Copy your research description from the previous weeks. Make sure you've edited it according to the comments that you've got from your peers. Be sure that there are a number of relevant, academic references at this stage. Also start to describe "limitations" of your study (it will develop further throughout the study).

Background and problem statement:

Cyber bullying and online harassment threaten the prosperity and existence of online communities [1]. According to a 2014 study conducted by Pew Research, 40% of internet users have faced harassment and 73% of users have seen others get harassed [2]. Human content moderation is neither scalable nor effective, which increases the need for automatic detection of abusive language. Deep Learning methods are reported to have the best performance on this task [5]. However, these methods can easily be fooled by systematically modified syntax, known as adversarial examples. A model that assigns a toxicity score of 90% to the sentence 'you're an idiot' may give a score as low as 20% to its adversarial counterpart 'you're an i d i o t'. More generally, adversarial sentences are highly toxic sentences receiving a low score because of a non-semantic modification of their text [3].

This project aims to explore methods to increase text-classifiers' resistance to such examples. Previous studies attempted to identify the types of text transformations most likely to affect toxicity scores ([3]). Other studies outlined general solutions to overcome this issue ([7]), some involving human intervention ([6]). This study will test the effect of one solution, adversarial training, on improving the robustness of deep learning models. If time permits, the study can also extend to test the effect of character-based embedding on the same task. The base model for this study would be a Recurrent Neural Network with attention mechanisms, as it is reported to have a state-of-the-art performance on non-adversarial datasets [4]. The dataset used in the study is a set of comments from Wikipedia's edit discussions, published by Conversation AI through a Kaggle competition in 2018. The project would report on any achieved results and point to areas for improvement to guide future research on this topic.

Project aim:

Improve the resistance of automatic toxicity detection models to adversarial attacks.

Objectives:

- Use adversarial training to increase the robustness of an attention RNN model for toxicity detection.
- Make the model and training procedure available as an open-source project for future research.
- Report on the approach(es) explored for adversarial training and the achieved performance (measured in accuracy of prediction).

References:

[1] O'Brien, D., & Kayyali, D. (2015, January 22). Facing the Challenge of Online Harassment. Retrieved January 23, 2020, from https://www.eff.org/deeplinks/2015/01/facing-challenge-online-harassment

[2] Bottino, S., Bottino, C., Regina, C., Correia, A., & Ribeiro, W. (2015). Cyberbullying and adolescent mental health: systematic review. Cadernos De Saúde Pública. doi: 10.1590/0102-311X0003611

[3] Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017, February 27). Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv:1702.08138

[4] Kohli, M., Kuehler, E., & Palowitch, J. (2018). Paying attention to toxic comments online.

[5] Georgakopoulos, S. V., Vrahatis, A. G., Tasoulis, S. K., & Plagianakos, V. P. (2018). Convolutional Neural Networks for Toxic Comment Classification. In SETN '18: Proceedings of the 10th Hellenic Conference on Artificial Intelligence. New York, NY: Association for Computing Machinery.

[6] Dinan, E., Humeau, S., Zhang, B., Chintagunta, B., & Weston, J. (2019, August 17). Build it Break it Fix it for Dialogue Safety: Robustness from Adversarial Human Attack. arXiv:1908.06083

[7] Mehdad, Y., Tetreault, J. (2016, September). Do Characters Abuse More Than Words? Conference: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. doi: 10.18653/v1/W16-3638

Other potential references:

[8] Chakrabarty, N. (2012). A Machine Learning Approach to Comment Toxicity Classification

[9] Mehdad, Y., Tetreault, J. (2017, September). Deeper Attention to Abusive User Content Moderation. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. doi: 10.18653/v1/D17-1117

Limitations:

Limitations of this study include access to a diversified dataset for comment classification. The only one we found so far with enough use is Conversation AI's. The applicability of our study may also be limited to English speaking environments, since we do not intend to achieve similar results in other languages for now. We also face time limitations due to the relatively short span of this course/project, which is why we tried to narrow down our research to a single model and a single counter-adversarial method. Finally, the lack of experience of our team in this topic may also be a hindrance to the timely development of the project.

Types of information needed:

1) A formal definition and/or taxonomy of toxic language, cyberbullying and online harassment.

2) An overview of text-classification models and a literature-supported justification of the chosen model (Recurrent Neural Network with Attention Mechanisms).

3) Dataset of online comments for training and testing the model.

Data collection methods:

Information needed in 1) and 2) can be collected through literature review in the field of NLP, particularly in tasks relevant to the detection of toxic language. This type of data would be qualitative, and its use would depend on the understanding and interpretation of the researchers. Since the data is available through published reports, our data collection method in this case is *document studies.*

For the information mentioned in 3), we can use *observation studies* or *document studies* [1.b]. Observation studies would allow us to collect data about the online behavior of a group of participants in a non-invasive study environment. Document studies would be existing online comments collected from the public profiles of consenting participants. Since we are solely interested in natural language texts generated online, and not in the behavior of participants in such situations, we believe document studies to be more suitable for our research. It is also notably easier to organize this data collection scheme given our constraints in time and resources.

One variant of document studies we can use is collecting data (online comments) directly from a group of participants, then having said data labeled by another (larger) group of participants based on the level of toxicity of the text. This process is time consuming and requires resources we may not have access to (especially for the labeling step). We therefore opted for a ready-made publicly available dataset, collected by Conversation AI, and published by Kaggle through a data science competition in 2018. This dataset has been used extensively in research like ours (see [4], [6] and [8] from the reference list).

We chose a quantitative data analysis method, since our research uses statistical methods to assign numerical scores to online texts based on their level of toxicity. [2.b]

References for research methods:

[1.b] Johannesson, P., & Perjons, E., (2012, p.30). A design Science Primer. CreateSpace Independent Publishing Platform.

[2.b] Johannesson, P., & Perjons, E., (2012, p.31). A design Science Primer. CreateSpace Independent Publishing Platform.