Project Proposal: Comparative Study of Pre-Trained Models for Swiss German Dialect Classification

Björn Brauer, Justin Verhoek, Marion Andermatt, Thomas Joos

1 Introduction

The goal of this project is to conduct a comparative study on the effectiveness of different pre-trained models for Swiss German dialect classification. The project will leverage the SwissDial Dataset presented in the paper by Dogan-Schönberger et al. (2021), which provides a labeled corpus of audio and text for eight major Swiss German dialects. This dual-modality dataset enables us to examine models on both text or speech data and leads to our central research question: Which type of model or pre-training strategy domain-specific, general-purpose, or speech-based is most effective for a downstream, low-resource language task like dialect classification?

The study will explore either one of the two classification approaches, depending on the feasibility: one using text-based models or one using speech-based models. This framework will allow for a robust comparison and provide a clear answer to the research question.

2 Text-Based Classification

This approach frames dialect identification as a purely Natural Language Processing (NLP) task, using the textual transcripts of the SwissDial Dataset. Written text captures lexical, morphological, and orthographic differences between dialects, which are valuable cues for classification.

The project will compare several pre-trained models that illustrate a range of architectural paradigms and pre-training strategies. These include SwissBERT (Vamvas et al., 2023), a domain-specific BERT variant; German BERT (Chan et al., 2020) as a general-purpose baseline; XLM-R (Conneau et al., 2019) as a multilingual transformer; ByT5 (Xue et al., 2022), a token-free encoder-decoder model; and fastText (Joulin et al., 2017), a lightweight embedding-based baseline. This diversity enables us to investigate whether performance is driven more by model architecture, training data, or efficiency trade-offs. An overview with more information about the chosen models is presented in Table 1.

3 Speech-Based Classification

This approach addresses the speech processing component of the project by building systems that can classify dialects directly from audio signals. Unlike the text-based setting, speech data captures phonetic, prosodic, and acoustic cues that are often central to dialectal variation. The SwissDial Dataset offers high-quality recordings for this purpose, enabling the exploration of whether these speech-specific features facilitate better dialect discrimination than textual transcripts alone.

The project will compare several pre-trained models that represent different philosophies of speech modelling. These include Wav2Vec2 (Baevski et al., 2020), a transformer trained directly on raw waveforms; AST (Gong et al., 2021), a spectrogram-based model leveraging vision transformer architectures; and Whisper (Radford et al., 2023), a versatile encoder-decoder model trained on diverse multilingual audio. Each of these approaches offers a distinct method for leveraging pre-training for downstream classification, allowing us to evaluate whether representations based on raw audio, spectrograms, or multitask training provide the greatest benefit for dialect discrimination. An overview of the models is given in Table 2.

4 Data: The SwissDial Dataset

The project uses the SwissDial Dataset, provided by the Media Technology Center at ETH Zurich. The dataset is publicly available under a *Creative Commons Attribution-NonCommercial 4.0* license and should be cited in academic work (Dogan-Schönberger et al., 2021). It contains approximately 2’700 voice recordings, each lasting be-

Table 1: Overview of Text-Based Models for Swiss German Dialect Classification

Model	Architecture Type	Pre-training Data & Domain	Key Advantage
General German BERT	Masked Transformer (encoder)	Massive German-language corpus (12GB)	Serves as a strong general-purpose baseline.
SwissBERT	BERT-style, domain-specific	Swiss German news, web texts, and tweets	Highly specialized linguistic knowledge of Swiss German.
XLM-R (XLM-RoBERTa)	Multilingual masked transformer	Large scale (100 languages)	Provides multilingual capability for generalization.
ByT5	Token-free (byte-level) encoder-decoder transformer	General text corpus	Offers robustness to spelling variations and noise.
fastText	Shallow word/subword embeddings + classifier	General text corpus	Extremely fast and efficient, with subword capability.

Table 2: Overview of Speech-Based Models for Swiss German Dialect Classification

Model	Architecture Type	Pre-training Data & Domain	Key Advantage
Wav2Vec2	Encoder-only Transformer	Massive, unlabeled audio corpus	Learns a rich, universal representation of speech.
AST	Vision Transformer	Audio Spectrograms	Alternative approach using spectrograms for analysis.
Whisper	Encoder-Decoder	Diverse multilingual audio	Versatile, multitasking model for comparative analysis.

tween 3 and 10 seconds, across eight major Swiss German dialects: Aargau, Bern, Basel, Lucerne, St. Gallen, Graubünden, Valais, and Zurich.

In addition to the audio recordings, the dataset provides transcripts in all eight Swiss German dialects and in Standard German, stored in JSON format.

5 Evaluation and Final Deliverables

To ensure the study is academically rigorous, the final report must include a detailed evaluation of the models. The models will be judged not just on raw accuracy but on a more nuanced set of metrics, including Precision, Recall, and F1-Score, which are standard for classification tasks. Confusion Matrices could also be used to visually represent and analyse where each model is succeeding and where it is making errors. This evaluation will allow us to draw clear conclusions about which type of pre-training—domain-specific, general-purpose, or speech-based—is most effective for Swiss German dialect classification.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. *arXiv preprint arXiv:2010.10906*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken swiss german. *arXiv preprint arXiv:2103.11401*.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. Swissbert: The multilingual language model for switzerland. *arXiv preprint arXiv:2303.13310*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.