Thesis project

September 12, 2023

Project idea: The project aims to inspect the algorithm's interpretability within natural language processing (NLP). The first step to do this is to consider an ensemble of texts in which the polarization of sentiment is less as possible (i.e. a very long contract). Each sentence of the chosen texts will be considered one element of a time series: the first sentence t0, the second t1 an so on. For each sentence, a sentiment output will be associated with a number (positive or negative) through an NLP algorithm. The time series obtained in this way will represent the sentiment baseline (T0). The text will be modified by injecting some polarised terms with a certain probability according to a particular probability distribution (perhaps a sinusoidal one). In this way, another time series will be produced (T1). Our first goal will be to compare the time series obtained from the subtraction T1-T0 with the probability distribution used to inject the terms: we expect these two are quite similar, but we would quantify how much. If everything goes well, we will consider the time series associated with the neurons' weights.

First steps:

- Rapid inspection of the literature about XAI for NLP and if something similar was done (this part will be useful for the first part of your thesis)
- Selection of an ensemble of text sufficiently long that is less polarized as possible
- Choose a ML algorithm (or perhaps a set) that, given a sentence, returns a sentiment
- Implement it with Python (perhaps with COLAB)

Update meeting: Every Wednesday at 15.00 (generally). Small presentations or small report are requested in order to have a record of the work. A laboratory notebook in which you record your work is suggested (but not requested).