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

Tu My Doan (✉)[0000−0002−9440−5847], Benjamin Kille[0000−0002−3206−5154], and Jon Atle Gulla[0000−0002−9806−7961]

Norwegian University of Science and Technology, Trondheim, Norway
{tu.m.doan, benjamin.u.kille, jon.atle.gulla}@ntnu.no

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. Concretely, we introduce SP-BERT, a model trained with parliamentary speeches in Norwegian, Swedish, Danish, and Icelandic. To show its utility, we evaluate its ability to predict the speakers' party affiliation and explore language shifts of politicians transitioning between Cabinet and Opposition.

Keywords: SP-BERT · Scandinavian LM · Political Text Mining

1 Introduction

Political texts are pervasive. They are available in many forms from political manifestos, over political speeches, and debates, to news articles. They constitute an important resource for social and political study. Analysing political texts raises challenges when dealing with large amounts of data. The complexity of political languages and their nuances make the task even more challenging, especially for those lacking political background. Political domain is also known to be complex, and hard to analyze. This holds also for Norwegian politics. The use of language, both written and spoken, plays a crucial role in shaping political discourse and decision-making.

Large Language Models (LLMs) are proven to be powerful tools in the field of Natural Language Processing (NLP). Pre-trained LLMs capture the language's complexity and represent texts. However, LM resources are rare for Norwegian politics. To fill in the gap, we introduce SP-BERT—a pre-trained BERT LM for Scandinavian Politics in four languages: Norwegian, Swedish, Danish and Icelandic. As a use case, we use SP-BERT to identify the shifts in Norwegian politics by learning text representation. We also analyse the changes in word choices by politicians when they switch positions (Opposition vs. Cabinet). We aim to gain a deeper understanding of the current state of Norwegian politics through linguistic strategies used by politicians and political parties. Furthermore, language models' ability to efficiently process vast amounts of data enables faster analysis for political domain than traditional methods. We define two research questions:

- Does the LM trained exclusively on political text outperform the multilingual BERT and/or language-specific BERT model on the task of classifying Norwegian and Swedish political text? (RQ_1)
- Does being in government/opposition change how politicians express their views in Norwegian politics? In which way? (RQ_2)

The rest of the paper is structured as follow: Section 2 conveys related work for LM in Politics, and political analysis. Details about SP-BERT model are described in Section 3. Section 4 discusses shifts in Norwegian political speeches. Section 5 concludes the paper with suggestions for future research.

2 Related Work

This section presents related work concerning language models for politics and analyzing political texts.

2.1 Language Models for Politics

Since its introduction, BERT [4] has been used for various tasks in NLP. Especially for English, BERT achieves state-of-the-art results. Working with languages other than English represents a challenge. Researchers either rely on multilingual models, such as mBERT [4], or gather data and pre-train a model dedicated to their target language. Training is both costly and time-consuming.

In Northern Europe, we find NB-BERT [11] and NorBERT [12] for Norwegian, and models for Swedish [17], Danish [9], Finish [23], and Icelandic [20]. To capture as much of the language as possible, researchers tend to include as much text as possible. This leads to generic models. In contrast, for English we observe more specialized models. Liu et al. [15] trained a language model for political texts that focuses on ideology and stance detection. Hu et al. [8] presented ConfliBERT dedicated to deal with texts concerning political conflict and violence. To the best of our knowledge, there is not yet a model that deals with politics for smaller-scale language such as the language family in Northern Europe.

2.2 Political Analysis

Political Science relies on analysing texts. Abercrombie and Batista-Navarro [1] study the semantic changes in UK Parliamentary Debates. Chen et al. [3] focus on analysing the political bias and unfairness of news articles at different levels in the US. Maronikolakis et al. [18] analyse the political parody in social media for US, UK, and the rest of the world. Walter et al. [24] analyse the ideological bias for German parliamentary proceedings. Magnusson et al. [16] analyse the Swedish parliamentary debates.

The numerous examples of text analyses for Political Science demonstrates the need for tools to process more texts. Hence, we build a large pre-trained language model specifically for political texts to support political research. Shared culture and language in Northern Europe let us assume that the model will provide a benefit to all political scientists in that region.

3 SP-BERT Language Model

We introduce SP-BERT, a pre-trained language model for political text in Norway, Sweden, Denmark, and Iceland. The section describes the data sources, pre-processing, the training procedure, and an evaluation set up.

3.1 Corpora

We focus on data sources for Norwegian, Swedish, Danish, and Icelandic that relate to politics. Parliamentary speeches fulfil that requirement.

Norwegian datasets: we obtain parliamentary speeches from three sources:

- The Talk of Norway (ToN) [13]: a rich annotated dataset containing 250 373 speeches from the Norwegian Parliament in the period from 1998 to 2016.
- Norwegian Parliamentary Speech Corpus (NPSC) [21]: a speech dataset with recordings from Norwegian Parliamentary meetings from 2017 to 2018. The dataset has 64 531 sentences from about 9722 speeches.
- Due to low number of speeches in Norwegian Parliamentary, we decided to crawl more data from the Norwegian Parliamentary website using their API [1]. We have collected data from 01 Jan 2019 up to end of February 2023. As a result, we obtained 3158 additional speeches.

Swedish dataset: Data come from the ParlSpeech(V2) dataset [19] — a full-text corpora of 6.3M parliamentary speeches from nine European countries. Data were collected from 02 October 1990 to 21 December 2018. The corpus contains 355 059 speeches. In addition, we obtained more recent data from the Swedish Parliament website [2].

Danish dataset: The Danish parliamentary speeches also come from the same source as the Swedish [19]. This corpus contains 455 076 Danish speeches from 07 October 1997 to 20 December 2018.

Icelandic dataset: We use data coming from the IGC-Parl corpus [22]. There are 388 650 speeches in the time span from 1911 to 2020.

3.2 Data Pre-processing

We pre-processed the texts collected from all sources. Typically, speeches start off with a short reference to the parliament's president. We employed a set of regular expression to remove those. Further, we eliminated redundant white spaces and removed markup. We removed speeches with less than 60 tokens as these were frequently questions or answers and not speeches. As a result, we obtained a data set of about 1.44 million speeches. They split into 16 % Norwegian, 32 % Danish, 25 % Swedish, and 27 % Icelandic. We took part of the data to form two evaluation sets (see Table 1), and kept the rest for training.

[1] https://data.stortinget.no/om-datatjenesten/bruksvilkar/
[2] https://data.riksdagen.se/data/anforanden/

Table 1: Evaluation datasets summary. The numerical values represent the labels we use in fine-tuning classifiers. For Task 1, we labelled data based on political leaning of the party. In Task 2, data was labelled based on political party.

Task 1: Political Leaning Classification			
Label (Political Leaning)	# items	Label (Political Leaning)	# items
0: Left	1300	0: Left	18 802
1: Right	1471	1: Right	18 308
Sum	2771	Sum	37 110
Task 2: Party Affiliation Classification			
Label (Party Name)	# items	Label (Party Name)	# items
0: Right	539	0: Moderate Rally	6418
1: Left	256	1: Sweden's Social Democratic Labor	11 897
2: Labor	604	2: The Christian Democrats	2981
3: Center	334	3: Left	3296
4: Progress	473	4: Green	3504
5: Christian People's	203	5: Liberals	2053
6: Socialist Left	362	6: Center	2601
		7: Sweden Democrats	4360
Sum	2771	Sum	37 110

3.3 Experimental Setup

Pre-training Setup To train SP-BERT[3], we follow the approach by original BERT paper [4]. Since removing Next Sentence Prediction (NSP) loss helps improve the performance in downstream tasks [14], we only keep the Masked Language Model (MLM) training objective. The architecture of mBERT model [4] is similar to the original BERT. It has 12 layers, 12 attention heads, and 768 hidden dimensions. As we want to build a pre-trained LM for multi-languages, mBERT serves as a good starting point [4]. This model has already been trained on more than 100 languages including Norwegian (Bokmål and Nynorsk), Swedish, Danish, and Icelandic. We first train the model for 1M steps on batch size 128, sequence length 256, learning rate 1×10^{-4}, with Adam optimizer [10] on NVIDIA A100 80GB GPUs. Later, we change the sequence length to 512, batch size 64 and continue the training for another 0.5M steps to learn better context.

Fine-tuning Setup To find the best hyper-parameters for each task, we experiment with batch size of 64, learning rate $\in$ {5e-5, 4e-5, 3e-5}, sequence length of 512 for max 10 epochs. To evaluate the performance, we use *Accuracy* and F_1 score. Besides doing full fine-tuning the classifier, we also explore using class weight[5] when dealing with imbalanced dataset following the work done in [5].

[3] https://huggingface.co/tumd/sp-bert
[4] https://huggingface.co/bert-base-multilingual-cased
[5] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

3.4 Evaluation

We are interested in the performance for politics-specific tasks. Consequently, we define two evaluation tasks to assess the model's performance. There is no public data set for this purpose which is why we created two datasets. One of them is in Norwegian while the other is in Swedish. We fine-tune a classifier for political text. Table 1 shows the distribution of speeches.

- **Task 1 (Binary Classification)** The task concerns classifying the speech's political leaning. We matched the parties to either left or right following discussions with experts.
- **Task 2 (Multi-label Classification)** We define a multi-class problem in predicting the speaker's party affiliation. We omit parties that have very few items in each language (see Table 1).

3.5 Results and Discussions

Table 2 shows results on both binary and multi-label task. The left column shows the model. We consider three models for each language. First, the multilingual BERT model serves as a baseline. Second, we use a language-specific BERT model. Third, we present our proposed model. Besides, we illustrate a fine-tuned version of each model which puts more weights on the minority classes. For the binary classification, we present the accuracy for each model. SP-BERT achieves the best accuracy for Norwegian. The Swedish BERT model performs slightly better for Swedish.

Concerning the multi-class problem, we show the accuracy, macro F_1 score, as well as the accuracy for the best and worst class. For Norwegian, we observe a similar picture. SP-BERT performs best. The Swedish BERT model performs somewhat better for Swedish. Looking at the column with the worst class-specific score, we notice that the weighting improves the performance. Either the score for the class improves or a different class becomes the worst performing.

In our analysis, we follow the work in [6] who define a method to compare two binary classifiers. Suppose, we compare two binary classifiers A and B. Both supply a set of predictions for L given texts $\{\hat{y}_A\}_{\alpha=1}^{L}$ and $\{\hat{y}_B\}_{\alpha=1}^{L}$. We distinguish three cases: $\hat{y}_A = \hat{y}_B$ indicating that both classifiers agree; $\hat{y}_A = y, \hat{y}_B \neq y$ indicating that classifier A predicts correctly whereas B fails; $\hat{y}_A \neq y, \hat{y}_B = y$ indicating that classifier A fails whereas B predicts correctly. Goutte and Gaussier [6] show that having counted the number of instances for the three cases, we can approximate the distributions with a Dirichlet:

$$\Pr(\pi|Z,\alpha) \sim \text{Dirichlet}(N_1+\alpha_1, N_2+\alpha_2, N_3+\alpha_3), \tag{1}$$

where π refers to the probability of each case, Z refers to the counts (N_1, N_2, N_3), and α captures the prior information. In our evaluation, we consider $\alpha = \frac{1}{2}$. We employ Markov Chain Monte Carlo (MCMC) with NUTS [7] and four chains to generate 100 000 samples from the posterior distributions in each comparison.

Table 2: Results of the sequence classification for Task 1 and 2. We report the Accuracy for both tasks, and the best and worst class performance with F_{1macro} for Task 2. The values are highlighted with gradient colors with incremental intensity from low to high. Best values in the group have more intense color.

Model	Task 1	Task 2					
	Acc	Acc	(best)		(worst)		F_{1macro}
Norwegian							
bert-base-multilingual-cased [a]	0.627	0.360	0.515	(3)	0.208	(5)	0.358
nb-bert-base [b]	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 [a]	0.822	0.651	0.825	(1)	0.440	(5)	0.593
bert-base-swedish-cased [c]	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

[a] https://github.com/google-research/bert
[b] https://github.com/NBAiLab/notram
[c] https://huggingface.co/KB/bert-base-swedish-cased

Figure 1 shows the posterior distributions for two comparisons (more comparisons omitted due to space limitations). Each plot presents two densities. The densities express the probability that one language model performs better than the other.

4 Identifying Shifts in Norwegian Politics

We investigate the shifts in the position of politicians and parties.

4.1 Politicians' position shifts with SP-BERT representations

First, we take the ToN data [6] and check whether Principal Component Analysis (PCA) projects speeches into discernible sub-spaces. Concretely, we take speeches of politicians about the same topic if they gave them both as member of the government and opposition. Subsequently, we apply PCA to their embeddings obtained with SP-BERT model. Specifically, we fine-tuned SP-BERT on

[6] Due to space limitation, we omit the detailed pre-processing steps

classifying party position (Opposition/Cabinet) to help model learn about the party position differences, thus, improve visualization of obtained embeddings.

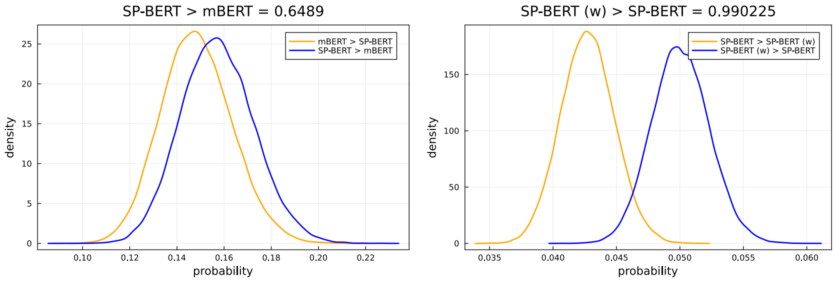


Fig. 1: Exemplary comparison of SP-BERT versus mBERT (Norwegian), and weighted SP-BERT versus SP-BERT (Swedish).

To capture the changes in political position, we filter a list of politicians with fairly balanced number of their given speeches for specific topics per party. First, we visualize the embeddings with PCA and then use k-means clustering to verify whether the text representations and the clusters are aligned.

We analyze seven political parties and a number of politicians. We learn that not all cases, there are significant shifts in the way politicians speak when they change their political position. Fortunately, some cases show distinctions quite well. We only show here some plots that we think are interesting (see Figure 2). The PCA projects all speeches onto two dimensions. Both plots show a noticeable grouping between speeches of Cabinet or Opposition. The k-means clustering confirms the insight from PCA with few exceptions.

4.2 Parties' position shifts in word choices with POS Tagging

To understand politicians' word choices better when they change political roles, we utilize a part-of-speech (POS) tagger. We lemmatize nouns, verbs, and adjectives with SpaCy [7]. We focus on words of more than four characters to remove less expressive terms. We compute the frequencies with which politicians use these words and distinguish between cabinet and opposition. We obtain two values between 0.0 and 1.0 for each term. We plot these values onto a two-dimensional scatter plot labelled with the term. With a log-scale, we avoid cluttering. The diagonal line expresses that both usages are equal. In other words, the politicians use the term as often in cabinet speeches as in opposition speeches. On the other hand, terms that are far from the line are either used predominantly in cabinet (lower) or opposition (upper). Figure 3 shows two examples of politicians' word choices when changing roles. Figure 3a illustrates the usage of nouns by members of the Center Party. The colour shows the output of a

[7] https://spacy.io

(a) PCA visualization (b) K-Means visualization

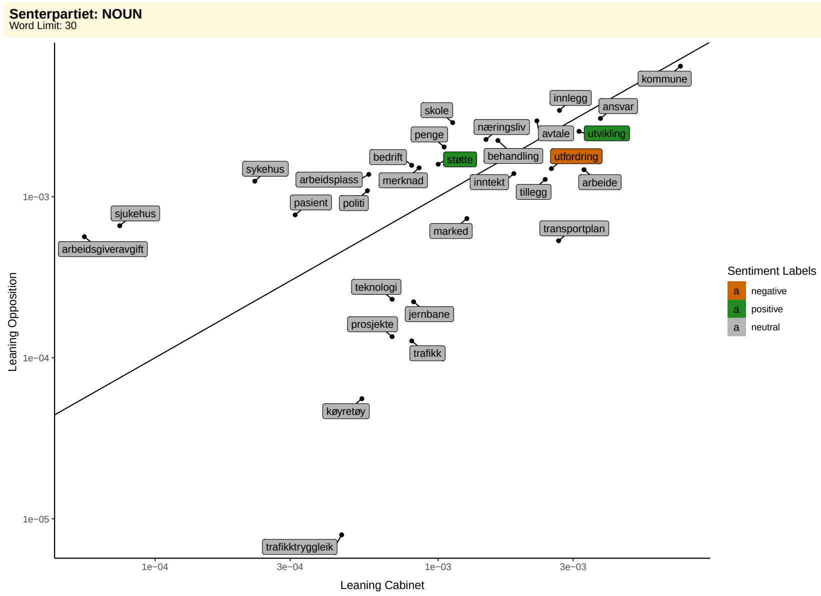
Fig. 2: Visualization of speeches given by Anders Anundsen (Progress Party) about Ministries. In (a), we notice two clusters produces by the PCA. In (b), K-Means confirms with clustering with a few mis-classified speeches at the intersection of both clusters.

sentiment lexicon [2]. The plot shows the 30 most frequently used nouns. We omit the plots for the remaining parties and verbs, and adjectives due to space limitations. We observe that terms related to infrastructure are more commonly used in cabinet speeches. Terms related to health-care such as Hospital (sykehus) appear more often in opposition speeches.

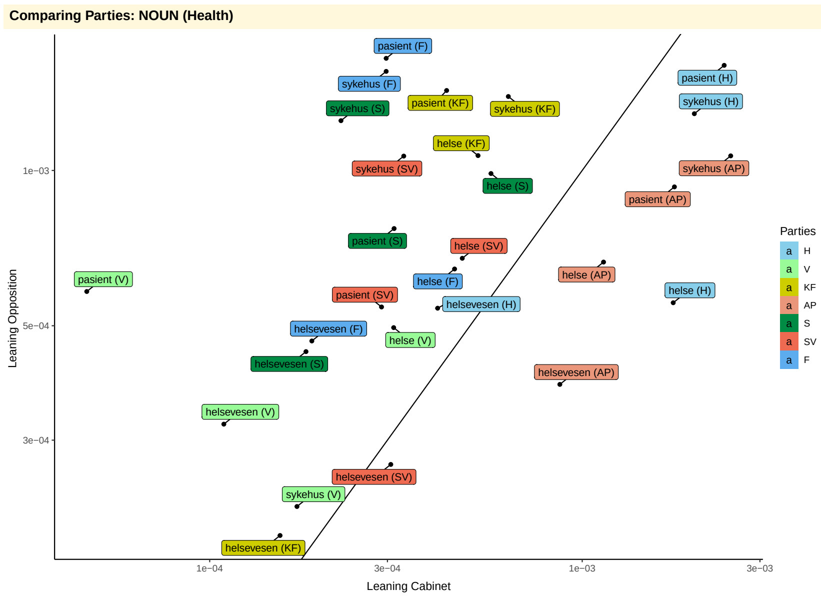
We also compare the usage of nouns related to health for all parties. Figure 3b uses the same display but shows all parties for a selection of health-related terms. Parties are colour-coded. The terms are in Norwegian. Due to space restrictions, we omit the plots for other topics such as transport, or education. We can clearly see that some parties cover health-care more when they are in government whereas others cover it in opposition. For instance, the Labour Party (Arbeiderpartiet, AP) covers health topics predominantly in government. Conversely, the Center party (Senterpartiet, S) talks more about it in opposition.

5 Conclusion and Future Work

We introduced SP-BERT, a language model for political texts in Scandinavian languages. The model will support political and social research in that researchers will be able to adequately represent texts and process them more automatically. Our investigation finds clear differences in the use of nouns concerning health in Norwegian politics. Results show that the model helps classifying texts more accurately. This is just one exemplary case that highlights the use of automated text processing for political sciences using language models. Language-specific model can perform similarly well or better in some instances.



(a) Center Party (Nouns)



(b) Health-related (Nouns)

Fig. 3: Plots depicting the word usage by Norwegian parliamentarians in cabinet and opposition.

We will add language models pre-trained for politics that can generate and transform texts. We want to apply SP-BERT to additional problems in the scope of politics, such as automatic viewpoints identification and sentiment analysis.

Acknowledgements 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. Abercrombie, G., Batista-Navarro, R.: Semantic Change in the Language of UK Parliamentary Debates. In: Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pp. 210–215, ACL (2019), https://doi.org/10.18653/v1/W19-4726
2. Barnes, J., Touileb, S., Øvrelid, L., Velldal, E.: Lexicon information in neural sentiment analysis: a multi-task learning approach. In: Proc. of the 22nd Nordic Conference on Computational Linguistics, Turku, Finland (2019)
3. Chen, W.F., Al Khatib, K., Wachsmuth, H., Stein, B.: Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. In: Proceedings of the 4th Workshop on Natural Language Processing and Computational Social Science, pp. 149–154, ACL (2020), https://doi.org/10.18653/v1/2020.nlpcss-1.16
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL'19, pp. 4171–4186, ACL (2019), https://doi.org/10.18653/v1/N19-1423
5. Doan, T.M., Kille, B., Gulla, J.A.: Using Language Models for Classifying the Party Affiliation of Political Texts. In: NLDB, pp. 382–393, Springer (2022), https://doi.org/10.1007/978-3-031-08473-7_35
6. Goutte, C., Gaussier, E.: A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: European conference on information retrieval, pp. 345–359, Springer (2005)
7. Hoffman, M.D., Gelman, A., et al.: The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research **15**(1), 1593–1623 (2014)
8. Hu, Y., Hosseini, M., Skorupa Parolin, E., Osorio, J., Khan, L., Brandt, P., D'Orazio, V.: ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence. In: Proc. of NAACL'22, pp. 5469–5482, ACL (2022), https://doi.org/10.18653/v1/2022.naacl-main.400
9. Hvingelby, R., Pauli, A.B., Barrett, M., Rosted, C., Lidegaard, L.M., Søgaard, A.: DaNE: A named entity resource for danish. In: Proc. of the 12th Language Resources and Evaluation Conference, pp. 4597–4604 (2020)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Kummervold, P.E., De la Rosa, J., Wetjen, F., Brygfjeld, S.A.: Operationalizing a national digital library: The case for a Norwegian transformer model. In: Proc. of NoDaLiDa'21, pp. 20–29 (2021)

12. Kutuzov, A., Barnes, J., Velldal, E., Øvrelid, L., Oepen, S.: Large-Scale Contextualised Language Modelling for Norwegian. In: Proc. of NoDaLiDa'21, pp. 30–40, Linköping University Electronic Press, Sweden (2021)
13. Lapponi, E., Søyland, M.G., Velldal, E., Oepen, S.: The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. Language Resources and Evaluation pp. 1–21 (2 2018), ISSN 1574-020X, https://doi.org/10.1007/s10579-018-9411-5
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
15. Liu, Y., Zhang, X.F., Wegsman, D., Beauchamp, N., Wang, L.: POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. In: Findings of the Association for Computational Linguistics: NAACL 2022, pp. 1354–1374, ACL (2022), https://doi.org/10.18653/v1/2022.findings-naacl.101
16. Magnusson, M., Öhrvall, R., Barrling, K., Mimno, D.: Voices from The Far Right: a Text Analysis of Swedish Parliamentary Debates (2018)
17. Malmsten, M., Börjeson, L., Haffenden, C.: Playing with Words at the National Library of Sweden - Making a Swedish BERT. CoRR (2020)
18. Maronikolakis, A., Sánchez Villegas, D., Preotiuc-Pietro, D., Aletras, N.: Analyzing Political Parody in Social Media. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4373–4384, ACL (2020), https://doi.org/10.18653/v1/2020.acl-main.403
19. Rauh, C., Schwalbach, J.: The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies (2020), https://doi.org/10.7910/DVN/L4OAKN
20. Snæbjarnarson, V., Símonarson, H.B., Ragnarsson, P.O., Ingólfsdóttir, S.L., Jónsson, H., Thorsteinsson, V., Einarsson, H.: A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. In: Proceedings of the 13rd Language Resources and Evaluation Conference, pp. 4356–4366, European Language Resources Association, Marseille, France (Jun 2022)
21. Solberg, P.E., Ortiz, P.: The Norwegian Parliamentary Speech Corpus. arXiv preprint arXiv:2201.10881 (2022)
22. Steingrímsson, S., Barkarson, S., Örnólfsson, G.T.: IGC-parl: Icelandic corpus of parliamentary proceedings. In: Proceedings of the Second ParlaCLARIN Workshop, pp. 11–17, European Language Resources Association (2020), ISBN 979-10-95546-47-4
23. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: BERT for Finnish. arXiv preprint arXiv:1912.07076 (2019)
24. Walter, T., Kirschner, C., Eger, S., Glavaš, G., Lauscher, A., Ponzetto, S.P.: Diachronic Analysis of German Parliamentary Proceedings: Ideological Shifts through the Lens of Political Biases. In: 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 51–60 (2021), https://doi.org/10.1109/JCDL52503.2021.00017