Multi-label Scandinavian Language Identification (SLIDE)

Mariia Fedorova*, Jonas Sebulon Frydenberg*, Victoria Handford*,
Victoria Ovedie Chruickshank Langø*, Solveig Helene Willoch, Marthe Løken Midtgaard,
Yves Scherrer, Petter Mæhlum, David Samuel
Department of Informatics, University of Oslo
{mariiaf,jonassf,vlhandfo,victocla,solvehw,marthemi,
yvessc,pettemae,davisamu}@ifi.uio.no

Abstract

Identifying closely related languages at sentence level is difficult, in particular because it is often impossible to assign a sentence to a single language. In this paper, we focus on multi-label sentence-level Scandinavian language identification (LID) for Danish, Norwegian Bokmål, Norwegian Nynorsk, and Swedish.[1] We present the Scandinavian Language Identification and Evaluation, SLIDE, a manually curated multi-label evaluation dataset and a suite of LID models with varying speed–accuracy tradeoffs. We demonstrate that the ability to identify multiple languages simultaneously is necessary for any accurate LID method, and present a novel approach to training such multi-label LID models.

1 Introduction

Correctly identifying the language of a short piece of text might seem like a simple (and possibly already solved) task. While differentiating between two distant languages might be straightforward, we show that, when focusing on a group of closely related languages, this task becomes substantially more challenging. This is especially true when we consider the fact that language identification (LID) tools have to be fast and efficient, as they are often used for preprocessing large quantities of texts.

In this paper, we focus on the four closely related Scandinavian languages: Danish, Norwegian Bokmål, Norwegian Nynorsk, and Swedish. In order to accurately differentiate within this group, we move away from the standard single-label (multi-class) language identification and instead treat this problem as multi-label classification task, allowing for the identification of multiple languages simultaneously as illustrated in Figure 1. Sentences valid in multiple Scandinavian languages are fairly common—they account for about 5% of our evaluation dataset and 16% of the sentences shorter than 6 words. If not accounted for, these examples can skew evaluation of existing systems. The three main contributions of SLIDE (Scandinavian Language Identification and Evaluation), are as follows:

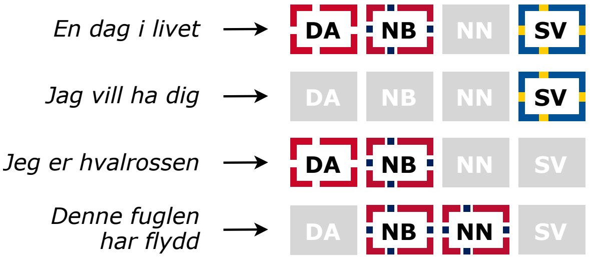

Figure 1: **Scandinavian similarity** Accurate language identification has to necessarily be multi-label when discriminating between closely related languages.

1. **A multi-label evaluation dataset** We have created a manually corrected multi-label LID dataset for four Scandinavian languages. We present two evaluation methods using this dataset: one designed for a more accurate evaluation of traditional multi-class LID methods, and a second for assessing the performance of multi-label methods.

2. **A suite of LID models** We train a family of language identification models of varying complexities. The best performing models are

[*] Equal contribution.

[1] While acknowledging that the term *Scandinavian* in English sometimes also includes Icelandic and Faroese, we use the term Scandinavian in the sense of *Mainland Scandinavian*, in accordance with established and legal usage of the term in these languages. We also consider Swedish as a single language, overlooking the nuances between Finland-Swedish and Sweden-Swedish.

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 179–189
March 2, 2025 ©2025 Association for Computational Linguistics

based on fine-tuned BERT models and smaller, substantially faster models based on FastText embeddings. The source code, datasets and models are released at https://github.com/ltgoslo/slide.

3. **A novel multi-label LID method** Manual creation of a clean multi-label LID dataset is costly. Instead, we present a novel method of silver-labeling such a dataset by utilizing existing machine translation models.

2 Related work

Language identification The task of identifying the language of a text is an "old" NLP task dating back to the 1960s. Simple but relatively powerful tools have been available since the 1990s (Jauhiainen et al., 2019).

In recent years, the main focus of NLP research has shifted towards large language models, and especially towards extending their coverage to an increasing number of languages. As training data for underrepresented languages is mostly found in web crawls, reliable LID systems covering a large number of languages are more important than ever. While the earliest LID systems were restricted to a dozen languages, recent systems cover hundreds (Joulin et al., 2017; Grave et al., 2018; Burchell et al., 2023; Jauhiainen et al., 2022a) and even thousands (Kargaran et al., 2023) of languages.

In terms of methods, simple linear classifiers with character-level and word-level features have often outperformed more sophisticated neural models (Jauhiainen et al., 2019). Most currently available large-coverage LID models are based on the FastText architecture (Joulin et al., 2017), a multinomial logistic regression classifier with character n-gram embeddings as input features. These include FastText-176 (Joulin et al., 2017; Grave et al., 2018), NLLB-218 (NLLB Team et al., 2022), OpenLID (Burchell et al., 2023) and GlotLID (Kargaran et al., 2023). Different approaches are used by HeLI-OTS (Jauhiainen et al., 2022b), which bases its decisions on a combination of character n-gram and word unigram language models, and `gpt2-lang-ident`[2], which is a fine-tuned decoder-only model (Radford et al., 2019).

In practice, LID is most often applied to individual sentences, even though the tools can work with longer or shorter segments of text.

LID for closely related and Nordic languages To our knowledge, the only publication focusing specifically on LID for Nordic languages is Haas and Derczynski (2021). They compile a dataset for the six languages (including both Norwegian standards) from Wikipedia and evaluate a range of LID models on it. They find that the languages mostly cluster into three groups: Danish–Bokmål–Nynorsk, Swedish, and Icelandic–Faroese. Their models were not available online as of writing this paper. Besides this, de la Rosa and Kummervold (2022) present two FastText-based LID models: one containing only the 12 most common languages of the Nordic countries (including several Sámi languages, Finnish, and English), and one with an extended coverage of 159 languages.

Futhermore, the previously mentioned off-the-shelf LID systems (NLLB-218, OpenLID, GlotLID, HeLI-OTS) cover all six Nordic languages, with the exception of FastText-176, which does not include Faroese.

Multi-label language identification Most existing LID training and evaluation corpora are not manually labeled. Instead, they are based on the assumption that the language is determined by the source it is retrieved from. If a sentence is retrieved from a Danish newspaper, it is assumed to be only Danish. But when dealing with closely related languages, it is often the case that an instance cannot be unambiguously assigned to a single language (Goutte et al., 2016; Keleg and Magdy, 2023).

Recent proposals address this issue by framing LID between similar languages as a multi-label task (e.g., Chifu et al., 2024; Abdul-Mageed et al., 2024) and by manually annotating the evaluation data (e.g., Zampieri et al., 2024; Miletić and Miletić, 2024). However, these works do not include studies of Scandinavian languages.

3 Data

One of the main contributions of this paper is the release of manually and automatically annotated multi-label datasets. In Section 3.1, we introduce the sources from which we compile our datasets. We then present our manually annotated multi-label evaluation dataset (Section 3.2). Next, we describe a way to obtain multi-label annotations automatically for the larger training set in Section 3.3. Lastly, we outline different approaches to data augmentation in Section 3.4.

[2] https://huggingface.co/nie3e/gpt2-lang-ident

3.1 Data sources

As a starting point, we use the Universal Dependencies 2.14 treebanks (UD; Nivre et al., 2016, 2020), keeping their train/dev/test splits intact.[3] For each of the four languages, we associate each sentence in the treebank with the language tag corresponding to that treebank's language. This results in a foundational single-label dataset with the following language tags: Danish (DA), Norwegian Bokmål (NB), Norwegian Nynorsk (NN), and Swedish (SV). We further incorporate examples labeled as other, drawing random samples from other UD treebanks to represent other languages.

As the UD treebanks are manually annotated, we assume that the texts accurately reflect their corresponding languages. Additionally, the treebanks cover multiple genres, improving the robustness of the models to different text varieties. However, while the resulting dataset is clean, it is not disambiguated. For example, a sentence labeled as Nynorsk is almost guaranteed to be in Nynorsk, but it could also be a valid Bokmål sentence.

3.2 SLIDE dataset: manually multi-labeled evaluation data

Manual inspection To identify multi-label instances in the validation and test splits, we performed a combination of automatic filtering and manual annotation. Automatic filtration was done by removing frequent words that unambiguously define a language (e.g. 'ikkje' is only valid in Nynorsk; the full list is to be found in our Github repository).

After filtering, we split the remaining instances among a group of annotators to manually check for cases of multilingual acceptability. All annotators were native or near-native Norwegian speakers. Annotation tasks were delegated depending on the speakers' knowledge and exposure to Swedish and Danish (all native speakers have received education in or about other Scandinavian languages through the public curriculum or university classes).

Unclear instances Most cases of multilingual acceptability involved short sentences with proper names, numbers, or words that are acceptable in multiple Scandinavian languages. Instances consisting of only proper names were annotated with all Scandinavian languages, even if more common in one language than another. Numerical values were treated similarly as they are universally acceptable across the languages.

Language	Train split	Validation split	Test split
Bokmål	23 120	2 543	2 098
Danish	5 977	563	677
Nynorsk	21 587	2 031	1 628
Swedish	6 911	553	1 250
Other	8 360	1 124	1 745
Total	61 406	6 433	6 950

Table 1: **Dataset sizes** Number of sentences per language. Multi-label samples are reported once for each language, while the summary row shows total number of unique sentences.

Non-Scandinavian instances Sentences from other languages that are not valid in the Scandinavian languages retain the other label, and we set restrictions on when this label is used. This distinction is crucial as it ensures that the other label exclusively identifies non-Scandinavian sentences, setting it apart from the potential multi-label nature of the remaining labels. For example, this instance from the Danish treebank, "- Gerne.", is labeled as only Danish, despite it also being acceptable in German. This approach allows us to evaluate a model's ability to handle ambiguity and focus on the sentences that could belong to multiple Scandinavian languages, without having to consider all possible languages.

Punctuation errors We found several sentences that were orthographically identical in Danish and Bokmål, where commas were the sole distinguishing factor. When a subordinate clause occurs in the first position of a sentence, both languages include a comma at the end of the clause. However, if the subordinate clause does not occur in the first position, Danish can include a comma before that clause[4], whereas Norwegian cannot[5]. The optional comma, in this case, means that Danish can follow the same punctuation rules as Norwegian but does not have to, making differentiation difficult.

Such a sentence is shown in example (1) from the Danish treebank. The words in this sentence are

[3] Specifically, we use the following UD treebanks: no_bokmaal, da_ddt, no_nynorsk, and sv_talbanken.

[4] https://ro.dsn.dk/?type=rulesearch&side=49

[5] https://sprakradet.no/godt-og-korrekt-sprak/rettskriving-og-grammatikk/kommaregler/

written the same in Danish and Bokmål however, the comma introducing the subordinate clause *at hun skulle havne på et teater* is technically not allowed in Norwegian.

(1) Der stod ingen steder i Mai Buchs eksamenspapirer, at hun skulle havne på et teater.

It said nowhere in Mai Buch's exam papers that she would end up in a theater.

We decided to annotate such sentences as both Danish and Bokmål, thereby focusing on lexical information rather than punctuation. This is due to Norwegians' challenges with following comma rules in general (Michalsen, 2015, pp. 37-39), perhaps due to Norwegian earlier having Danish comma rules (Papazian, 2013). We also find 29444 examples of a comma preceding *at* 'that' in the Norwegian LBK corpus, keeping in mind that some of these might be examples of other usage (Fjeld et al., 2020).

Code switching There were also sentences in the dataset that included more than one language. One such example is:

(2) Låten heter "The spirit carries on."

The song is called "The spirit carries on."

For these sentences that include non-Scandinavian words, we annotated them for the Scandinavian languages only. In cases where a sentence had words from different Scandinavian languages, e.g. a Nynorsk quote in a Bokmål sentence, we made small changes to make the sentence monolingual.[6]

Number of multi-label instances The statistics of the validation and test sets are shown in Table 1. The resulting shares of multi-label instances in the validation and test sets are 6% and 5% respectively.

3.3 Automatically multi-labeled training data

As there is no available multi-labeled training dataset for any subset of the Scandinavian languages, and manually annotating a large-enough dataset would be out-of-scope for this project, we decided to silver-label the UD training split automatically. To do so, we converted the task of machine translation into the task of language identification. This conversion then allows us to utilize existing high-quality resources for multi-label language identification.

Alterations	Loose accuracy	Exact-match accuracy
Augmentation + Regex	98.6	**96.4**
Augmentation	98.4	96.3
Regex	98.4	96.2
NER	**98.7**	95.5
Base	98.3	96.2

Table 2: **Ablation study** Impact of data augmentation and regular expression normalization on SLIDE-base measured by test set performance. "Augmentation" refers to punctuation augmentation, "Regex" refers to regular expression normalization, "NER" refers to named entity swaps and "Base" is neither of the above.

Machine translation conversion The method relies on our observation that machine translation models tend to stay conservative and minimize the changes between the source and target texts. Thus, if the translation of a sentence does not lead to any changes, we label it as a valid sentence of the target language. This means that the machine translation model can only add additional language labels to a sentence as a result; we do not use the translated sentences in any other way.

Specifically, we use `NorMistral-11b` to perform the translation (Samuel et al., 2024). While this large language model is able to translate in a zero-shot manner, we increase its reliability by fine-tuning it on the small high-quality Tatoeba evaluation set (Tiedemann, 2020) in all translation directions between Bokmål, Danish, Nynorsk and Swedish.

3.4 Data augmentation

Punctuation augmentation To prevent our models from relying too much on punctuation, we augment the training data with random punctuation. This is especially important for disassociating punctuation from the `other` tag, for which the training data exhibits punctuation noise to a higher degree than the Scandinavian language examples. We randomly add either (i) a period, an exclamation point, or a question mark to the end of the sentence or (ii) a hyphen, dash or comma at the beginning of the sentence. Additionally, there is a 1/3 chance of including an intervening space. This augmentation scheme is chosen to try to mimic punctuation

[6] There were few instances of this, however, it is important to mention that there is not a complete 1-to-1 correlation between the source material and our dataset.

variance that is present in sentence-level (parallel) corpora.

This method is only applied to instances not labeled as `other` and is performed on about 7.5% of the training data. This value is heuristically chosen.

Regular expression normalization We normalize URLs, email addresses, and numbers into the following special symbols: ⟨`URL`⟩, ⟨`mail`⟩ and ⟨`num`⟩. These elements are not informative for language identification, and we do not want a model to associate them with a certain language.

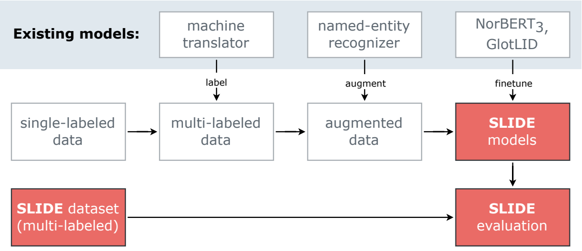

Figure 2: **Training pipeline** A diagram that illustrates the flow of the full training pipeline. We start with a high-quality, single-labeled training dataset, then extend it with multi-label annotations using a strong machine translation model. The dataset is further augmented by randomly swapping named entities identified by existing NER models and through other rule-based augmentations. We use the (augmented) data to fine-tune strong tranformer-based models from a family of NorBERT$_3$ models (Samuel et al., 2023), a fast model from the GlotLID static word embeddings. Finally, the manually-annotated multi-label dataset is used to evaluate the resulting models.

Alphabet variations The alphabet of the four Scandinavian languages differs by the usage of the letters *ä, ö* (in Swedish) and *æ, ø* (in Danish and Norwegian). To ensure that the model does not learn to associate the presence of these letters solely with their corresponding languages, we augment the training data by adding Swedish sentences containing the Danish–Norwegian letters and Danish and Norwegian sentences containing the Swedish letters (e.g., in proper names and in the context of quotations).

We use the NPK parallel corpus[7] containing translations of news texts from Bokmål to Nynorsk to extract texts containing *ä* and *ö*. For Swedish, we use the EU Bookshop corpus (Skadiņš et al., 2014) to extract Swedish sentences containing *æ* and *ø*. Together, this yielded 10,262 sentences, which are included in Table 1.

Named entity swaps We also want to prevent a model from associating named entities with a given language. Although named entities are unequally distributed across languages, they are not necessarily language-dependent. We perform named-entity recognition (NER) on the training data using the spaCy[8] to identify and extract persons, organizations, locations, and miscellaneous entities. We randomly swap the recognized entities with other entities from the same category to try to break up any connection between entity name and a given language.

4 SLIDE evaluation

We introduce two evaluation metrics in our comparison: *loose* and *exact-match* accuracy.

Loose accuracy This evaluation metric is designed for models that output only one language label per input, which is common for off-the-shelf classifiers like FastText and NLLB. According to this metric, a prediction is considered correct if the single predicted label is among the gold labels. This metric is unreliable for multi-label models, since a model that always predicts all four languages would get 100%.

Exact-match accuracy This evaluation metric is more strict and requires an exact match between the predicted and gold labels sets, making it more appropriate for models capable of predicting multiple labels.

Per-language scores Additionally, we report the F$_1$-score for each individual language to measure the quality of classifications for each of the four languages separately. Here, a true positive prediction happens if and only if the respective language is present both in the set of predicted labels and in the set of gold labels.

5 SLIDE training methodology

In this section, we present our approach to training the SLIDE models. We explore two main direc-

[7] `https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-80/`

[8] `https://spacy.io/` pipeline. We use the large language-specific models, where the Norwegian model is used for Bokmål and Nynorsk.

tions: transformer-based models that achieve high accuracy but require more computational resources, and a fast model based on static word embeddings that trades accuracy for faster inference times.

5.1 Transformer models (SLIDE x-small, small and base)

Fine-tuned masked language models are nowadays the most popular sequence classification solution for problems that require accurate solutions and reasonable inference time (Devlin et al., 2019).

Selection of BERT family We assessed massively multilingual, Scandinavian, and Norwegian BERT-like models with comparable number of parameters in order to choose a model to focus on for further optimizations.

We test two massively multilingual models: *XLM-RoBERTa-base* (Conneau et al., 2020), which is trained on a corpus containing 100 languages (including the Scandinavian languages) and has a total of 278M parameters, as well as *DistilBERT-multilingual-base* (Sanh et al., 2019), which is a distilled version of the multilingual BERT base model trained on Wikipedia data from 104 languages (including all the Scandinavian languages) with 135M parameters. The Scandinavian model we use is called *ScandiBERT* (Snæbjarnarson et al., 2023); it is a BERT-like model with 125M parameters trained on Icelandic, Danish, Norwegian, Swedish and Faroese data. Finally, *NorBERT$_3$-base* (Samuel et al., 2023) is a masked language model trained mostly on Norwegian data.

Preliminary experiments showed that the NorBERT$_3$ models performed the best on our dataset, as shown in Table 3. We thus use the NorBERT$_3$ models for further experiments and consider the following sizes from this family of models: *xs* (15M parameters), *small* (40M parameters), and *base* (123M parameters). This allows us to train SLIDE models of varying accuracy-to-speed trade-offs.

Training details Fine-tuning is done using the `transformers` library (Wolf et al., 2020) and the `PyTorch` framework (Ansel et al., 2024). We use binary cross-entropy as the loss function to train the model for multi-label classification.

To find our final hyperparameters, we perform a simple grid search. The models are fine-tuned with a learning rate of $5 \cdot 10^{-5}$, a batch size of 64, 1% warmup steps with a linear scheduler together with the AdamW optimizer. We train the models for 3 epochs (2,877 steps) and load the best checkpoint at the end based on metric performance (weighted multi-label accuracy). Model evaluation is performed on the validation set every 100 training steps. We fine-tuned the three NorBERT$_3$ models in this way and release them as SLIDE-xs, SLIDE-small and SLIDE-base.

Model	Loose accuracy	Exact-match accuracy	Macro F_1
XLM-RoBERTa-base	96.8	94.6	95.4
DistilBERT-base	96.5	94.5	95.2
ScandiBERT	97.6	95.9	96.6
NorBERT3-base	**98.6**	**96.4**	**97.0**

Table 3: **Base model selection** We made our choice based on the validation data split, the metrics in this table, given in percent, are for the test split. F_1 is per-language exact match. NorBERT3 refers to the same model as SLIDE.

Various training set compositions were evaluated; the best model was trained on the multi-label UD dataset combined with the 'alphabet variations' dataset using the punctuation augmentation approach and regular expression normalization described in Section 3.4. We also observe that lowercasing the training set leads to slightly better performance. Therefore, we applied lowercasing to all the training data. While performance typically improves with more training data, this was not observed on our validation set. The final training set has a skewed label distribution: 35% Bokmål, 33% Nynorsk, 13% other, 11% Swedish, and 9% Danish. The validation and test sets reflect similar skews (see Table 1). We briefly tested both upsampling and downsampling to balance labels, but the multi-label nature of the data made this challenging, and it ultimately yielded no improvement.

5.2 Static-word-embedding model (SLIDE-fast)

Since our dataset is smaller than that used to train baseline FastText models, we train a tiny multi-label model instead of concentrating efforts on pre-training a model on our dataset. The model is based on GlotLID sentence embeddings and has 20.9k parameters, not counting the input embeddings. It uses a feed forward network with 1 hidden linear layer of size 64 and a ReLU activation function between it and the output linear layer, and is trained with a regular binary cross-entropy loss. We selected the 0.5 sigmoid threshold to accept a class based on the validation data split. The `other` class is selected only if all other classes are below the threshold. Reducing number of classes from 2,102 to 4 explains faster inference (Table 4) than that of original GlotLID.

Additional Scandinavian data Since a SLIDE-fast model trained on the same training dataset as the larger model does not correctly discriminate Bokmål from Nynorsk and Danish sentences, we enhance the training dataset with additional Bokmål, Nynorsk, Danish and Swedish sentences from the Tatoeba evaluation dataset (automatically labeled in the same way as the UD-based training dataset). NER, punctuation augmentation and regular expression normalization are not applied to the resulting training split.

6 Experiments

We evaluate our SLIDE models against several established LID baselines, comparing both prediction accuracy and speed. Our evaluation focuses on two key aspects: performance on our manually annotated multi-label test set, and generalization to out-of-domain data. We first describe the baseline models used for comparison, then present our main results and the results of our out-of-domain experiments.

6.1 Baselines

We compare against LID models available at the time of writing that support the four Scandinavian languages: FastText-176 (Joulin et al., 2017), NLLB-218 (Grave et al., 2018), NB-Nordic-LID (de la Rosa and Kummervold, 2022), OpenLID (Burchell et al., 2023), GlotLID (Kargaran et al., 2023); Heliport, a faster version of HeLI-OTS (Jauhiainen et al., 2022b)[9], and `gpt2-lang-ident`.

While top-k prediction with confidence scores is possible for the FastText and GPT2-based models, we observe that the confidence scores are unreliable, i.e. there is no consistent threshold value that improves performance, and for all baseline models, except Heliport, the best results are achieved when they are used as single-label classifiers.

6.2 Main results

Table 4 presents the main results of our experiments on the manually-annotated SLIDE test set. We report loose accuracy and exact-match accuracy as overall metrics, along with per-language exact-match F_1 scores for each of the four languages and the ’other’ category. Additionally, we measure inference speed in milliseconds per sentence, averaged over three runs[10].

6.3 Out-of-domain test set

Haas and Derczynski (2021) provide two test sets with single-label annotations, extracted from Wikipedia. In order to evaluate our models on an out-of-domain dataset and compare them with previous work, we use their two test splits containing 3 000 and 14 960 samples respectively and map Icelandic and Faroese to the ‘other’ label. We present the results on these test sets in Table 5.

7 Discussion

Performance of baseline models The baseline models exhibit varying levels of performance, see Table 4 for detailed metrics. These results demonstrate that, while most FastText-based models offer speed advantages, they fall short in accuracy for closely related languages such as Norwegian Bokmål and Norwegian Nynorsk. GlotLID, though slower (0.51 ms/sentence), provides the best performance among the baseline models, with Heliport being a close contender while being significantly faster (0.02 ms/sentence). `gpt2-lang-ident`, originally pretrained as a monolingual English model, fails to tell Danish and two Norwegian languages from each other, while being able to detect Swedish and ‘other’, which again highlights the importance of a dataset focused on Scandinavian languages.

Performance of SLIDE models Our three BERT-based LID models SLIDE-xs, SLIDE-small and SLIDE-base perform the best on our test set, with the `base` version reaching an exact-match accuracy of 96.4%, while the `small` and `xs` both reach 95.7%. This comes at the cost of significantly longer runtimes compared to the static embedding models. These models are suitable when high accuracy is of most importance. However, it is worth noting that we measured inference speed solely on a CPU, one sentence at a time, to ensure a fair comparison with the faster baseline models intended for CPU usage. Using a GPU with larger batch sizes would result in significantly faster runtimes for the transformer models.

[9] `https://github.com/ZJaume/heliport`

[10] Measured on an AMD EPYC 7702 CPU, with a batch size of 1.

Model	Loose accuracy	Exact-match accuracy	NB F_1	DA F_1	NN F_1	SV F_1	Other F_1	Runtime ms/sample
BASELINES								
gpt2-lang-ident	61.2	58.9	47.0	24.0	36.9	83.6	86.2	52.07
FastText-176*	80.5	77.7	72.6	66.0	55.7	92.7	93.5	**0.01**
NLLB-218*	95.3	91.6	93.0	85.9	89.0	96.8	93.6	0.08
NB-Nordic-LID*	83.3	80.6	85.0	67.0	84.8	89.7	70.2	0.02
OpenLID*	94.2	90.2	91.5	82.6	88.7	95.7	93.3	0.08
GlotLID*	97.2	93.4	93.5	89.5	89.4	97.9	98.1	0.51
Heliport (HeLI-OTS)	96.5	92.6	90.9	89.0	91.2	97.6	97.2	0.02
OUR MODELS								
SLIDE-fast	95.7	93.4	94.5	90.2	92.4	97.5	96.4	0.16
SLIDE-x-small	97.8	95.7	97.5	90.4	96.2	98.0	98.7	13.22
SLIDE-small	98.1	95.7	97.7	89.9	96.3	98.0	99.1	19.70
SLIDE-base	**98.6**	**96.4**	**98.1**	**92.0**	**97.1**	**98.6**	**99.4**	38.41

Table 4: **Detailed results on the manually-annotated multi-label SLIDE test split** The best result for each metric is typeset in bold; higher values are always better, except for the runtimes. * shows which baselines use FastText.

While our SLIDE-fast model reaches the same exact-match accuracy as GlotLID, 93.4%, it performs better on Nynorsk, Bokmål and Danish, with Nynorsk performance increasing by 3%.

Overall, performance on Danish is consistently the lowest—the best model reaches 92% F_1. Our models have been trained on more Bokmål than Danish data, and we observe a slight tendency to predict only Bokmål instead of both Bokmål and Danish for multilingual samples. We do, however, notice the same trend with lower Danish performance across all evaluated models, see Table 4.

As seen in Table 2, the punctuation augmentation led to minor performance improvements. The main motivation behind this approach, however, is increased robustness to noisy data. While the model trained with named entity swapping (see Section 3.4) gained the highest loose accuracy performance, 98.7%, it performed poorly on exact-match accuracy, 95.5%. We therefore decided not to include this in the final SLIDE models.

Error analysis Common error sources are proper names (half of ‘other’ instances misclassified as Scandinavian contains proper names (e.g. ‘kruvi: *Karl Marx*’), instances in English (30% of ‘other’ instances misclassified as Scandinavian), and loanwords (‘- Ta avisa *Kommersant*.’, ‘Server med pastasalat med bakte grønsaker og *tsatsiki* til’, ‘Men Anne Linnet - *oh la la*.’) Bokmål and Nynorsk are confused most often. If a sentence valid both in Bokmål and Nynorsk contains irregular Bokmål spelling like ‘høg’ instead of ‘høy’, and ‘tjuvfiske’ instead of ‘tyvfiske’, it is likely to be misclassified as Nynorsk only. Some errors imply that particular tokens influence the prediction more than a sentence representation as a whole: ‘høyre’ is a valid word both in Nynorsk (‘hear’) and Bokmål (‘right’), but the sentence ‘I alle år har vi fått *høyre* at med dagens forbruk er det olje nok for mange tiår.’, which is Nynorsk because of ‘høyre’ used as a verb, is misclassified as Bokmål, while a both Bokmål and Nynorsk sentence ‘I den nye designen er *høgre* og venstre spalte på framsida til nettavisa fjerna.’ is misclassified as only Nynorsk because of the spelling. Additionally, some ‘other’ instances containing subwords matching those in Scandinavian are misclassified, although the whole sentence semantics does not make any sense: ‘Va shiaulteyr *er ny* skeabey harrish boayrd.’ (Manx).

Out-of-domain evaluation In order to ensure that we do not overfit to the UD data, we evaluate our models on the out-of-domain test set presented in Section 6.3, which was the only LID dataset specific for Scandinavian languages available at the

Model	3K test split	15K test split
SLIDE-base	92.7	95.3
SLIDE-fast	85.4	88.5
GlotLID	**93.0**	**95.7**

Table 5: **Performance on an out-of-distribution single-labeled datasets** Accuracy on the test sets from Haas and Derczynski (2021). As this dataset is single-label, we consider a prediction to be correct, if one of the predicted languages is correct.

time of writing. While SLIDE-base reaches lower performance than GlotLID on this test set, we must add that this dataset is heavily preprocessed: lowercased and stripped out of numbers, punctuation signs and some accented characters. We also noticed a fair amount of mislabeled sentences in the dataset, with sentences like "ou di be t aatm ne en wadi", "atahualpa yupanqui" and "tromssan ruijansuomalainen yhdistys" being labeled as Swedish, Danish and Nynorsk, respectively. Furthermore, this dataset contains Icelandic and Faroese as the `other` languages, which are similar to Nynorsk in many cases. In short, we cannot draw confident conclusions from this result, but it hints at the worst-case performance of our models on out-of-distribution inputs.

8 Conclusion

We release a novel multi-label LID dataset for Danish, Norwegian Bokmål, Norwegian Nynorsk and Swedish with manually annotated validation and test splits. Using machine translation for creating a silver multi-label training dataset from a single-label one has proved to be efficient.

Although fine-tuning models for a specific data source may be helpful to obtain high performance on a selected test set, such models (especially the FastText-based ones) may be not robust towards the test dataset change. Also, excessive training data preprocessing may lead to performance degradation on data from unknown domains compared with training without any preprocessing.

Limitations

We limit ourselves to the larger Scandinavian languages, and include neither the other closely related Nordic languages Faroese and Icelandic (also known as Insular Scandinavian), nor the smaller Scandinavian varieties with a limited written tradition, such as Scanian, Elfdalian and Bornholmsk. We also do not look at other sources of variation, e.g., dialectal, diachronic or otherwise different varieties found in literature or social media.

Another limitation is that while all Norwegians generally understand Swedish and Danish well, as these languages are a compulsory part of the public curriculum, and also teaching languages of Norwegian universities, their productive capabilities are much lower, and there might be cases of mislabeling.

Acknowledgments

We would like to thank Helene Bøsei Olsen and Karoline Sætrum for their work on annotating the initial version of the test set. Some computations were performed on resources provided by Sigma2 – the National Infrastructure for High-Performance Computing and Data Storage in Norway.

References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings* *of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

Ruth Vatvedt Fjeld, Anders Nøklestad, and Kristin Hagen. 2020. Leksikografisk bokmålskorpus (LBK) – bakgrunn og bruk. *Oslo Studies in Language*, 11(1):47–59.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

René Haas and Leon Derczynski. 2021. Discriminating between similar Nordic languages. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 67–75, Kiyv, Ukraine. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022a. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 3912–3922, Marseille, France. European Language Resources Association.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022b. HeLI-OTS, off-the-shelf language identifier for text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3912–3922, Marseille, France. European Language Resources Association.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

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.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Bård Borch Michalsen. 2015. *Komma. Kommategnets personlighet, historie og regler*. Juritzen forlag.

Aleksandra Miletić and Filip Miletić. 2024. A gold standard with silver linings: Scaling up annotation for distinguishing Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 36–46, Torino, Italia. ELRA and ICCL.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Eric Papazian. 2013. Moltke moe og norsk språknormering fram til 1907. *Språklig samling*, pages 69–104.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Javier de la Rosa and Per Egil Kummervold. 2022. NB-Nordic-LID.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a benchmark for Norwegian language models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

David Samuel, Vladislav Mikhailov, Erik Velldal, Lilja Øvrelid, Lucas Georges Gabriel Charpentier, and Andrey Kutuzov. 2024. Small languages, big models: A study of continual training on languages of norway.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Mahesh Bangera. 2024. Language variety identification with true labels. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10100–10109, Torino, Italia. ELRA and ICCL.