Ethics & Bias in AI Coursework

Nam Le
jssb25

I. Introduction (648 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, as well as security [1].

A. 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 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].

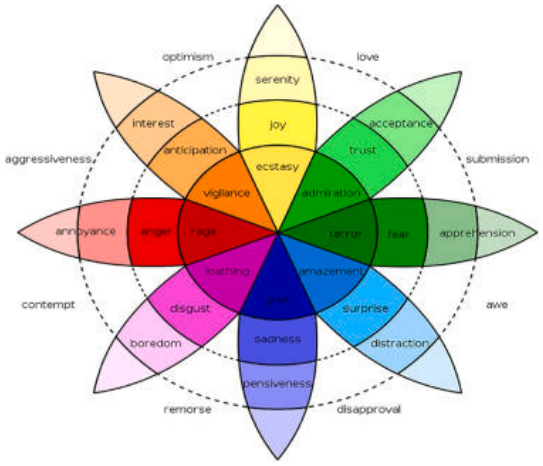

Figure 1: Plutchik's wheel of emotions, with the base emotions as well as their amplified/attenuated versions. Intensity increases towards the center and vice versa [2].

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. 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 a single class output, can be used to calculate a confidence score, is easily transformed into a positive/negative/neutral output and is easy to do (softmax layer). 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.

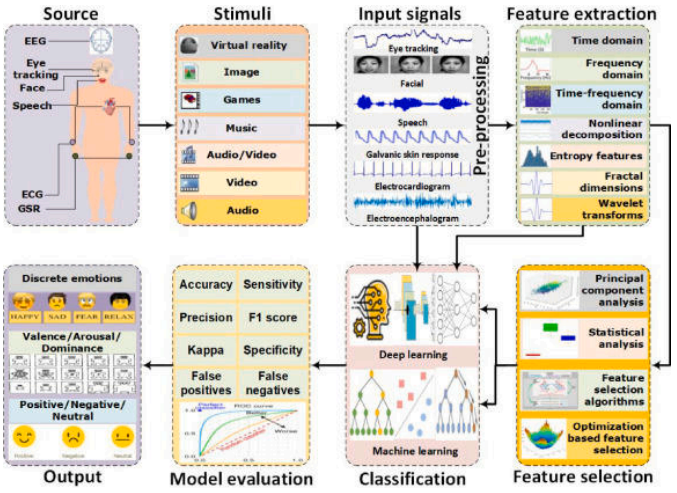

Figure 2: Process of training an emotion recognition model to be used for our model from choosing a dataset to outputting a verdict [1].

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 3, Figure 4, and Figure 5.

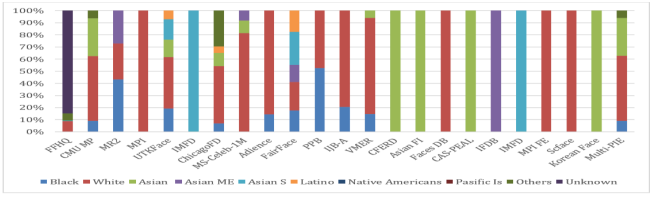

Figure 3: Racial composition in common facial emotion recognition datasets with the "White" majority class [3].

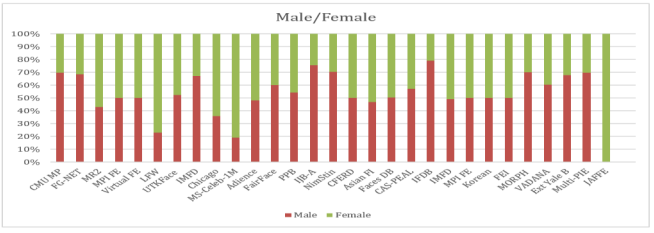

Figure 4: Gender composition in common facial emotion recognition datasets with the the "Male" majority class [3].

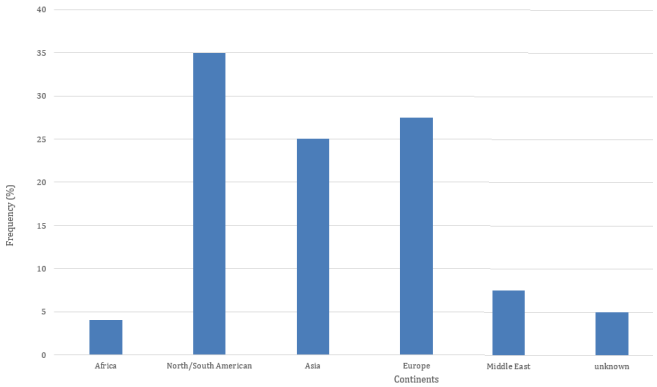

Figure 5: Source distribution of facial emotion datasets. North/ South America, Asia and Europe make up 87.5% of all frequencies [3].

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. The latter is more difficult, but is the best way to ensure that the model is not biased, and so we will therefore do this.

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 [4], whereas AI bias refers to "computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals" [5]. Therefore that an ethical impact aims to root out any potential biases in the model, 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" [6], 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 in a medical clinic where it can be used as a healthcare surveillance system to monitor the mental health of patients. 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. Developments in this area can be seen in Marwan Dhuheir's works [7].

An issue in this is consent. 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 (potentially) not mentally well enough to make decisions? Issues like this is beyond our scope of responsibility, but it is important to consider them.

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] 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] 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.

[4] 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

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

[6] 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

[7] 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