Ethics & Bias in AI Coursework

Nam Le
jssb25

I. Introduction (528 words)

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

A. Description of tasks

This is a classification task which predicts the emotion of a person from an image of their face, and outputs the emotion(s) associated with it.

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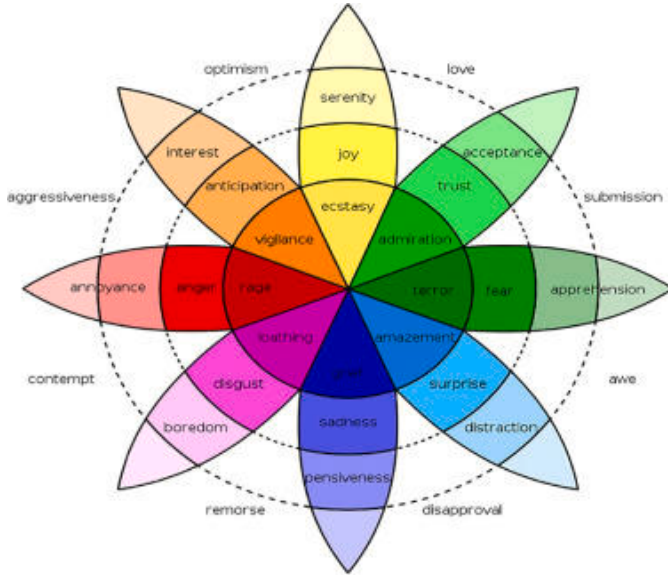

Figure 1: Plutchik's wheel of emotions, with the base emotions as well as their amplified/attenuated versions. Intensity increases towards the center and vice versa [2].

We do not use the full range of emotions as it is not practical to train a model to recognize all of them. The base emotions are sufficient for most applications.

For the output format, we want a probability distribution over the classes. This has the benefit of giving more information than single class outputs, can be used to calculate a confidence score, is easily transformed into a positive/negative/

neutral output and is easily done (softmax layer). With this, we allow for more model flexibility, which is crucial for deployment to various organization and/or be made accessible via open sourcing.

The dataset used for training is custom-made from online datasets, and the process of producing it will be elaborated in section III.

B. Ethical impact, Value Sensitive Design, and use case

AI ethics refers to a set of values and principles that guide the responsible use of AI technologies [3], whereas AI bias refers to "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals" [4]. An ethical impact assessment aims to root out any potential biases in the model by considering stakeholders' values/principles, allowing for fair outcomes for all.

To do this, we employ the use of Value Sensitive Design (VSD). Value in VSD refers to value, defined as "what a group of people consider important in life" [5], hence VSD is a methodology which considers the values of both direct and indirect stakeholders in the design of technology.

VSD involves three investigation steps.

1. Conceptual: Identifies stakeholders, how they are affected, discuss trade-offs between values,...
2. Empirical: Using quantitative/qualitative methods to expand on the concepts found in the previous step.
3. Technical: Analysis of existing technological mechanism on, and proactive design supporting of human values.

A potential use case of this model is for healthcare surveillance system [6]. In hospitals, the model could be used to detect signs of depression or anxiety in patients, and alert the clinic to administer medicine. Compared to traditional methods, such as human observation, the model could be less prone to error and be more available to patients. This technology could even be used in a public setting to detect signs of mental illness in a population.

In the hospital example, we must make sure the model is only used on images of patients who have given explicit consent, but how do we get the consent of people who are not mentally well enough to make decisions? Issues like this is beyond our scope of responsibility, but it is important to consider them.

II. Ethical impact assessment informed by VSD

In section II and III, we focus solely on healthcare use cases.

Stakeholders	Values	Potential risks/ harms
Healthcare providers (Direct)	Nonmaleficence, respect for human anatomy - The model's decisions should not be absolute, and the model's limits needs to be recognized	A doctor should not be punished with medical negligence and have their medical license revoked for not following the model's (incorrect) decisions
Patient (Indirect)	Informed consent, privacy - Hospitals expected to not record, store and process data unless given explicit consent	Personal data stored in hospital databases could be breached and leaked. Data could also be exposed to third parties
Medical institution (Indirect)	Human welfare, freedom from bias - Casualties must be minimized to maintain reputation so the existing system must work well for all	Biased diagnosis and preventable deaths could lead to lawsuits, loss of trust, tarnished reputation, and funding cuts (if government funded facility).
Pharmaceutical Company (Direct)	Accuracy - Production and distribution chains could be disrupted if output data has too many incorrect predictions	Impacted distribution could cause sending too much/little resources to hospital, leading to shortages/ wastage
Insurance companies (Indirect)	Explicability - The outcome of the model must be clear to prevent any ambiguity as to whether or not insurance must cover for a client	Disputes over coverage would deteriorate the relationship between the hospital and the insurance company, leading to a cancelled contract
Government institution, public health enforcer (Direct)	Human welfare, universal usability -	
Government institution, regulators and policymakers (Indirect)	Courtesy, trust, accountability -	
General public, (Indirect)	Privacy, autonomy -	

Table 1: Ethical impact assessment using VSD in a hospital environment

- Talk about the book [7] when discussing methodology of getting the values of stakeholders
- Values are from [5], [8]

The values used in Table 1 are from both traditional [5] and AI-specific [8] sources. We use an emergent methodology which focuses more on empirical investigations as opposed to conceptual investigations as the latter introduces researcher bias and tends to misrepresent the stakeholders' values [7].

III. Recommendations & Considerations

Choosing a dataset is the most important part of the process, especially in avoiding bias. The dataset should be representative, but most facial emotion recognition datasets simply are not. This can be seen in Figure 2, Figure 3, and Figure 4.
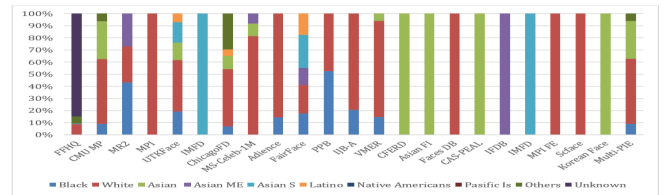

Figure 2: Racial composition in common facial emotion recognition datasets with "White" being the majority class [9].
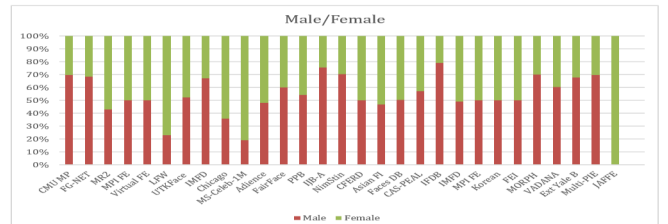
Figure 3: Gender composition in common facial emotion recognition datasets with "Male" being the majority class [9].
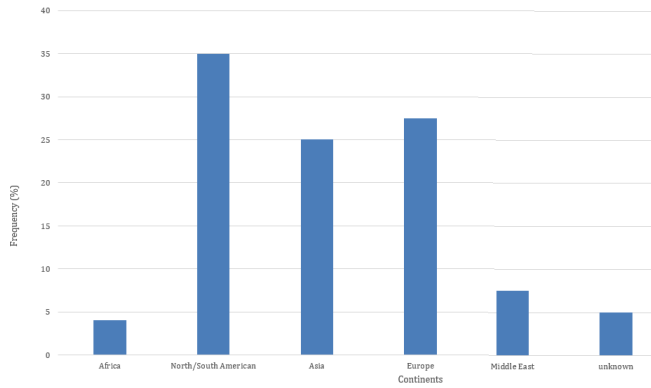

Figure 4: Source distribution of facial emotion datasets. North/South America, Asia and Europe make up 87.5% of all frequencies [9].

To deal with this issue, we can either choose the least biased dataset from the available ones, or aggregate multiple datasets to create a more representative one. We chose the latter despite its difficulty since thats the best way to ensure that the model is unbiased. An algorithm in [9] will be used to measure the bias in our new dataset.
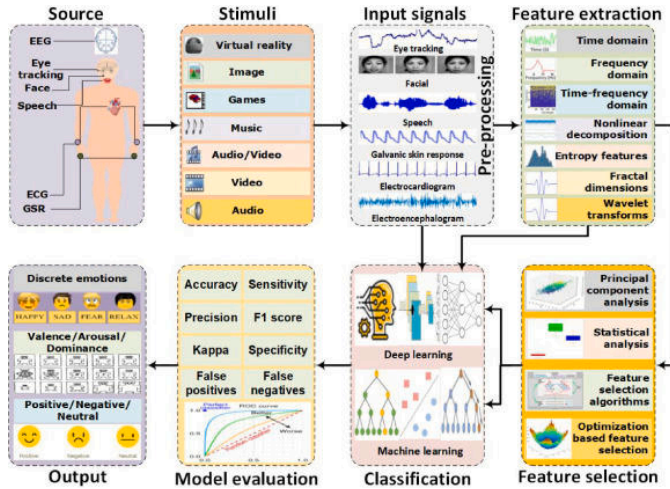

Figure 5: Process of training an emotion recognition model to be used for our model from choosing a dataset to outputting a verdict [1].

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] 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

[3] 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

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

[5] B. Friedman, P. Kahn, and A. Borning, "Value Sensitive Design and Information Systems". 2008. [Online]. Available: https://link.springer.com/chapter/10.1007/978-94-007-7844-3_4

[6] 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

[7] J. Davis and L. P. Nathan, *Value Sensitive Design: Applications, Adaptations, and Critiques*. 2015. [Online]. Available: https://doi.org/10.1007/978-94-007-6970-0_3

[8] S. Umbrello and I. van de Poel, "Mapping value sensitive design onto AI for social good principles". 2021. [Online]. Available: https://doi.org/10.1007/s43681-021-00038-3

[9] A. M. Udefi, S. Aina, A. R. Lawal, and A. I. Oluwafemi, "An Analysis of Bias in Facial Image Processing: A Review of Datasets". 2013.