Bias Mitigation for Depression Identification

The word ‘fairness’ in machine learning is used in various ways. We distinguish between the unintended biases in a machine learning model and the potential unfair applications of the model. Every machine learning model is designed to express a bias. Existing classifiers unintentionally become biased when labeling conventional comments as depressed due to using identity keywords representing sensitive information, e.g., sadness, crying, etc. Language models also misclassify non-depressed sentences as depressed.

The ML model is not intended to discriminate due to the usage of specific keywords in a comment - so if the model does so, we call that unintended bias. We contrast this with fairness, which we use to refer to a potentially negative impact, mainly when normal individuals are treated as depressed. Initial versions of text classifiers trained on clinical data showed problematic trends for particular statements. Non-depressive statements containing specific identity terms, such as “My mother is sad,” can be given unreasonably high depressive scores. We call this false positive bias. The source of this bias was the disproportionate representation of identity terms in the training data: terms like “sad” were so frequently used in depressive comments that the models overgeneralized and learned to associate disproportionately those terms with the depressive labels.

To illustrate this distinction, consider a model for depression that has unintended bias at a given threshold. For instance, the model may give comments that contain the word ‘sad’ scores above the threshold independently of whether the comment is depressive. If clinicians use such a model to identify patients through online-delivered psychotherapy interventions, we might speculate that the model will harm society. Thus, we might say the model’s impact is unfair (to patients who wish to express their feelings). However, if the model is used to sort and review all comments before they are classified; then, the model’s unintended bias may not cause any unfair impact on clinicians.

[1,2] are among the first works to mitigate the unintended bias effects in such classifiers by balancing the training dataset. In this thesis, we will identify and mitigate unintended model bias using existing bias removal mechanisms.

Contact: **Sameen Mansha** <sameen@kth.se>
Distributed Computing Group, SCS Division, KTH
Examiner: **Prof. Vladimir Vlassov** <vladv@kth.se>
Distributed Computing Group, SCS Division, KTH

Technical Requirements:

- Familiarity with Python libraries, Text Classifiers, HuggingFace Transformers, etc.

References:

- Dixon,L.,Li,J.,Sorensen,J.,Thain,N.,Vasserman,L. **Measuring and mitigating unintended bias in text classification**. In Proceedings of the ACM AAAI Conference on AI,Ethics, and Society, pp.67–73(2018).
- Bailey,A.,Plumbley, M.D. **Gender bias in depression detection using audio features**. In IEEE European Signal Processing Conference (EUSIPCO),pp.596–600(2021).