Zero-Shot Classification of Learning Materials for a Large German Publisher

Freddy Jeufack Fotsop, Mehrdad Ebrahimian, Ernst Mehlmann, Ray Man Ng,
Nils Kahlscheuer, Vincent Stolle, Nadine-Christine Wessel & Betta Younes

Project & Methods

Background & Objectives

Searching and selecting appropriate learning material manually is highly time-consuming for students and teachers. Automated labelling with Zero-Shot Classification (ZSC) during the creation of educational content by editors enables customized creation of learning material pools, serving as a promising opportunity to facilitate editors' daily work.

The objectives of this project are defined as follows:
- Automated categorization (labelling) of learning materials based on ZSC
- Reutilization of teaching content and learning material
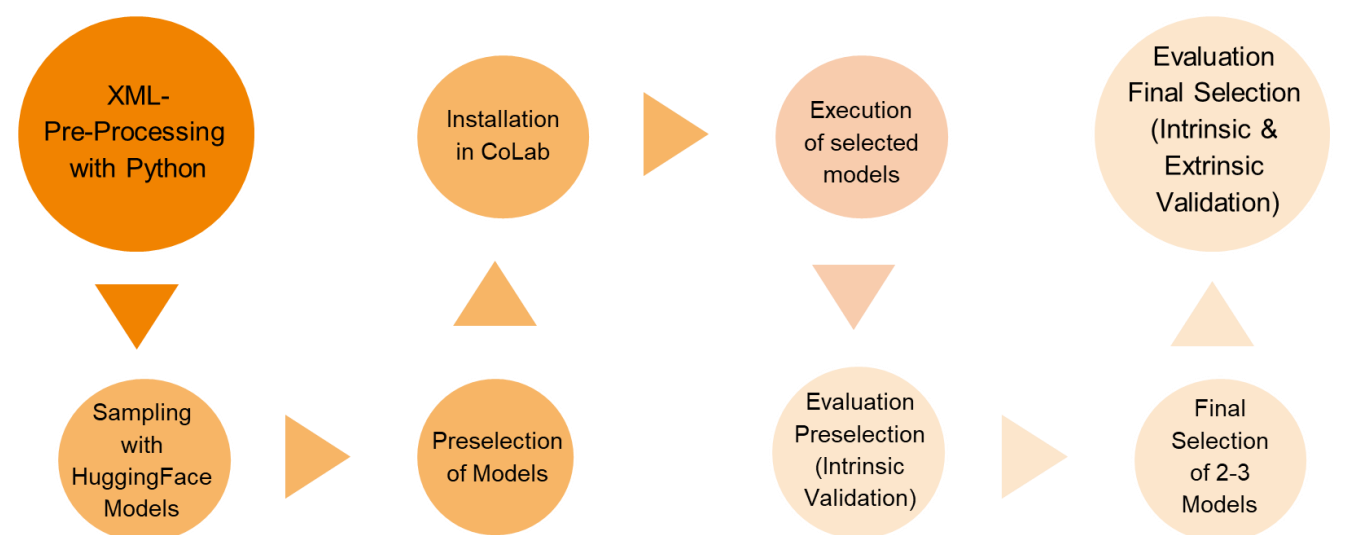- Sketch of Application to search, list and correct, if applicable, learning material

Fig. 1: Proto-Persona Editors [4]

NLP-Pipeline Applied

Fig. 2: Project Pipeline

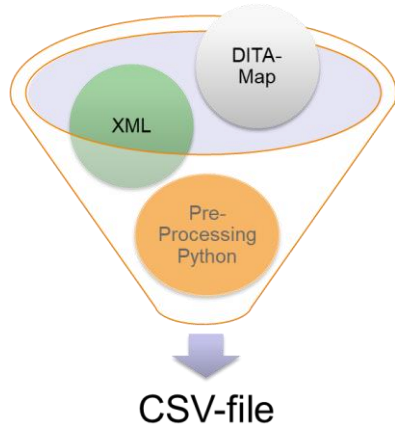## Data

DITA-Map Documents as Input

CSV-file
Fig. 3: Pre-Processing

The XML pre-processing is needed because the learning content is stored using the DITA Learning & Training format [4]. Data input consists of two DITA-Maps i.e., libraries which are made up of several hundred of documents (XML-files). These files can contain titles, vocabulary lists or other learning materials. Only the files with exercises are relevant for this project. Text within paragraphs (<p> .. </p>) and titles (<title> .. </title>) are extracted for further data processing.

Manual Multi-Labelling of 100 Files
(randomly-selected, two assessments per data set)

Topics	Skills
Photo	Reading
Poetry	Writing
Gender & History	Speaking
Technology	Listening
Cultures	
Music & Arts	
Politics	
Future	
Tradition	

Tab. 1: Labelling Categories

Inter Annotator Agreement
(based on Krippendorff's Alpha [1])

Topics	Skills
0.38	0.47

Fig. 4: Results Krippendorff's Alpha

Sampling with 10 HuggingFace Models

Model	Language	Matching Score
distilbert-base-uncased-mnli	English	13%
bart-large-mnli-yahoo-answers	English	64%
squeezebert mnli	English	23%
mobilebert uncased mnli	English	36%
paiintent	English	18%
covid-twitter-bert-v2-mnli	English	68%
xlm-roberta-large-xnli	Multi	50%
roberta-large-mnli	Multi	45%
German_Zeroshot	Multi	64%
xlm-roberta-large-xnli-anli	Multi	59%

Tab. 2: Sampling Evaluation (Intrinsic Validation)

The sampling included all HuggingFace ZSC-models which cover the English language [2]. Final models have been selected based on a comparison of the score of correct classification between the tested HuggingFace models and a predefined manually-labelled set of 11 exercises. The test of HuggingFace models was computed online by filling in the mask of model with the extracted case data.

Results & Conclusion

Evaluation

Model / Intrinsic Validation	Bart-large-mnli-yahoo-answers		Covid-twitter-bert-v2-mnli		German Zeroshot	
	Topics	Skills	Topics	Skills	Topics	Skills
Correct Classification (X out of 10)	8	6	8	7	8	6
Hamming Loss [3]	0.16	0.32	-	-	0.19	0.36
Runtime (per 100 files) / Ratio	1h / 1:4:1		4 h / 1:4:1		1h / 1:4:1	

Tab. 3: Final ZSC-Models Intrinsic Evaluation

Categorization Quality:
- Applicable label sets for editors, depending on skills or topics needed, requires correction of few misallocations

Batch Process Time:
- Time-consuming, but no supervision or high computation power is necessary

Search and Selection Process Time:
- After labelling all tasks the time for searching and selecting is insignificant

Tab. 4: Extrinsic Evaluation of Business Case

Final Remarks

➢ Integration of section or chapter titles could facilitate the manual labelling process for learning material, which includes interactive or visual content, e.g. audio files or images.

➢ ZSC is a time-consuming computational process. Hence, ZSC execution shall preferably be done outside working hours.

➢ It has been observed that working with a smaller number of possible labels ZSC turns out to be conducive for this project.

➢ A next step would be the integration of ZSC into existing applications used by editors, e.g. a content management system for editing and assembling learning material.

References:
[1] Hayes, A. F., & Krippendorff, K. (2007): Answering the call for a standard reliability measure for coding data. Communication Methods and Measures, 1, 77-89.
[2] HuggingFace (2021): HuggingFace Models [Search mask: select Zeroshot Classficiation as task and Language en]. Available at URL: Link to HuggingFace (Accessed: 30 Jun 2021).
[3] Norouzi, M.; Fleet, D. & Salakhutdinov, R. (2012). Hamming Distance Metric Learning. Advances in Neural Information Processing Systems. University of Toronto.
[4] Mural (2021): Persona und UI. [in German language]. Available at URL: Link to Mural Collaborative Board [Accessed: 20 Jun 2021].
[4] OASIS (2018): Darwin Information Typing Architecture (DITA) Version 1.3 Part 3: All-Inclusive Edition [online]. Available at URL: Link to DITA Learning & Training Format (Accessed: 30 Jun 2021).
[5] Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media.