CUFE@VarDial 2025 NorSID: Multilingual BERT for Norwegian Dialect Identification and Intent Detection

Michael Ibrahim
Computer Engineering Department, Cairo University
1 Gamaa Street, 12613
Giza, Egypt
michael.nawar@eng.cu.edu.eg

Abstract

Dialect identification is crucial in enhancing various tasks, including sentiment analysis, as a speaker's geographical origin can significantly affect their perspective on a topic, also, intent detection has gained significant traction in natural language processing due to its applications in various domains, including virtual assistants, customer service automation, and information retrieval systems. This work describes a system developed for VarDial 2025: Norwegian slot and intent detection and dialect identification shared task (Scherrer et al., 2025), a challenge designed to address the dialect recognition and intent detection problems for a low-resource language like Norwegian. More specifically, this work investigates the performance of different BERT models in solving this problem. Finally, the output of the multilingual version of the BERT model was submitted to this shared task, the developed system achieved a weighted-F1 score of 79.64 for dialect identification and an accuracy of 94.38 for intent detection.

1 Introduction

Norwegian dialects represent a rich tapestry of linguistic diversity that reflects the historical, geographical, and social nuances of Norway. The country is home to a multitude of dialects, often categorized into four primary groups: Northern Norwegian (nordnorsk), Central Norwegian (trøndersk), Western Norwegian (vestlandsk), and Eastern Norwegian (østnorsk). These dialects are not merely regional variations; they embody unique grammatical structures, vocabulary choices, and phonetic characteristics that can vary significantly even within short distances. For instance, a speaker from Bergen may find it challenging to understand someone from Oslo due to the distinct phonetic and syntactic features of their respective dialects. This variation poses considerable challenges for linguistic studies and applications in natural language processing (NLP).

The ability to identify and differentiate between these dialects is crucial for various applications, including speech recognition systems, language learning tools, and sociolinguistic research. A study found that Norwegians are generally less adept at identifying their own dialects compared to speakers of other languages, such as Dutch (Gooskens, 2005). This suggests that while Norwegians are exposed to multiple dialects throughout their lives, the cognitive mechanisms underlying dialect identification may not be as finely tuned as previously thought. Furthermore, the role of intonation in identifying these dialects has been highlighted as particularly significant.

In recent years, advancements in machine learning and NLP have opened new avenues for addressing these challenges. Among these advancements, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a powerful tool for various language understanding tasks. BERT utilizes a transformer architecture that allows it to capture contextual relationships between words in a sentence effectively (Devlin, 2018). This capability is particularly valuable for dialect identification, where subtle differences in word usage or syntax can indicate distinct regional affiliations.

In this paper, the fine-tuning of multiple BERT-based models for identifying Norwegian dialects and detecting the intent of Norwegain text was explored. Different pre-trained models, including a Norwegian-specific BERT variant and multilingual BERT models were investigated to measure their efficacy in dialect identification and intent detection tasks. By leveraging transfer learning and fine-tuning strategies, the model's understanding of Norwegian dialects and text intent was improved, even in the context of a relatively under-resourced language, like Norwegian.

The rest of the paper is organized as follows:

Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 220–223
January 19, 2025. ©2025 Association for Computational Linguistics

Section 2 provides an overview of related research. Section 3 describes the dataset used for training and validation. Section 4 outlines the system, and Section 5 concludes the paper.

2 Related Work

BERT: Bidirectional Encoder Representations from Transformers (Devlin, 2018) has revolutionized the field of natural language processing by providing a pre-trained model that effectively captures bidirectional context in text. Unlike earlier models like word2vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) which generate static word embeddings, BERT uses a deep bidirectional transformer architecture (Vaswani, 2017) to produce dynamic representations of words based on their context. This pre-training, performed on large corpora like the English Wikipedia, allows BERT to excel across a wide range of downstream tasks through fine-tuning, including question answering, named entity recognition, and text classification.

Norwegian BERT: With the growing interest in NLP applications for low-resource languages, the development of Norwegian-specific transformer models has been pivotal. The NoTraM project has created a Norwegian transformer model that outperforms multilingual BERT (mBERT) on various classification tasks, including intent detection (Kummervold et al., 2021). This model demonstrates that fine-tuning language models specifically on Norwegian text can yield significant improvements in performance compared to generic models.

Norwegian BERT Application: BERT has revolutionized NLP by enabling models to understand context in a bidirectional manner, making it particularly effective for tasks involving nuanced language understanding. (Mæhlum et al., 2022) demonstrated the potential of a Norwegian BERT model for morphosyntactic analysis, highlighting its capacity to handle the complexities of dialectal variations. The model's architecture allows it to capture contextual relationships between words, which can be used for distinguishing between different dialects that may share similar vocabulary but differ significantly in usage.

The NoMusic Corpus: The introduction of the NoMusic corpus (Mæhlum and Scherrer, 2024) represents a significant advancement in resources available for Norwegian dialect identification and intent detection. This corpus consists of translations from the xSID dataset (van der Goot et al., 2021) into standard Norwegian Bokmål and eight dialects from three major Norwegian dialect areas. This corpus represents the first evaluation dataset focusing on non-standard Norwegian varieties, allowing researchers to analyze linguistic variations across different dialects systematically.

BERT for Dialect Identification: Mawdoo3 AI has developed a Multi-Dialect Arabic BERT model specifically for country-level dialect identification. This model was trained on a dataset comprising 21,000 labeled tweets from all 21 Arab countries and achieved a micro-averaged F1-score of 26.78% in the NADI shared task (Talafha et al., 2020). The success of this model highlights the efficacy of fine-tuning pre-trained transformer models for specific dialectal tasks.

BERT for Intent Detection: BERT has been effectively utilized for citation intent classification within academic texts. A study analyzed various BERT models fine-tuned on labeled datasets to classify citation intents and sentiments, revealing that BERT's contextual capabilities enhance its performance in understanding nuanced academic language (Visser and Dunaiski, 2022). While this application is not directly focused on conversational intent detection, it underscores BERT's versatility across different domains and its potential for enhancing understanding in specialized contexts like academic discourse.

3 Dataset

A subset of the Norwegian Multi-Dialectal Slot and Intent Detection Corpus (NoMusic) (Mæhlum and Scherrer, 2024) was used for training and testing this system. The NoMusic corpus was created by translating the xSID dataset, an evaluation dataset for spoken language understanding (slot and intent detection) (van der Goot et al., 2021) to eight Norwegian dialects and Norwegian Bokmål. For dialect identification, the development dataset consists of 3300 sentences, and the testing dataset consists of 5500 sentences, representing the four main dialects in Norway: North Norwegian, Trøndersk, West Norwegian and Bokmål. The data distribution among the four dialects is summarized in table 1

For intent detection, the development dataset consists of 3300 sentences and the testing dataset consists of 5500 sentences, representing 16 intents. The data distribution among the sixteen intents is

Language	Development	Testing
North Norwegian	600	1000
Trøndersk	900	1500
West Norwegian	1500	2500
Bokmål	300	500

Table 1: VarDial 2025 Norwegian slot and intent detection and dialect identification task - dialect identification data split statistics.

Language	Dev.	Testing
Add To Playlist	209	374
Book Restaurant	286	473
Play Music	264	429
Rate Book	165	352
Search Creative Work	209	363
Search Screening Event	253	407
alarm/cancel alarm	242	341
alarm/modify alarm	11	0
alarm/set alarm	264	319
alarm/show alarms	110	209
alarm/snooze alarm	22	33
alarm/time left on alarm	0	44
reminder/cancel reminder	110	198
reminder/set reminder	143	407
reminder/show reminders	132	209
weather/find	880	1342

Table 2: VarDial 2025 Norwegian slot and intent detection and dialect identification task - intent detection data split statistics.

summarized in table 2. Even though that all the intents in table 2 are available in the English, and the Norwegian-translated train dataset, you can notice that some intents are present in the development and missing from the testing set like “alarm/modify alarm”, and some intents are available in the testing set and missing from the development set like “alarm/time left on alarm” which makes it impossible for transformer-based models like BERT to detect a class like “alarm/modify alarm” when trained on the development dataset only.

4 Methodology

For the dialect identification, three versions of BERT were considered: “NbAiLab/nb-bert-base” (Kummervold et al., 2021), “ltgoslo/norbert” (Kutuzov et al., 2021), and “bert-base-multilingual-cased” (Devlin et al., 2018). Each BERT classifier was fine-tuned on 80% of the development data for 10 epochs with a learning rate of $2e-5$ and a batch size of 32, then the accuracy of each of those models was calculated based on the remaining 20% of the development set, the performance of those models is summarized in table 3.

Model	Accuracy
NbAiLab/nb-bert-base	74.61
ltgoslo/norbert	73.33
bert-base-multilingual-cased	79.91

Table 3: Performance of the fine-tuned BERT models for dialect identification on the development set.

Due to the difference in the performance between the “bert-base-multilingual-cased” model and the remaining BERT models, the multilingual BERT model used for final submission.

For the dialect identification task, the shared task organizers provided the baseline model from the ITDI shared task (Aepli et al., 2023), which employs a Support Vector Machine (SVM) classifier with TF-IDF-weighted features of character 1-to-4-grams. The baseline achieved a weighted F1-score of 77.42, whereas the developed multilingual BERT model surpassed it with a weighted F1-score of 79.64.

Similarly, for the intent dectection, three versions of BERT were considered: “NbAiLab/nb-bert-base” (Kummervold et al., 2021), “ltgoslo/norbert” (Kutuzov et al., 2021), and “bert-base-multilingual-cased” (Devlin et al., 2018). Each BERT classifier was fine-tuned on 80% of the development data for 10 epochs with a learning rate of $2e-5$ and a batch size of 32, then, the accuracy of each of those models was calculated based on the remaining 20% of the development set, the performance of those models is summarized in table 4.

Due to the difference in the performance between the “bert-base-multilingual-cased” model and the remaining BERT models, the multilingual BERT model was used for the final submission.

The developed BERT model significantly outperforms the baseline BERT model provided by the shared task organizers. The baseline utilizes an mBERT encoder with two separate decoder heads: one for slot detection and another for intent classification. While the baseline achieved an accuracy of 84.15% on the intent detection task, the developed multilingual BERT model achieved an accuracy of 94.38%.

Model	Accuracy
NbAiLab/nb-bert-base	92.30
ltgoslo/norbert	91.03
bert-base-multilingual-cased	95.88

Table 4: Performance of the fine-tuned BERT models for intent detection on the development set.

5 Conclusion and Future Work

In this paper, we have explored the efficacy of multiple BERT models in the tasks of Norwegian dialect identification and intent detection. The multilingual version of BERT produced the best results on the development data. Finally, the output of the multilingual version of the BERT model was submitted to this shared task, and it achieved a weighted-F1 score of 79.64 for dialect identification and an accuracy of 94.38 for intent detection on the test datasets.

Future work will focus on improving the BERT model by leveraging Norwegian-translated dataset to address the challenges posed by missing intents in the development dataset. Missing intents can impede the model's ability to learn comprehensive patterns for intent recognition, leading to suboptimal performance. By augmenting the development dataset with the Norwegian-translated dataset, we can introduce linguistic diversity and contextual richness, compensating for the gaps in the development dataset.

References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the vardial evaluation campaign 2023. *arXiv preprint arXiv:2305.20080*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Charlotte Gooskens. 2005. How well can norwegians identify their dialects? *Nordic Journal of Linguistics*, 28(1):37–60.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for norwegian. *arXiv preprint arXiv:2104.06546*.

Petter Mæhlum, Andre Kåsen, Samia Touileb, and Jeremy Barnes. 2022. Annotating norwegian language varieties on twitter for part-of-speech. *arXiv preprint arXiv:2210.06150*.

Petter Mæhlum and Yves Scherrer. 2024. Nomusic-the norwegian multi-dialectal slot and intent detection corpus. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 107–116.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Yves Scherrer, Rob van der Goot, and Petter Mæhlum. 2025. VarDial evaluation campaign 2025: Norwegian slot and intent detection and dialect identification. In *Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein T Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. *arXiv preprint arXiv:2007.05612*.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanovic, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked-language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Mexico City, Mexico. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Ruan Visser and Marcel Dunaiski. 2022. Sentiment and intent classification of in-text citations using bert. In *Proceedings of 43rd Conference of the South African Insti*, volume 85, pages 129–145.