TED4LAT

1862
RĪGAS TEHNISKĀ UNIVERSITĀTE

Study of Impact of Global Trends, Culture and Positive Practice Influences in Latvian Learning Materials

E-Learning Technologies and Management

Yelingyun Zhang
Scientific advisers:
Assoc. prof. Atis Kapenieks; Prof. Marina Platonova

15.07.2025

TED4LAT, No. 101079206

The Problem: Tracing Linguistic Influence in Educational Texts

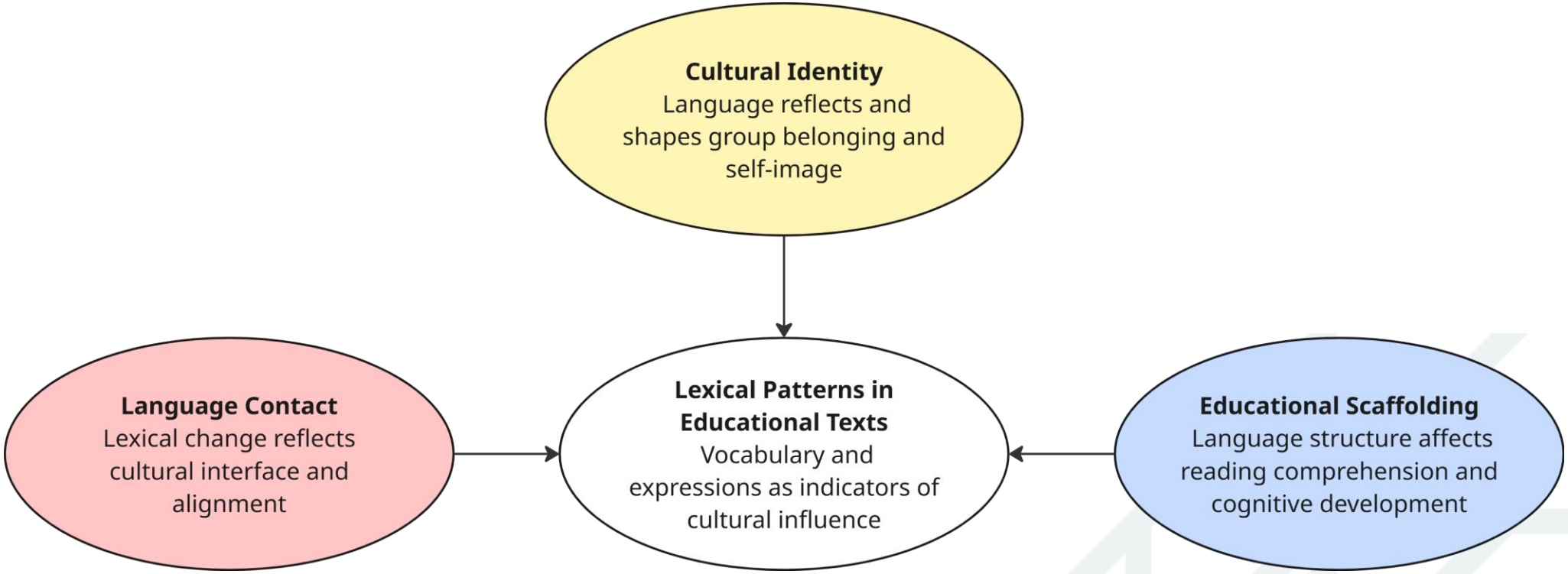
- **Research Focus**
  – Latvian educational materials (Wikipedia, textbooks)
  – Global trends, foreign elements, positive linguistic practices
  – Focus on lexical-level patterns: loanwords, calques, metaphors
- 70% of Latvian corpus is translated language
- Borrowed terms dominate key domains (AI, technology)
- Risk of identity weakening and cognitive overload in learners
- Lack of tools to trace linguistic influence in small-language education

Methodological Innovation & Societal Relevance

- Fine-tuned mBERT for Latvian loanword detection in a low-resource setting
- Developed a monolingual approach to loanword detection, avoiding reliance on bilingual or parallel corpora
- Integrated language-contact theory into a modular binary classification pipeline
- Extended detection beyond phonetic borrowings to include structural borrowing (calques) based on linguistic patterns
- Built an annotated dataset based on linguistic categories (loanwords, calques)
- Promotes awareness of language–culture alignment in material design
- Supports language preservation policy by visualizing lexical borrowing patterns
- Provides tools for scaffolded reading environments
- Builds open-access methods for monitoring linguistic change in under-resourced languages

Theoretical Foundations

- Three perspectives on how language forms carry culture and shape learning.



Theoretical sources: Matras (2009); Hammond et al. (2001); Kramsch (1998); Givón (2005); Carneiro et al. (2020).

Object and Subject of the Study

Object of the Study

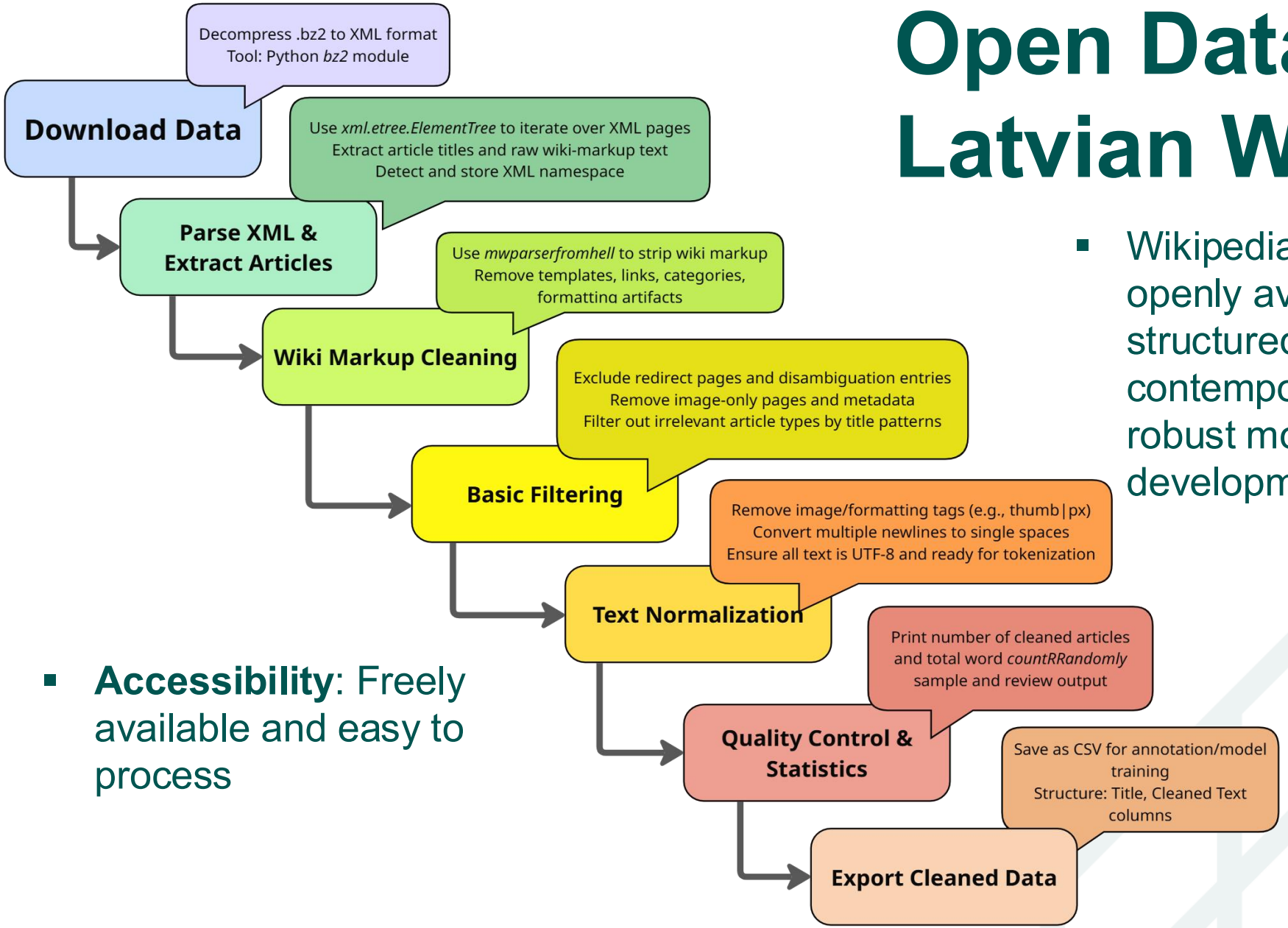
- Language use in Latvian educational materials
- Focus on how lexical borrowing reflects global cultural influence

Subject of the Study

- A machine learning–based framework for detecting borrowed lexical elements
- Focused on loanwords and calques in Latvian Wikipedia as a proxy for educational discourse

Open Data Corpus: Latvian Wikipedia

- Wikipedia provides a large, openly available, and well-structured corpus of contemporary Latvian, enabling robust model pre-training and development.

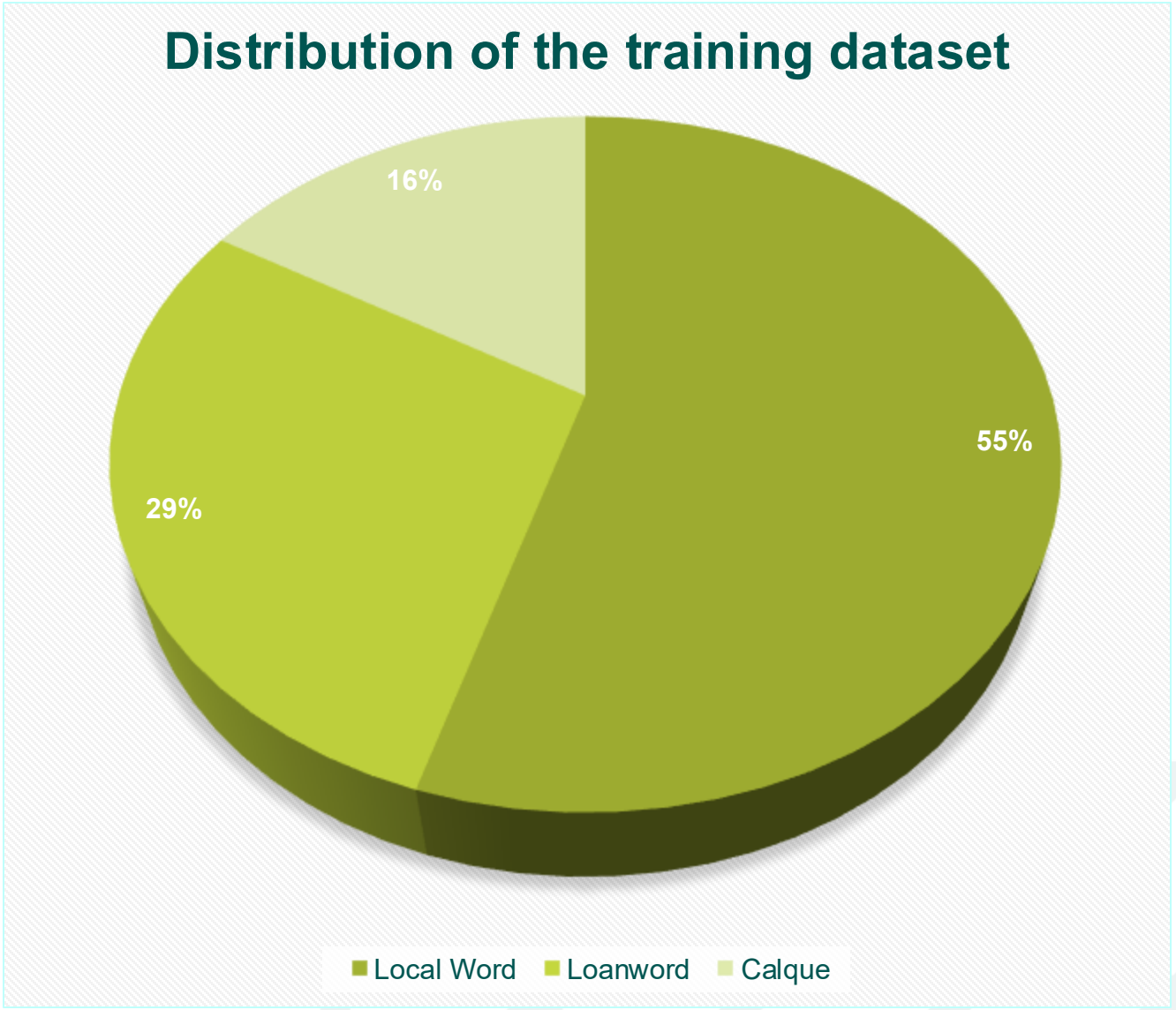


- **Accessibility**: Freely available and easy to process

Methods

Model: Multilingual BERT (mBERT), pre-trained on >100 languages, including Latvian

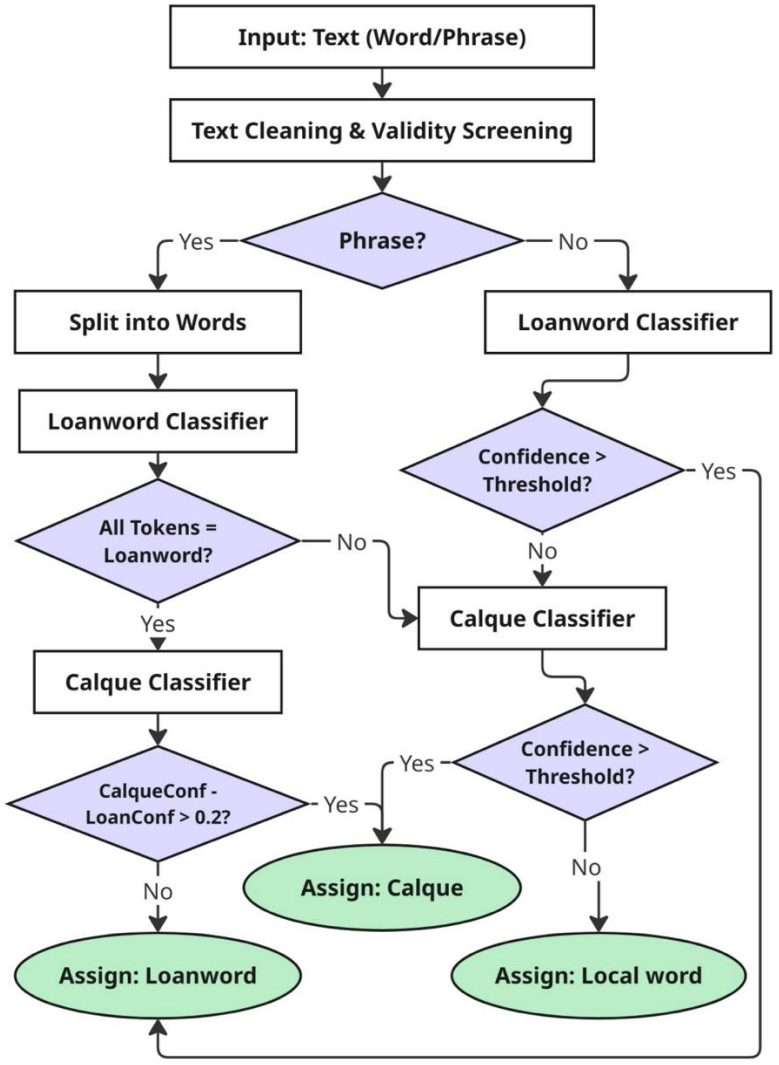
- **Task:** Detecting borrowing in Latvian texts
- **How**: Fine-tune using a smaller, manually labeled Latvian dataset
- **Weighted cross-entropy** assigns a higher penalty to mistakes on underrepresented classes, encouraging the model to pay more attention to them.



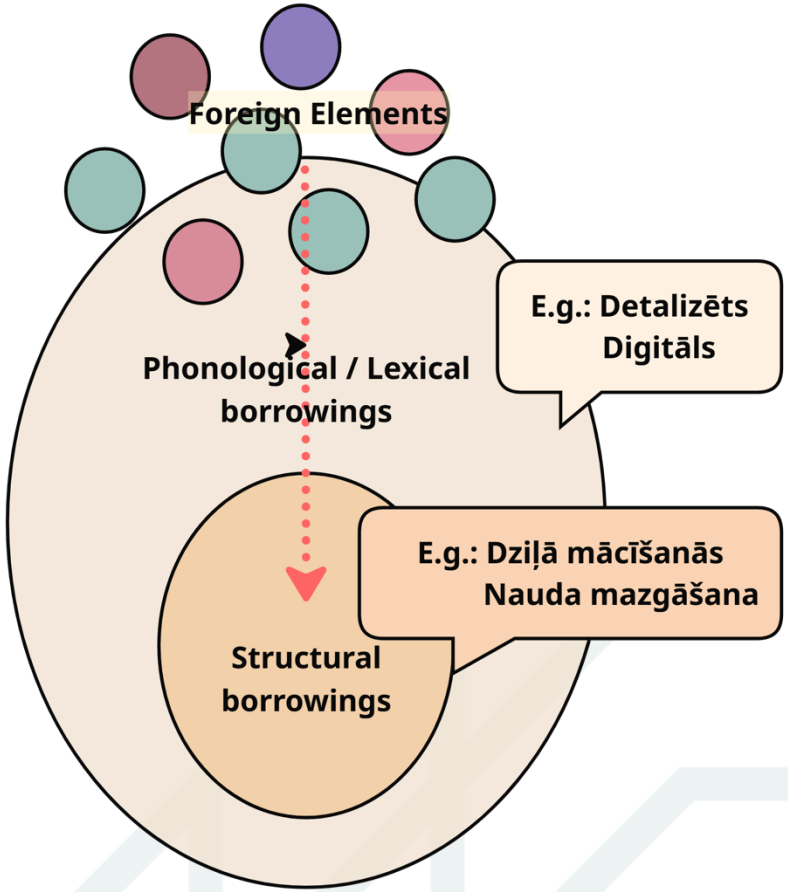
Method: Semi-supervised Learning with Pseudo-labels

- **Pseudo-labels:**
- Unlabeled examples are assigned predicted labels by a model trained on real data. These "pseudo-labeled" samples are then included in further rounds of training as if they were real.
- **Feasibility & Cautions**
- Expands effective training data:
  - Pseudo-labeling leverages a large pool of unannotated text, reducing reliance on expensive manual annotation.
- Improves generalization:
  - The model learns broader patterns, potentially boosting accuracy and recall on rare or complex cases.
- Risks:
  - If the model's initial accuracy is low, adding many pseudo-labeled examples may amplify noise and introduce systematic errors.

Method: Modular Binary Classification Pipeline



- Decompose the multi-class task into a sequence of binary classification steps.
- Breaking a challenging multi-class problem into two-stage binary decisions reduces confusion and error propagation—especially important when certain categories have overlapping features.
- The staged binary process mirrors the natural stratification

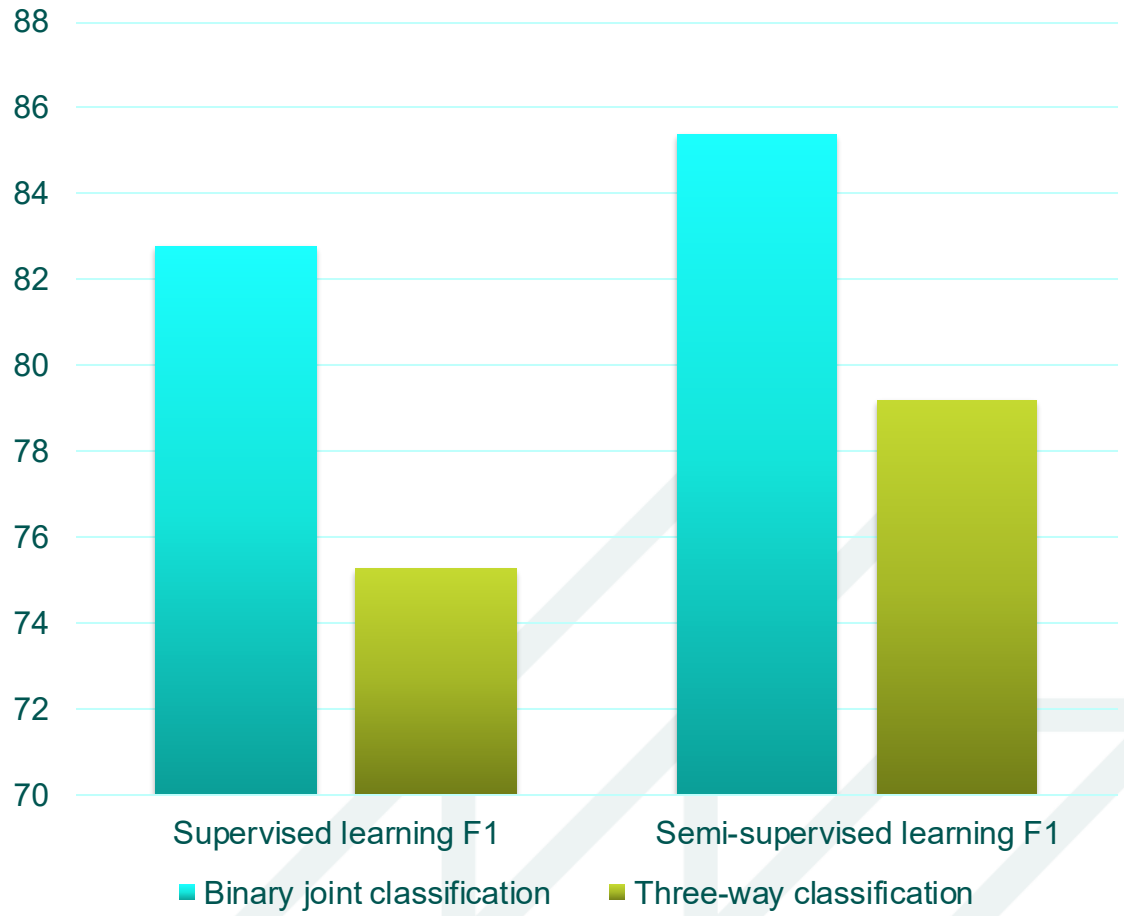


- Based on Matras, Y. (2009). Language contact. https://doi.org/10.1017/cbo9780511809873

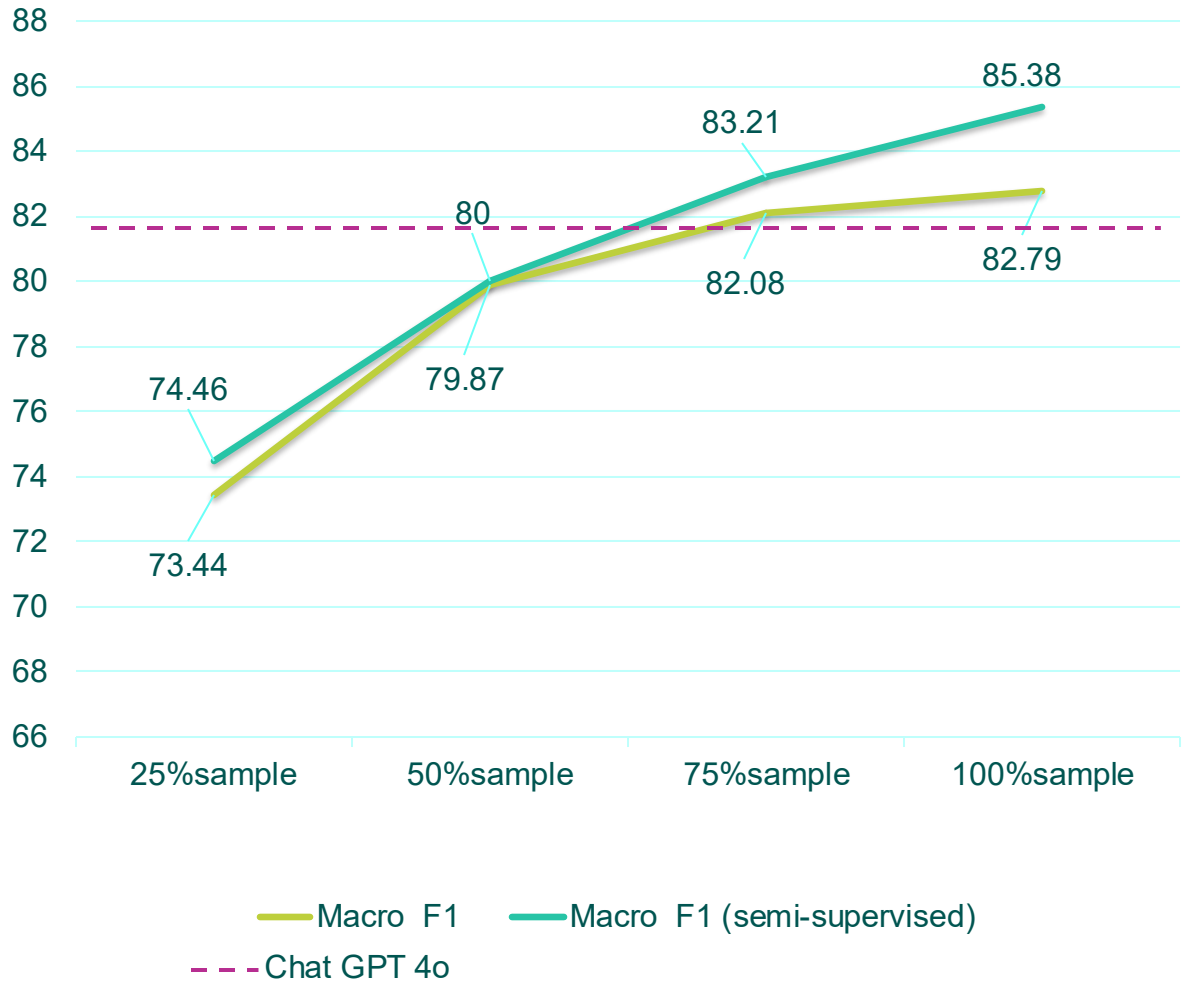
Ablation Study

Why Perform Ablation Studies?

- To identify the optimal balance between labeled and pseudo-labeled data for training;
- To determine how much model performance improves as more annotated data becomes available;
- To compare the effectiveness of different model architectures.
- Both architectures improve with semi-supervised training
- Joint binary pipeline consistently outperforms the three-way classifier



Main Findings



- Supervised model performance increases with data, but shows diminishing returns beyond 75% sample
- Semi-supervised training becomes increasingly effective as base model improves (pseudo-label quality rises)
- At 75% data scale, model performance surpasses GPT-4o benchmark on lexical borrowing detection
- Calque classification remains more challenging due to limited signals and fewer training examples
- Main classification errors occur in borderline cases such as place names, personal names, and media titles

Conclusions & Future Work

- Monolingual training showed strong performance, suggesting that native and borrowed patterns retain learnable structural differences
- The binary classification pipeline reflects not only computational efficiency but also cognitive stages of perception and category separation
- Structural borrowing has long been considered difficult to detect due to its localized surface form, but reliable identification is possible, even in low-resource settings.
- Open data enabled scalable, low-barrier model development and promotes future reproducibility in different language settings
- Next steps: extend domain coverage to real educational texts and analyze underlying pattern distinctions that support model decisions

Reference

- Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011
- Carneiro, P., Lapa, A., Reis, J., & Ramos, T. (2020). Testing pragmatic inferences: The impact of language and culture. Psicológica, 41(1), 1–20. https://doi.org/10.2478/psicolj-2020-0001
- Givón, T. (2005). Context as other minds: the pragmatics of sociality, cognition and communication. http://ci.nii.ac.jp/ncid/BA72853578
- Matras, Y. (2009). Language contact. https://doi.org/10.1017/cbo9780511809873
- Nath, A., Saravani, S. M., Khebour, I., Mannan, S., Li, Z., & Krishnaswamy, N. (2022, October 1). A generalized method for automated multilingual loanword detection. ACL Anthology. https://aclanthology.org/2022.coling-1.442/
- Hammond, J. (2001). *Scaffolding: Teaching and learning in language and literacy education*. Primary English Teaching Assoc., PO Box 3106, Marrickville, New South Wales, 2204, Australia.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kramsch, C. J. (1998). Language and culture. https://ci.nii.ac.jp/ncid/BA38537597
- Veisbergs, A. (2017). TRANSLATION LANGUAGE: THE MAJOR FORCE IN SHAPING MODERN LATVIAN. Vertimo Studijos, 2(2), 54. https://doi.org/10.15388/vertstud.2009.2.10603

Acknowledgement

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101079206.

Thank you!