Bias or Stereotype Detection in Online Humorous Content (SSB)

Humour is a prevalent aspect of human communication that can be found in a wide variety of written materials. As a result, humour classification and generation tasks have attracted significant academic attention in recent years. However, existing corpora used for these tasks often lack detailed labelling of the specific elements that contribute to humour. Additionally, humour is highly subjective, making it difficult to analyse how hateful or offensive messages are softened through humour to gain acceptance. Nonetheless, with the rise of social media and a plethora of textual content, it becomes increasingly important to address detection of stereotypes and hate speech in order to keep these forums safe and respectful across users of all demographics. To this end, we aim to build a text classification system that is able to detect the use of stereotypes or biased representations in humorous content extracted from a collection of jokes in English publicly available on Kaggle. This is done with an objective to gain a more robust understanding of online humor.

**Starting materials:**

* Datasets of jokes in English available on Kaggle annotated with three sets of labels: type of joke, target and rhetorical device used to deliver the underlying humor

**Initial Deliverables:**

* Produce a labelled dataset with a newly proposed annotation structure
* Build a text classification model via finetuning or prompt engineering that is able to identify the type and intended target of textual humor.

We want the performance of our model indicate the feasibility of our adopted annotation framework.