TU Wien

Data Science

# Applied Deep Learning

**- Course Project -**

*Authors*
Johannes Bogensperger

October 29, 2019

# 1 Overview

For international research and intelligence the Austrian Armed Forces employ multiple Full-time equivalent(FTE) soldiers/conscripts which label articles of interest. The goal is to enable officers to receive an overview of current developments in relevant categories e.g. cyber-incidents. In order to achieve this, the mentioned soldiers label articles with predefined keywords out certain categories. To this moment there are approximately 140.000 articles in a variety of languages labelled.

The purpose of the purpose of the project is to build a deep learning model to label articles of various international papers with keywords in order to present the current development of multiple domains of interest. This supervised classification task will executed using a innovative natural language processing (NLP) framework from Google which is called Bidirectional Encoder Representations from Transformers (BERT). BERT enables transfer-learning for NLP tasks and outperformed the state-of-the-art models on eleven natural processing tasks [2].

BERT is used for a variety of text classification and NLP task of which these cited papers are used in a highly similar field of application [1] [3].

The final model could be embedded in the labelling process as recommender system for ease of use or in bulk loading and labelling existing sets of articles.

The type of project would classify as a **NLP-Project** and in particular a Text classification problem. The dataset is not publicly available and is not preprocessed or prepared for this application. Therefore it classifies as **bring your own data**.

## 1.1 Data Source

The full dataset consists out of 140k of classified articles in a variety of languages in XML format. Each item contains:

- Language: the language in which the original article was published. But the texts itself are all either german or english.
- Title: the title of the article
- Content: The full text of the article which is either english or german.
- Label: The labels added manually by miltary personnel e.g. Cyber Crime, Gesellschaft, Wirtschaft, Sicherheitslücke or similiar keywords which are either german or an anglicism.
- various: there are multiple additional fields as geolocation, date, time, region, author which will be excluded.

# 2 Project schedule

| Description | Completion Date | Hours |
|---|---|---|
| Understanding of the Basics of BERT | 14.11.2019 | 8 |
| Meetings with AAF to gather all necessary data and understand the domain | 14.11.2019 | 4 |
| Parse and preprocess XML files | 21.11.2019 | 8 |
| Investigate and set-up optimal model architecture and run initial training runs | 05.12.2019 | 10 |
| Fine tune Model, consider possible improvement possibilites and evaluate performance | 12.12.2019 | 8 |
| Deploy model and integrate it in a basic webfrontend | 09.01.2019 | 16 |
| Documenting the final solution + Readme | 16.01.2019 | 2 |
| Craft final report and presentation | 22.01.2019 | 10 |
| | Sum | 66 |

Sidenote: I know this number of hours might not reflect the 45 hours planned in TISS for this course, but I consider it unlikely that I will be able to produce a result which satisfies my expectation/AAF needs in this time.

# References

[1] Saadullah Amin, Günter Neumann, Katherine Dunfield, Anna Vechkaeva, Kathryn Chapman, and Morgan Wixted. Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert. 09 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Jihang Mao and Wanli Liu. Factuality classification using the pre-trained language representation model bert. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)*, 2019.