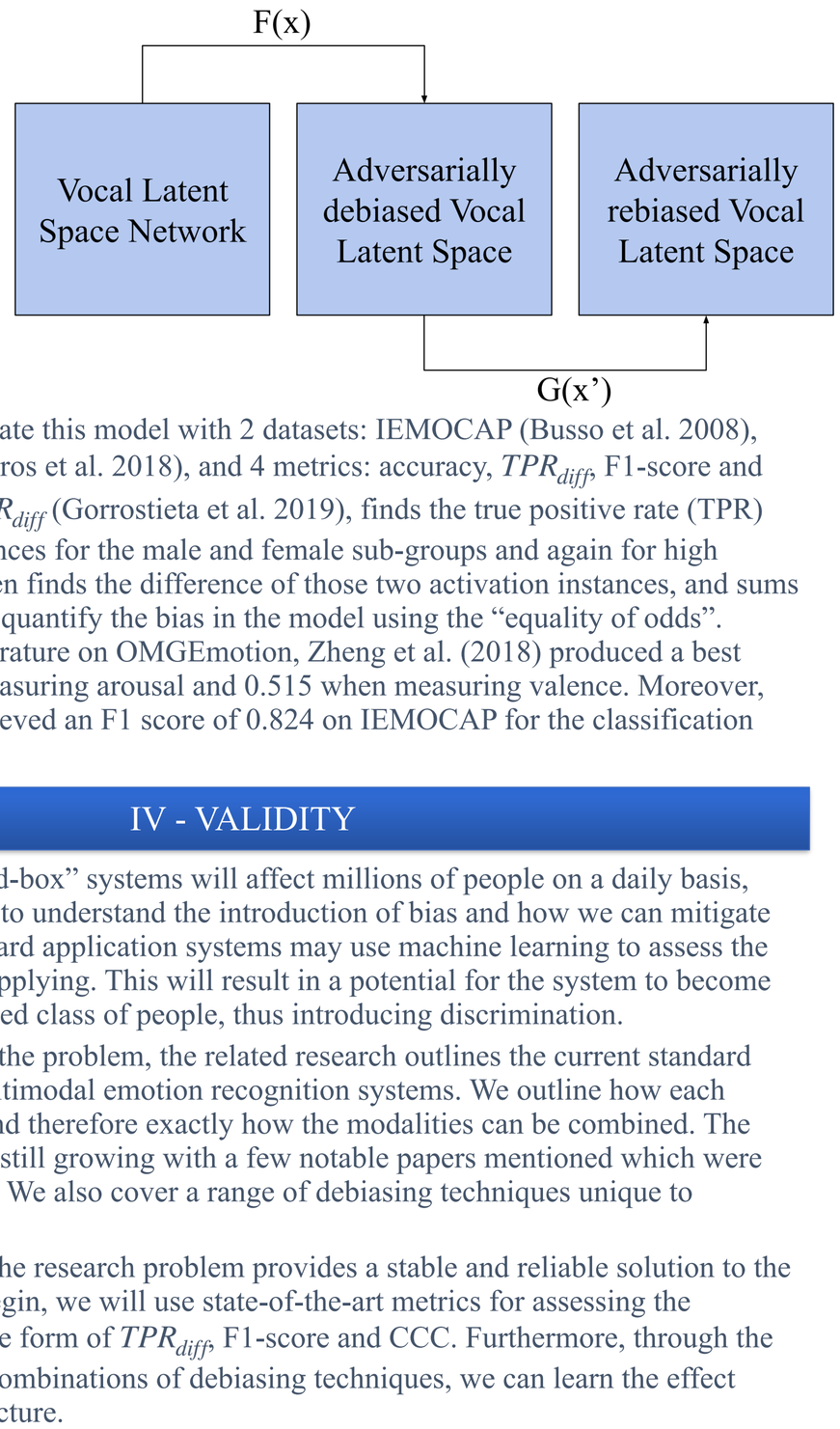
$$\nabla_W \mathscr{L}_{\mathrm{P}} - \mathrm{proj}_{\nabla_W \mathscr{L}_A} \nabla_W \mathscr{L}_{\mathrm{P}} - \alpha \nabla_W \mathscr{L}_{\mathrm{A}}.$$

After training, the predictor architecture should converge on the equality of odds definition. For the vocal data, we add an equality loss function:

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

In other words, this finds the recall of male and female classes to ensure that, over time, they are the same for low and high activations. (Gorrostieta et al. 2019). Finally, to help maintain the overall accuracy, each loss function will include a low weighted RMSE value.

Developing the adversarial debias method, we use a cycle consistency method to rebias the latent space and visualise the effect of the adversarial approach.



Evaluation - We evaluate this model with 2 datasets: IEMOCAP (Busso et al. 2008), and OMGEmotion (Barros et al. 2018), and 4 metrics: accuracy, TPR_{diff}, F1-score and CCC. Most notably TPR_{diff} (Gorrostieta et al. 2019), finds the true positive rate (TPR) for low activation instances for the male and female sub-groups and again for high activation instances; then finds the difference of those two activation instances, and sums them. This allows us to quantify the bias in the model using the "equality of odds". When comparing to literature on OMGEmotion, Zheng et al. (2018) produced a best CCC of 0.397 when measuring arousal and 0.515 when measuring valence. Moreover, Mittal et al. (2020) achieved an F1 score of 0.824 on IEMOCAP for the classification problem.

IV - VALIDITY

The bias in these "closed-box" systems will affect millions of people on a daily basis, therefore it is important to understand the introduction of bias and how we can mitigate it. For example, credit card application systems may use machine learning 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.

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 exactly how the modalities can be combined. The research for this field is still growing with a few notable papers mentioned which were released in March 2021. We also cover a range of debiasing techniques unique to specific modalities.

The process of solving the research problem provides a stable and reliable solution to the overarching issue. 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.

REFERENCES

[1] - Pinterest. 2021. Personalised Soundwaves Custom sound wave printCustom | Etsy in 2021 | Soundwave art, Sound waves design, Sound waves. [online] Available at: <https://www.pinterest.com/pin/76983474871823220/> [Accessed 30 April 2021].

[2] - Lenker, M., 2016. Leonardo DiCaprio's Oscar Speech Focuses on Climate Change - Variety. [online] Variety.com. Available at: <https://variety.com/2016/film/news/leonardo-dicaprio-oscar-speech-climate-change-1201717970/> [Accessed 30 April 2021].

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.

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.

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

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.

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

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.

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.

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