Perspectives on Forests and Forestry in Finnish Online Discussions - A Topic Modeling Approach to Suomi24

Telma Peura, Attila Krizsán, Salla-Riikka Kuusalu and Veronika Laippala
School of Languages and Translation studies,
University of Turku
tejpeu, attila.krizsan, srkbos, mavela
@utu.fi

Abstract

This paper explores how forests and forest industry are perceived on the largest online discussion forum in Finland, Suomi24 ('Finland24'). Using 30,636 posts published in 2014–2020, we investigate what kind of topics and perspectives towards forest management can be found. We use BERTopic as our topic modeling approach and evaluate the results of its different modular combinations. As the dataset is not labeled, we demonstrate the validity of our best model through illustrating some of the topics about forest use. The results show that a combination of UMAP and K-means leads to the best topic quality. Our exploratory qualitative analysis indicates that the posts reflect polarized discourses between the forest industry and forest conservation adherents.

1 Introduction

The importance of forests as carbon sinks has been globally recognized as part of climate change mitigation (IPCC, 2023). In Finland, where forests have a significant socio-economic role, the issue has received increased attention and created tensions across different economic and political views (Makkonen et al., 2015; Kellokumpu, 2022; Blattert et al., 2023). In fact, around 75% of Finnish land area is covered by forests of which only 12.9% is partially or totally conserved from industrial forest management (Ministry of Agriculture and Forestry, 2024).

Perspectives of the forest industry have also been prominent in the media. Analyses of Finnish newspapers show that despite emerging multi-objective discourses, the positive framing of the forest industry still seems to dominate (Näyhä and Wallius; Takala et al.). However, computational approaches to forest discourses have not, to our knowledge, been applied.

While analyzing the representation of forests in the mainstream media is valuable, the voices of common citizens cannot be overlooked. In fact, around 60% of Finnish forests are owned by private individuals (Karppinen et al., 2020). The right of public access and the high percentage of private forest ownership make public opinion critical to understanding how forest-related issues are perceived and debated.

To set light on the perspective of forest owners and users and understand their attitudes towards forest management, we used data from Suomi24 (translated as 'Finland24'). Suomi24 is the oldest and largest online forum in Finland and has been called a pool of Finnish public opinions (Ylisiurua, 2024).

We applied topic modeling to cluster documents and to identify forest-related themes in our dataset. Recent advances in machine learning and large language models have led to the development of new topic modeling tools (Abdelrazek et al., 2023). In particular, Bidirectional Encoder Representations from Transformers (BERT) have been found to be powerful in many NLP tasks (Wijanto et al., 2024; Devlin et al., 2019). BERTopic presented by Grootendorst (2022) has proved to perform well in many topic modeling tasks and was also adopted in our work. The modular approach of BERTopic allowed us to build several different models. The different combinations were compared through computational and qualitative measures.

In this paper, our first aim is to evaluate the performance of different BERTopic models and demonstrate how topic modeling can be used to identify relevant topics about the use of forests in Finland. Second, we aim at characterizing how forest management and industry are discussed in our Suomi24 dataset.

Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025), pages 10–15
March 2, 2025 ©2025 University of Tartu Library

The model evaluation results showed that there was great variance in the model quality. However, a comparison of topic keywords showed that all of them captured similar topics that can give valuable insights into Finnish forest discourse. The qualitative exploration suggested that pro-forestry discourses dominate over pro-nature discourses, but the distinction between these two is not always clear. Finally, we briefly discuss how this analysis can be extended in the future.

2 Data and Methods

Our methodology combined quantitative exploration and closer qualitative analysis of selected topics. The design allowed to compare the performance of different topic models on unlabeled data. The steps of the workflow are described in this section.

2.1 Dataset Preparation

The Suomi24 corpus was gathered and made openly available by the Language Bank of Finland[1] (Lagus et al., 2016). Overall, it contains discussions from 2001 to 2020, amounting to over 480,000,000 tokens (City Digital Group, 2021). In our study, we use posts beginning from year 2014, when the Forest Act providing a legislative framework for forest management in Finland was amended (Ministry of Agriculture and Forestry, n.d.). Following Lehti et al. (2020), we curated a list of search words to collect posts that were potentially relevant for our study. The list contained terms related to forest industry, forest conservation and the recreational use of forests. In addition, Word2Vec was applied to expand the list with semantically similar words in the same corpus[2]. This was done to reduce the subjectivity of the search words and to make the resulting dataset more comprehensive. Next, we removed duplicates and filtered out short documents (under 7 tokens). Upper-case words were lowercased. The final dataset consisted of 30,636 documents, when 10% of the total data was retained as a test set for later use.

2.2 Topic Modeling

We selected BERTopic (Grootendorst, 2022) as our topic modeling approach. Based on pre-trained language models, BERT can generate contextual vector embeddings of text documents (Wijanto et al., 2024). BERTopic relies on the assumption that semantically similar documents have similar embeddings, and the pipeline consists of the following steps: First, documents are converted into BERT embeddings with a pre-trained language model. In our experiment, we compared the performance of a multilingual sentence transformer, 'paraphrase-xlm-r-multilingual-v1' (Reimers and Gurevych, 2019), and the Cased Finnish Sentence BERT model, specifically trained for Finnish language [3]. Next, to optimize the clustering performance, the dimensionality of the embeddings is reduced. By default, the framework employs UMAP (McInnes et al., 2020), but some experiments have obtained superior results with principal component analysis (PCA) (Wijanto et al., 2024). Thus, both algorithms were tested.

For topic creation, we used two different clustering algorithms, HDBSCAN, and K-Means. The advantage of HDBSCAN is that it assigns the label -1 to documents considered noise (Grootendorst, 2022), and it can automatically determine the number of topics (McInnes et al., 2017). In contrast, the number of topics for K-Means has to be predetermined. To estimate an optimal number of topics, the elbow method (Cui, 2020) and silhouette scores (Shutaywi and Kachouie, 2021) were used.

Finally, BERTopic uses a class-based variant of term frequency-inverse document frequency (c-TF-IDF) to produce topic representations from the clusters. Instead of a classical TF-IDF that extracts words important for a document, the proposed c-TF-IDF procedure extracts words that have importance for the whole topic (Grootendorst, 2022).

2.3 Evaluation Methods and Qualitative Analysis

We evaluated the models in two ways. As a computed metric, we chose the coherence score C_v that has been found to correlate well with human ratings (Röder et al., 2015). Moreover, a member of the research team reviewed the topic keywords (20 per topic) of all models and rated their quality as good, satisfactory or unsatisfactory. For a good topic, all keywords had to be co-

[1] https://www.kielipankki.fi/corpora/suomi24/

[2] The similarity was calculated using http://epsilon-it.utu.fi/wv_demo/]

[3] For documentation, see https://turkunlp.org/paraphrase.html

Embedding model	**Dimensionality reduction**	**Clustering**	**Topics**	**Coherence score, C_v**	**Avg. topic size (nr of -1 docs)**	**Quality topics**
Finnish	UMAP	HDBSCAN	175	0.49	52 (21 623)	
	PCA	HDBSCAN	35	0.45	18 (30 013)	
	UMAP	K-means	175	0.47	175	**99**
	PCA	K-means	200	**0.54**	153	52
Multilingual	UMAP	HDBSCAN	175	0.49	60 (20 145)	
	PCA	HDBSCAN	32	0.48	32 (29 659)	
	UMAP	K-means	175	0.47	175	93
	PCA	K-means	150	0.50	204	41

Table 1: An overview of trained models and evaluation results. For the models using HDBSCAN, the size of the 'noise' cluster (nr of -1 docs) is reported along with the average topic size.

herent, the label 'satisfactory' allowed for 2-3 outliers, and the label 'unsatisfactory' was used for mixed or incomprehensible keywords. The number of good-quality topics was used as an indicator of model performance. Only good-quality topics (represented as 'Quality topics' in Table 1) were considered in the further qualitative analysis.

Since K-means forces all documents into some clusters, the documents with a low topic probability were filtered out. A good threshold was found experimentally to be at M_{topic} - SD_{topic} where *M* is the mean probability of the assigned topic and *SD* the respective standard deviation per topic cluster.

After this, relevant topics were identified on the basis of topic keywords. The relevance was determined by the following criteria: the topic was of good quality and the keywords were related to forestry and forest management. Consequently, e.g. recreational forest activities such as berry-picking and hiking, were not considered in this paper. A member of the research team read a sample of 20 documents from each potentially relevant topic to validate the selection.

The relevant documents were grouped into broader thematic categories, and posts from these thematic categories were used in the preliminary close reading analysis.

3 Results and Analysis

The combinations of different algorithms and the evaluation results are shown in Table 1. The results point to a discrepancy between the computational and human-annotated measures of topic coherence, as the columns 'Coherence score, C_v' and 'Quality topics' show. While the amount of good-quality topics was highest for the models using UMAP and K-means, the models with PCA yielded a better coherence score. It indicates that the coherence measure C_v is not well adapted for BERTopic.

Moreover, the HDBSCAN algorithm labeled most of the documents as 'noise', while a closer look at the discarded documents showed that many of them were relevant to forest discussion, and the 'noise' category keywords contained several forest-related terms. Due to this, the HDBSCAN models were not included in further evaluation of topic quality and qualitative analysis.

Although the performance of the models varied, we observed that all of them produced topics with similar keywords. This reinforced our confidence in the reliability of the generated topics.

The Finnish sentence embedding model performed slightly better than the multilingual one, but the choice of the dimensionality reduction and clustering algorithms had a greater effect on the result. Overall, UMAP was the most suitable dimensionality reduction algorithm for our dataset and K-means functioned well for topic clustering.

As Table 1 shows, the combination of Finnish BERT model, UMAP and K-means yielded the highest amount of good-quality topics. Since the difference from the multilingual model was relatively small, we analyzed the hierarchical topic structure[4] of these two models and inspected a sample of 10 documents from 15 randomly selected topics. This check confirmed that the Finnish model performed best with our data, and it was selected for further analysis[5].

[4] The hierarchy was produced with BERTopic's in-built hierarchical topic modeling function.

[5] The topic assignments are provided on: https://github.com/TurkuNLP/forest-in-s24.

Topics	Example keywords	Theme	Nr of posts
30, 80, 99, 112, 136, 141	kasvatus ('forestry'), raivaus ('clearing'), taimikko ('seedling stand'), omistaja ('owner'), kemera ('a forestry funding'), metsuri ('logger')	Forestry	1,185
2, 10, 43, 72, 173	luonnonsuojelija ('environmentalist'), linkola (a Finnish ecologist and nature activist, Pentti Linkola), vihreät ('The Greens'), luonnonsuojelualue ('nature reserve'), biologia ('biology')	Nature conservation, environmentalists	1,111
42, 73, 106, 133	ostaja ('purchaser'), hinta ('price'), m^3, pystykauppa ('stumpage sale'), kuitupuu ('pulpwood'), osake ('share'), hakkuukone ('harvester')	Forest and timber trade	1,017
24, 88, 151	Avohakkuu ('clearcutting'), puupelto ('forest field'), päätehakkuu ('regeneration felling'), metsä ('forest'), puu ('wood')	Clearcutting	678
34, 64, 79, 127	CO_2, ilmasto ('climate'), turve ('peat'), hiilinielu ('carbon sink'), päästöt ('emissions'), energiantuotanto ('energy production')	Climate change	662
120, 165	metsänhoitoyhdistys, mhy ('Forestry management association, FMA'), jäsenyys ('membership'), palvelu ('service')	Forestry management associations	508

Table 2: A table of relevant topics with example keywords and topic size.

The topic annotation and evaluation showed that various forest-related themes were discussed, and 41 of the quality topics were considered relevant from the perspective of forestry and forest industry. The most prominent of these are listed in Table 2. All translations to English are done by the authors. A comparison of topic sizes indicates that topics related to forest management and trade (2 880 posts) dominate over topics about forest conservation and climate change (1 773 posts).

However, the distinction between the themes is not always clear. For instance, the proponents (example 1) and opponents (example 2) of clearcutting both appeal to the health of the forest:

> (1) In Finland, forest management aims to ensure that forests only have healthy growing trees. No thickets or rotten wood.

> (2) Forest fields and pine trees struggling along ditched banks are not forests. Forests exist only in nature reserves and among the few landowners who think with their own brains.

The term 'forest field' is frequently evoked by the opponents. Example 2 also shows how the intelligence of the forest owners is questioned. Similarly, the proponents of clearcutting rely on their expertise and criticize their opponents for not knowing the field. A typical view is shown in example 3.

> (3) Finland has university-level forestry education and, even on a global scale, Finland is one of the most competent and professional forestry countries. It is sad and stupid to see how eagerly people who live in cities and know almost nothing about forests discuss forest management and take strict positions on, for example, this issue of clearcutting.

Overall, the exploratory close reading suggested that the issue of clearcuttings is polarized with few negotiating voices in the discussions.

4 Discussion and Future Work

In this paper, we presented a framework that combined topic modeling and qualitative exploration to investigate how forest-related issues are addressed in a Finnish online forum, Suomi24. We compared different BERTopic models, and the evaluation results showed that its default clustering algorithm, HDBSCAN, did not function well with our data. Based on our observations, numerous relevant posts were discarded by these models.

Best results were obtained by combining Finnish sentence BERT, UMAP, and K-means.

As Abdelrazek et al. (2023) point out, the parameters of a neural topic model are often difficult to interpret and hence it is hard to diagnose why the HDBSCAN model did not work. K-means was found to produce topics of better quality, but as the method forces all texts into some clusters, we needed to filter the resulting topics to discard irrelevant documents.

Even the best models contained several topics of low quality, which is due to several reasons. First, we collected the dataset from the Suomi24 corpus using a list of search words, which means that it potentially contained several texts not related to the themes of interest. Misspellings and colloquial language in the posts introduced noise in the data, leading to suboptimal sentence embeddings and reduced model accuracy. While the Finnish sentence transformer outperformed the multilingual one, Finnish is still a lower-resource language, which may show in the performance of the models.

The best model could be improved by changing the number of topics. In addition, we did not test different hyperparameters for the used algorithms, so our final model could be improved through finetuning the UMAP and K-means modules.

We noted a striking difference in the computational and human annotated results of quality evaluation. Moreover, the coherence measure C_v is usually used with LDA models, and it measures coherence based on the co-occurrence of the given topic keywords in a corpus. Since BERTopic generates topics through embeddings, not words, this approach does not fully capture the semantic coherence of the generated topics. These observations remind us that quality in topic modeling is dependent on several aspects (Abdelrazek et al., 2023) and computational performance measures can be misleading. Thus, human evaluation is crucial when the resulting topics are used for qualitative analysis. Overall, the evaluation of neural topic models calls for new measures.

Many topics shared a common broader theme, and this overlap suggests that the number of topics could be further reduced. However, close reading the posts showed that different topics offered diverse viewpoints and reflected distinct discourses on the same theme. For instance, the theme related to nature conservation and environmentalists could have been further divided into political, activist, and other perspectives on the theme. Although the scope of this paper did not allow us to delve deeper into these differences, it was an interesting observation for future studies.

The exploratory qualitative analysis showed that opinions on forestry and forest management tend to be polarized. In the future, we aim to expand the analysis of such polarization by studying texts in selected topics (e.g., clearcuttings) by applying methods of 'making strange', close-reading (Gasper, 2022) and analyses of topic chains following Li (2004) and Li and Thompson (1981).

Acknowledgments

This work has been supported by Kone Foundation, grant number 202302300 for *Seedling*.

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Clemens Blattert, Mikko Mönkkönen, Daniel Burgas, Fulvio Di Fulvio, Astor Toraño Caicoya, Marta Vergarechea, Julian Klein, Markus Hartikainen, Clara Antón-Fernández, Rasmus Astrup, Michael Emmerich, Nicklas Forsell, Jani Lukkarinen, Johanna Lundström, Samuli Pitzén, Werner Poschenrieder, Eeva Primmer, Tord Snäll, and Kyle Eyvindson. 2023. Climate targets in european timber-producing countries conflict with goals on forest ecosystem services and biodiversity. *Communications Earth Environment*, 4(1):1–12.

City Digital Group. 2021. Suomi24-korpus 2001-2020, VRT-versio.

Mengyao Cui. 2020. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, 1(1):5–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. (arXiv:1810.04805).

Des Gasper. 2022. 'making strange': Discourse analysis tools for teaching critical development studies. *Progress in Development Studies*, 22(3):288–304.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

IPCC. 2023. Summary for policymakers. In *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1–34. IPCC.

Heimo Karppinen, Harri Hänninen, and Paula Horne. 2020. *Suomalainen metsänomistaja 2020*. Luonnonvarakeskus.

Ville Kellokumpu. 2022. The bioeconomy, carbon sinks, and depoliticization in finnish forest politics. 5(3):1164–1183. Publisher: SAGE Publications Ltd STM.

Krista Hannele Lagus, Minna Susanna Ruckenstein, Mika Pantzar, and Marjoriikka Jelena Ylisiurua. 2016. Suomi24: muodonantoa aineistolle.

Lotta Lehti, Milla Luodonpää-Manni, Jarmo Harri Jantunen, Aki-Juhani Kyröläinen, Aleksi Vesanto, and Veronika Lappala. 2020. Commenting on poverty online : A corpus-assisted discourse study of the suomi24 forum. 37. Accepted: 2021-03-05T10:02:31Z Publisher: Suomen kielitieteellinen yhdistys.

Charles N Li and Sandra A Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.

Wendan Li. 2004. Topic chains in chinese discourse. *Discourse Processes*, 37(1):25–45.

Marika Makkonen, Suvi Huttunen, Eeva Primmer, Anna Repo, and Mikael Hildén. 2015. Policy coherence in climate change mitigation: An ecosystem service approach to forests as carbon sinks and bioenergy sources. 50:153–162.

Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. (arXiv:1802.03426). ArXiv:1802.03426 [stat].

Ministry of Agriculture and Forestry. 2024. Forest resources in finland.

Ministry of Agriculture and Forestry. n.d. Forest act.

Annukka Näyhä and Venla Wallius. Actors, discourses and relations in the finnish newspapers' forest discussion: Enabling or constraining the sustainability transition? 169:103331.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Meshal Shutaywi and Nezamoddin N. Kachouie. 2021. Silhouette analysis for performance evaluation in machine learning with applications to clustering. *Entropy*, 23(66):759.

Tuomo Takala, Ari Lehtinen, Minna Tanskanen, Teppo Hujala, and Jukka Tikkanen. The rise of multi-objective forestry paradigm in the finnish print media. 106:101973.

Maresha Caroline Wijanto, Ika Widiastuti, and Hwan-Seung Yong. 2024. Topic modeling for scientific articles: Exploring optimal hyperparameter tuning in bert. *International Journal on Advanced Science, Engineering Information Technology*, 14(3).

Marjoriikka Ylisiurua. 2024. *Online swarm dynamics at Suomi24 discussion: Turning points and their stimulation*. Ph.D. thesis, Helsingin yliopisto.