Adversarial NLI

Addressing Dataset Biases
PoC: Douwe Kiela

The Adversarial NLI dataset (https://github.com/facebookresearch/anli) has been collected via a novel adversarial human-and-model-in-the-loop procedure, where annotators were tasked with fooling models. The dataset has made quite a splash already, but there is a lot of room for improvement. In particular, the collection procedure means that the training data may be a lot more noisy than for a usual statically collected dataset. The validation and test sets were properly validated, but this wasn't done on the training set because it would be too expensive. This capstone project would be around filtering training data to maximize performance on the validation and test sets. A good starting point is the ANLI repository itself, which has a codebase that shows how you can train various kinds of state-of-the-art models on the datasets. The hope of this project would be that we can figure out good ways to filter adversarially collected training data automatically. A good paper in the literature to start from would be AFLite https://arxiv.org/pdf/2002.04108.pdf, but simple heuristics like avoiding duplicates might already help.