NTNU

Norwegian University of Science and Technology

Department of Computer Science

TDT4310 - Project Report

NOU Hearing Response Text Analysis

Authored by:

Nicolai Thorer Sivesind & Henrik Haug Larsen

21st April 2024

Abstract

This study investigates the application of natural language processing (NLP) techniques in analyzing submissions from a public hearing on Norway's climate policy towards 2050. Focusing on the responses to the NOU 2023: 25 document, we built a dataset labeling the different responses into 7 different industry sectors. We further applied in-context learning and Latent Dirichlet Allocation for argument mining of the dataset, alongside multi-label classification to identify the communications patterns of the various actor types. Despite limitations in argument extraction performance, the classification results were promising, demonstrating the ability of NLP to recognize unique linguistic styles among various actor types. Our findings indicate that NLP can significantly enhance the analysis of extensive textual data in public policy contexts, suggesting a potential shift in how researchers may approach text analysis in social sciences. This preliminary study lays some exploratory groundwork for further research into refining NLP techniques for more effective deployment in social science research and beyond.

Contents

List of Figures

List of Tables

1 Introduction

Natural language processing has in recent years seen massive advancements through the development and optimization of large language models. These advancements have opened new alleys into many segments of society for their application, both in terms of task complexity and performance. A group of researchers at the Institute of Social Science at the University of Oslo are now looking toward large language models for establishing new techniques for aiding in full-text analysis, which is an integral, but time-exhastive task. Until now, this has typically relied on well-established, but limited computational linguistic techniques. As part of a preliminary exploration before potentially launching a major interdisciplinary project in early 2025, they have tasked us, Nicolai Thorer Sivesind and Henrik Haug Larsen, with conducting a small study, to see how exploration of the problem-domain may be approached from the perspective of master students with competence in computer science.

1.1 Research Purpose

In this project, we explore the use of natural language processing techniques to argument-mine submissions from a public hearing on Norway's climate policy toward 2050. Specifically, our study focuses on responses to the NOU 2023: 25 document, utilizing a combination of in-context learning, LDA-topic modeling for mining arguments, and comparing computational linguistic classifiers to large language model classifiers. Our objective has been to identify arguments presented by different actors/parties, such as private businesses, municipalities, and voluntary organizations, amongst others, and classify the authors' industry sectors based on their argumentative textual structures.

1.2 Overview of Process

To summarize our work had the following process:

1. *Build NOU Hearing dataset:*
   (a) Scrape NOU-hearing response documents from NOU 2023: 25 webpage
   (b) Clean retrieved data
   (c) Label data points by actor type according to national entity register, Brønnøysundregisteret
2. *Argument mine dataset using various approaches:*
   - Splitting by paragraph (metric baseline)
   - LDA-topic sentence similarity
   - Zero-shot in-context learning using LLAMA2 75B
3. *Train and compare various multi-label classifiers for actor type classification across original and argument-mined datasets:*
   - Naive-Bayes classifier
   - Support Vector Machine Classifier
   - NB-Bert Sequence Classifier

2 Theory

2.1 Argument Mining

Topologically, an argument contains one or more premises and one conclusion, but one may further be decomposed into more fine-grained components. Argument mining is the process of identifying, decomposing, and extracting these textual structures which ultimately forms such argumentative structures in natural language (Lawrence and Reed, 2020). There are many applications in which argument mining may provide beneficial insights. Examples of these are to aid in argumentative writing, support qualitative analysis of argumentative articulation, or automatically scoring school essays based on the presence of various argumentative structures (Stahl et al., 2024).

2.2 In-Context Learning

In-context learning (ICL) is a method leveraging the generalization abilities of autoregressive large language models at inference-time by providing such a model with task-specific information typically consisting of a natural language prompt (the instruction) and an input text (the data to be processed). Unlike techniques that adjust the model's weights, such as fine-tuning, ICL operates solely through prompting, relieving the typical need for large datasets of labelled text and extensive computational resources.

ICL can be implemented in several forms: zero-shot, one-shot, and few-shot learning. Collectively, these are often termed *k-shot learning*, where *k* denotes the number of labelled examples included in the prompt (Shah, 2024).

2.3 Latent Dirichlet Allocation and Coherence Scores

Latent Dirichlet Allocation (LDA) is a generative probabilistic model used for identifying underlying topics within a corpus. Such topics are typically not explicitly defined in human-readable terms but are instead inferred from the distribution of words across the documents. Each topic is characterized by a set of words or tokens, each associated with a certain confidence score, indicating the relevance of the word to the topic (Blei, 2003). Once such a model is built, it can be used to categorize segments of text into one or more topics.

Coherence scores is a measure of how well the topics generated by an LDA-model fit a set of reference corpus. One of the algorithms to calculate coherence is C_V which calculates pairwise similarities between the top terms within a topic (Pedro, 2022).

2.4 Naive-Bayes Classifier

Naive-Bayes Classifiers are a set of algorithms based on Bayes' Theorem. It is a simple and fast classifier used in many use-cases such as sentiment detection or rating classification. The algorithm is well suited for text as it can handle high-dimensional data such as numerically embedded texts (GeeksForGeeks, 2017).

Multinomal Naive-Bayes (MNB) is the most relevant to the task of multi-label classification. This variant is designed to classify data based on discrete features, such as the frequency of words in a document, compared to the Gaussian Naive-Bayes, which is made for continuous features. The formula for MNB can be seen below in equation 1 (Ratz, 2022)

$$Pr(W|C_k) = \frac{(\sum_{i=1}^{n} w_i)!}{\prod_{i=1}^{n} w_i!} \times \prod_{i=1}^{n} p_{k_i}^{w_i} \tag{1}$$

2.5 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression. It works by maximizing the margin between two classes in a multidimensional plane. A benefit of SVM compared to algorithms like Naive-Bayes is that SVM can handle non-linearly separable data, in addition to being both efficient and robust against overfitting (IBM, 2024).

2.6 NB-BERT

Nb-BERT is an LLM created by the AI-Lab at the National Library of Norway based on a multilingual BERT model, and further pre-trained on their digital collection *The Norwegian Colossal Corpu* at the National Library containing texts from both Bokmål and Nynorsk from the last 200 years (NationalBiblioteket, 2024).

BERT stands for *Bidirectional Encoder Representations from Transformers* and uses a transformer-based neural network consisting of 42 million to 1.3 billion parameters to understand human-like language and perform natural language inference. BERT utilizes a masked-sentences approach to perform analytical NLP tasks. This characteristic has made it excel in tasks such as text-classification. As it is pre-trained, it can be efficiently fine-tuned for down-stream tasks such as domain-specific classification, using a relatively small amount of labeled data compared to an untrained version (GeeksForGeeks, 2020).

3 Related Works

3.1 Argument mining with Large Language Models

The steep performance-improvement curve, and subsequently rise in popularity of large language models has led to new studies on how they can be used for argument mining, as highlighted by Chen et al. (2024). This research tested the abilities of these models to identify claims, find evidence, determine stances, detect arguments, and generate responses to counter arguments. The models used in this study were GPT-3.5-Turbo, Flan-UL2, and Llama-2-13B, all of which are autoregressive models.

To test these capabilities, Chen et al. used standardized prompts that are specific to argument mining tasks, applying both zero-shot and few-shot learning approaches. The models were tested on a dataset prepared by IBM, which includes arguments pulled from Wikipedia articles.

Their results showed that some models, like GPT-3.5-Turbo and Flan-UL2, performed very well. However, the LLaMA-2-13B model, which was one of the smaller models tested, did not perform as well, particularly in identifying claims and evidence. This shows that there's a significant difference in how various LLMs can handle complex tasks like argument mining.

3.2 Argument using Latent Dirichlet Allocation

Lawrence, Reed et al. (2014) tested argument mining using latent dirichlet allocation (LDA) topic modeling by looking at topics similarity of close-proximity sentences. In their study, they found that if there were overlapping LDA topic predictions of the previous sentence then there existed high likelihood of a pairwise argumentative link between them. Despite lack of evidence supporting the hypothesis of topical relations with manual analysis of the data, their automated results supported the hypothesis with a precision of 0.72 and recall of 0.77 when comparing the *resulting structure* to a manual analysis. However, they highlighted that while there is a clear inference relationship between two propositions, the directionality of this inference, whether one proposition supports or contradicts another, was not determined by this method (Lawrence and Reed, 2020).

3.3 Effectiveness of Fine-tuned BERT for text-classification

Bilal and Almazroi (2023) compared a fine-tuned BERT model of different sequence lengths to the established bag-of-word models like the K-nearest neighbour, Naive-Bayes and SVM. The dataset they used contained *helpful* and *not helpful* comments on Yelp, and the task of each model was to classify the comments. Their results were positive and the BERT model with a sequence length of 320 scored the best with an accuracy of 0.707, with other BERT versions and SVMs following close behind. It has been reported that SVM does not perform well on larger datasets and requires extensive preprocessing. They concluded that the BERT model works well in classifying reviews, but it was only tested on short texts in their study (Bilal and Almazroi, 2023).

Other works supporting the effectiveness of fine-tuning of BERT-models for text-classification is the work of Sivesind and Winje (2023), where they discriminated the authorship of GPT-3.5-genereated research abstracts from human-writtens ones, with accuracy, recall, precision and subsequently F1 at the 98th percentile and above. Their work substantiated the effectiveness of BERT as a classification model compared to substantially larger unidirectional language models such as LLAMA-13B which received similar results on in-domain domain data, but fell behind on cross-domain evaluation compared to their BERT-based models. The BERT-based models also achieved decent evaluation scores in considerably fewer training-steps compared to the unidirectional models (Sivesind and Winje, 2023).

4 Methods

This section introduces the process of building the dataset for training and performance-evaluation, our approaches to argument mining and text-classification.

4.1 Dataset

The dataset used in this report has been built by scraping responses to the Norwegian Official Reports (NOU) hearing, *Høring - NOU 2023: 25 Omstilling til lavutslipp - Veivalg for klimapolitikken mot 2050 - rapport av Klimautvalget 2050* (Regjeringen, 2023). It contains hearing responses from 213 distinct actors from distinct fields in Norway, such as municipalities, academic institutions, private individuals, amongst others. Each hearing response was then labeled by setting up an script for retrieving their industry sector label from the Norwegian national entity register, *Brønnøysundregisteret* (Brønnøysundregistrene, 2024). Following a manual inspection, and the various actors present in the dataset along with their responses were sorted into 7 different classes/actor labels:

- Interessegruppe - Interest Group
- Offentlig forvaltning - Public Administration
- Bedrift - Business
- Privatperson - Private Individual
- Politisk parti - Political Party
- Frivillig organisasjon - Non-profit Organization
- Akademisk institusjon - Academic Institution

The dataset was retrieved by initially converting HTML and PDFs of hearings into two different CSV files. This was done by using python framework *BeautifulSoup* to retrieve the hearings and further extract the correct HTML classes. Regarding the PDF responses, these files were each downloaded and manually exported into a separate .txt-file and cleaned manually since there was no easily identifiable common structure for retrieving these files.

All retrieved response-texts where stored in an initial CSV-file along with their author, date and actor label, before being further split into a new sub-response CSV-file where each row each held a single paragraph from a response along with the same supportive data mentioned above.

4.2 Argument Extraction

To perform the task of extracting arguments, we have employed two approaches *zero-shot in-context learning* and *Latent Dirichlet Allocation topic modelling*.

4.2.1 Zero-shot In-context Learning Argument Mining

Our initial approach to argument extraction was to apply zero-shot in-context learning with the Llama2 70B large language model through an API. As with any in-context learning, the model both takes inputs and outputs in natural language, requiring prompt-engineering for inputs and extraction-scripting to optimally clean the outputs of redundant text provided by the model. The prompt for mining arguments were experimented with to optimize results. The final prompt used for the performing the fully-automated argument extraction is displayed in figure 1.

```
"prompt": "I have a hearing from an organisation and
            want to extract the arguments in this text, write the
            arguments without any paraphrasing and summarization.
            Write the output in Norwegian: {Inserted paragraph}"
```

Listing 1: LLaMA-2-70B argument prompt.

For extracting the desired argument from the output-text, additional cleaning included removing padding as it usually contained a greeting or some other introductory text. The text was checked whether the arguments found by LLaMA existed in input-text by identifying overlapping segments. Most English arguments were also excluded from the list to mitigate potential outliers. Some texts were translated to English by the model, and this was attempted to be cleaned using a SpaCy language detector. Despite these efforts, some English-translated text remains in the dataset. Finally, the arguments were saved as a CSV-file dataset which we will refer to as *ICL-Arguments*.

4.3 Latent Dirichlet Allocation Argument Mining

Our second approach to mining arguments were based on the approach presented by Lawrence, Reed et al. (2014), by building a latent dirichlet allocation (LDA) model from the *nou_hearings* dataset. Words that are present in 50% of the documents have been removed from the LDA corpus. To find the appropriate amount of LDA topics across the corpus, a separate model was built for each number of topics on the interval of $[2, 25]$ with 500 passes each, along with calculating a coherence score using *c_v measure* for each model. The optimal model was selected based on the highest coherence score above 2 topics, which was at approximately 0.38 calculated from the model of 18 topics. Despite the 2-topic model scoring minimally higher than the 18-topic model, this model was discarded as the corpus was inferred to hold more topics than this through manual analysis. The coherence value of each of models are presented in figure 1.

To perform the actual argument extraction, the *nou_hearings* dataset was first split into sentences using a SpaCy NLP-model with it's Norwegian Bokmål medium-sized language package (largest available). Furthermore, each sentence was assigned topics using the selected LDA-model with 18 topics. Finally, all sentences were iterated through, and any sequence of sentences (two or more) with overlapping topics were extracted as a single argument, aligning with the approach of Lawrence, Reed et al. No additional cleaning was deemed necessary following a manual inspection. We refer to this dataset as *LDA-Arguments*.

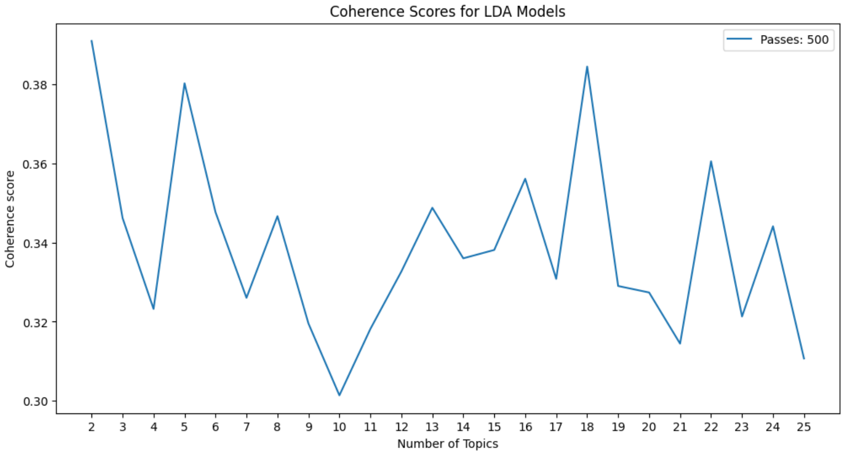


Figure 1: Coherence graph of the number of topics, where 18 topics were optimal

4.3.1 Evaluation of mined argument quality

To evaluate the quality of the arguments mined in the two aforementioned approaches, various datasets were explored to train an evaluator model, both Norwegian datasets and English. The Norwegian dataset $norec_{fine}$ by Velldal et al. (n.d.) containing polar expressions, opinion holders and opinion targets, was explored but concluded insufficient for our intended application of evaluating the accuracy of argument mining. Further, the English dataset *TACO – Twitter Arguments from COnversations* by Feger and Dietze (2023) held desirable features such as labeling of argument and non-argument texts. Despite this, methods for translating a subset of this dataset for our intended use on Norwegian text would both be costly and too time-consuming, taking into account the time frame of this exploratory project.

Efforts have therefore instead been shifted towards building text-classification models across the various datasets to analyze and evaluate how different classification architectures perform, and how the argument-mined data affects the performance of these models.

4.4 Text-Classifiers

To perform the classification of actor type (industry sector) in the various datasets, we have implemented three approaches to text classification for comparison: *Multinomal Naive-Bayes Classifier*, *Support Vector Machine using Grid Search optimization* and *Fine-Tuned Sequence Classification using a Bidirectional Transformer*. As the datasets consist of 7 labels, this is subsequently a multi-label classification task. In general for all three approaches: labels have been converted from textual labels into integers for classification and then converted back to textual labels for human interpretation. For testing we have applied an 80-20 training-split of the dataset, in order to evaluate the true generalization of the model.

4.4.1 Multinomal Naive-Bayes Classifier

The Naive-Bayes Classifier has been used as a base model for actor classification. We have compared two vectorization algorithms to find the optimal for our application - TfidfVectorizer (Tf-IDF-V) and a CountVectorizer (CV). Tf-IDF-V is based on the statistical significance of a term within a document and the document collection and is fundamentally bound to its frequency. Similarily, CV is based on pure term frequency within each document but excludes document collection frequency (Saket, 2024). We have also looked into Doc2Vec to see if it would perform better or worse than the more classical vectorizers.

4.4.2 Support Vector Machine with grid search optimization

The SVM classifier was used as a second base model to see how it would compare to the Naive-Bayes classifier and the larger language models. We used the same preprocessing and vectorization techniques as for the Naive-Bayes classifier. A grid search was also applied to find the optimal hyperparameters as shown below in listing 2.

```
hyperparameters: {
                  'C': 1,
                  'gamma': 'scale',
                  'kernel': 'linear'
                 }
```

Listing 2: Hyperparameters for SVM model.

4.4.3 Fine-Tuned Sequence Classification using a Bidirectional Transformer

Our final approach involves fine-tuning a pre-trained transformer model, more specifically the *Nb-BERT-Large model*, a multi-lingual bidirectional transformer model which has been further pre-trained on the Norwegian Colossal Corpus (Kummervold et al., 2021), and consists of 356-million parameters (*NbAiLab/nb-bert-large* 2024). This has been done using Huggingface's python framework *Transformers*. Relevant parameters in listing 3.

```
hyperparameters: {
                   padding='True',
                   truncation='True',
                   max_length='512',
                   auto_find_batch_size='True',
                   optim='adamw_torch',
                  }
```

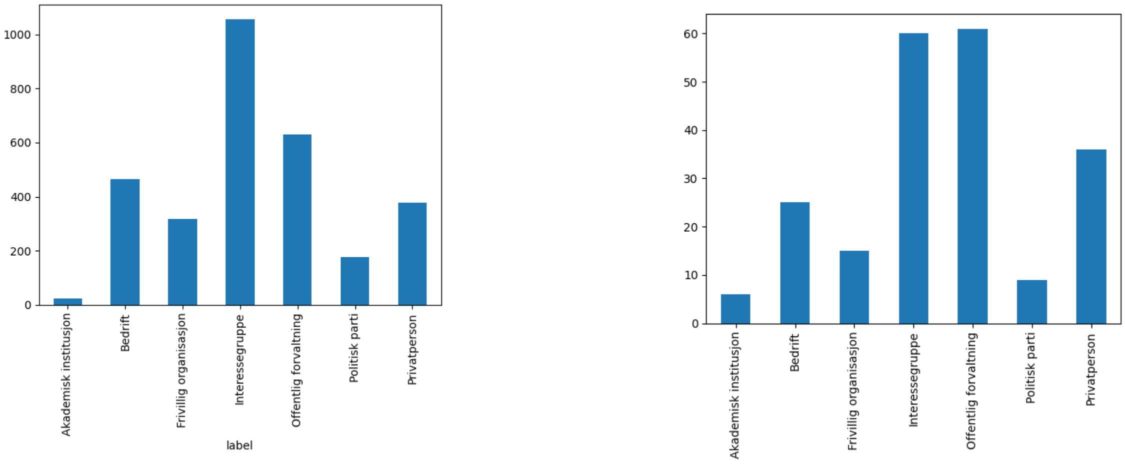
Listing 3: Relevant hyperparameters for fine-tuning Nb-Bert

5 Results

This section presents the results of the building of our dataset and the various approaches to argument mining and multi-label classification of Norwegian text.

5.1 Building dataset

The result of our dataset-building was two datasets, one where each data point held an entire response-text *nou_hearings_full_text* holding 213 records, and one where each data point held a single paragraph *nou_hearings* holding 3048 records. The paragraph frequency distribution organized by actor label is displayed in figure 2.



(a) The number of paragraphs in each class.

(b) The number of documents in each class.

Figure 2: Label distributions represented in actor classes.

Figure 2 illustrates the distribution of actor classes across our two raw datasets, *nou_hearings* and *nou_hearings_full_text*. A notable observation is that the *Offentlig forvaltning (Public Administration)* and *Akademisk institusjon (Academic institution)* classes tend to produce shorter responses compared to other classes. These classes form a large part of the document dataset but a relatively smaller part of the paragraph dataset. Similar characteristics can be observed for the *Privatperson (Private person)* class to a smaller degree. In contrast, the remaining classes show proportionate distributions between the two datasets, indicating they write responses of consistent length across both documents and paragraphs. This consistency suggests a uniform style of response from a global perspective, unlike the variability seen in classes like Public Administration, Private Persons, and Academic Institutions.

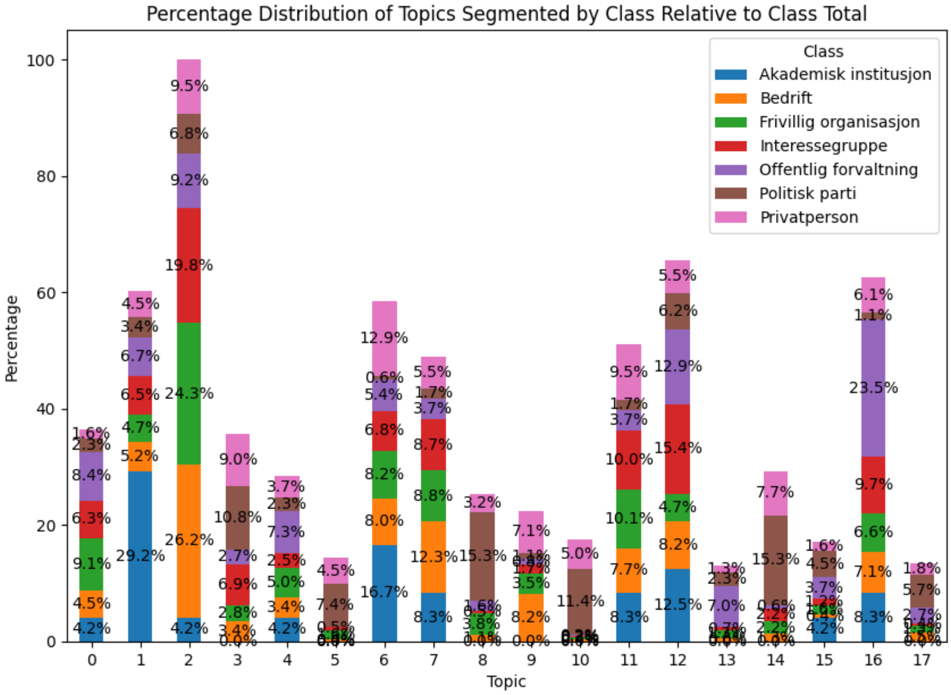


Figure 3: Stacked histogram displaying the topic distribution of all paragraphs, separated into actor labels. Paragraph frequency of each actor has been normalized to the percentage of their total class paragraph frequency to more clearly visualize actor-specific topic allocation rather than frequency. Each paragraph is only classified with their top-confidence topic.

In figure 3, we can see the allocation of paragraph topics. Wordclouds describing these topics can be found in appendix A.1. Most notably is topic 2, which has a large portion of allocation. Its most defining words are *nature*, *increase*, *greenhouse gass emissions* and *loss*. Interestingly, Academic institutions dominate topic 1, the most defining words of this class is ”support” which may refer to *financial support*, and other defining words are *krav/requirements* and *virkemiddel/instrument/means of action*.

5.2 Argument mining

5.2.1 In-Context Learning Argument Mining

The results of the in-context learning argument mining approach resulted in a dataset that holds the same amount of records as the parent dataset *nou_hearings* of 3048, ultimately providing one argument for each paragraph in the hearings. The arguments extracted from each paragraph using LLaMa-2-70B have varied in quality, as some paragraphs in the model extracted sentences that did not contain the minimal conditions to be called an argument, such as a claim and a supportive premise. The model also sporadically responded in English making it challenging to improve the prompt for all cases. As our fine-tuned classification approach is based on an initially multi-lingual model, we have continued using this data and fine-tuned a BERT model to see how it would perform compared to the other models, and if language has a comparable impact.

```
text: "Akademikerne støtter utvalget i at det er behov for
        styrkede rammer og et system som hjelper oss til å tenke
        mer helhetlig og langsiktig om omstillingen
        til et lavutslippssamfunn.
        Det er behov for langsiktighet og systematikk i hvordan klima
        og natur ivaretas.
        Grunnlaget må være brede ambisiøse klimaforlik.",
actor: "Akademikerne",
label: 0
```

Listing 4: Example of an argument using ICL mining

5.2.2 Latent Dirichlet Allocation Argument Mining

The result of the latent Dirichlet allocation (LDA) mining approach reduced the parent dataset *nou_hearings* of 3048 records to *1040*, considerably omitting all sentences that did not have any overlapping topics with its neighbour sentences. Similar to the ICL argument mining approach, LDA returned varying results, and extracted texts are a mixture of arguments and statements.

```
text: "Den pekte igjen på Helsedirektoratets rapport
        Samfunnsgevinster av å følge Helsedirektoratets kostråd
        fra 2016. Beregningene der viste helserelaterte
        samfunnsgevinster på 250-300 mrd.
        2023-kroner pr år dersom hele befolkningen fulgte kostholdsrådene.
        Gevinstene fordelte seg på flere leveår og bedret livskvalitet,
        reduserte helsetjenestekostnader og redusert produksjonstap som
        følge av redusert sykefravær, uførhet og tidlig død.
        Vi ønsker her å påpeke den reduserte belastningen
        på helsetjenestene.",
actor: "Arbeidsgiverforeningen Spekter",
label: 0
```

Listing 5: Example of an argument using LDA mining

5.3 Multi-label Text-Classification

In this sub-chapter, we quickly present the results from the classification models, before going into the key takeaways of the base classification models and finetuned BERT-classifier.

Raw Paragraph Dataset

	Accuracy	Precision	Recall	F-measure	Evaluation size
NBC Count	0.62	0.58	0.45	0.47	610
SVM Tf-IDF	0.66	0.53	0.46	0.48	610
Raw NOU Classifier	**0.79**	**0.65**	**0.66**	**0.65**	610

Table 1: Performance Metrics using the macro average score. The best scores are highlighted in bold.

ICL-Arguments Dataset:

	Accuracy	Precision	Recall	F-measure	Evaluation size
NBC CV	0.51	0.49	0.37	0.38	610
SVM Tf-IDF	0.51	**0.58**	0.36	0.37	610
ICL NOU Classifier	**0.64**	0.53*	**0.50**	**0.50**	610

Table 2: Performance Metrics using the macro-average score. The best scores are highlighted in bold. **Some fields in the dataset were not labeled setting the precision to 0, reducing the macro precision metric.*

LDA-Arguments Dataset

	Accuracy	Precision	Recall	F-measure	Evaluation size
NBC CV	0.52	0.54	0.32	0.34	208
SVM Tf-IDF	0.53	0.68	0.42	0.48	208
LDA NOU Classifier	**0.72**	**0.58**	**0.57**	**0.57**	208

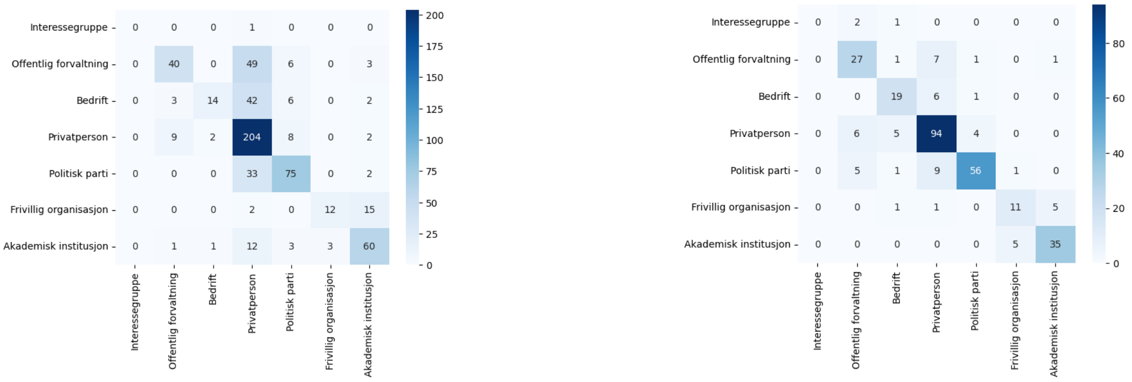
Table 3: Performance Metrics using the macro-average score. The best scores are highlighted in bold

5.3.1 Baseline Classifier Models

The baseline models proved to be fast and efficient learners and also performed decent across all datasets. For the *nou_hearings* dataset (*Raw-Paragraph*) both baseline models performed decently in terms of accuracy, both between the 60th and 70th percentile, and somewhat lower in terms of precision and recall. For the *ICL-Arguments* and *LDA-Arguments*, we see a slightly reduced performance at approximately 50% accuracy and proportionally reduced precision and recall compared to *Raw-Paragraphs*. As a reference, random classification will achieve an accuracy of approximately 33% due to unbalanced label dataset distributions.

We compared multiple vectorizations for every dataset. CV performed best for the Naive-Bayes classifier for every case, and Tf-IDF-V performed best for the SVM model.

The confusion matrices for the three datasets with Naive-Bayes and SVM are shown in appendix A.2. In general, we see that these models tend to incorrectly predict *private person* as the class, especially when the true label is *public administration*. The two best confusion matrices from the BERT classifier and the best base model are shown as confusion matrices in figure 4.



(a) SVM model

(b) Fine-tuned BERT model

Figure 4: Confusion matrices on the NOU paragraph dataset.

5.4 Finetuned NB-BERT classifier

The fine-tuned NB-BERT model performs considerably better across all datasets. It dominates the metrics in every domain, except for the precision in the ICL-model. For this specific dataset, some data points are given a precision of 0 despite correct classification for unknown reasons, ultimately reducing its overall precision in this domain. Its top performance is in the *Raw-Paragraph* dataset, at an accuracy of 79%. The remaining metrics were similar to the baseline models; slightly lower, and maintained a proportional relationship to the accuracy across all three domains. Despite being considerably slower to train in comparison to the baseline models, it learned the general features of each class using relatively few training steps, starting to show diminishing returns at around 500 training steps and converging towards an accuracy of about 80 %, which it peeked above at an evaluation at about 1300 training steps. The learning curve for the models in terms of accuracy can be seen in figure 5. Similar behavior is displayed for the other evaluation metrics which is provided in appendix A.3. These results show that these models are able to significantly differentiate how the different industry sectors articulate their NOU-responses.

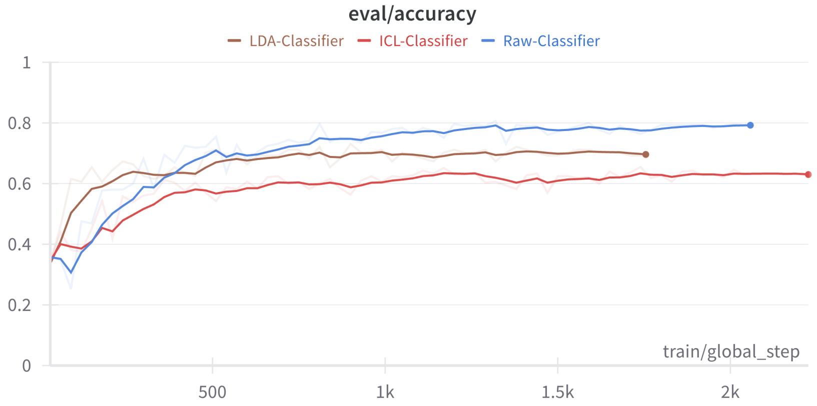


Figure 5: Graph showing the accuracy over time during fine-tuning of the three models: Raw-, ICL-, and LDA-Classifier. Smoothing of 0.5 applied.

6 Discussion

6.1 Dataset actor distribution discrepancies

The dataset analysis of data point distribution discrepancies in section 5.1 highlights potential communication characteristics among different actor classes in NOU-Hearing responses. Notably, the *Public Administration* and *Academic institution* classes tend to produce shorter responses, suggesting a preference for concise communication in formal settings. This brevity might reflect the structured communication norms typical of these sectors. Other factors may be the motivation behind a comprehensive response. In contrast, other classes show more extensive textual contributions, indicating a more detailed and exploratory communication approach, which may be due to increased motivation for responding, such as the feeling of having a unique opportunity of significant impact, having strong opinions on the subject, which potentially could be typical of *interest organizations* working towards specific political agendas, or simply a need to address complex issues more comprehensively.

These reflections are, given the limited size of the data, highly speculative, and a much deeper analysis of linguistic patterns must be conducted to be able to draw any conclusions regarding the subject at hand. Further investigation into the factors influencing these differences could enhance our understanding of communication characteristics and underlying factors across various industry sectors.

6.2 Argument Mining

In our argument mining efforts, despite not being able to perform a statistical evaluation, surface-level observations of our results suggest that these approaches did not perform optimally and are not yet sufficient for real-world applications. Our in-context learning model LLaMa2-70B, translated some texts to English despite explicitly being instructed to respond in Norwegian. This is one of many well-known unfortunate traits of these newly released generative large language models (LLMs). However, given the steep curve of advancements within this field, performance is likely to increase over time.

Our in-context learning mining results align well with the findings of Chen et al. (2024), which concluded that LLaMa performed the poorest of the autoregressive models in their study of using LLMs for argument mining. The model was in our case chosen purely due to the substantially increased costs of competitor APIs.

The LDA Model proved to be a more consistent approach to argument mining, but no true value was found other than reducing isolated sentences and subsequently reducing the dataset. In addition to this, despite the approach of Lawrence, Reed et al. (2014) using small LDA-models and achieving decent result for argumentative linking, LDA-models are typically trained over a massive corpus of data. Building a LDA-model at such a small scale provides inconsistent allocation of topic words even with hundreds of passes, ultimately being very inconsistent.

Evaluating argument-mining approaches requires comprehensively labeled datasets, which is a complex task demanding proper competence in linguistic analysis and a comprehensive understanding of argumentative structures. In our search for datasets to use for training argument evaluation models, we found only the English *TACO* dataset to be of satisfactory level, despite argument mining being an established research field for many years. This may substantiate the somewhat currently limited approaches to efficient argument mining and highlight the need for more resources being put into data labeling to see substantial advancements in the field. With the ever-evolving large language modeling, this could potentially be a task such a model could aid in, for the future of argument mining.

6.3 Classification significance

Despite our diminished results in argument mining, our results in classifying the various industry sector actors are considered to be a success. Both baseline models, and our fine-tuned LLM-classifier displayed a significant ability to differentiate the actors considering the baseline of guessing being at approximately 33% in our dataset. This shows that there are differences in linguistic characteristics present in these NOU-hearing responses, which has been the motivation behind this study to begin with. These results indicate that there is value in further investigating into applying NLP techniques for analyzing linguistic characteristics of industry actors in NOU-hearing responses. Such research efforts could reveal valuable insights within the field of social science. Taking it a step further, developing more comprehensive pipelines for aiding in full-text analysis using state-of-the-art language models could potentially improve both the quality and efficiency of the work being this research domain.

7 Conclusion

This study explored the application of natural language processing (NLP) techniques for analyzing submissions to public hearings on Norway's climate policy. By constructing and analyzing datasets of full texts and individual paragraphs, we identified distinctive communication styles among different actors, such as Public Administration and Academic Institutions. Although the argument mining methods employed—specifically in-context learning and Latent Dirichlet Allocation did not perform optimally, they highlighted the challenges and potential of NLP in extracting meaningful arguments from policy texts.

The classification experiments demonstrated that NLP could effectively differentiate between submissions from various actors, confirming the presence of unique linguistic patterns that could be identified to a significant degree. These findings suggest that NLP holds promise for enhancing text analysis in the field of social science, pointing towards the need for further development of these techniques to improve their accuracy and applicability in real-world applications. Developing more comprehensive pipelines for aiding in full-text analysis using state-of-the-art language models could potentially improve both the quality and efficiency of the work conducted in social science, ultimately providing valuable insights for researchers, and subsequently the society as a whole.

7.1 Future work

Future research should consider utilizing larger language models, such as GPT-4 or a multilingual version of LLaMA-3, for in-context learning argument mining to potentially achieve desirable results. Additionally, the development of a labeled Norwegian argument dataset would be crucial for fine-tuning models dedicated to argument extraction.

As an additional last-minute effort, we attempted to train a sentiment analysis regression model using two different datasets, to determine the positions of various actors regarding the hearing. However, challenges in achieving satisfactory performance within a limited time-frame despite training on two distinct labeled Norwegian datasets led us to omit these efforts from this report. Future studies could revisit this approach on our dataset with better tools, training data, models, and a larger time perspective.

Bibliography

Lawrence, John and Chris Reed (Jan. 2020). 'Argument Mining: A Survey'. en. In: *Computational Linguistics* 45.4, pp. 765–818. ISSN: 0891-2017, 1530-9312. DOI: 10.1162/coli_a_00364. URL: https://direct.mit.edu/coli/article/45/4/765-818/93362 (visited on 20th Apr. 2024).

Stahl, Maja et al. (Apr. 2024). *A School Student Essay Corpus for Analyzing Interactions of Argumentative Structure and Quality*. arXiv:2404.02529 [cs] version: 1. URL: http://arxiv.org/abs/2404.02529 (visited on 20th Apr. 2024).

Shah, Deval (2024). *What is In-context Learning, and how does it work: The Beginner's Guide — Lakera – Protecting AI teams that disrupt the world.* URL: https://www.lakera.ai/blog/what-is-in-context-learning (visited on 17th Apr. 2024).

Blei, David M (Jan. 2003). 'Latent Dirichlet Allocation'. en. In.

Pedro, João (Jan. 2022). *Understanding Topic Coherence Measures*. en. URL: https://towardsdatascience.com/understanding-topic-coherence-measures-4aa41339634c (visited on 21st Apr. 2024).

GeeksForGeeks (Mar. 2017). *Naive Bayes Classifiers*. en-US. Section: Python. URL: https://www.geeksforgeeks.org/naive-bayes-classifiers/ (visited on 17th Apr. 2024).

Ratz, Arthur V. (Apr. 2022). *Multinomial Naive Bayes' For Documents Classification and Natural Language Processing (NLP)*. en. URL: https://towardsdatascience.com/multinomial-na%C3%AFve-bayes-for-documents-classification-and-natural-language-processing-nlp-e08cc848ce6 (visited on 17th Apr. 2024).

IBM (2024). *What Is Support Vector Machine? — IBM*. en-us. URL: https://www.ibm.com/topics/support-vector-machine (visited on 19th Apr. 2024).

NationalBiblioteket (2024). *NB BERT-base*. nb-NO. URL: https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-72/ (visited on 18th Apr. 2024).

GeeksForGeeks (Apr. 2020). *Explanation of BERT Model - NLP*. en-US. Section: AI-ML-DS. URL: https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/ (visited on 18th Apr. 2024).

Chen, Guizhen et al. (Mar. 2024). *Exploring the Potential of Large Language Models in Computational Argumentation*. arXiv:2311.09022 [cs] version: 2. URL: http://arxiv.org/abs/2311.09022 (visited on 11th Apr. 2024).

Lawrence, John, Chris Reed et al. (June 2014). 'Mining Arguments From 19th Century Philosophical Texts Using Topic Based Modelling'. In: *Proceedings of the First Workshop on Argumentation Mining*. Ed. by Nancy Green et al. Baltimore, Maryland: Association for Computational Linguistics, pp. 79–87. DOI: 10.3115/v1/W14-2111. URL: https://aclanthology.org/W14-2111 (visited on 20th Apr. 2024).

Bilal, Muhammad and Abdulwahab Ali Almazroi (Dec. 2023). 'Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews'. In: *Electronic Commerce Research* 23.4, pp. 2737–2757. ISSN: 1572-9362. DOI: 10.1007/s10660-022-09560-w. URL: https://doi.org/10.1007/s10660-022-09560-w.

Sivesind, Nicolai Thorer and Andreas Bentzen Winje (2023). 'Turning Poachers into Gamekeepers: Detecting Machine-Generated Text in Academia Using Large Language Models'. eng. Accepted: 2023-07-11T17:30:32Z. Bachelor thesis. NTNU. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3078096 (visited on 20th Apr. 2024).

Regjeringen (Oct. 2023). *Høring - NOU 2023: 25 Omstilling til lavutslipp - Veivalg for klimapolitikken mot 2050 - rapport av Klimautvalget 2050*. nb-NO. Horing. Publisher: regjeringen.no. URL: https://www.regjeringen.no/no/dokumenter/horing-nou-2023-25-omstilling-til-lavutslipp-veivalg-for-klimapolitikken-mot-2050-rapport-av-klimautvalget-2050/id3009052/ (visited on 15th Apr. 2024).

Brønnøysundregistrene (2024). *Forsiden*. nb. URL: https://www.brreg.no (visited on 15th Apr. 2024).

Velldal, Erik et al. (n.d.). 'NoReC: The Norwegian Review Corpus'. en. In: ().

Feger, Marc and Stefan Dietze (Aug. 2023). *TACO: Twitter Arguments from COnversations*. en. DOI: 10.5281/ZENODO.8030026. URL: https://zenodo.org/doi/10.5281/zenodo.8030026 (visited on 21st Apr. 2024).

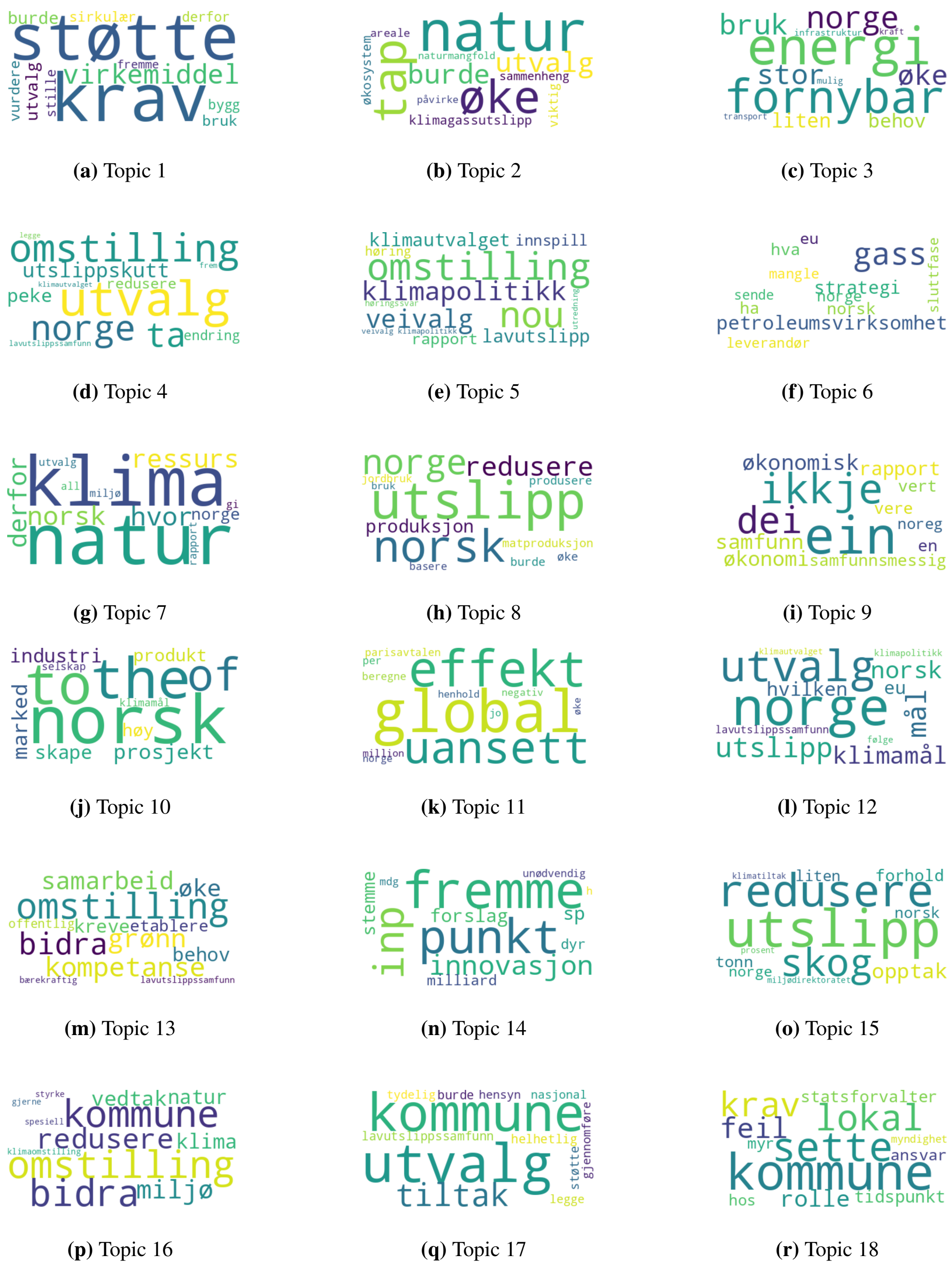
Saket, Sheel (2024). *(9) Count Vectorizer vs TFIDF Vectorizer — Natural Language Processing — LinkedIn*. URL: https://www.linkedin.com/pulse/count-vectorizers-vs-tfidf-natural-language-processing-sheel-saket/ (visited on 18th Apr. 2024).

Kummervold, Per E et al. (2021). 'Operationalizing a national digital library: The case for a norwegian transformer model'. In: *Proceedings of the 23rd nordic conference on computational linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, pp. 20–29. URL: https://aclanthology.org/2021.nodalida-main.3.

NbAiLab/nb-bert-large (Feb. 2024). URL: https://huggingface.co/NbAiLab/nb-bert-large (visited on 21st Apr. 2024).

A Appendix

A.1 Wordclouds LDA



(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

(e) Topic 5

(f) Topic 6

(g) Topic 7

(h) Topic 8

(i) Topic 9

(j) Topic 10

(k) Topic 11

(l) Topic 12

(m) Topic 13

(n) Topic 14

(o) Topic 15

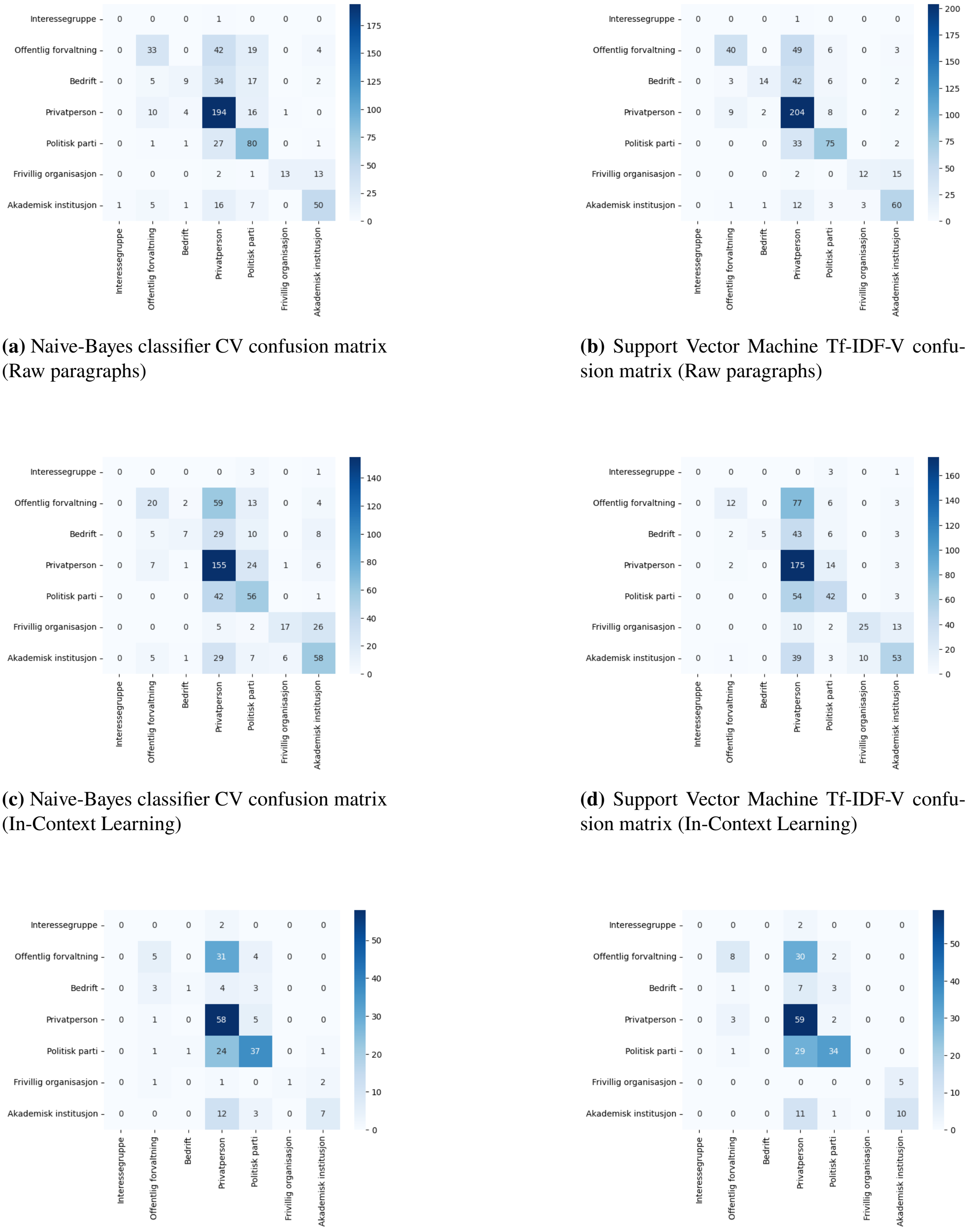
(p) Topic 16

(q) Topic 17

(r) Topic 18

Figure 6: Wordsclouds for the 18 topics created

A.2 Confusion matrices



(a) Naive-Bayes classifier CV confusion matrix (Raw paragraphs)

(b) Support Vector Machine Tf-IDF-V confusion matrix (Raw paragraphs)

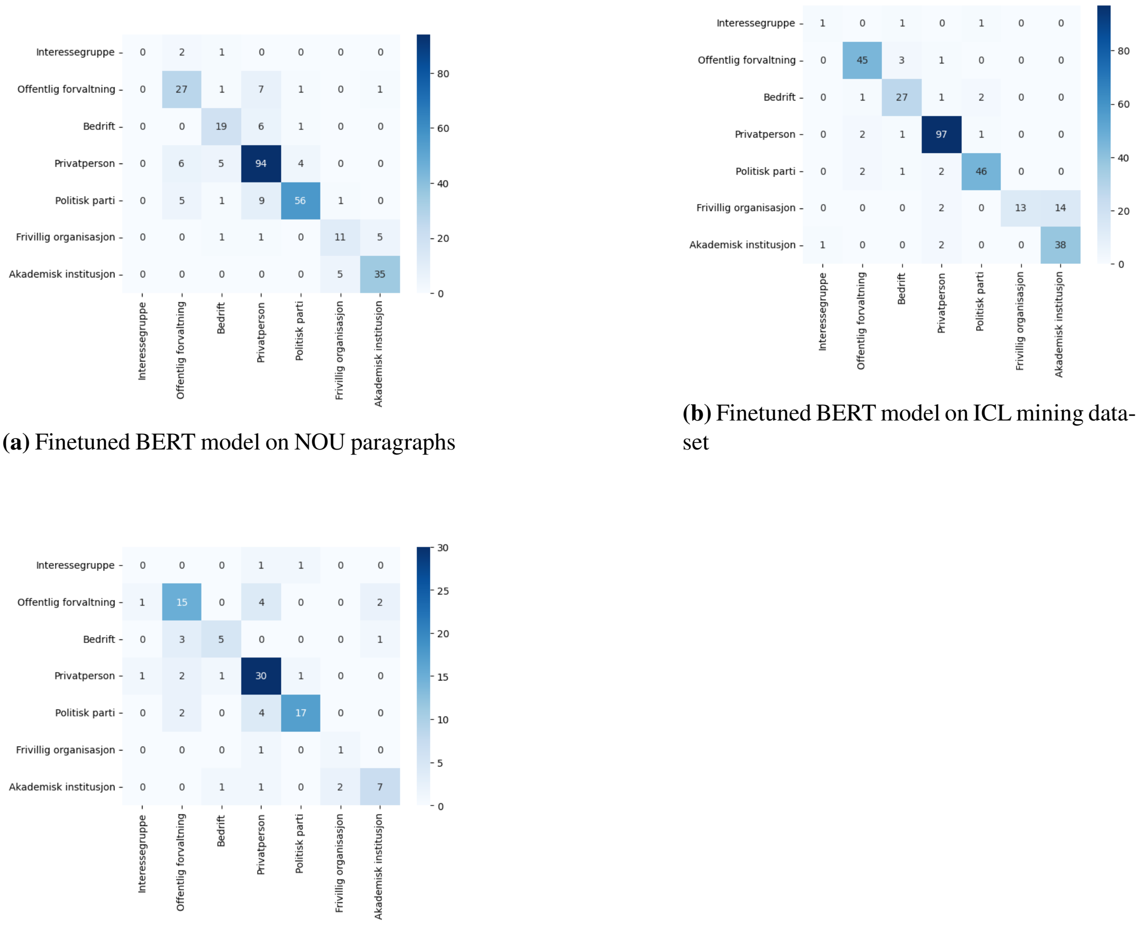
(c) Naive-Bayes classifier CV confusion matrix (In-Context Learning)

(d) Support Vector Machine Tf-IDF-V confusion matrix (In-Context Learning)

(e) Naive-Bayes classifier CV confusion matrix (LDA Mining)

(f) Support Vector Machine Tf-IDF-V confusion matrix (LDA Mining)

Figure 7: Combined confusion matrices for Naive-Bayes classifier and Support Vector Machine across the different datasets



(a) Finetuned BERT model on NOU paragraphs

(b) Finetuned BERT model on ICL mining dataset

(c) Finetuned BERT model on LDA Mining

Figure 8: Combined confusion matrices and BERT model performance across the different datasets

A.3 Nb-BERT additional Training Metric Time-series

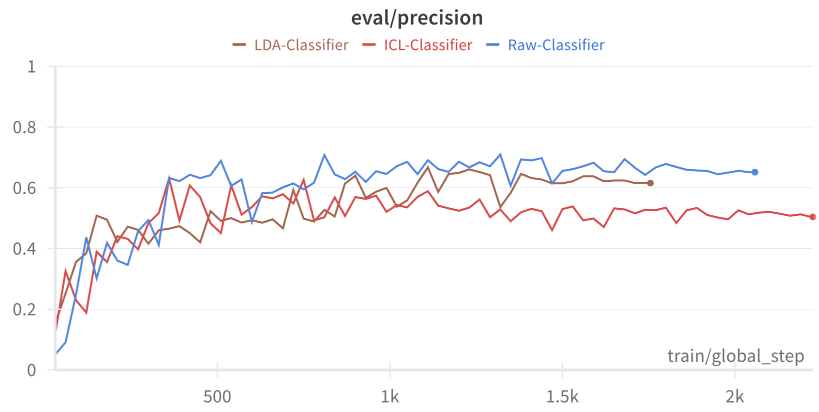


Figure 9: Graph showing the Precision over time during fine-tuning of the three models: Raw-, ICL-, and LDA-Classifier.

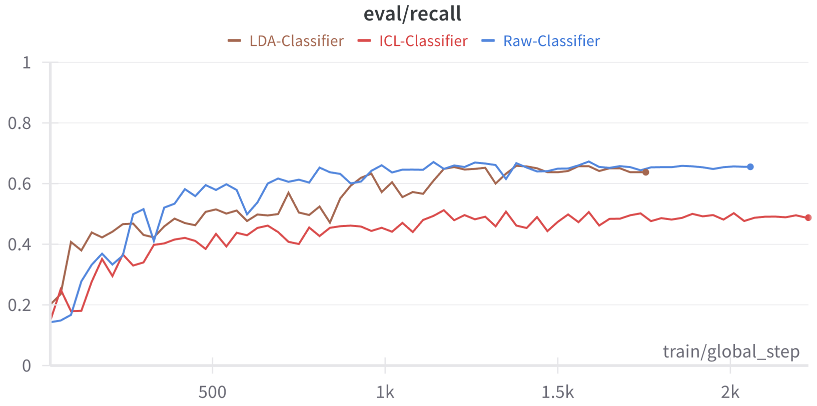


Figure 10: Graph showing the Recall over time during fine-tuning of the three models: Raw-, ICL-, and LDA-Classifier.

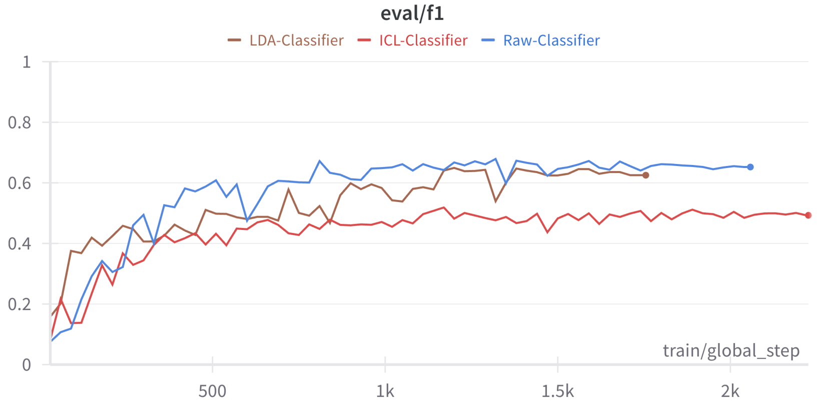


Figure 11: Graph showing the F1-Score over time during fine-tuning of the three models: Raw-, ICL-, and LDA-Classifier.