# Project proposal
# IPTC News Categorization

Jan Wojtas

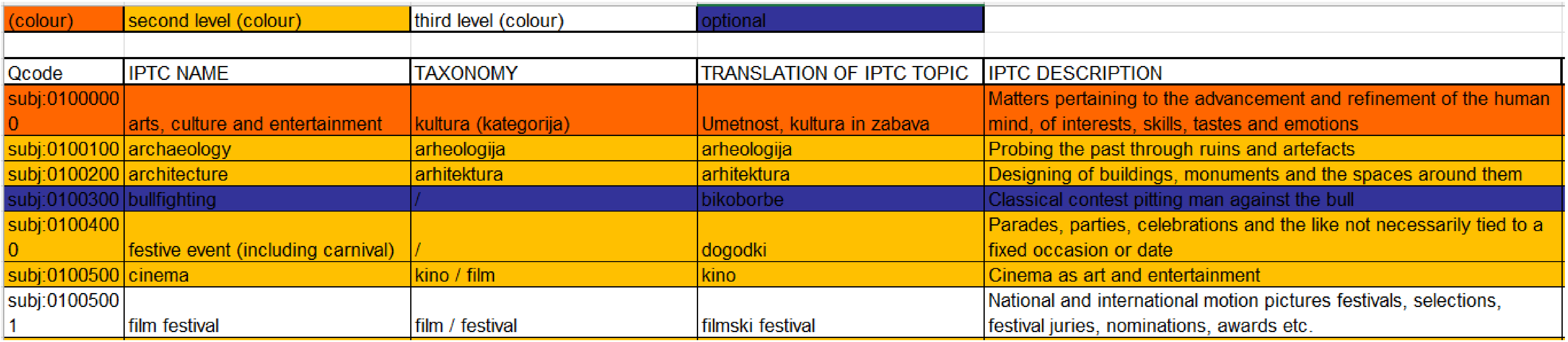
Paulina Szymanek

Łukasz Zalewski

Mikołaj Zalewski

# What is the IPTC taxonomy?

The **IPTC taxonomy** (International Press Telecommunications Council) is a standardized set of categories, codes and definitions used in the field of journalism, media, and publishing to categorize and tag news content and media assets.



| | | | |
|---|---|---|---|
| (colour) | second level (colour) | third level (colour) | optional |

| Qcode | IPTC NAME | TAXONOMY | TRANSLATION OF IPTC TOPIC | IPTC DESCRIPTION |
|---|---|---|---|---|
| subj:01000000 | arts, culture and entertainment | kultura (kategorija) | Umetnost, kultura in zabava | Matters pertaining to the advancement and refinement of the human mind, of interests, skills, tastes and emotions |
| subj:0100100 | archaeology | arheologija | arheologija | Probing the past through ruins and artefacts |
| subj:0100200 | architecture | arhitektura | arhitektura | Designing of buildings, monuments and the spaces around them |
| subj:0100300 | bullfighting | / | bikoborbe | Classical contest pitting man against the bull |
| subj:01004000 | festive event (including carnival) | / | dogodki | Parades, parties, celebrations and the like not necessarily tied to a fixed occasion or date |
| subj:0100500 | cinema | kino / film | kino | Cinema as art and entertainment |
| subj:01005001 | film festival | film / festival | filmski festival | National and international motion pictures festivals, selections, festival juries, nominations, awards etc. |

# Structure

The IPTC taxonomy is structured in a hierarchical manner with multiple levels, typically consisting of four levels:

- the top-level,
- the category,
- the subcategory,
- and the specific code,

allowing for a detailed categorization of news and media content.

# Project goals

- Scientific Goal: Categorization of news articles according to the IPTC taxonomy
- Research question: Can state-of-the-art NLP techniques effectively automate the categorization of news articles in line with the IPTC taxonomy?

# Significance and justification

Need for a scientifically sound approach to automate news categorization that can keep pace with the rate of information production.

# Impact of Project Results

Increasing the accuracy and consistency of categorization, improving news discoverability.

Contribution to academic research in NLP, offering insights into the application of machine learning in real-world text classification tasks.

The methodology could be adapted to other domains that require text categorization.

# Specific Research Goals

- Establishing a baseline for IPTC news article categorization using traditional machine learning models.

- Assessing the influence of embeddings for the overall evaluation.

- Investigating and implementing advanced deep learning techniques for improved classification performance.

- Evaluating and comparing the effectiveness of different NLP models in the context of IPTC taxonomy.

# Risk Analysis

- Data scarcity, labeling, quality issues
- Challenges in interpretability
- Technical risks with model integration
- Underestimation of the score
- Not satisfactory results
- Human errors

# Data

STA News Dataset – Slovenian and English articles.

8778 English articles from 2023.

IPTC Taxonomy – mapping IPTC categories to articles

Other: AG News, DBpedia14

# State-of-the-art

- LLM Embeddings (OpenAI Ada, ST5, etc.)
- Mask-guided BERT
- XLNet
- Seed-guided methods: SeededLDA, CatE

# Popular methods in literature

# Seed-Guided methods

- Utalize concept of a seed – a unigram or a phrase under which a set of terms that form a coherent topic may be found. Those terms can be a unigram or a phrase as well.
- More accurate word semantics learning for topic discovery than "bag-of-words" assumption.
- E.g. SeekTopicMine, SeededLDA, CatE.

# XLNet

- AR (autoregressive) models: predict next token based on the preceding token sequence
- BERT: predict [MASKED] tokens based on context
- XLNet: permute the token sequence. Predict next token based on the previous tokens with regard to the given permutation.

# XLNet c.d.

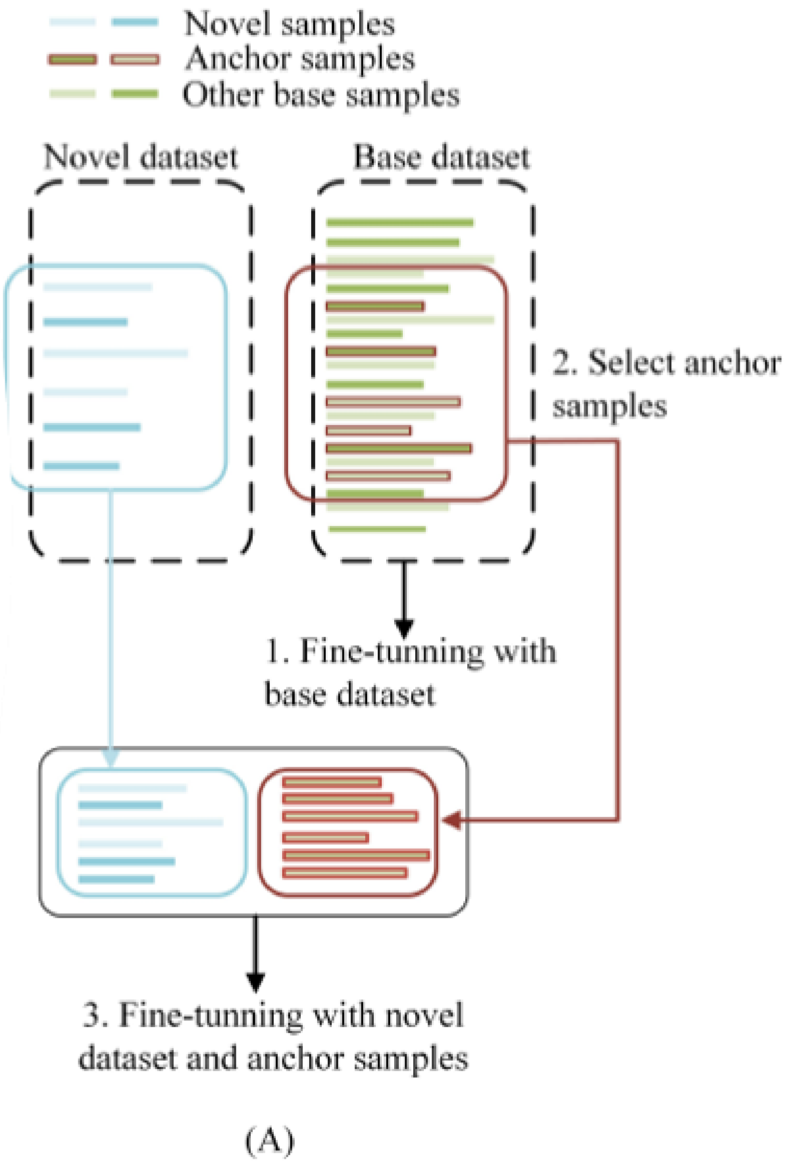
**Advantage over AR modelling**: Capture bidirectional context

**Advantage over BERT:** No corruption of tokens with masks, does not introduce pretraining-fine-tuning discrepancy

Beats BERT on many benchmark datasets, e.g. **RACE**, **SQuAD** (reading comprehension), **QNLI** (does context sentence have answer to question).

# Mask-guided BERT

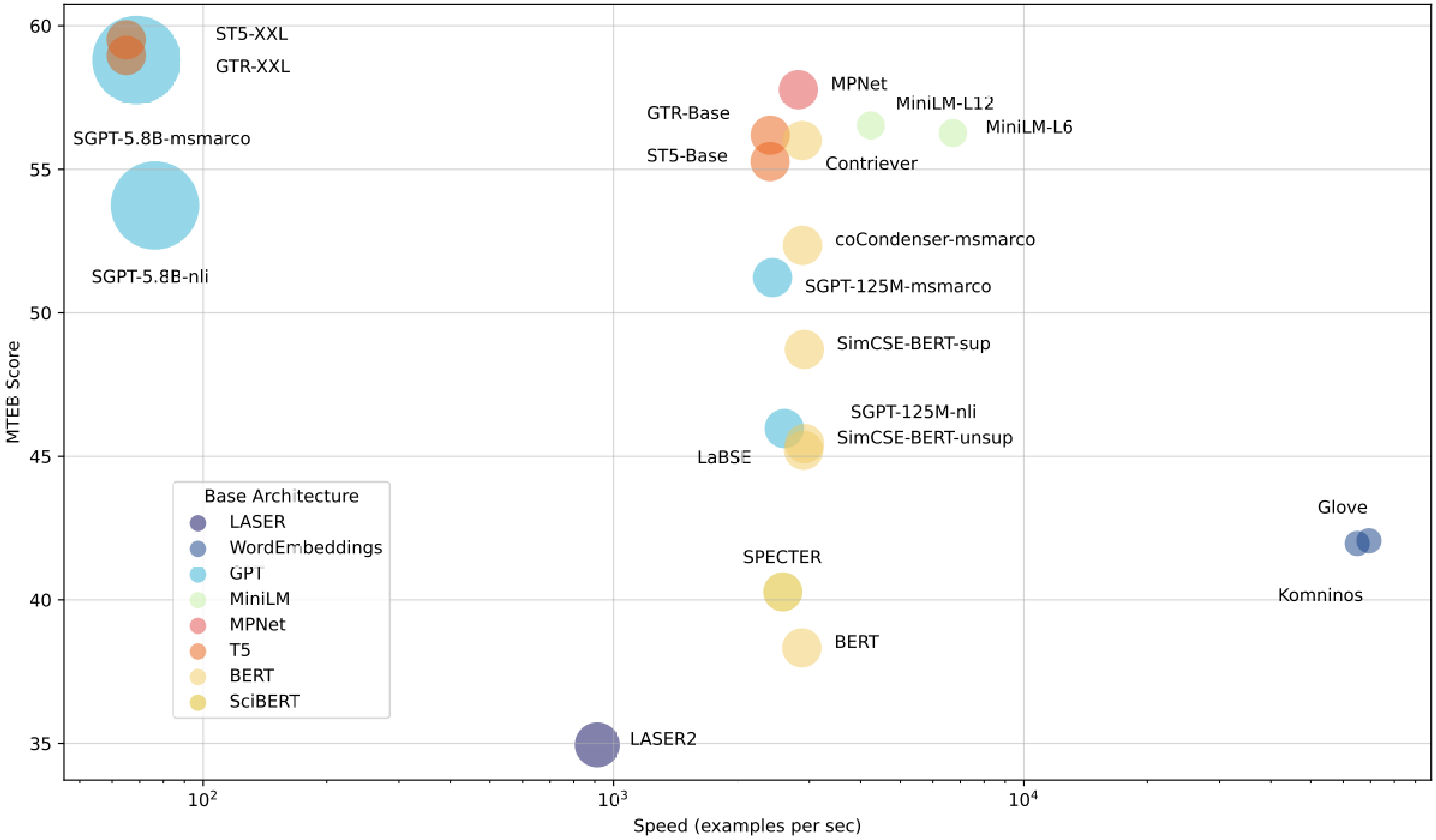
Framework for few-shot learning



Liao, Wenxiong, et al. "Mask-guided bert for few shot text classification." *arXiv preprint arXiv:2302.10447* (2023).

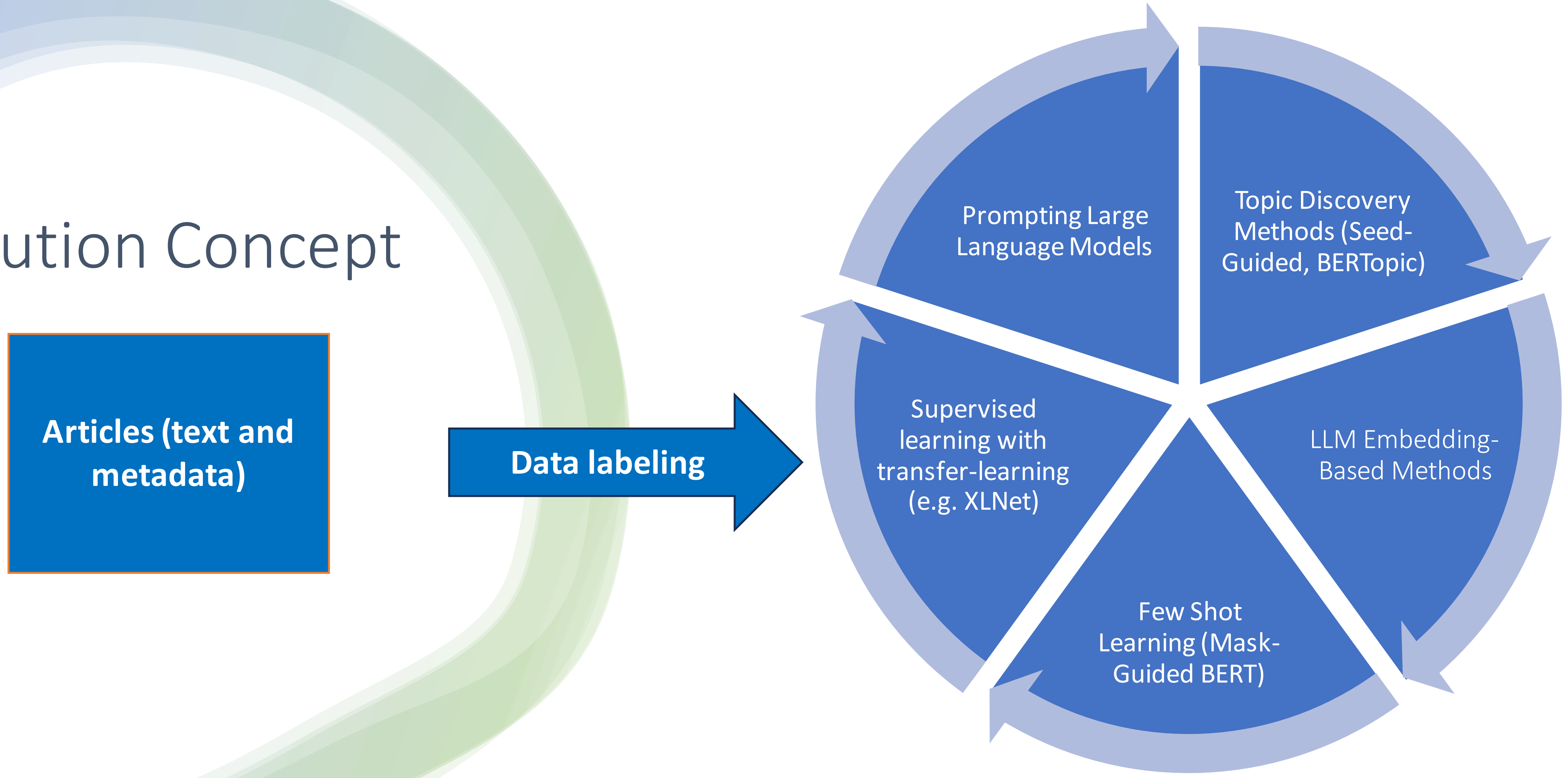
# LLM Embeddings

- The paper ”Massive Text Embedding Benchmark (MTEB)” by Muennighoff et al. 2022 has benchmarked 33 language models and finds that no single text embedding method is superior across all tasks.
- The benchmark found ST5 models dominate the multilingual classification task across most datasets. ST5-XXL has the highest average performance, 3% ahead of the best non-ST5 model - OpenAI's Ada.
- There is a significant trade-off between model performance and speed.



ST5-XXL
GTR-XXL
SGPT-5.8B-msmarco
SGPT-5.8B-nli
MPNet
MiniLM-L12
MiniLM-L6
GTR-Base
ST5-Base
Contriever
coCondenser-msmarco
SGPT-125M-msmarco
SimCSE-BERT-sup
SGPT-125M-nli
SimCSE-BERT-unsup
LaBSE
Glove
Komninos
SPECTER
BERT
LASER2
Base Architecture
LASER
WordEmbeddings
GPT
MiniLM
MPNet
T5
BERT
SciBERT
MTEB Score
60
55
50
45
40
35
Speed (examples per sec)
$10^2$
$10^3$
$10^4$

# Solution Concept



Articles (text and metadata)

Data labeling

Prompting Large Language Models

Topic Discovery Methods (Seed-Guided, BERTopic)

LLM Embedding-Based Methods

Few Shot Learning (Mask-Guided BERT)

Supervised learning with transfer-learning (e.g. XLNet)

# Our main focus: embedding-based classification

Measuring similarity of news embedding to category description embedding

Training supervised classifiers on top of embeddings

## References

Bogery, Raghad et al. (2022). “Automatic Semantic Categorization of News Headlines using Ensemble Machine Learning: A Comparative Study”. In: *Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University*.

Dai, Zihang et al. (2019). “Transformer-xl: Attentive language models beyond a fixed-length context”. In: *arXiv preprint arXiv:1901.02860*.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Liao, Wenxiong et al. (2023). “Mask-guided bert for few shot text classification”. In: *arXiv preprint arXiv:2302.10447*.

Muennighoff, Niklas et al. (2022). “MTEB: Massive text embedding benchmark”. In: *arXiv preprint arXiv:2210.07316*.

Yang, Zhilin et al. (2019). “Xlnet: Generalized autoregressive pretraining for language understanding”. In: *Advances in neural information processing systems* 32.

Zhang, Yu et al. (2023). *Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts*.

# QUESTIONS?