Plan of work (appended to thesis contract)  
**Name of student** *Michiel Anthonie van Nederpelt*  
**Degree programme** *Text Mining*

**student number** *2684581*  
**Name of supervisor(s**)/ *Ilia Markov*  
**Chair** ......................................................................................  
**Second reader** ......................................................................................  
**Chair** ..... .........................................................................................  
**Working title** *F\*ck your !ntelligence- on our way to general intelligence*  
**Topic *F\*ck your !ntelligence*** *– on our way to general intelligence/opening pandora’s black box of Neural Networks- a case study in state of the art in- and cross-domain offensive language detection model performance with checklists*

**Aim and relevance** *an overview of artificial intelligence in society/business/research with a case study in state-of-the-art offensive language detection models with templates*  
**Problem definition** *How are state-of-the-art language models performing in-domain and cross-domain datasets***Data-collection** *data for hate-speech provided by Ilia Markov:*

* **Dataset 1: OLIDv1** dataset for offensive language detection:
* Data: <https://github.com/idontflow/OLID>
* Paper: <https://aclanthology.org/N19-1144.pdf>

This dataset was used in the SemEval 2019 shared task on offensive language detection ([OffensEval 2019](https://aclanthology.org/S19-2010.pdf)).

Focus on **Subtask A** (identify whether a tweet is offensive or not). The dataset is  preprocessed so that label ‘1’ corresponds to offensive messages (‘OFF’ in the dataset description paper) and ‘0’ to non-offensive messages (‘NOT’ in the dataset description paper).

* **Dataset 2: HASOC** dataset: <https://dl.acm.org/doi/pdf/10.1145/3368567.3368584>

The dataset was pre-processed (hasoc.csv) in the same way as the OLIDv1 dataset: label ‘1’ corresponds to hateful/offensive tweet and ‘0’ to non-hateful/non-offensive tweet.

* **Another dataset (instead of HASOC)?** E.g., <http://nlp.uned.es/exist2022/>

*Data and opinions for the AI overview through interviews and published papers*

**research method** *colab for hate speech perfomance*  
**Provisional organization of chapters**

* *Machine learning in business (strategy, decision making, Porter’s framework)*
* *Natural language processing*
  + *Focus on deep learning*
  + *Data  (short section)*
    - *Sexism in data*
    - *“old” society norms*
    - *Quality of data*
  + *Hate speech as a “deep dive” for one of the ethical components of NLP*
    - *How is big tech dealing with hate speech on their platforms*
    - A (small) NN system design for a (basic) analysis of a dataset
      * Run your notebook on [colab](https://colab.research.google.com), which has (limited) free access to GPUs.
      * You need to enable GPUs for the notebook:
      * navigate to Edit → Notebook Settings
      * select GPU from the Hardware Accelerator drop-down
      * Install the [simpletransformers library](https://simpletransformers.ai/): *!pip install simpletransformers*
      * (you will have to restart your runtime after the installation)
      * Follow the [documentation](https://simpletransformers.ai/docs/usage/) to load a pre-trained BERT model: ClassificationModel('bert', 'bert-base-cased')
    - Challenges and ethical discussion on platform censorship
  + *Write an article for the general public (in a newspaper/blogpost or similar) to make overall NLP less of a black box for the general public*
* *Interview multiple people in large companies through my network and make some sort of overview of challenges in Data Science and/or algorithmic processing. In an ideal situation get input from a (large) variety of companies/governmental organisations. From Quantum black (McKinsey) to the Dutch intelligence services (AIVD), from the Nederlandse Bank to a MIT professor.*
* *Robotics, the future- possibilities and dangers-*
* *Artificial intelligence in business, society and universities (maybe challenges for universities vs. big tech like google (Quantum computing, resource availability etc.))*
* *The future of artificial intelligence -> towards general intelligence (start from “what is intelligence” to the overall timeline and challenges in general intelligence*

**Provisional book list**   
**(appendix)**

*MIT course – “artificial intelligence- implications for business and strategy”*

*Articles and supplementary resources for AI overview:*

* Gardner, Howard. 1983. *Frames of Mind: The Theory of Multiple Intelligences.*New York: Fontana Press.
* Kiron, David. 2017. “What managers need to know about artificial intelligence.” *MIT Sloan Management Review,* January 25, 2017. http://sloanreview.mit.edu/article/what-managers-need-to-know-about-artificial-intelligence/
* <https://towardsdatascience.com/how-does-machine-learning-work-6dd97f2be46c>
* <https://towardsdatascience.com/an-introduction-to-deep-learning-af63448c122c>
* <https://towardsdatascience.com/deep-learning-for-computer-vision-c4e5f191c522>
* <https://towardsdatascience.com/gangogh-creating-art-with-gans-8d087d8f74a1>
* <https://www.theverge.com/2018/10/16/17985168/deep-learning-revolution-terrence-sejnowski-artificial-intelligence-technology>
* <https://www.youtube.com/watch?v=rm_SP9bzILk> - A.I. and the Future of Work, Daniela Rus - WGS 2018
* <https://www.seattletimes.com/business/technology/tech-giants-pile-up-voice-recordings-so-computers-can-listen-and-learn/>
* <https://edition.cnn.com/2019/03/07/tech/ai-voices-sound-human/index.html>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/>
* <https://www.amazon.science/publications/tico-19-the-translation-initiative-for-covid-19>
* <https://www.techrepublic.com/article/grammys-go-high-tech-with-ibm-and-artificial-intelligence/>
* <https://mitsloan.mit.edu/ideas-made-to-matter/why-finance-deploying-natural-language-processing>
* <https://www.automationanywhere.com/company/blog/product-insights/cognitive-automation-4-keys-to-success>
* <https://www.uipath.com/rpa/robotic-process-automation>
* <https://www.assetfinanceinternational.com/index.php/technology/technology-archive/technology-articles/19778-use-cases-for-robotic-process-automation-rpa-in-equipment-and-auto-finance>
* <https://www.news24.com/news24/columnists/guestcolumn/opinion-service-delivery-issues-robotic-process-automation-can-help-20201124>
* <https://www.nice.com/guide/rpa/what-is-intelligent-process-automation-ipa/>
* <https://www.digitalcommerce360.com/2020/01/08/walmart-unveils-robot-run-warehouse-for-online-grocery-orders/>
* <https://www.roboticsbusinessreview.com/retail-hospitality/5-robots-grocery-stores-now/>
* <https://news.mit.edu/2018/soft-robotic-fish-swims-alongside-real-ones-coral-reefs-0321>
* <https://www.medicaldesignandoutsourcing.com/12-surgical-robotics-companies-you-need-to-know-2020/8/>
* <https://www.theverge.com/21283724/wearable-robotic-arm-supernumerary-research-sherbrooke>
* <https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/where-machines-could-replace-humans-and-where-they-cant-yet>
* <https://www.weforum.org/agenda/2020/09/short-history-jobs-automation/>
* <https://time.com/5876604/machines-jobs-coronavirus/>
* <https://www.visualcapitalist.com/charting-automation-potential-of-u-s-jobs/>
* <https://www.mckinsey.com/featured-insights/future-of-work/how-will-automation-affect-jobs-skills-and-wages#>
* <https://hbr.org/2017/04/thinking-through-how-automation-will-affect-your-workforce?referral=03759&cm_vc=rr_item_page.bottom>

For the case study:

* **Dataset 1: OLIDv1** dataset for offensive language detection:
* Data: <https://github.com/idontflow/OLID>
* Paper: <https://aclanthology.org/N19-1144.pdf>

This dataset was used in the SemEval 2019 shared task on offensive language detection ([OffensEval 2019](https://aclanthology.org/S19-2010.pdf)).

**2. Check list:**

* <https://homes.cs.washington.edu/~marcotcr/acl20_checklist.pdf>
* <https://github.com/marcotcr/checklist>
* <https://arxiv.org/pdf/2012.15606.pdf>

**Additional literature:**

Error analysis: <https://arxiv.org/pdf/1809.07572.pdf>

Typos: Spelling variations are sometimes used adversarially to obfuscate and avoid detection ([Vidgen et al., 2019](https://aclanthology.org/W19-3509.pdf))

Negation: Offensive language detection models have been shown to struggle with correctly classifying negated phrases such as “I don’t hate trans people” ([Rottger et al., 2021](https://arxiv.org/pdf/2012.15606.pdf))

**Timetable**  
• start date *April 4th*

• orientation phase *April 4th – 24th of April*  
• execution phase *April 24th –* *may 24th*   
(submission dates for chapters) ......................................................................................  
......................................................................................  
**• completion phase** *may 24th – end of june*  
**• planned thesis presentation date**.................................................................................  
**• planned graduation date** *August*  
**Remarks**

*Challenges: cross-domain, overestimated performance on held-out-date*

*Experimental setup:*

* *In-domain: Train OLID (train.csv) – test OLID (test.csv) + checklist*
* *In-domain: Train HASOC – test HASOC (same label distribution as in OLID test.csv) + checklist*
* *Cross-domain: Train OLID (train.csv) – test HASOC (test) + checklist*
* *Cross-domain: Train HASOC (train) – test OLID (test) + checklist*