Ethics & Bias in AI Coursework

Nam Le
jssb25

I. Introduction

This report focuses on a machine learning task for emotion recognition on human face images. Whilst the concept may not be novel, it is a task which is useful in a variety of applications, such as in marketing, human-robot interactions, healthcare, as well as security [1].

A. Ethical impact and Value Sensitive Design

AI ethics refers to a set of values and principles that guide the responsible use of AI technologies [2], where as bias in computer systems refers to "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals" [3]. It follows therefore that an ethical impact aims to find and root out any potential biases in the model, allowing for fair outcomes for all individuals.

Doing so, however, is not a simple task. As such, we employ the use of Value Sensitive Design (VSD) to guide the development of the model.

B. Description of tasks

This is a classification task which involves predicting the emotion of a person from an image of their face, and outputting the emotion(s) associated with said image.

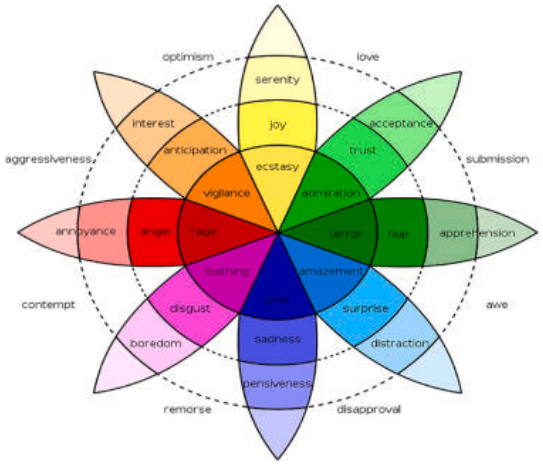

Figure 1: Plutchik's wheel of emotions, with the base emotions as well as their amplified/attenuated versions. Intensity increases towards the center and vice versa [4].

The range of possible emotions is typically decided by a dataset, though we aim to pick a dataset which contains the basic emotions of: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. According to Plutchik's wheel of emotions, which is a widely accepted model in discrete emotion theory, these are the basic emotions that are universally recognized [1].

In Figure 1 we do not use the full range of emotions as it is not practical to train a model to recognize all of them, and because the base emotions are sufficient for most applications.

Once we have decided on the classes, we need to decide on the output format of the model. We will use a softmax layer, which outputs a probability distribution over the classes. This has the benefit of giving more information than a single class output, can be used to calculate a confidence score, and is easily transformed into a positive/negative/neutral output. By choosing this output format, we allow for flexibility in the use of the model, which is crucial for a model that is to be deployed by various organization and/or be made available through open sourcing.

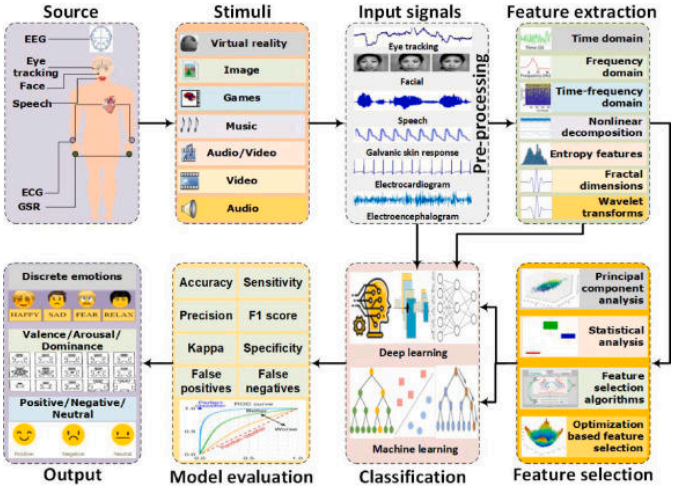

Figure 2: Typical process of training an emotion recognition model [1].

Figure 2 shows the framework we will use in training the model from start to end.

C. Use case and apparent issues

A potential use case of this model is in a medical clinic where it can be used as a healthcare surveillance system to monitor the mental health of patients. Specifically, the model could be used to detect signs of depression or anxiety in patients, and alert the clinic to administer medicine or provide support. Compared to traditional methods, such as observation from a nurse or doctor, the model could be more accurate,

less prone to human error and be more available to patients. Developments in this area can be seen in Marwan Dhuheir's works [5].

II. Paper

Emotions	Counts	% of total

Table 1: Counts of emotions after cleaning

III. Methods

References

[1] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations". 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523003354

[2] D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector". 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3240529

[3] B. Friedman and H. Nissenbaum, "Bias in Computer Systems". 1996.

[4] R. Plutchik and H. Kellerman, *Theories of Emotion, Vol. 1*. 2013. [Online]. Available: https://books.google.co.uk/books?id=Ie1GBQAAQBAJ&lpg=PP1&ots=MA0_sk5AGf&lr&pg=PP1#v=onepage&q&f=false

[5] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey". 2021. [Online]. Available: https://arxiv.org/abs/2107.05989