SP-BERT: A Language Model for Political Texts in Scandinavian Languages

Tu My Doan, Benjamin Kille, and Jon Atle Gulla

NorwAI Norwegian Research Center for AI Innovation

NTNU

Abstract

Language models are at the core of modern Natural Language Processing. We present a new BERT-style language model dedicated to political texts in Scandinavian languages. We introduce SP-BERT, a model trained with parliamentary speeches in Norwegian, Swedish, Danish, and Icelandic. We evaluate its ability to predict the speakers' party affiliations and explore language shifts of politicians transitioning between Cabinet and Opposition. We find a clear distinction on the use of nouns in health-care related speeches.

Motivation

- Transformer-based Language Models are useful tools
- Using a particular data set helps to work on a particular domain or task
- Politics represents a domain where language plays a major role
- Politicians communicate with citizens and voters
- **Research Question:** Can we build an LLM that reflects Scandinavian Politics?

Data

Norwegian datasets: We obtain parliamentary speeches from three sources:

- The Talk of Norway (ToN): a rich annotated dataset containing 250 373 speeches from the Norwegian Parliament in the period from 1998 to 2016.
- Norwegian Parliamentary Speech Corpus (NPSC): a dataset with 9722 speeches from Norwegian Parliamentary meetings from 2017 to 2018.
- We crawled 3158 more speeches (January 2019 to February 2023) from the Norwegian Parliamentary website using their API.

Swedish dataset: We obtained 355 059 speeches from the Swedish parliament from the ParlSpeech(V2) dataset and more recent speeches from the Swedish Parliament website.

Danish dataset: The 455 076 speeches from the Danish parliament also come from the ParlSpeech(V2) dataset.

Icelandic dataset: We obtain 388 650 speeches from the Icelandic parliament using the IGC-Parl corpus.

Training and Evaluation

SP-BERT Training Continued pre-training of mBERT: 1M steps with batch size 128, sequence length 256, and learning rate 1×10^{-4} on NVIDIA A100 80GB GPUs. Then, another 0.5M steps with sequence length 512, and batch size 64.

Model	Task 1	Task 2				
	Acc	Acc	(best)		(worst)	F_{1macro}
Norwegian						
bert-base-multilingual-cased	0.627	0.360	0.515	(3)	0.208 (5)	0.358
nb-bert-base	0.643	0.450	0.603	(3)	0.250 (5)	0.449
sp-bert-base (ours)	0.636	0.465	0.620	(4)	0.268 (2)	0.457
weight-bert-base-multilingual-cased	0.571	0.409	0.598	(4)	0.308 (6)	0.410
weight-nb-bert-base	0.602	0.418	0.542	(5)	0.283 (2)	0.426
weight-sp-bert-base (ours)	0.638	0.470	0.641	(4)	0.286 (1)	0.465
Swedish						
bert-base-multilingual-cased	0.822	0.651	0.825	(1)	0.440 (5)	0.593
bert-base-swedish-cased	0.877	0.692	0.856	(1)	0.370 (6)	0.625
sp-bert-base (ours)	0.871	0.681	0.858	(1)	0.399 (6)	0.626
weight-bert-base-multilingual-cased	0.855	0.603	0.716	(1)	0.400 (6)	0.554
weight-bert-base-swedish-cased	0.882	0.664	0.728	(1)	0.516 (2)	0.630
weight-sp-bert-base (ours)	0.878	0.663	0.757	(3)	0.465 (6)	0.619

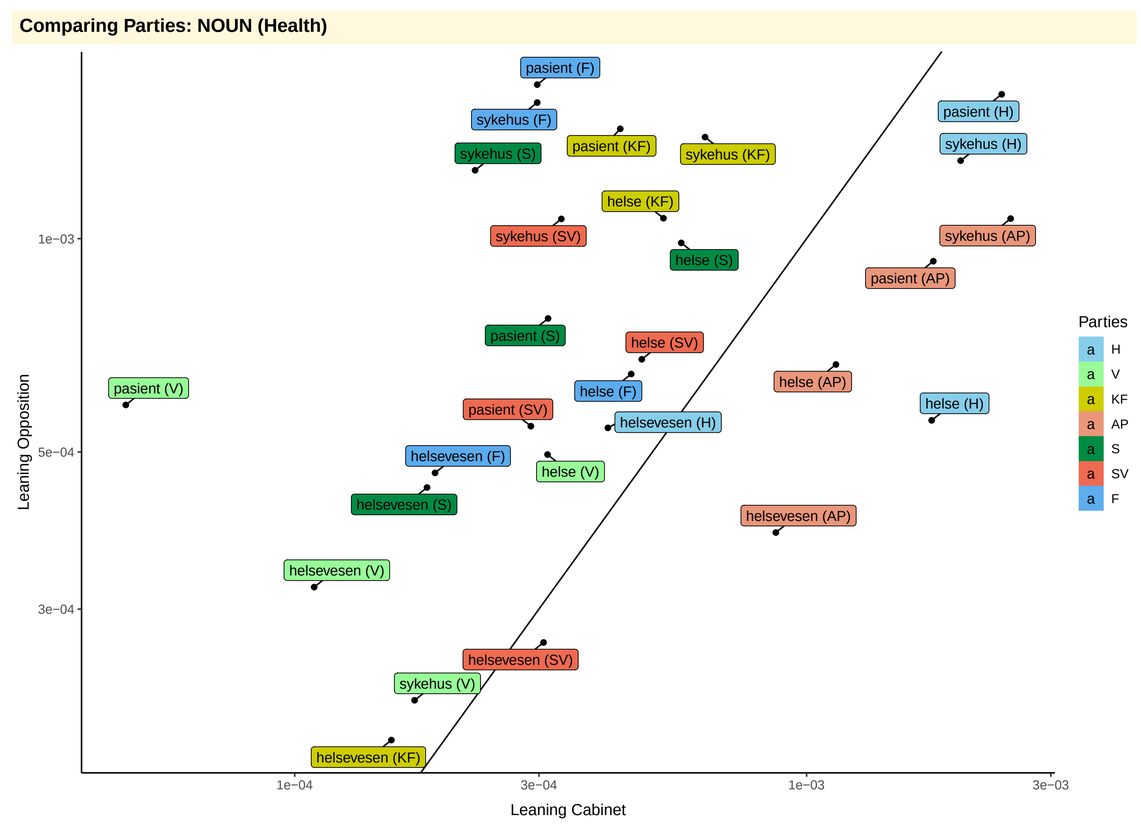
Evaluation (for Norwegian and Swedish Speeches)

- **Task 1 (Binary Classification)** The task classifies speech's political leaning. We matched the parties to either left or right following discussions with experts.
- **Task 2 (Multi-label Classification)** This is a multi-class problem to predict speaker's party affiliation. We omit parties that have very few items.

Baselines Multilingual Models (mBERT), language-specific models (nb-BERT, Swedish BERT).

Politicians' Word Usage

1. Part-of-Speech Tagger to obtain word classes
2. Lemmatizer to reduce word forms
3. Remove words with less than four characters
4. Compute frequencies of politicians' word usage in government and opposition
5. Chart both values in a two dimensional scatter plot
6. Add line for equal use of a term in government and opposition
7. Filter word classes and subjects



Conclusion

Large Language Models allow us to address many problems in Natural Language Processing. Training them with data specific to a domain and tasks yields performance improvements. In the future, these models can be used to better represent language and allow automated analysis of political speeches and texts. The results have been published at NLDB'23 [1].

Acknowledgement

This work is done as part of the Trondheim Analytica project and funded under Digital Transformation program at Norwegian University of Science and Technology (NTNU), 7034 Trondheim, Norway.

References

[1] T. M. Doan, B. Kille, and J. A. Gulla. SP-BERT: A Language Model for Political Text in Scandinavian Languages. In *International Conference on Applications of Natural Language to Information Systems*, pages 467–477. Springer, 2023.