# Part 3 multilingual modeling

First of all, the difference between cross-lingual and multilingual modeling should be made clear. Cross lingual models are trained in one language and then adapted to other languages. Which means that a model might be trained in the English language, but used for Dutch. Therefore, the model transfers knowledge from one language to the other language. This information regards structure and patterns that might apply across multiple languages. Multilingual models are trained on a dataset that contains multiple different languages and used on a dataset that contains one or more of these training languages. For example, a model is trained on Dutch and English and used to classify Dutch texts. This has been done in the notebooks as well.

To answer whether the methods will scale to German or French a deeper dive into the inner workings of deep learning and the machine learning model should be taken. The assumption is made that we use the model that is trained in English and Dutch to predict the OCEAN classes of text in French and / or German, making it a cross-lingual model.

From a theoretical perspective it is expected that the deep learning approach (transformers) can scale to languages like German or French, especially if the model is trained on Dutch and English. Without going into too much history, all languages are fairly related in terms of sentence and word structure. Dutch, English and German share an origin in Germanic languages with English being highly influenced by French in the 11th century. Therefore, with Dutch being closely related in terms of structure to German and English to French, this might mean that training a multi-lingual model on both Dutch and English can result in good results. The power of BERT is to capture general language patterns and semantics during (pre)training. The different language patterns and semantics are represented by the Dutch and English language. The transformers work with a vocabulary and map words, or parts of words, to a token. If a word is encountered that is not familiar to the model, it is not mapped. Therefore, very similar words are taken into account. It could be that the part that is mapped corresponds to the token of the other language, however, this is difficult to say.

The machine learning methods that are applied are XGBoost and SVM and are unlikely to scale from Dutch and/or English to French and German. XGBoost uses preprocessors like TF-IDF, Part Of Speech tagging, and Bag of Words to preprocess the text data. TF-IDF makes a term-document matrix which takes counts words literally and calculates its weight. Words that are language-independent might be captured by the model, but most of it will not. POS involves grammatical categories such as nouns, verbs and adjectives. The problem here is that these words are language specific and thus it assigns the labels on specific words. The same problem occurs with TF-IDF. Lastly Bag of Words counts the times a word occurs, thus making it difficult to make a model cross lingual. Again, language independent words make it through, but the rest is difficult to capture.

For both machine learning and deep learning methods, there are certain challenges to overcome when classifying multiple languages versus just English. First of all, the model has to learn multiple vocabularies, semantics and text structures. The model needs to capture language-specific nuances while having the same representation space. This is a challenge that should be overcome. Another challenge is that there is much more English text available online than any other language, it is easier to train a model on vast amounts of data than on small amounts.