Ethics & Bias in AI Coursework

Nam Le
jssb25

678 words

I. Introduction

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].

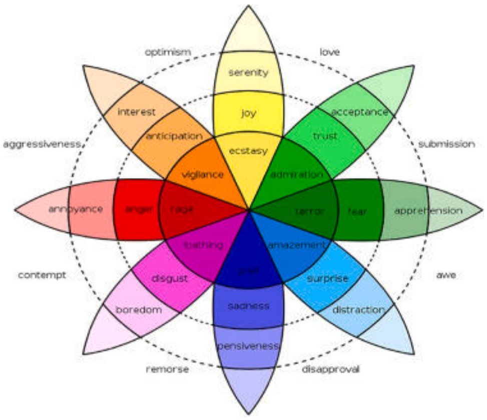

Figure 1: Plutchik's wheel of emotions, with the base emotions as well as their amplified/attenuated versions. Intensity increases towards the center and vice versa [2].

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, what values they hold, 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 depression of individuals 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.

In the following sections (section II and III), we focus solely on the healthcare use case in hospitals.

II. Ethical impact assessment informed by VSD

Stakeholders	Values	Potential risks/harms
Healthcare providers, such as a doctor or a nurse (Direct)	Respect for human anatomy, autonomy - • The model's decisions should not be absolute, and it's limits needs to be recognized • The model should act as an aid to help healthcare providers make their decisions, providing an additional point of view	A misdiagnosis might be given, in which case the healthcare provider will be forced to blindly follow the model's decision. If this results in harm to the patient, the healthcare provider could be held accountable for medical negligence, potentially leading to revocation of their medical license.
Patient (Indirect)	Informed consent, privacy - • A patient should have absolute authority over their data • Hospitals/medical institutions are expected to not record, store and process data unless given explicit consent	Personal data stored in hospital databases could be breached and leaked. In rare cases, institutions could (mis)use this data for their own benefit, such as selling it to third parties despite GDPR regulations. Bad actors could then use this data for nefarious purposes, such as identity theft or blackmail.
Medical institution, (Indirect)	Human welfare, freedom from bias - • Casualties must be minimized, and so the existing system must work well for all individuals despite their differences	Biased diagnosis, preventable deaths, and unequal access to the technology could lead to lawsuits and a tarnished reputation, and funding cuts (if it is government funded institution).
Pharmaceutical Company, model output data consumer (Direct)	Accuracy, environmental sustainability - • Production and distribution chains could be disrupted based on the model's output data. • Overproduction of medicine could also lead to resource wastage as well as environmental pollution	Bad quality data could cause the company to produce and distribute the wrong amounts/types of medicine, leading to a loss of profit. The company could also be held accountable for any environmental damage caused by overproduction.
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	A correct, but unexplainable diagnosis could lead to a refusal to cover the patient's medical expenses. Disputes over this would deteriorate the relationship between all parties involved.
Government institution, regulators and policymakers (Indirect)	Courtesy, trust, accountability - • A government's aim is to maintain the trust of its citizens, and so it must regulate use of the technology • The government must also be partially accountable for the technology's use and its consequences	Any technologies, in its inception, are prone to misuse [7]. If the government fails to regulate the technology, lots of inhumane Even with regulation, if the technology is misused, the government has a duty to take some responsibility for the consequences.

Table 1: Ethical impact assessment using VSD in a hospital setting.

The values used in Table 1 are from both traditional [5] and AI-specific [8] sources.

We use an emergent methodology to gather these opinions which focuses more on empirical investigations as opposed to conceptual investigations as the latter introduces researcher bias and tends to misrepresent the stakeholders' values [9].

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.

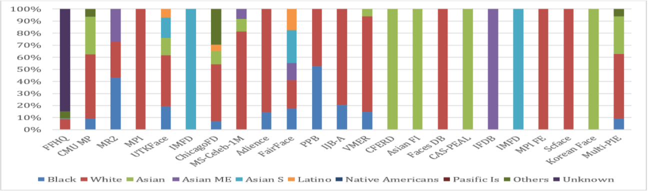

Figure 2: Racial composition in common facial emotion recognition datasets with "White" being the majority class [10].

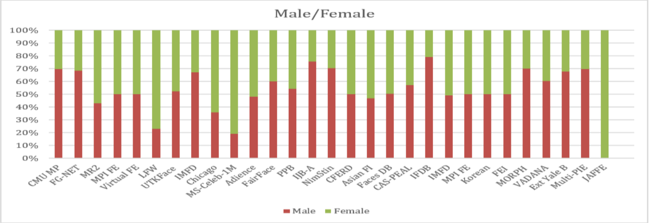

Figure 3: Gender composition in common facial emotion recognition datasets with "Male" being the majority class [10].

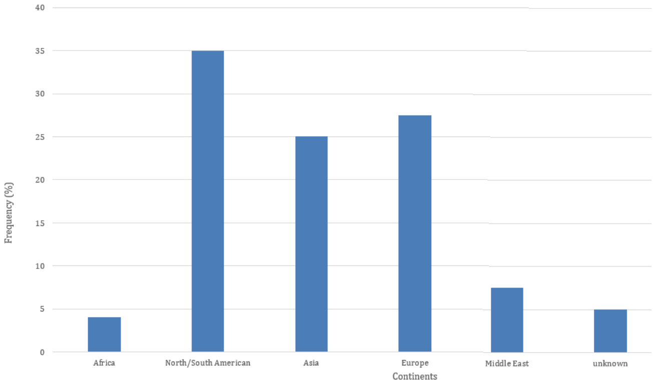

Figure 4: Source distribution of facial emotion datasets. North/ South America, Asia and Europe make up 87.5% of all frequencies [10].

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 [10] will be used to measure the bias in our new dataset.

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] T. Gebru *et al.*, "Datasheets for Datasets." 2018. [Online]. Available: https://arxiv.org/abs/1803.09010

[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] 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

[10] 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.