Cross-Domain Text Classification
in Natural Language Processing

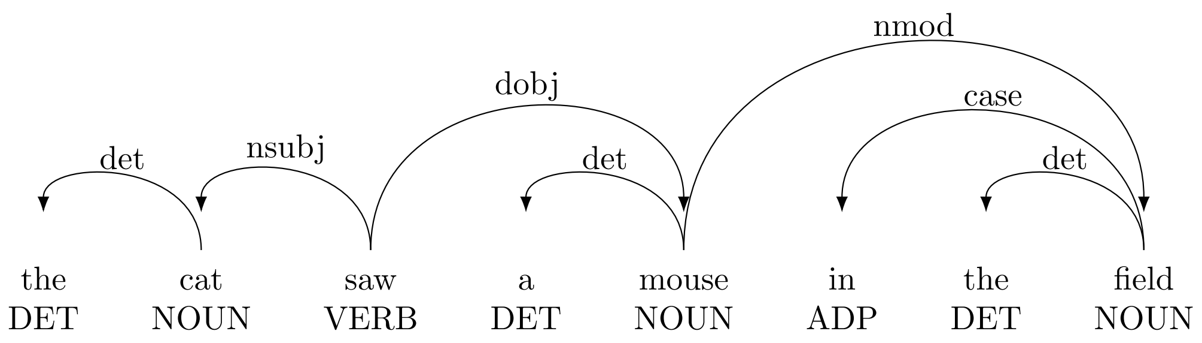
Overview

In Natural Language Processing, text classification is a task where properties of a document in natural language are determined by using machine learning techniques. Thereby, a model learns properties from training documents and then predicts previously unseen texts. If the training and testing documents are in different domains (e.g., contain a different type of text, or are written in different languages), the problem becomes "cross-domain" and therefore more difficult.

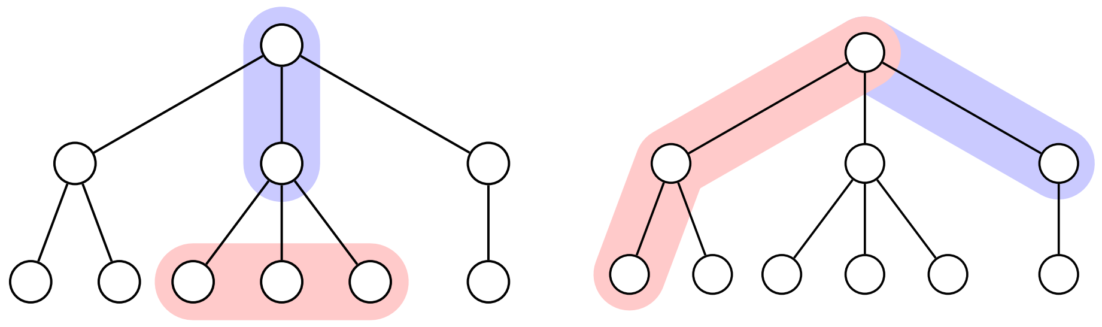
This thesis aims to categorize the field in general and find features and machine learning models suitable for these tasks. Thereby, **authorship classification** problems are focussed, which represent a subgroup of text classification problems where attributes of authorship of documents are analyzed.

Task: Cross-Language Authorship Attribution

... is a special classification task where the authorship of an unknown document is to be determined from a set of candidate authors, and the training documents are written in different languages than the ones to be predicted.

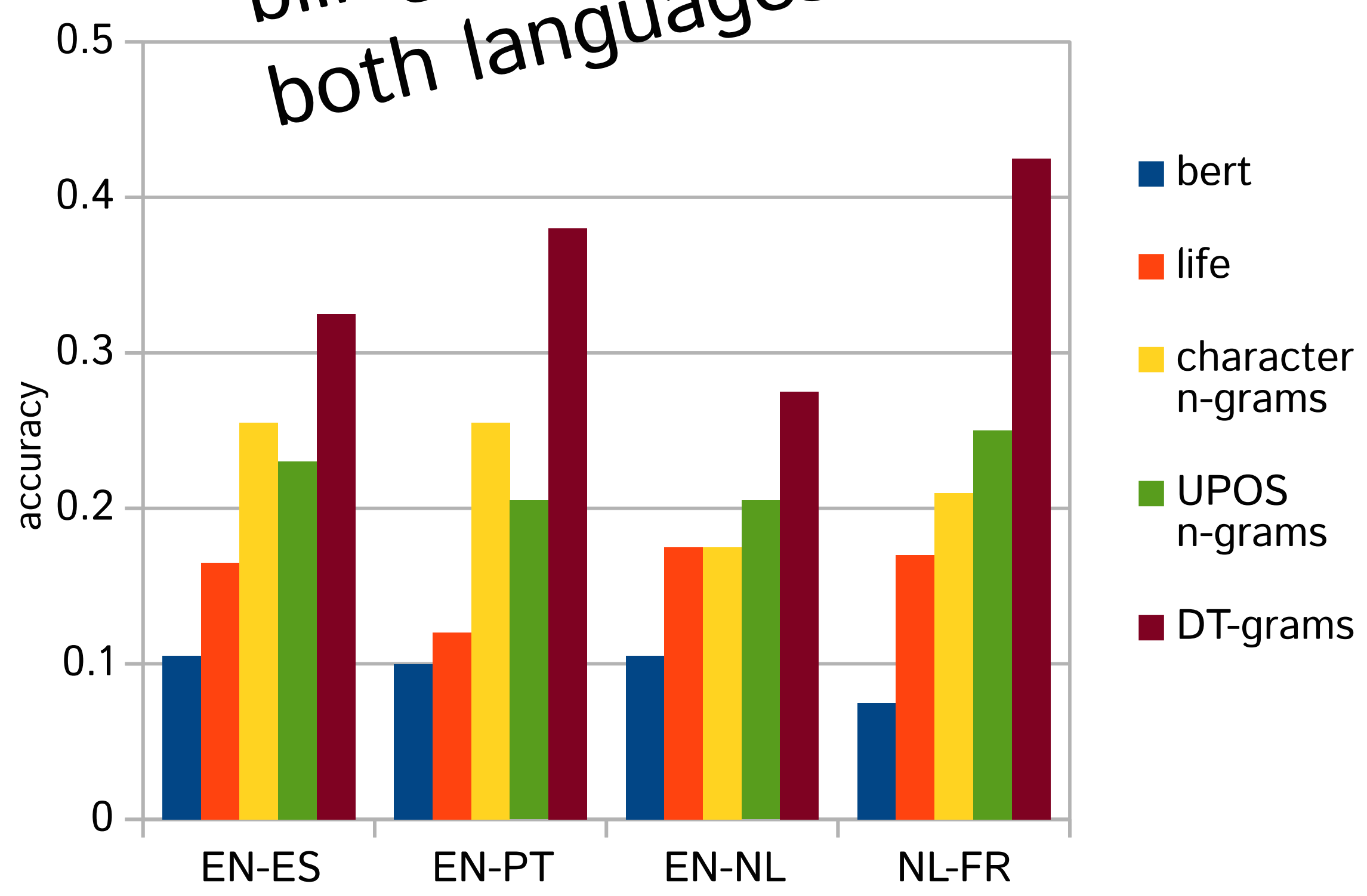


By using grammar based features which are language-independent, the language gap can be overcome



We use frequencies of substructures of dependency graphs (DT-grams) as features for machine learning models

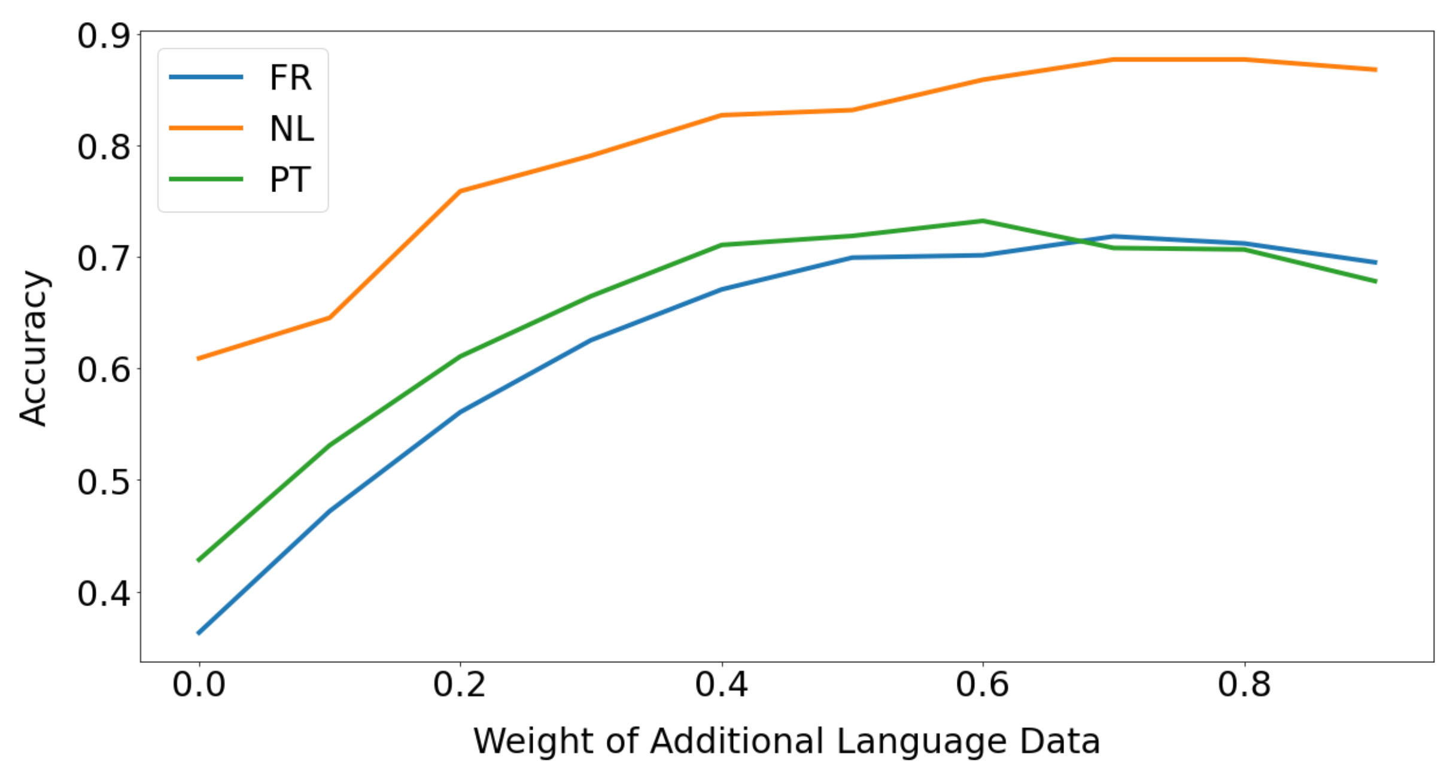
"Which distinguishing stylistic features do bilingual authors use in both languages?"



DT-grams show high performance on classifying bilingual authors

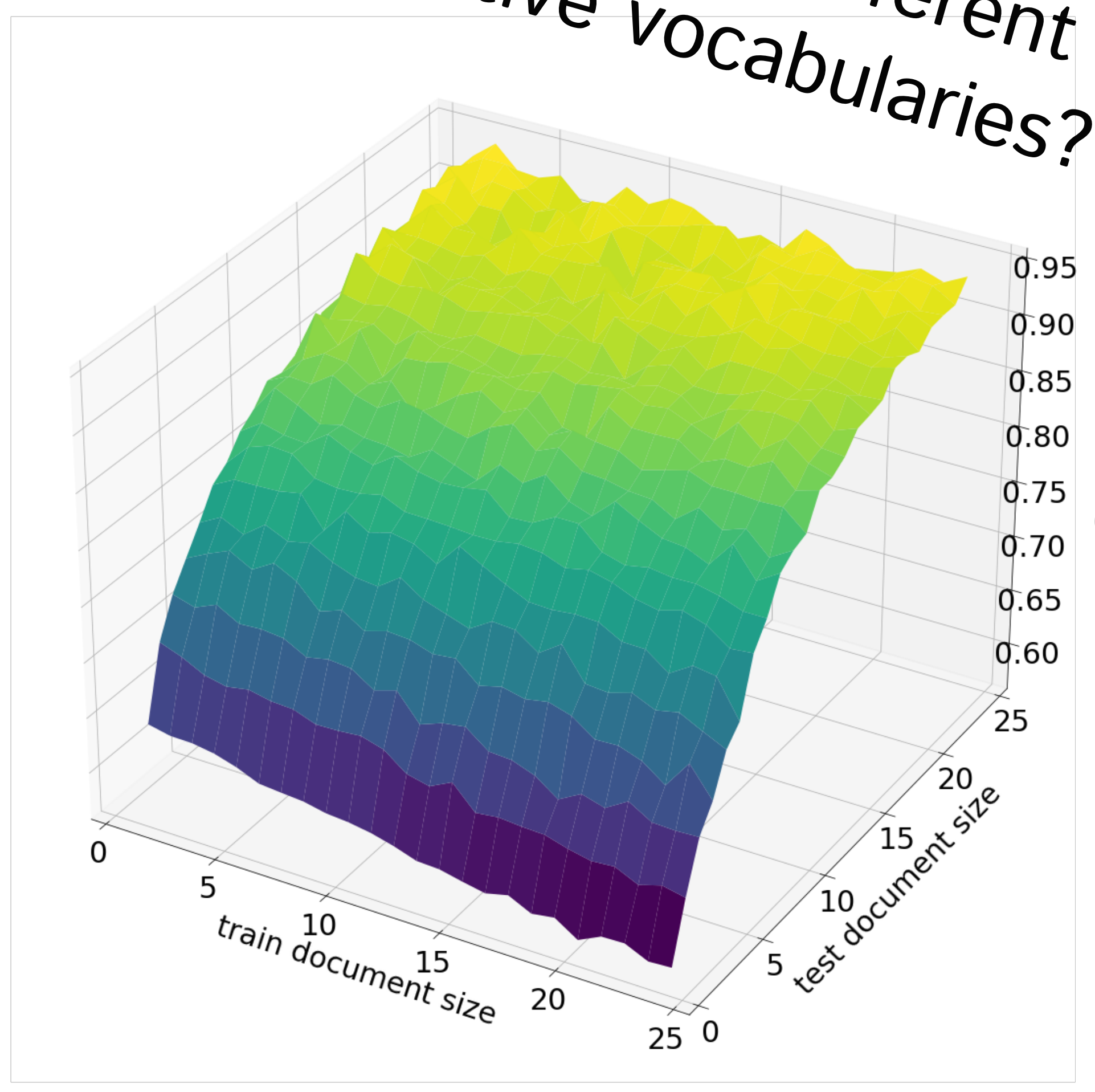
Task: Authorship Profiling

Predicting attributes of authors (like gender or age) of documents



Adding training data in a different language can be beneficial, but the influence depends on the language

Do authors of different ages use distinctive vocabularies?



We determine the influence of distributing the training samples into different sized chunks, which influences the feature frequency calculations.

Open Issues

Why do models that are state-of-the-art in many NLP fields (like GTP, Bert) perform badly in authorship classification?

In what ways can the domain similarity of training and testing documents be measured and how can this be used for applications?

Benjamin Murauer
benjamin.murauer@student.uibk.ac.at
Department of Computer Science - Databases and Information Systems Group
Supervisors: Prof. Dr. Günther Specht, Prof. Justus Piater, PhD.