Thesis Topic Example 3

Investigating Bias in Pre-trained NLP Models

Keywords: bias, NLP

Background:
Popular NLP models such as GPT-3 and BERT are primarily trained using large amounts of unlabelled text and self-supervised techniques. This training data will doubtless contain large amounts of biased and prejudiced information which will in turn be learned by the model. This bias will then likely appear in the solutions of downstream tasks created by these models eg. Question answering systems, automatic text summarisation algorithms.
This project would focus on looking for concrete examples of prejudice and bias in these models, and investigations as to the precise effect of these biases on the complex systems built on top of these to solve downstream tasks. The project will also involve looking at possible solutions and strategies to avoid or correct these problems.

Aim:
The outcomes of this project are several concrete examples of clear bias present in these pre-trained models. Ideally these should appear as part of a large system implemented by the student, for example a question answering system built using BERT.
Mitigation strategies should also be included, ideally implemented by the student, but could also be discussed and collated in the final report.

The student is expected to gain a good practical knowledge of NLP and the various pre-trained models which are used nowadays, BERT, GPT-3 etc.
The student is expected to implement their own system to solve a complex NLP task and to display how this system might manifest biases present in the pre-trained model from which it is built.
The student should also investigate, and potentially implement, possible solutions for removing or reducing bias in NLP models.