Create SuperMBERT

- **Supervisor**: Ayyoob Imani
- **Examiner**: Prof. Schütze
- **BSc, MSc, Open**: MSc/BSc
- **Summary**: MBERT (multilingual BERT) is the latest progress on multilinguality, where all languages are represented using one model in one vector space. For example, if we take the vectors for the English word ‘apple’ and the German word ‘Apfel’ from their corresponding unilingual BERT, there will be no relationship between the vectors. But the vectors MBERT generates for 'apple' and 'Apfel' are similar since they both correspond to the same concept but in different languages. This has revolutionized cross-lingual transferability, i.e. when we finetune BERT to solve a task (like sentence classification) for one language (e.g. English), we can use the same finetuned model to solve that task in any other language, with no need for training data in other languages. MBERT currently supports around 100 languages, which is a small portion of all the languages that exist. In this project, we aim to increase the number of supported languages to more than 1000. This may sound trivial since we can basically do the same thing that we did for those 100 to the other languages. But unfortunately, this is not the case since the same amount of resources does not exist for all other languages. Therefore we need to come up with innovative ways to deal with the lack of resource problem. The biggest challenge is that we don't have enough resources for these languages, and to overcome it, we have to make the best out of the little __but precious__ amount of resources we have. We have to make a few changes to BERT's architecture to make its training more efficient and also come up with better training procedures to make our goal achievable.
- **Prerequisites**: enthusiasm, Good programming background (preferably python), basic knowledge of NLP, DL, and Pytorch