Group 2 - Project Proposal

M. Garzoni di Adorgnano (4866168), O. A. Kirschstein Schafer (1164783), G. Kooij (6608779), R.E. Lucas (6540384) & M. Pieke (8474752)

**Problem Description & Project Idea**

When black box models outperform humans, it can be useful to understand clearly which features are most relevant in order to assist humans to make more effective decisions in the future. For this purpose, we aim to compare human attention maps with machine generated attention maps to find out the extent to which the two systems pick up on the same features for their decisions. By attention maps, we mean maps which indicate which features play an important role in the final decision of either human or model. Specifically, we plan to use clustering techniques to find out on which groups of data the algorithms outperform humans and based on which aspects of the data this happens.

**Dataset**

We will use the data collected by Sen et. Al, which concerns restaurant reviews on Yelp. This is composed of 4 columns:

1. *Input.label*: here are the original Yelp labels for each review, converted to 0 (negative review) or 1 (positive review) - one and two-star reviews are mapped to 0, four and five-star reviews are converted to 1
2. *Input.text:* this related to the original review from the Yelp dataset
3. *Answer.Q1Answer*: this is the sentiment label collected from the annotator
4. *Answer.html\_output:* this is the human attention map (HAM) collected from the annotator

**Methods & Experimental Setup**

We will be using the input labels and text to train different classifiers (if we use this dataset, then we will train variants of RNNs). We will then use the same similarity metrics developed by Sen et. al (2020) to compare the machine generated attention maps with the human attention maps. Furthermore, we will segment the dataset based on its labels. For each of these segments, we will apply a clustering algorithm to find separate groups of semantically homogeneous groups/clusters. For each unique cluster, we will analyse the difference in performance between human annotators and SML algorithms using the metrics developed by Sen et.al. The clusters that show a significant performance difference will be analysed more thoroughly to get insights into the differences in the decision process.

**Biggest risk**

One of the problems that could arise is that we do not find a subset in our dataset where the model outperforms humans. This would mean that we do not have any data to analyse and explain the model for. If this happens, we will switch the attention to the opposite case, I.e., where humans outperform the model. If we cannot find such a subset, we will look at a different problem.

**Reference**

Sen, C., Hartvigsen, T., Yin, B., Kong, X., & Rundensteiner, E. (2020). Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? *Proceedings of the 58th Annual Meeting of the Association for*   *Computational Linguistics*, 4596–4608. https://doi.org/10.18653/v1/2020.acl- main.419