Assignment 4 – Data Strategy

Data collection strategy:

The first thing that is important to get clear from the beginning, is that our samples will not be retrieved from us personally, instead we will use samples that have been already collected. Since we are trying to improve the capability of neural networks for the detection of adversarial toxic comments, collecting a huge amount of conversations would be impossible for us, and expect that in those conversations there are good examples of toxic comments would be even more optimistic. We plan to use a dataset of comments collected from Wikipedia's edit discussions. The dataset was prepared by Google's Conversation AI and was made available to the public through a Kaggle competition in 2018. This notebook provides a statistical overview of the dataset. We consider our sampling frame to be all edit comments available on Wikipedia.

Our sample will be representative, since we expect that with enough amount of conversations we will cover, of course not every situation of any context, but at least a great quantity of them, making the NN capable of understanding toxic comments in any different possibility and so representing as much as possible any conversation. Due to this our sample frame must match only two specific pre-requisites, apart of course from being English speakers, since our research is mainly done for English NLP, it should contain a decent amount of toxic conversations.

As I said the samples have already been collected (due to this , the time for data collection at least for us is minimum), but indeed inside this sample we expect to find some level of toxicity, therefore our data strategy for selecting the samples will be purposive sampling [2.c, p 1] since we are not going to discriminate inside the data with any specific filter, and any member of the population will have the same chances of being selected for the research, but this dataset has been selected by us due to its specific situation since it's known that it contains toxic comments that have been pre-rated by humans.

Between probability sampling and non-probability sampling following [1.c, p 15], our choice is non-probability sampling. The main reason extracted from this source is "it is not feasible to include a sufficiently large number of examples in the study", the cost would be too high and we wouldn't have any confidence on our dataset that it contains enough toxic comments. And for the last, as we said on "data collection methods" we have chosen our informants/dataset since it has been specifically developed for tasks of this topic (detection of toxic comments).

Finally, we identify topics of discussion as a potential bias of our dataset. Since our data is collected from a single source, we expect toxic sentences used to be relevant to particular topics (e.g., political differences). This may affect the generalization of our model especially in production environments. However, since this dataset has been used extensively in research ([4], [5], and [6] are only few examples), we believe this bias does not pose a serious threat to generalization and is thus of minimal importance.

References for data collection strategy:

[1.c] Denscombe, M. (2014). The good research guide: for small-scale social research projects. McGraw-Hill Education (UK).

[2.c] Kandace Landreneau. Sampling strategies:
http://www.natco1.org/research/files/SamplingStrategies.pdf

[3.c] Marshall, M. N. (1996). Sampling for qualitative research. Family practice, 13(6), 522-526.

Dataset

Replace sampling with 'data collection strategy' in the questions below:

- How exactly will the proposed sampling strategy relate to the aims of this research study? (It is important to make sure that this kind of sampling will be suitable for meeting the aims.)
- If the sampling is claimed to be "random", how is randomness ensured? (It is important to understand that "random" has a special meaning in sampling terminology, different from its ordinary meaning.)
- If the sampling is claimed to be "representative", who / what population is it representative of?
- What is the name of the exact sampling technique that is going to be used? (The text should also have a reference to an exact page number in methodology literature (Denscombe, Randolph, Creswell, etc.) where that technique is explained.)
- If there are exact numbers in the sampling strategy (e.g. "at least 7 people will be interviewed"), is that number just made up or is it justified in some way? (No "made up" numbers should exist.)
- If the study uses automated collection of instrument data (e.g. software inputs/outputs), what kinds of biases are there, if any? (Very often the kinds of inputs planned for the test already bias the test so that it is not fair.)
- How long do you expect the data collection will take if done like this?
- If the study is a qualitative study, what kinds of criteria does the author set for data saturation?
- Looking at the purpose of the sample (representative / exploratory), how suitable is the actual sampling technique for that purpose?
- What biases has the author identified with the selected sampling strategy, and what other biases are there?