**Meeting 15/10/20 Notes**

* Look at language models in different areas e.g. financial/biomedical (look at links sent)
* Combine datasets – fine tune transformers, filter out non-toxic…
* Toxicity subtypes have a lot in common but very subjective (annotators may not agree if comment is insult/personal attack/offensive)
* Think about solution design – how used/where and impact (embed feature on Twitter, target feature for vulnerable people?)
* Demographic info not widely available (especially location), might be able to collect data based on level of education (more educated means more exposure to different subgroups and more likely to promote diversity) – can include paragraph in paper on how demographics info on annotators could help understand unconscious bias
* Look at relationship between offensiveness and constructiveness – constructiveness could dilute the likelihood of an attack and increase sensitivity of classifier – look at when personal attacks develop
* Create table summarizing main ideas and paragraph on feasibility (is the data available, could we get the data if needed?), look at if addressed in literature (indirectly) – see if can add to literature and merge datasets – compensate for disadvantages, what’s most impactful, what is the most interesting – decide on idea next week
* What is needed to make ideas more feasible – some funding for annotations for test set?
* (Optionally – continue looking at more papers)
* Send ethics form