UNIVERSITY OF GENEVA

MASTER'S IN COMPUTER SCIENCE THESIS

Automated Scoring System for Assessing Swiss Municipalities Sustainability.

Author:
Muhammad Azeem ARSHAD

Supervisor:
Dr. Alexandre DUPUIS

A thesis submitted in fulfillment of the requirements for the degree of MSc in Computer Science

in the

Faculty of Science
Centre Universitaire d'Informatique (CUI)

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

January 27, 2025

UNIVERSITY OF GENEVA

Abstract

Faculty of Science

Centre Universitaire d'Informatique (CUI)

MSc in Computer Science

Automated Scoring System for Assessing Swiss Municipalities Sustainability.

by Muhammad Azeem ARSHAD

This thesis presents a research project to develop and implement an automated scoring system to assess the Environmental, Social, Governance (ESG) factors across Swiss municipalities. Initially designed as a framework applicable to all municipalities in Switzerland, we apply it to certain medium-sized cities: Nyon, Rolle, and Vevey. By fine-tuning large language models based on the transformers architecture, we classified municipal council transcripts into ESG categories and subsequently applied sentiment analysis to derive ESG ratings. Our methodology included creating a new balanced dataset translated into French from a collection of English article headlines and fine-tuning CamemBERT models. Multiple models were produced using High-Performance Computing (HPC), and a weighted voting scheme was employed to combine and enhance classification accuracy.

The results indicated strong performance in identifying environmental, governance, and non-ESG-related content, with some challenges in distinguishing social aspects. Key findings revealed a significant representation of environmental topics and a notable increase in governance discussions. The social category may be underrepresented, possibly due to overlaps with the non-ESG category. Emerging trends showed stable ratings for Nyon and Vevey, while Rolle exhibited slightly more variability due to the lower number of available council transcripts.

The automated system demonstrated efficiency in analysing large volumes of municipal transcripts, providing valuable insights for policymakers. Limitations included occasional misclassification of social content and a focus on French-language data. By translating the dataset into German or Italian, the framework could be broadened to encompass the entire country. This work offers a novel approach for automatic ESG assessment in the public sector in Switzerland, facilitating more informed decision-making and setting a benchmark for future research in automated ESG classification.

Acknowledgements

I would like to express my sincere gratitude to Dr. Alexandre Dupuis for his continuous guidance and unwavering support throughout the journey of this thesis.

A heartfelt thank you to my family, whose unwavering support, understanding, and encouragement helped me stay focused and motivated during the more difficult times.

Contents

List of Tables

List of Abbreviations

Env	Environmental
Soc	Social
Gov	Governance
ESG	Environmental, Social, Governance
HPC	High-Performance Computing
CB-512	CamemBert-Base-512
CB-1024	CamemBert-Base-1024
CBL-512	CamemBert-Large-512
CBL-1024	CamemBert-Large-1024
TF	Term Frequency
IDF	Inverse Document Frequency
IQR	Interquartile Range

Chapter 1

Introduction

In recent years, sustainable development has emerged as a critical factor in both the private and public sectors. Various frameworks have been established to assess the sustainability efforts of companies and governments. The ESG framework, which incorporates Environmental, Social, and Governance factors, is one such approach. These criteria have become integral to investment strategies, significantly influencing decision-making processes.[Hvi17]

TABLE 1.1: MSCI ESG Government Ratings: ESG Risk Factors[Inc24]

Pillars	**Risk factors**	**Sub-factors (Exposure)**	**Sub-factors (Management)**
Env. risk	Natural resource	Energy Security Risk, Water Resources, Productive Land and Mineral Resources	Energy resource management, Resource conservation, Water resource management
	Env. externalities and vulnerability	Vulnerability to environmental events Environmental externalities	Environmental performance, Impact of environmental externalities
Social risk	Human Capital	Basic Human Capital, Higher Education and Technological Readiness, Knowledge Capital	Basic Needs, Human Capital Performance, Human Capital Infrastructure, Knowledge Capital Management
	Economic environment	Employment, Wellness	-
Gov. risk	Financial governance	Financial capital	Financial Management
	Political governance	Institutions, Judicial and penal system, Governance effectiveness	Political rights and civil liberties, Corruption control, Stability and peace

In order to evaluate governments, rating agencies have developed specific criteria to assess countries' exposure to and management of ESG factors. For example in table 1.1, MSCI identifies several risk factors associated with countries' ESG practices[Inc24]. They attribute a *Risk Exposure Score* that shows the extent to which a country is vulnerable to these same factors, and a *Risk Management score* that refers to the strategies and policies to mitigate the risks they are exposed to. These scores are finally combined with certain weights attributed to every factor before attributing a final ESG rating to a public entity.[LLC24] A study by the same company further indicates that ESG factors can also affect the sovereign risk of nations over the long term and can guide investment decisions in these countries [Inc16].

While a considerable amount of research has focused on applying machine learning to ESG scores in the private sector, the textual aspects, particularly in the public sector, have been less explored. The study [TK22] compiles into a table previous studies performed on corporate esg issues and ratings, by describing the methodologies and results. By doing so, the authors highlight the growing importance of ESG reporting and the associated challenges in standardising data due to inconsistencies across different reporting frameworks. They further demonstrate that transparent ESG practices can enhance trust among stakeholders, improve governance, reduce profit volatility, and attract long-term investors, thereby providing a competitive advantage.

Despite some advancements, there has been limited research on analysing textual resources such as meeting transcripts and other forms of documentation to evaluate companies with respect to ESG factors. Additionally, no recent studies have specifically focused on the French language within the ESG framework. The study by [PE22] pioneers an approach by fine-tuning transformer-based large language models to simultaneously interpret texts and assign ratings. They compile a dataset using US government reports and scores from rating agencies such as *S&P* Global, intentionally avoiding companies' internal reviews to mitigate biased positive tones. While they report promising results, the lack of detailed information about their dataset, their model fine-tuning process and a more detailed evaluation of their results makes it difficult to fully assess their findings. In our work, we will also attempt to classify and rate texts but will do so separately.

Finally, the recently published research (*December 2023*) [Sch+23] conducted similar works

to ours, by pre-training DistilBert [San+20] and RoBerta [Liu+19] models and further fine-tuning them with their datasets. For each ESG category, they fine-tune the models separately, thus obtaining models that perform binary classification for each category (*non-esg* or E/S/G).

In this thesis, we extended existing methodologies by fine-tuning a French transformer-based large language model to classify publicly available city council proceedings from recent years with respect to Environmental (Env), Social (Soc) and Governance (Gov) factors. We then scored the classified texts using sentiment analysis techniques and developed a systematic process to collect council records and assign ratings by municipality. This approach enabled us to generate spreadsheets for each transcript, detailing the ESG label for every sentence and uncovering significant insights into the representation of ESG factors over the past two years. By adding ratings to each text segment, our preliminary observations suggest that this approach could serve as a valuable tool for public policymakers, helping them save time and make more informed and sustainable decisions.

Chapter 2

ESG Classification

2.1 Dataset

To train a model for our classification task, we need to initially have a dataset that would suit our needs. Unfortunately, as of recent, there has been no publicly available dataset in French that classifies texts according to Env , Soc, or Gov categories. The manual compilation and classification of such a dataset would have been time-consuming, as it would require tens of thousands of sentences to adequately represent each category. To circumvent this, we propose leveraging existing English datasets with pre-classified sentences. We would then translate it into French for our needs. Among the available datasets online, the *gold_standard_corpus* [Fis+23] dataset stands out as one of the largest and the most consistent for our needs.

The dataset includes several columns, as outlined below:

TABLE 2.1: *gold_standard_corpus* dataset outline

headline	**guardian keywords**	**esg category**	**mentions company**
General Motors seeks to reassure Vauxhall on UK job losses	['job losses']	Soc	yes
SSE powers to 40% rise in retail profits despite losing 500,000 customers	['environment']	Env	yes
Facebook's cats are the new opium of the people | Kevin McKenna	['others']	non-esg	yes
McDonald's to scrap Luxembourg tax structure	['tax avoidance', 'corporate governance']	Gov	yes

For our study, which focuses on classifying the proceedings of certain Swiss municipalities, the relevant dataset columns are *headline* and *esg-category*. Due to class imbalance in the original dataset, we initiated our analysis by sampling an equal number of rows for each label, i.e. approximately 4000, aligning with the maximum number of headlines labeled as *"Governance"* in the raw dataset. We also take into account unique tags from the column *guardian_keywords* in the same raw dataset to help us obtain diverse samples.

We observed that the headline lengths are significantly shorter than the text segments of the municipal reports that we would like to classify. To mitigate this discrepancy, we extended each headline by retrieving the corresponding full articles online and subsequently generating concise summaries limited to five sentences. To obtain these summaries, we use python's *Newspaper3k* library, that summarises an article based on the relevance and importance of its sentences. It uses *NLTK*'s pre-trained English tokenizers, and attributes a score to each sentence w.r.t its title and keyword relevance, its sentence length and relative position in the article.

The texts were then translated into French. Few tools are available for freely translating large datasets, and the only effective method we found for translating over 20,000 rows of text was using *Google Sheets*. We uploaded the spreadsheet to Google Sheets and utilised the `GOOGLETRANSLATE` function for translation. Given that our research focuses on French-speaking cities in Switzerland, the dataset can also be translated into German and Italian for studying other regions. For example, the following headline gives us its respective article summary and translation:

- **Headline:** *Indoor carbon dioxide levels could be a health hazard, scientists warn.*
- **Summary:** *Indoor levels of carbon dioxide could be clouding our thinking and may even pose a wider danger to human health, researchers say. While air pollutants such as tiny particles and nitrogen oxides have been the subject of much research, there have been far fewer studies looking into the health impact of CO 2 . The team found a number of studies have looked at the impact of such levels on human cognitive performance and productivity. Any health impacts, they add, might be particularly problematic for children or those with health conditions that might exacerbate the effects. And even if the impacts are reversible, said Hernke, it would depend on people being able to access air with low levels of CO 2 .*

- **French translation:** *Les niveaux de dioxyde de carbone intérieur pourraient être un risque pour la santé, avertissent les scientifiques: Les niveaux intérieurs de dioxyde de carbone pourraient obscurcir notre réflexion et peuvent même représenter un danger plus large pour la santé humaine, selon les chercheurs. Bien que les polluants atmosphériques tels que les minuscules particules et les oxydes d'azote aient fait l'objet de nombreuses recherches, il y a eu beaucoup moins d'études sur l'impact sur la santé du CO 2. L'équipe a constaté qu'un certain nombre d'études ont examiné l'impact de ces niveaux sur les performances cognitives humaines et la productivité.Tout impact sur la santé, ajoutent-ils, pourraient être particulièrement problématiques pour les enfants ou ceux qui ont des problèmes de santé qui pourraient exacerber les effets. Et même si les impacts sont réversibles, a déclaré Hernke, cela dépendrait de la possibilité d'accéder à l'air avec de faibles niveaux de CO 2.*

2.1.1 Dataset structure

We get the following histogram in figure 2.1 representing the number of instances for each label, indicating that the classes are relatively balanced. The slight variations among the labels arise from the challenge in retrieving the URL or extracting the text from the URL using the aforementioned Python libraries.

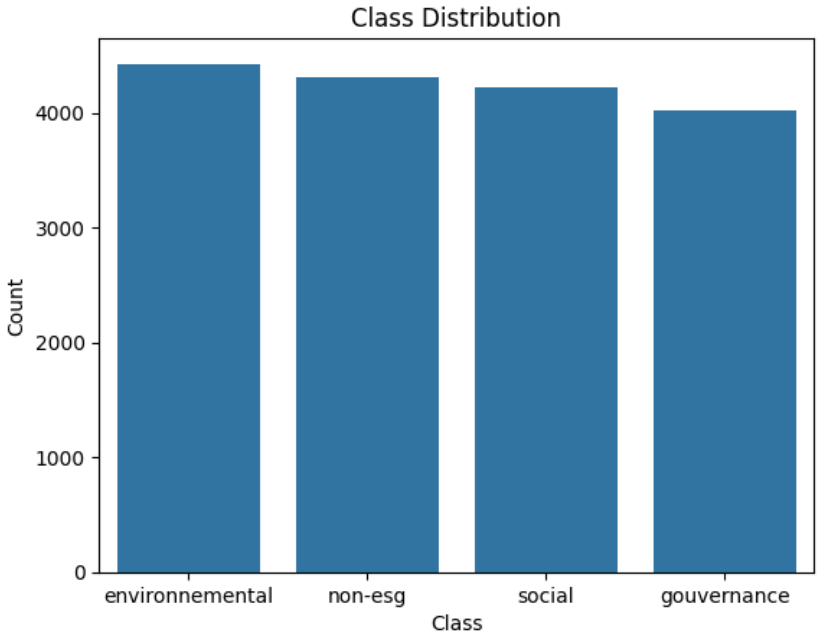


FIGURE 2.1: Dataset class distribution

The boxplot in figure 2.2 below displays the average text lengths (number of characters) for

each label. With a balanced dataset with respect to its labels, we observe minimal variability between the labels, suggesting that this will not introduce significant bias during training.

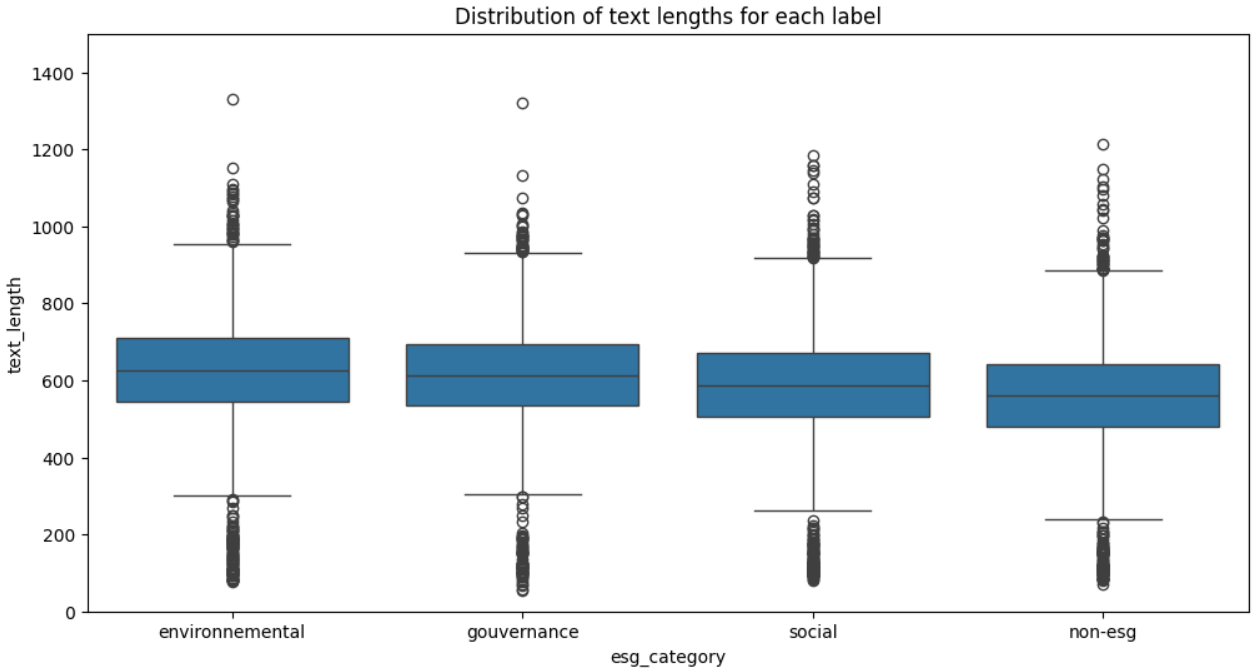


FIGURE 2.2: Dataset input text lengths' boxplot, in characters. The boxes correspond to the area where most of the data lies. The vertical lines are the whiskers which encompassmost of the data, while the dots represent the outliers.

We use the Term Frequency (TF)-Inverse Document Frequency (IDF) statistics to identify and evaluate the most important words in the dataset, based on its frequency across documents. The significance of a word in a document grows with its frequency of appearance, while being offset by the frequency of the word in the corpus. This enables us to see how reliable the dataset is and if no major bias is present in the dataset.

The *TF* tf_{ij} is the weight of a term t_i in a document d_j computed with

$$tf_{ij} = \frac{freq_{ij}}{\max_k freq_{kj}} \tag{2.1}$$

The *IDF* idf_{ij} measures the general importance of the term t_i. We obtain it by dividing the number of total documents N by the number of documents n_i containing the term t_i, before taking its logarithm

$$idf_i = \log\left(\frac{N}{n_i}\right)$$

And we finally obtain the term-weighting scheme defined as

$$w_{ij} = tf_{ij} \cdot idf_i$$

In our case, each document corresponds to a row in our dataset. We present the following table 2.2, which contains the top ten words for each label. We can observe that the top ten words are relevant to their respective labels, and the *Non-ESG*'s top *TF-IDF* words do not have a specific link to any other label, further reducing any potential bias in this context and reinforcing our decision to use this dataset for training. Certain words such as "*pouvoir*" that appear in the *non-ESG* category also appear in the three main categories, thus minimizing any such bias.

Environmental	**Social**	**Governance**	**non-esg**
climatique	femme	fiscal	déclarer
changement	travail	société	pouvoir
pollution	droit	entreprise	faire
déclarer	déclarer	livre	quil
pouvoir	travailleur	sterling	devoir
plastique	pouvoir	million	dernier
émission	faire	déclarer	dun
climat	noir	payer	cest
mondial	devoir	pouvoir	grand
faire	homme	actionnaire	trump

TABLE 2.2: Top TF-IDF words for each class

2.2 Methodology and Evaluation Metrics

In this study, we mainly employ macro-averaged metrics for *precision*, *recall*, and *F1-score* to evaluate the performance of our models. The F1-score is a harmonic mean of precision and recall. Precision is the proportion of true positive instances among all positive predictions

(i.e., the accuracy of positive predictions):

$$\text{Precision} = \frac{TP}{TP + FP}$$

where TP are True Positives and FP are False Positives.

Recall is the proportion of true positive instances among all actual positive instances (i.e., the model's ability to detect positive instances):

$$\text{Recall} = \frac{TP}{TP + FN}$$

where FN are False Negatives.

The F1-score is an overall measure of the model's accuracy that considers both precision and recall:

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

It is particularly useful when the data distribution is uneven (i.e., when the number of instances of different classes vary significantly). Macro-averaging is chosen to ensure that each class contributes equally to the final evaluation, regardless of its frequency in the dataset. To compute it, we take the arithmetic mean of the metric values across all classes to obtain the macro-averaged score. This gives an equal weight to each class and it gives a better insight for any imbalanced data too.

Our methodology involves the use of the CamemBERT model [Mar+20] and a multilingual large language model (LLM) designed for zero-shot classification as baseline models.These baselines provide a reference point to measure the improvement achieved through fine-tuning, with CamemBERT chosen specifically for its focus on the French language.

We will thus perform the following steps through this chapter:

1. **Baseline Evaluation**: Evaluate the performance of our pre-trained CamemBERT model and the multilingual zero-shot LLM on our dataset without any fine-tuning. This provides a benchmark for comparison.

2. **Fine-tune** the CamemBERT model with the curated french dataset above.

3. **Compare** the results based on the aforementioned metrics

4. **External Comparison:** Use the models from the study [Sch+23] that were fine-tuned for the similar task, and compare their results with our model. Since the referenced study trained a separate model for each category, we will evaluate each category-specific model from that study against our corresponding fine-tuned models.

For the comparative analysis, a sample from our training dataset will be used to evaluate and compare the results. This approach ensures the comparison is fair by using the same consistent data for evaluation.

2.3 Baseline

To assess the models we train, we establish a set of baseline models representing the current state-of-the-art in text classification for ESG aspects. These baselines consist of pre-trained large language models such as CamemBERT [Mar+20] and a fine-tuned multilingual BERT model for zero-shot classification.

We select CamemBERT due to its robustness and high performance in various NLP tasks in French, as noted by the authors of this model [Mar+20]. Additionally, we choose a fine-tuned version of DeBERTa [Lau+24] for zero-shot classification, as it ranks among the best in terms of accuracy for multilingual models.

Finally, we will assess the models developed by [Sch+23] to compare our findings with their results.

2.4 Fine-tuning

Bidirectional transformers, as employed in the CamemBERT models, leverage the architecture developed by [Dai+19] to process text by simultaneously considering information from both past and future contexts within a sequence [Dev+19]. This approach enables the model to capture a deeper understanding of the language structure, which can help improve the interpretation of texts. Our objective is to enhance the performance beyond what was achieved with the baseline models.

To minimize noise in our dataset, we preprocessed the data by lowercasing the text, lemmatizing, and removing the stopwords using the Python library *spaCy*. We divided the dataset into training, validation, and testing sets with ratios of 0.6, 0.2, 0.2, respectively.

In order to fine-tune, we use the base and large CamemBERT models. The base model was pre-trained on the *Oscar* corpus[Aba+22] and has 110M parameters, while the larger model hasa 335M parameters and was pre-trained on the CCNet corpus[Wen+19]. The difference in the number of parameters and the corpus used for training, led us to train and evaluate both models. While both datasets are derived from web crawled data, *CCNet* has an advanced text cleaning process to improve the text quality. On the other hand, *Oscar* focuses on having a broader dataset with a wide variety of content types. This can be beneficial for tasks that contain more varied data, while models trained on *CCNet* are preferred for tasks requiring precise language understanding and generation.

Considering the extensive lengths of the transcripts that we chose, we also developed a model class for each model, where we increased the maximum embedding length. The original model configuration limits the number of tokens it can process due to a predefined maximum sequence length of 512(`max_position_embeddings`). To address this, we duplicated the original position embeddings, thus enabling the pretrained model to handle longer sequences without the need for complete retraining, thereby saving considerable time.

The embedding length represents the maximum number of tokens (for each character) that the model can process in a single sequence. By extending this length from 512 to 1024 tokens, the model can capture more context from lengthy discussions from the transcripts. A longer embedding length allows the model to understand and retain more information within a single pass, improving its ability to generate accurate classifications and insights.

However, this practical workaround could introduce issues due to the assumption of cyclic position embeddings. The duplication assumes that position embedding patterns are cyclic, which may not always hold true, potentially leading to semantic inconsistencies.[Dai+19] We believe that the quantity of data provided during training, coupled with the average input text length of approximately 1000 characters, can mitigate this issue.

Thus, for training, we respectively have four model classes that we will fine-tune over CamemBERT-base and CamemBERT-large:

- CamemBert-Base-512 (CB-512)
- CamemBert-Base-1024 (CB-1024)
- CamemBert-Large-512 (CBL-512)
- CamemBert-Large-1024 (CBL-1024)

TABLE 2.3: Hyper-parameters the four models

(A) CB-512

Parameter	Value
Batch Size	64
Gradient Steps	8
Epochs	30
Learning Rate	*cosine*

(B) CB-1024

Parameter	Value
Batch Size	64
Gradient Steps	8
Epochs	35
Learning Rate	*cosine*

(C) CBL-512

Parameter	Value
Batch Size	64
Gradient Steps	8
Epochs	30
Learning Rate	*cosine*

(D) CBL-1024

Parameter	Value
Batch Size	64
Gradient Steps	8
Epochs	30
Learning Rate	*cosine*

The computations for training the models were performed at University of Geneva using Baobab HPC service.[1]. We use the following parameters of the cluster to train the four models:

```
#SBATCH --partition=shared-gpu
#SBATCH --time=0-12:00:00
#SBATCH --mem=0
#SBATCH --gres=gpu:1,VramPerGpu:30G
```

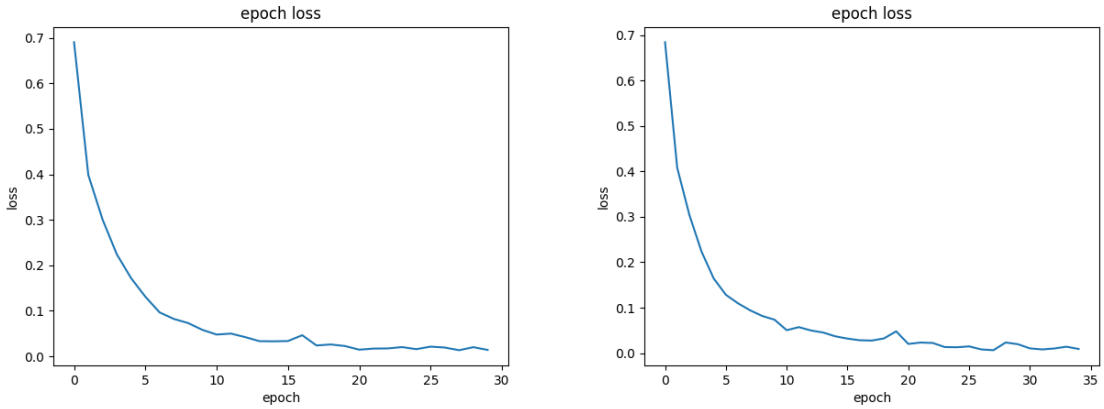
LISTING 2.1: SLURM Job Submission Script

The configuration line `VramPerGpu:30G` in the sbatch script 2.1 above was added to accommodate the training of CamemBert-Large models, as memory constraints were encountered with the GPU resources available by default.

[1]Documentation for the cluster can be found here: https://doc.eresearch.unige.ch/hpc/start

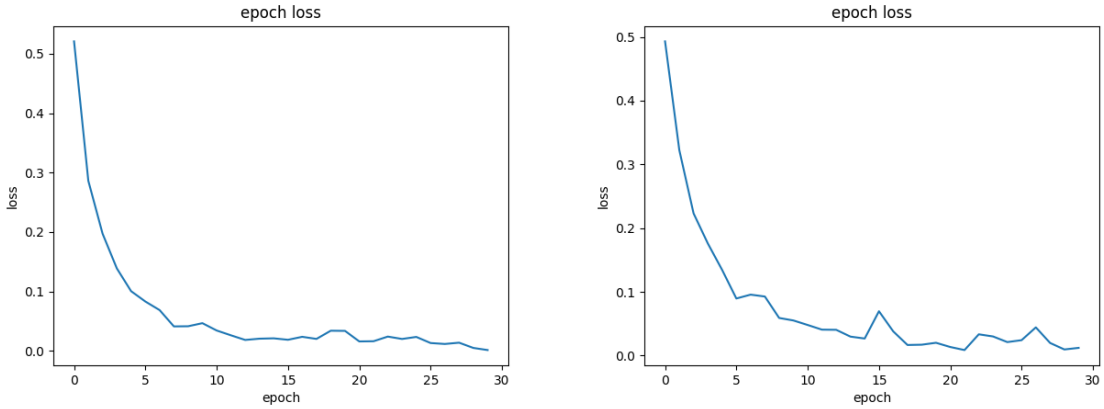
2.4.1 Results

Figures 2.3 and 2.4 illustrate the training losses of the models over a period of 30 epochs, highlighting the stable learning process. It reaches a certain plateau after 20 epochs.



(A) CB-512 (B) CBL-1024

FIGURE 2.3: Training loss for the fine-tuning of CamemBert-Base models over 30 epochs



(A) CBL-512 (B) CBL-1024

FIGURE 2.4: Training loss for the fine-tuning of CamemBert-Large models over 30 epochs

Table 2.4 presents the macro-averaged F1-scores for both training and testing, as well as the average training time, computed over five rounds of training. With the CamemBERT-Base models, we observe a slight improvement in scores with an increase in the embedding size. This improvement may be attributed to the presence of data rows containing texts longer than 512 characters. However, for the CamemBERT-Large models, the F1-scores remain

consistent regardless of the embedding size. Given that the scores do not decrease at the very least, the models with longer embedding sizes can also be considered for practical use.

TABLE 2.4: Macro-averaged F1-scores and Average Training Times. Run on the university of Geneva's HPC, with a GPU of upto 30GB of VRAM and 1 CPU core (default allocation).

Model	**f1-score (train data)**	**f1-score (test data)**	**Avg Training Time (*hrs*)**
CB-512	0.86	0.85	1.3
CB-1024	0.87	0.86	1.13
CBL-512	0.87	0.86	3.08
CBL-1024	0.86	0.86	2.91

Table 2.5 summarises the F1-scores obtained by our trained models and compares them with several baseline models. The baselines include the CamemBERT-base and large models used for zero-shot classification (classification of data into categories without any prior exposure or training on specific examples from those categories), and the DeBERTa-v3 model, which in this case was trained for zero-shot classification.

Additionally, we compare our results with those from the model trained by [Sch+23], which, like ours, is trained for the same purpose but in English. It comprises three separate models, each designed to perform binary classification for one of the ESG categories (Environmental/ Social/ Governance or *None*). Online they are each denominated as *ESGBert/environmentalBERT*[2], *ESGBert/SocialBERT*[3] and *ESGBert/GovernanceBERT*[4]. Thus, to obtain an accuracy score, we input the sentences categorised under each ESG label from the English version of our test dataset into the corresponding model. Specifically, sentences labelled as Environmental are input into the Environmental model, and the accuracy is computed. This process is repeated for the Social and Governance models, allowing us to compute the accuracies for each respective category.

[2]https://huggingface.co/ESGBERT/EnvironmentalBERT-environmental
[3]https://huggingface.co/ESGBERT/SocialBERT-social
[4]https://huggingface.co/ESGBERT/GovernanceBERT-governance

TABLE 2.5: Test dataset F1-scores obtained for each baseline models, the reference model (ESGBERT), and our fine-tuned models. The first three are baseline models directly obtained from *Huggingface* without fine-tuning. ESGBert is our reference model from [Sch+23] that is trained to perform the a similar task in English.

Model	non-esg	Env	Soc	Gov
CamemBert-Base	0.01	0.39	0.14	0.07
CamemBert-Large	0.24	0.32	0.30	0.30
deBert-v3-Large-0shot	0.27	0.77	0.46	0.67
ESGBert	-	0.91	0.71	0.05
CB-512	0.76	0.92	0.83	0.90
CB-1024	0.78	0.94	0.83	0.90
CBL-512	0.75	0.93	0.82	0.91
CBL-1024	0.77	0.92	0.83	0.91

After training, our models significantly outperform the baseline models. For environmental sentences, the classification performance is comparable to that of ESGBert. However, for the social category, our models show a relative improvement of 15%. In the governance category, *ESGBert-GOV* achieves a very low accuracy of only 0.05. This poor performance can likely be attributed to a substantial discrepancy between the training data provided to *ESGBert-GOV* model and the testing data.

2.5 ESG-Classifier Framework

We tested the four models on the municipality session records. When classifying a meeting record, we observed that each model predominantly predicted differently the segments of the transcript. This discrepancy is likely due to variations in the base models: The CamemBERT-base model was pre-trained on the OSCAR[Aba+22] corpus and has $110M$ parameters, while CamemBERT-Large has $335M$ parameters and was pre-trained on the CCNet [Wen+19] dataset.

To combine the predictions from each model, we employ a simple weighted voting approach. The weights are determined by evaluating the performance of the four models on a separate, manually annotated smaller dataset, that would not induce any bias. Due to time constraints, this dataset was constructed manually, with approximately 20 sentences per label. We follow the following steps to generate the weights and label selection:

1. For each model M_i, we generate the corresponding predictions $\hat{y}_i$.

2. we compute the $f1$ evaluation metric for each model w.r.t each label. We obtain a matrix $M^{4\times4}$:

$$\mathbf{M} = \begin{bmatrix} \text{metric}_{1,c1} & \text{metric}_{1,c2} & \text{metric}_{1,c3} & \text{metric}_{1,c4} \\ \text{metric}_{2,c1} & \text{metric}_{2,c2} & \text{metric}_{2,c3} & \text{metric}_{2,c4} \\ \text{metric}_{3,c1} & \text{metric}_{3,c2} & \text{metric}_{3,c3} & \text{metric}_{3,c4} \\ \text{metric}_{4,c1} & \text{metric}_{4,c2} & \text{metric}_{4,c3} & \text{metric}_{4,c4} \end{bmatrix}$$

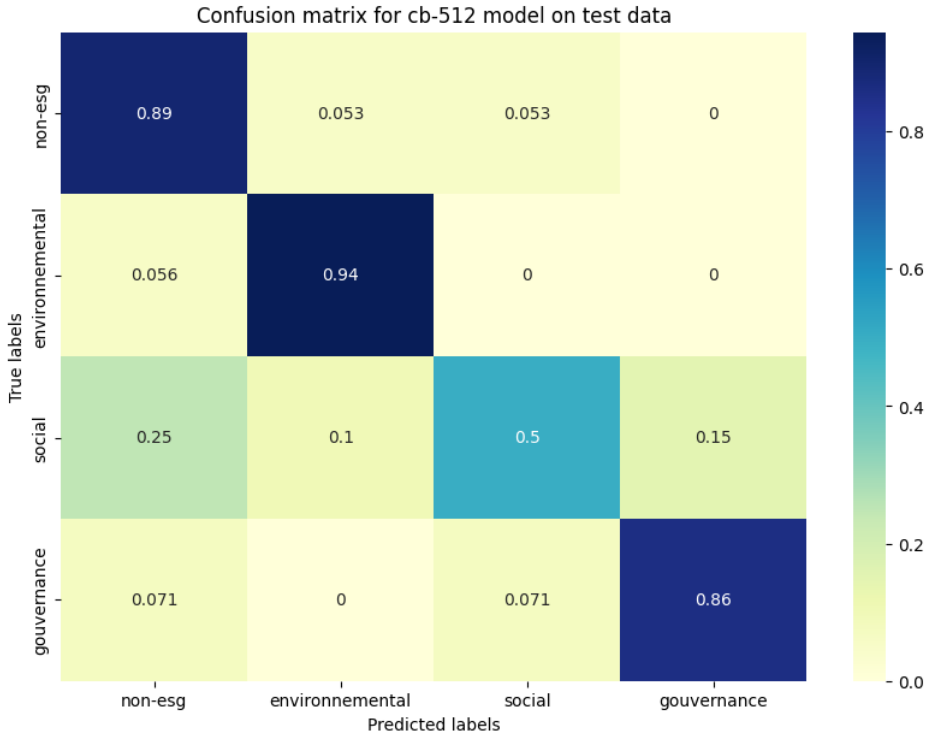
   where c_j represents each class (*non-esg*,*Env*,*Soc*,*Gov*)

3. we perform column-wise normalisation as we adjust the values, so that each category is comparable across different models, and we obtain the final weights matrix $\mathbf{W}$.

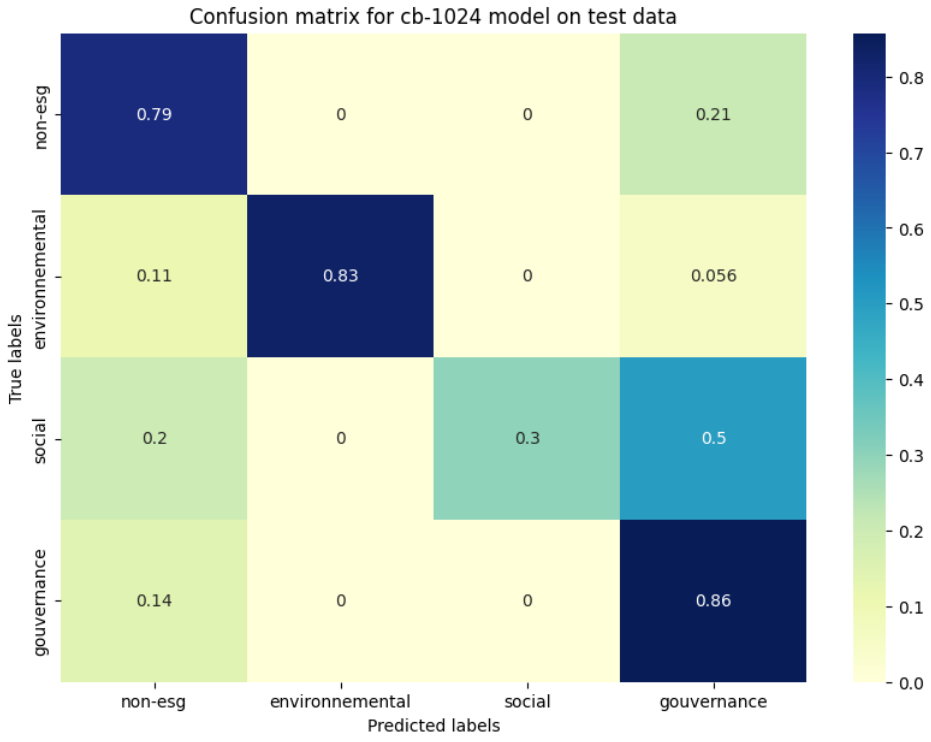
4. To select the label, we initialise a dictionary to accumulate the weighted scores for each class

5. For each model prediction, the corresponding weight from $\mathbf{W}$ is added to the score of the predicted class. The class with the highest accumulated score in the dictionary is selected as the final prediction

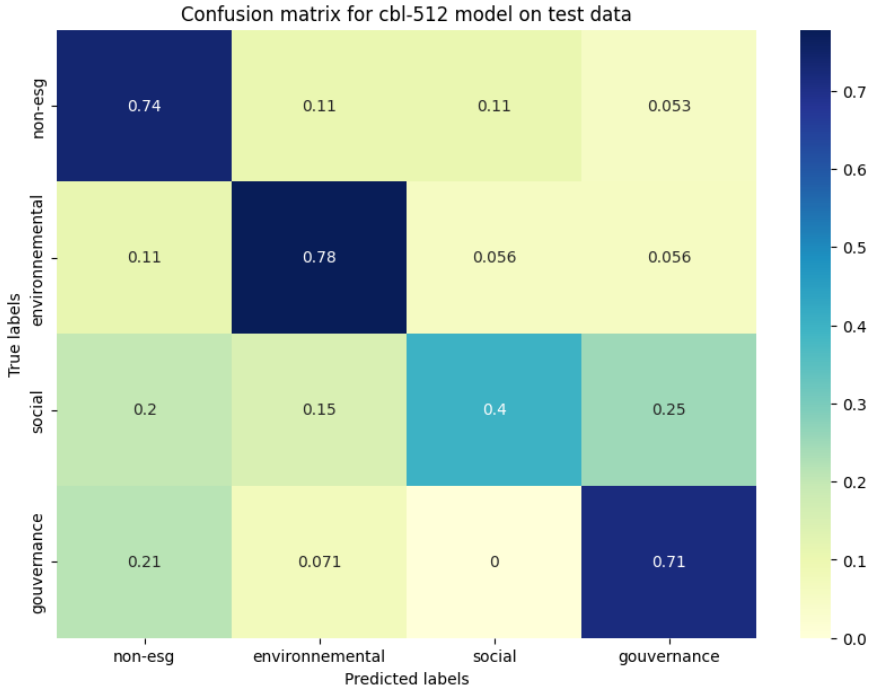
We obtain the following confusion matrices for each fine-tuned model on the smaller manually annotated dataset :



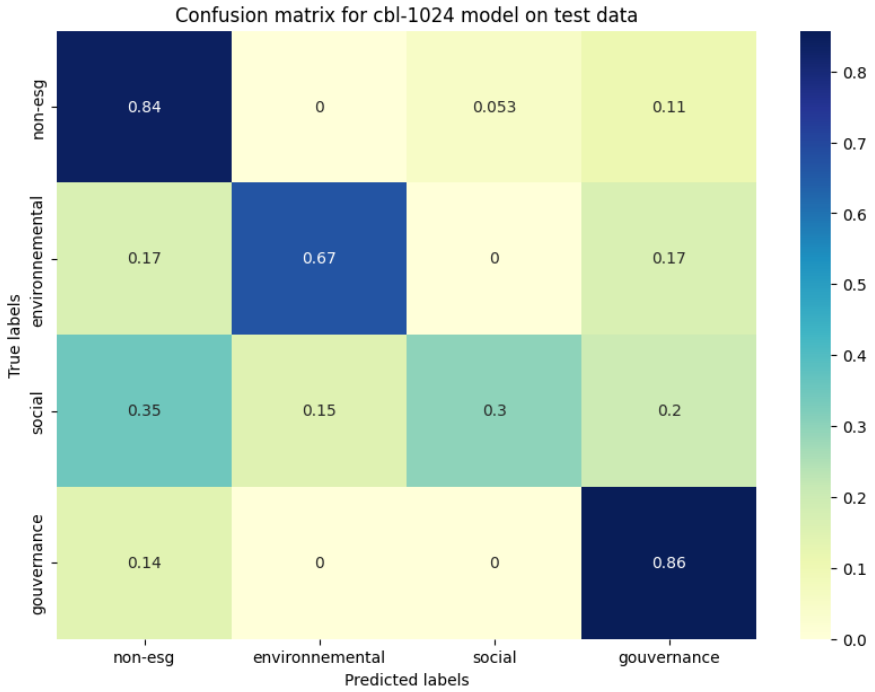
(A) CamemBert-Base-512 (CB-512)



(B) CamemBert-Base-1024 (CB-1024)



(C) CamemBert-Large-512 (CBL-512)



(D) CamemBert-Large-1024 (CBL-1024)

FIGURE 2.6: Confusion Matrices for each fine-tuned model

Overall, in the confusion matrices in the figures 2.6, we can observe that for the *non-ESG*, Env and Gov labels, the models overall perform quite well. However, for the social category, they struggle to differentiate with the *non-ESG* label. With the implementation of the above

weighting scheme, we obtain the following matrix:

$$\begin{bmatrix} 0.2636 & 0.2436 & 0.289 & 0.246 \\ 0.2344 & 0.2593 & 0.2350 & 0.2345 \\ 0.2564 & 0.2593 & 0.2485 & 0.3139 \\ 0.2459 & 0.2377 & 0.2270 & 0.2052 \end{bmatrix}$$

And we obtain the following confusion matrix:

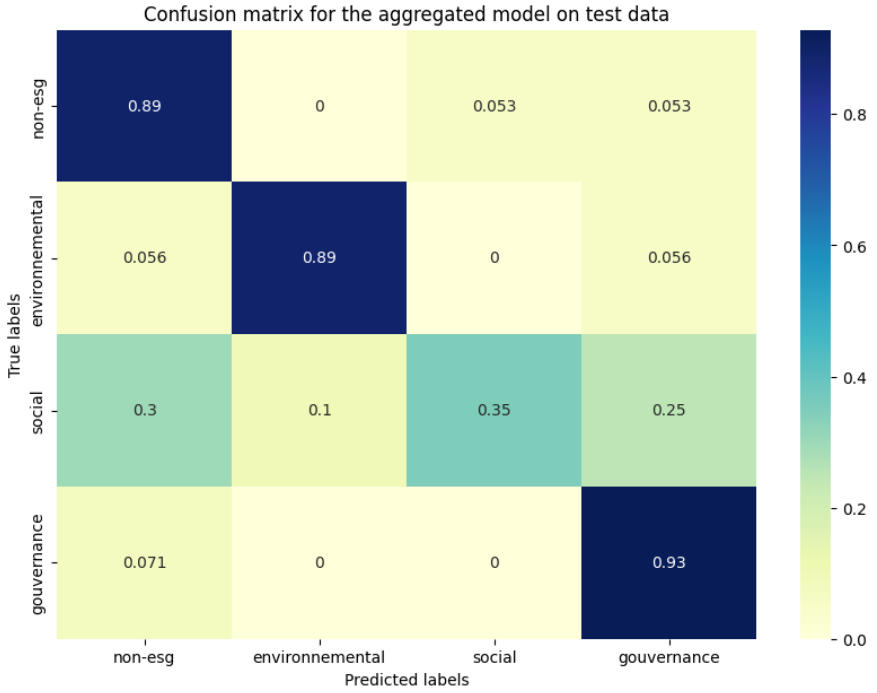


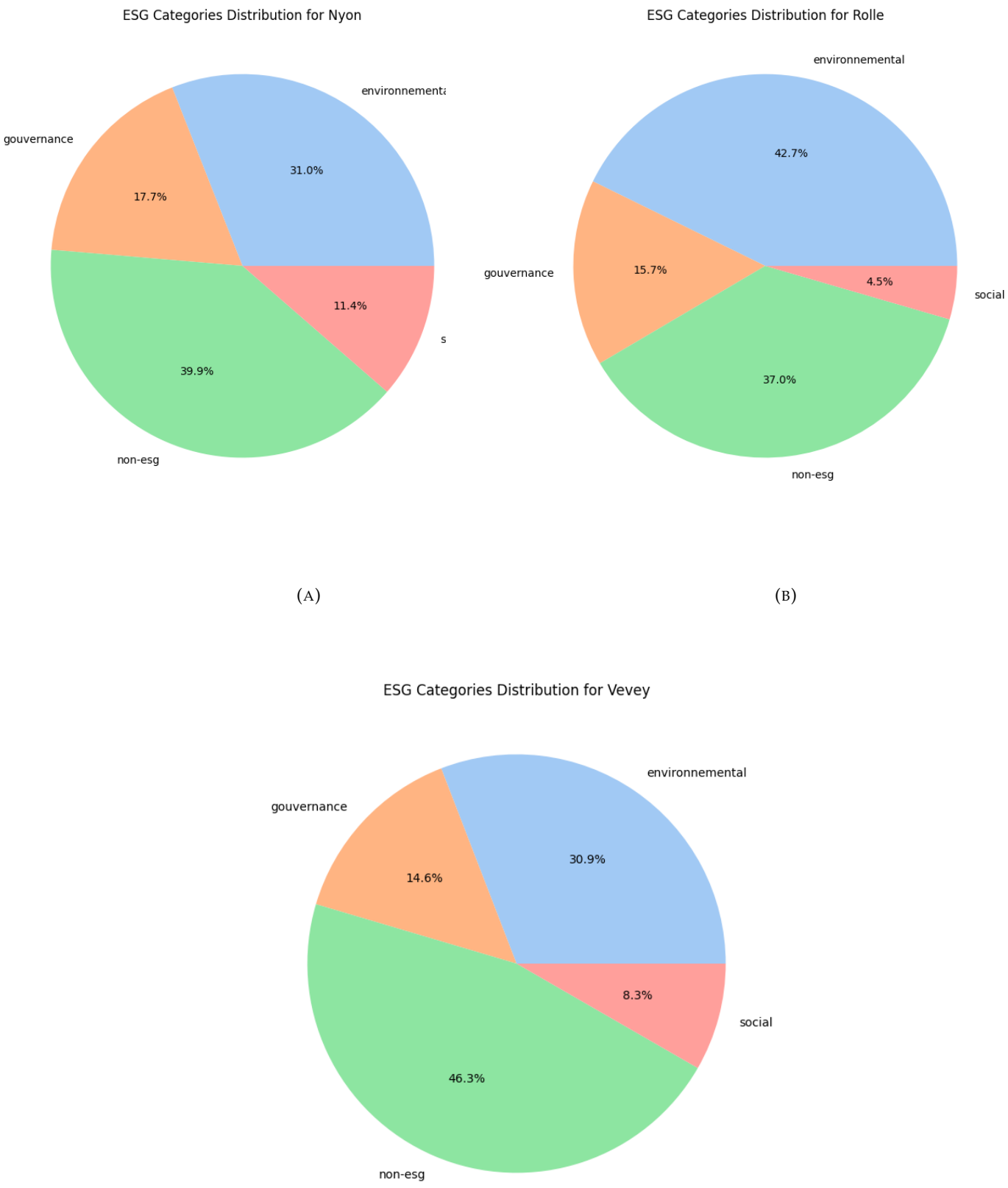
FIGURE 2.7: Confusion matrix with the aggregated model on the smaller, manually annotated dataset

The matrix values are quite close, but they indeed make a certain difference on the final results. Indeed, the accuracies for the *non-ESG* and *env.* labels remain close to 0.90, while the governance's category results increase to from 0.86 to 0.93. Although the accuracy for the social label slightly decreases, we will still use the aggregated model's prediction for classifying the transcripts, as it allows us to leverage the combined strengths in of the semantic interpretation for the four models.

2.6 Municipal Joint sessions' transcripts Classification

In this section, we report the results of the analysis of transcripts from the municipalities of Nyon, Rolle and Vevey; selected due to their geographical proximity. To capture emerging trends, we analysed the transcripts of 2022 and 2023. Our goal is to analyse and report any discernible patterns that could emerge from the dataset. Finally, we will exhibit a global overview of our classification results, followed by a detailed analysis and attempt to find any key pattern that could emerge.

We firstly obtain the following distribution of ESG categories over both years for each city:



(A) (B)

(C)

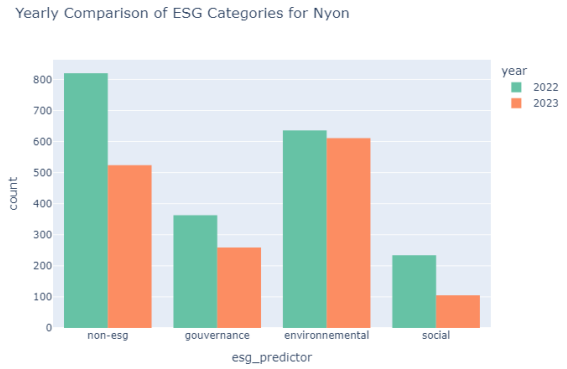
FIGURE 2.8: ESG categories distribution for (A) Nyon, (B) Rolle and (C) Vevey

We can first observe that a large proportion of the transcript is classified as *non-ESG*. This classification can be interpreted in two ways. On one hand, it might indicate that certain segments that could potentially belong to one of the three ESG categories were not classified as such. On the other hand, this can be seen positively because having false negatives is preferable to having false positives, as the latter could distort the final rating.

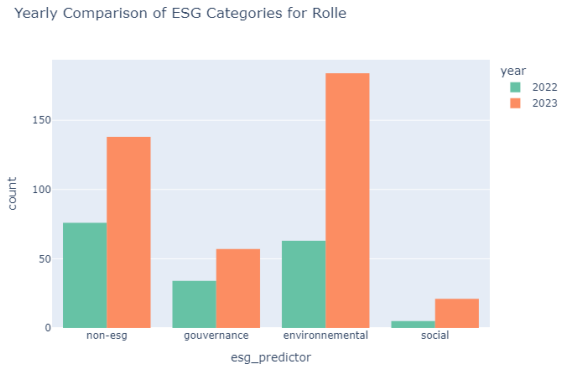
For clarity, a false positive is an incorrect classification where a text segment is mistakenly identified as belonging to one of the ESG categories when it does not, while a false negative would be a text that belongs to an ESG category but is not classified as such.

In the following histograms 2.9a, 2.9b and 2.9c, we can observe the yearly trends for each category. For the cities of Nyon and Vevey, no significant differences can be observed between the three categories, except that there are more discussions about environmental factors compared to the other two categories. Whereas, for the city of Rolle, there is limited talk regarding social factors, and a notable spike from 2022 to 2023 for the environmental category.

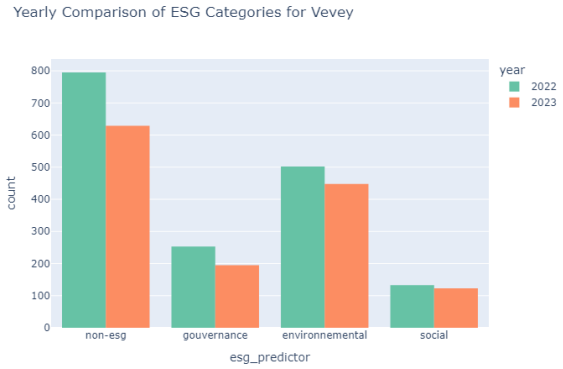
Overall, A possible reason for the social category being small could be that a certain amount of segments labeled as non-ESG should have been categorised under the social label.



(A) Nyon



(B) Rolle



(C) Vevey

FIGURE 2.9: Yearly comparison of ESG categories for Nyon, Rolle and Vevey

2.6.1 Word Clouds

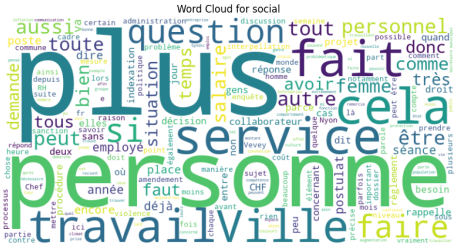
By observing the word clouds in the figure 2.10 below, we can further visually evaluate the classification results too. We take into account the data for all the cities together. Although the largest words in the images are recurrent french words that were not taken into account by the *Stopwords* from the `NLTK` library, a significant amount of words in the categories' respective word clouds indicate the correct classification. For example, in the wordcloud for the Governance category, abbreviated terms as specific as *COFIN*("Commission des Finances") and *COGES* (Commission des gestion) were picked up by the model, hence showing a certain efficacy in the semantic understanding of the sentences linked to the category. Similarly, for every category, numerous terms that represent well each category can be found in the word cloud for each category,

However, it is noticeable that the non-ESG category also includes words that could belong to the three main labels, suggesting that some sentences may have been mislabeled. However,

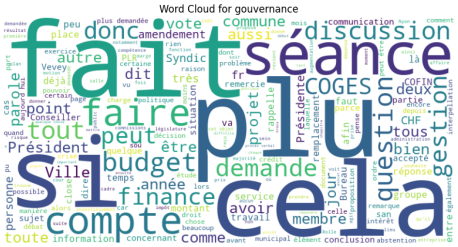
as previously mentioned, it is preferable to have sentences in the non-ESG category rather than misclassifying them under the three main labels.

(A) Non-ESG

(B) Environmental



(C) Social

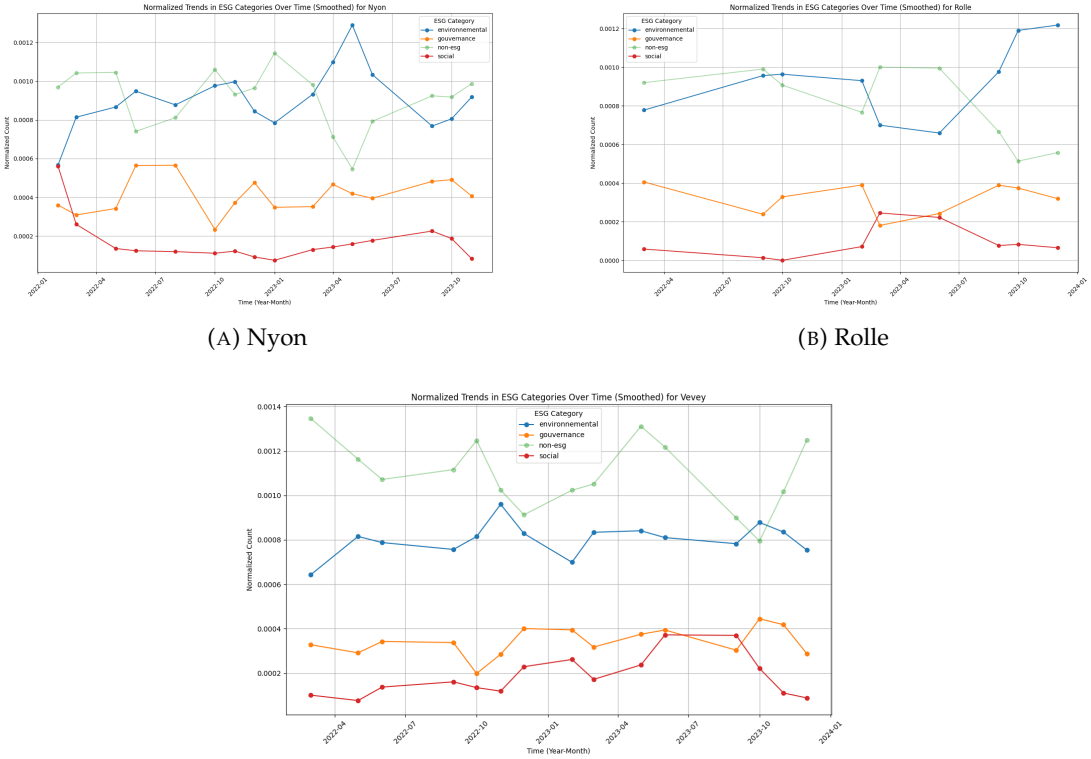


(D) Governance

FIGURE 2.10: Word clouds for all the classified transcript, encompassing all three cities.

2.6.2 Seasonal trends

Finally, to further analyse the classification results, we can also examine the trends in the frequency of each label in each transcript, over the two years. Given that some transcripts can be longer than others, we decide to normalise the frequency by the document length, thus making each trend from every city comparable.



(A) Nyon

(B) Rolle

(C) Vevey

FIGURE 2.11: Normalised trends in ESG categories for (A) Nyon, (B) Rolle and (C) Vevey.

Excluding the *non-ESG* factor, we can observe that environmental factors are the most frequently discussed category among the three cities. Following this, Governance and Social factors are nearly equal in prominence for Rolle and Vevey. However, in the city of Nyon, the Governance factor slightly surpasses the Social factor.

Based on the trends in figure 2.11 and the histograms in Figure 2.9, it is apparent that the social category is quite underrepresented in the meeting records. An explanation for this could be that the model is confusing the *non-ESG* and *social* labels. Additionally, upon examining the word clouds, we can notice that many words in the *non-ESG* word-cloud could

also belong to the governance or social categories. This overlap likely contributes to a certain number of mislabeling of text segments that discuss social risks/factors as *non-ESG*.

2.7 Conclusion

In this chapter, we outlined an approach to classifying texts based on the Environmental, Social, and Governance (ESG) criteria. Given the absence of French datasets specific to this domain, we created a new dataset by translating and extensively modifying an existing large corpus of article headlines from *The Guardian* to suit our needs. This process ensured a balanced and diverse dataset across all ESG categories.

Using this dataset, we fine-tuned CamemBERT base and large models available on the Huggingface platform. These models were trained to classify texts related to ESG, demonstrating notable improvements in classification performance, especially for texts linked to environmental factors. However, some challenges were observed, such as the model occasionally confusing the social and non-ESG categories.

Using a smaller manually annotated dataset, we were able to combine the models' prediction by implementing a weighted voting scheme. Using this framework, we classified transcripts of 2022 and 2023 from three nearby municipalities in Switzerland: Nyon, Rolle and Vevey. The transcripts were processed and divided into rows of texts as *.csv* files, and finally provided to the framework for classification. We were able to visually evaluate the classification through the word-clouds, before yearly comparing the frequencies and studying the trends of each category over the past two years.

Chapter 3

ESG Rating

PREVIOUSLY, we explored the methodologies and importance of classifying texts, specifically transcripts of public joint sessions, to identify content related to ESG domains. This classification was crucial as it allowed us to outline the discussions and themes that aligned with the ESG dimensions, thereby saving considerable time from manually reading every transcript and enabling a more focused analysis of government practices and policies.

We will hence develop an automatic rating system to these classified texts, through the application of sentiment analysis. Known as opinion mining, it involves determining the emotional tone behind a text or document. Models trained to perform this task are usually trained on datasets made from reviews available online or classified twitter posts with their respective sentiment. The dataset of reviews can be from product selling sites such as Amazon [Rez21], or movie reviews [PLV02]. As of recent, no dataset can be found to be specifically rating texts linked to ESG. Thus, in an attempt to rate the meeting records, we leverage multiple pre-trained sentiment analysis models to assess the contents' sentiments within the transcripts. In order to do this, we will employ and combine the results from three different models:

- ***BERT-base-uncased-sentiment*** [NLP23]: This multilingual model is one of the best-performing multilingual models available today, and trained on product reviews in six languages, including French. The table 3.1 presents the accuracies obtained by the authors for each language.

TABLE 3.1: Training data size and accuracies obtained by the authors, on 5000 separate product reviews.

Language	Training data size (# product reviews)	Accuracy (exact)	Accuracy (off by 1)
English	150*K*	67%	95%
Dutch	80*K*	57%	93%
French	**140K**	**59**%	**94**%
German	137*K*	61%	94%
Italian	72*K*	59%	95%
Spanish	52*K*	58%	95%

Accuracy (exact) refers to the percentage of correct predictions, while *Accuracy (off-by-one)* denotes the percentage of reviews where the model's prediction differed by only one star. We selected this model due to the large training dataset and the high accuracy obtained for the French language. Finally, it outputs a rating classification between 1 and 5.

- ***distilCamemBERT-base-sentiment*** [cma23]; [DA22]: This model was fine-tuned using two extensive French datasets: *Amazon reviews* [Keu+20] and *Allociné* [Bla20]. According to the authors, leveraging these large datasets would help minimize the bias in the model. Due to the large size of the datasets combined, the pre-trained model *DistilCamemBERT* model was used to fine-tune upon as it retains a certain level of performance but is as computationally efficient. The authors report the following evaluation results globally and for each class:

TABLE 3.2: Performance metrics for the distilcamembert-base-sentiment model.[DA22]

class	exact accuracy (%)	top-2 acc (%)	support
global	61.01	88.80	9,698
1 star	87.21	77.17	1,905
2 stars	79.19	84.75	1,935
3 stars	77.85	78.98	1,974
4 stars	78.61	90.22	1,952
5 stars	85.96	82.92	1,932

When examining class-specific accuracies, the model achieves decent results, with an overall improvement in the percentage of *top-2 accuracies*. The higher accuracies for

the 1-star and 5-star classes suggest that the model is particularly effective at identifying strong positive and negative sentiments. However, it may struggle with texts that are neutral or ambiguous. Despite the modest global accuracy, the high top-2 accuracy (also known as off-by-one accuracy) and the analysis of class-specific accuracies indicate that this model, along with the other two models we have chosen, can significantly aid our task.

Additionally, as mentioned the model's training on the distilled version of CamemBERT (which reduces the number of layers in the model's transformer architecture by half) results in faster prediction speeds. Despite this reduction, the model remains just as effective, achieving an accuracy of 97.57% compared to the model fine-tuned on the original CamemBERT, which achieves 95.74%[DA22].

- ***Finance-sentiment-fr-base*** [Bar23]:This final model was fine-tuned on the CamemBERT-base model with the translated version of the *Financial Phrase Bank* dataset [Mal+14]. We chose this model as it was specifically trained for the financial context, making it particularly suitable for our needs. Many sections of the transcripts discuss the finances, such as budgets, tax rates, etc... of their respective municipalities. Therefore, this model could enhance the accuracy of rating predictions for these sections. According to the authors, the model achieves an accuracy of 0.971. The model outputs three labels: negative, neutral and positive.

Given that the first two models provide a rating out of 5 and the third model classifies the data into three categories (positive, neutral, and negative), we standardise the output of the third model by mapping each label to a corresponding numerical rating as follows:

Negative: 1 **Neutral:** 3 **Positive:** 5

We then aggregate the predictions of the three models by averaging their outputs:

$$\text{agg\_sentiment} = \frac{\sum_{i=1}^{N} M_i}{N} \tag{3.1}$$

where M_i is the i^{th} model's rating output.

3.1 Sentiment Analysis of Municipal Joint Sessions Transcripts

Using the same classified transcripts from the previous chapter, we input them into the rating's sentiment analysis models to derive sentiment scores, which are then visualised and analysed for a detailed understanding.

3.1.1 Rating distribution over the years for each city

In the overall distribution of ratings in the figure 3.1, regardless of the labels, we can observe a broader spread in the ratings for the city of Nyon and Vevey. In contrast, the ratings for the city of Rolle exhibit a narrower spread, which can be attributed to the smaller number of transcript rows available for Rolle [see figure 3.2].

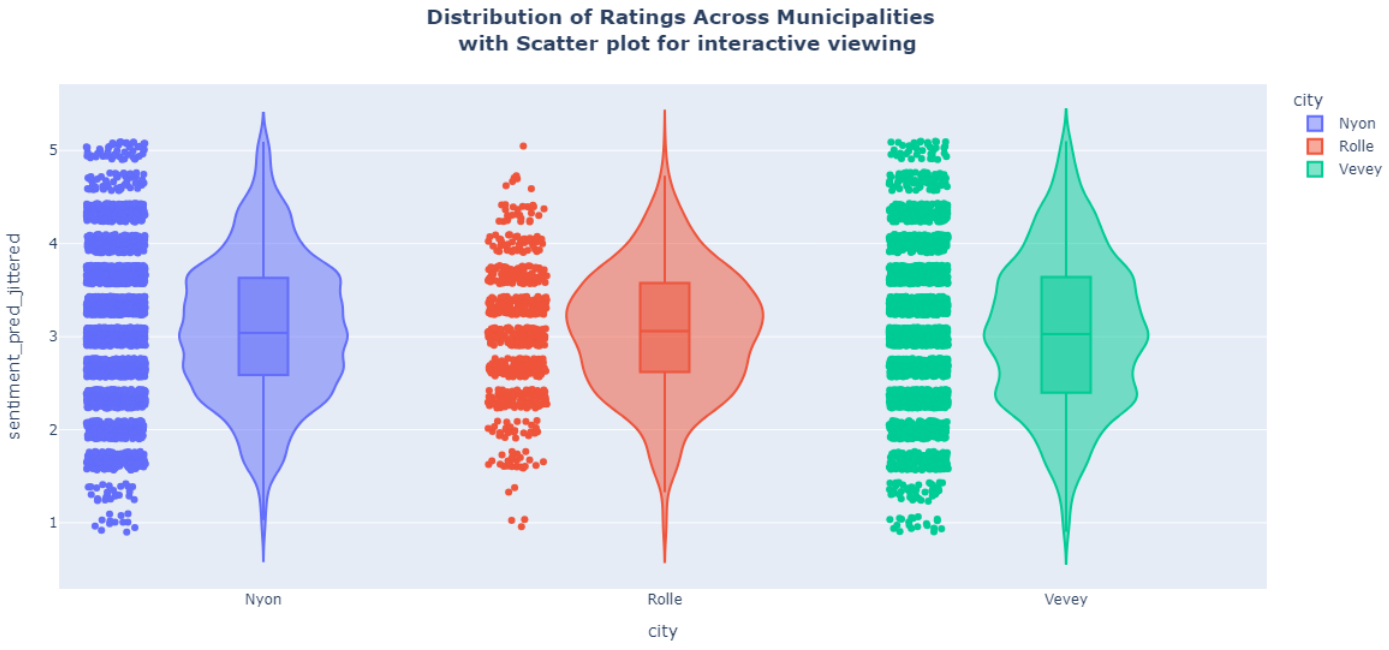


FIGURE 3.1: Distribution of Ratings Across Municipalities with Scatter plot for interactive viewing

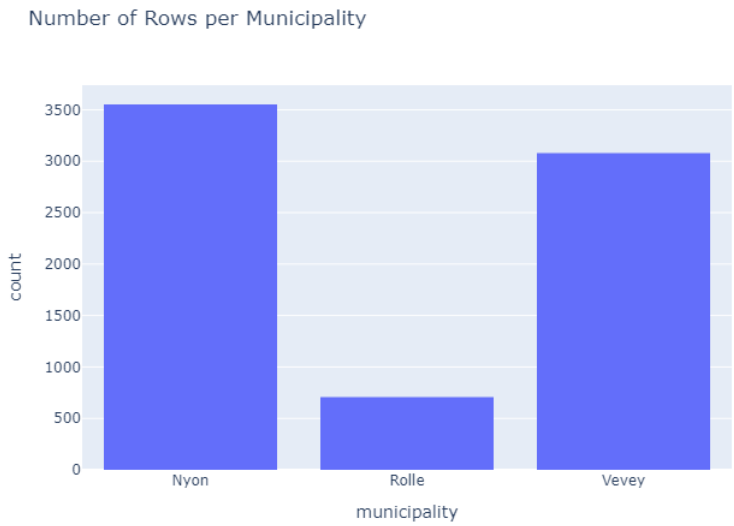


FIGURE 3.2: number of rows per municipality in figure

However, we observe that the ratings roughly follow a Gaussian distribution, with the median corresponding to the neutral rating of 3. This can be explained by the sentences in the transcriptions tend to keep a neutral speech.

Additionally, we included box plots in the visualisation to show the median and the different quartiles. Given the interactivity of the plot, hovering over the scatter plot to view a text segment and its rating helped us understand its position relative to the overall distribution of ratings and the data's specific quartile values (see fig 3.3).

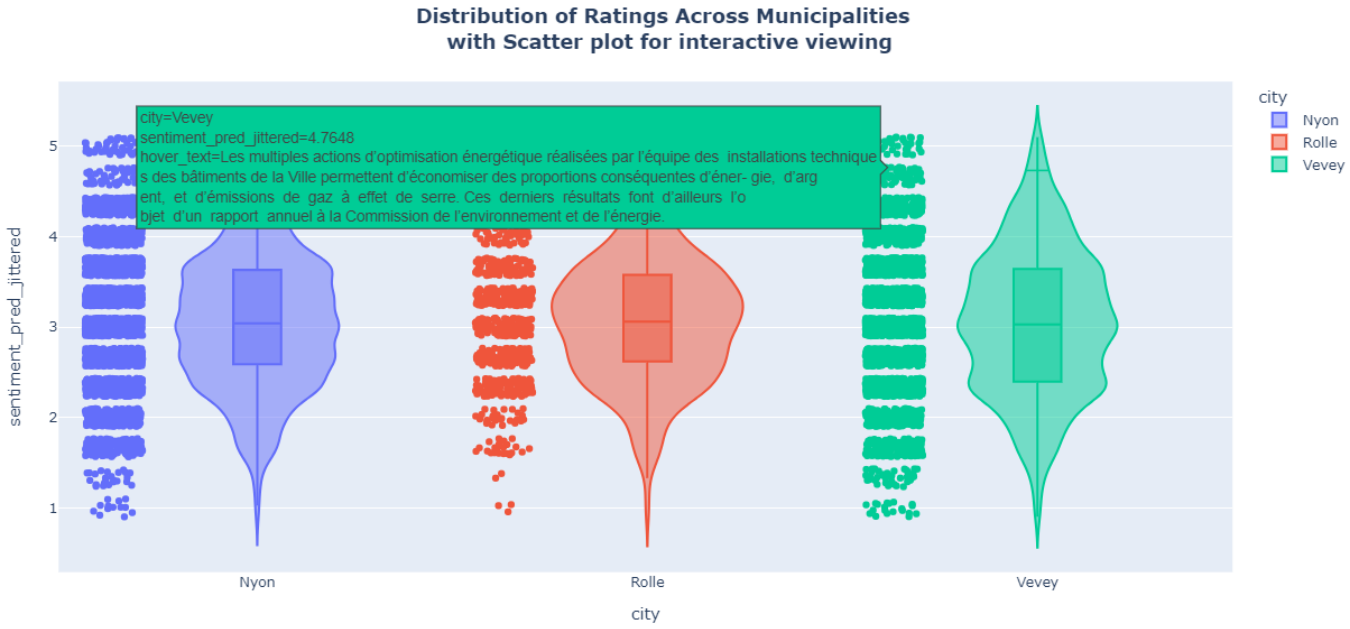


FIGURE 3.3: Interactive view of a text segment's rating in the Distribution of Ratings Across the Municipalities. A slight jitter was added to the scatter plot to ease the view of different samples

3.1.2 Rating distribution category wise

For each year, we get the following box-plot distributions in figures 3.4 and 3.5. Overall, as previously mentioned, the ratings for each category are centered around the median value of 3. For the environmental category, we can also observe that the rating slightly decreases for the three cities between 2022 and 2023, from $\sim$ 3.4 to 3.0.

Finally, we observe a variation in the third quartile among different cities. An anomaly is observed in the third quartile for the social category in the city of Rolle for 2022. This exception can be attributed to the very low number of data points for that specific category.

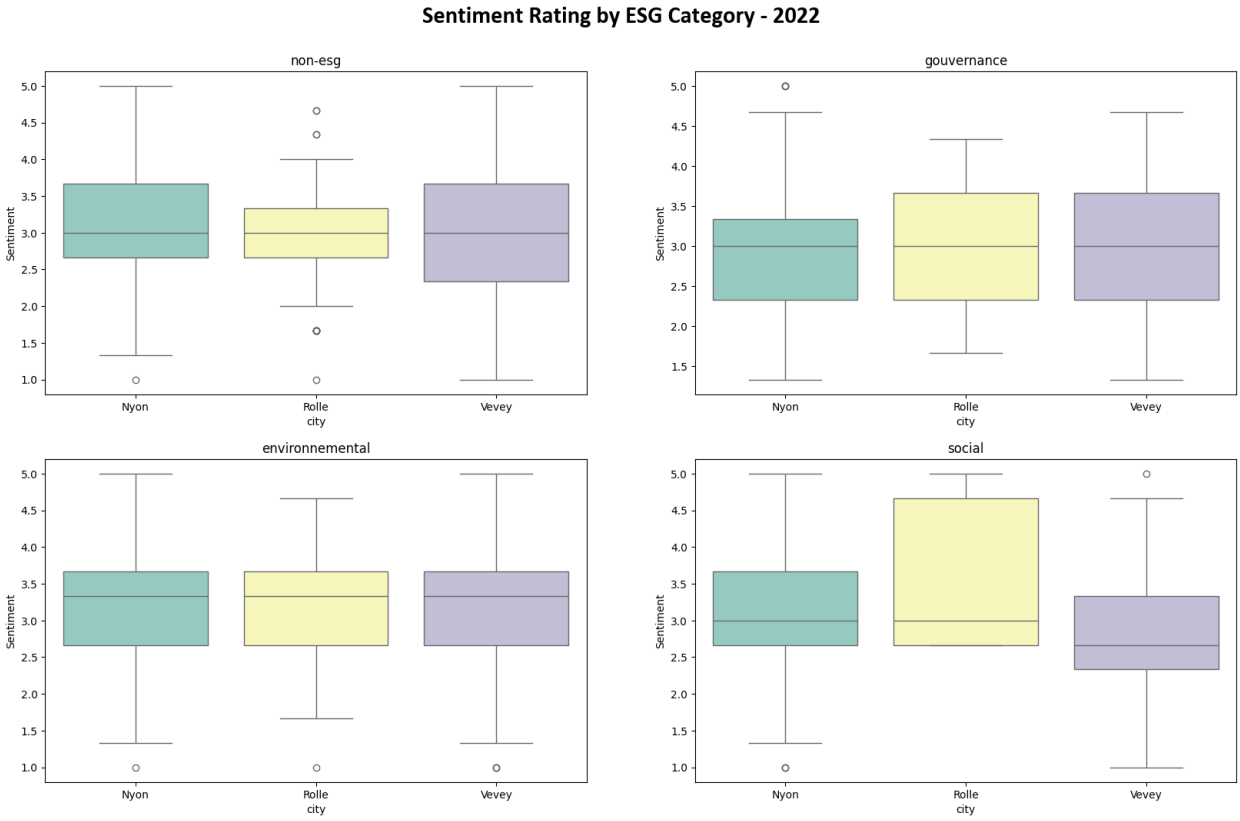


FIGURE 3.4: **Sentiment Rating by ESG Category and City for 2022**. The colored boxes represent the interquartile range (IQR), which contains the middle 50% of the data. The edges of each box indicate the 1^{st} (Q1) and 3^{rd} (Q3) quartiles. The line inside each box represents the median sentiment rating (Q2). The vertical lines (whiskers) extend to the minimum and maximum values within 1.5 times the *IQR* from the quartiles, with the dots indicating the outliers. Except for the city of Rolle, most of the data ranges from strongly negative to strongly positive sentiments. The smaller range of sentiment ratings for Rolle may be due to the lower number of text segments available, which also results in the presence of outliers.

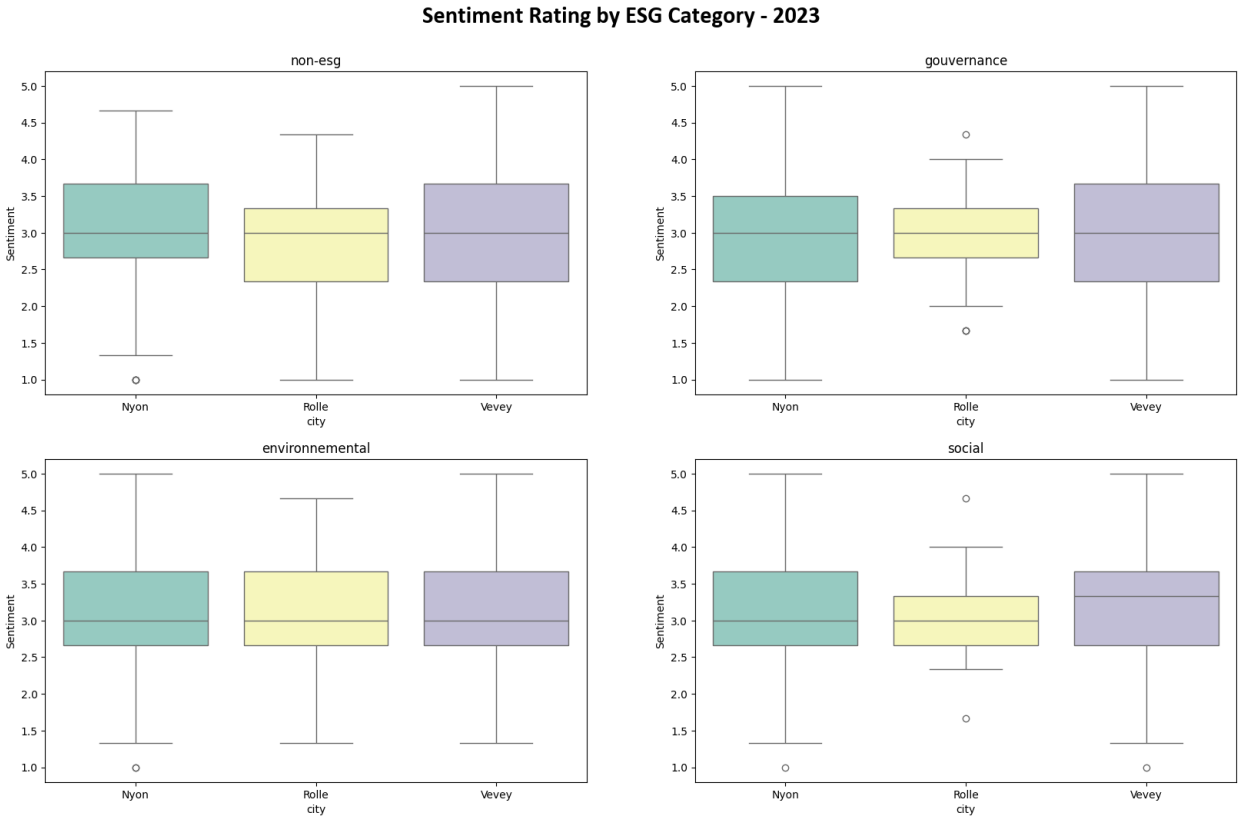


FIGURE 3.5: 2023 Sentiment rating w.r.t the city - Category wise

3.1.3 Interactive Data visualiser

To enhance the visualization of our data, we utilized the Python libraries *Dash* and *Plotly* to develop a dynamic and interactive graph viewer. This tool enables us to examine the data points (text and their ratings) by category and period. For example, in figure 3.6, we chose to view the social data for May 2023. In this month, only the transcripts of Nyon and Vevey were available (Rolle had sessions only in June and not in May).

The figure 3.6 demonstrates this functionality. Similar to Figure 3.3, we can hover over the points in the scatter plot to view the texts. This feature is particularly useful when comparing trends in ratings between each city over two years and studying patterns observed over several months, as discussed in the next section.

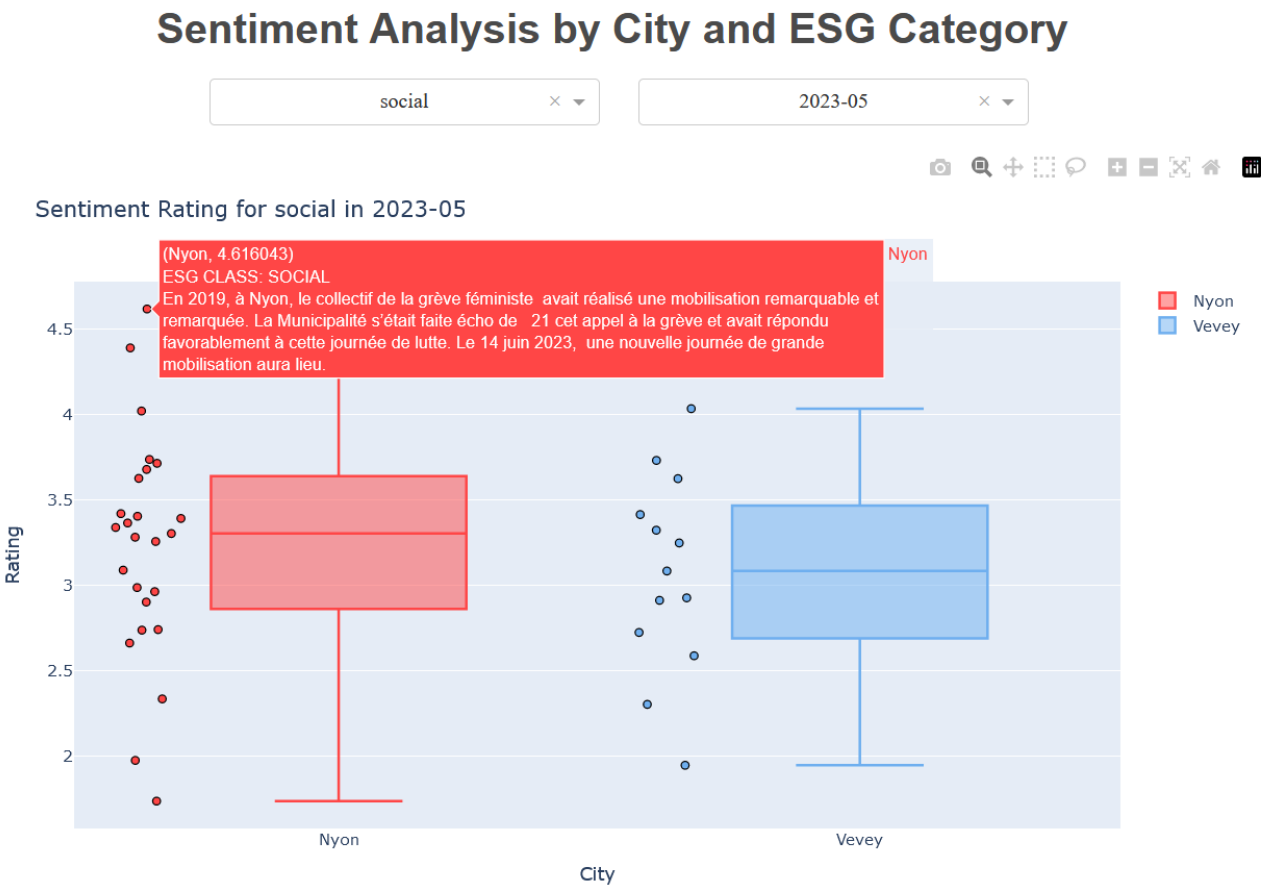


FIGURE 3.6: Interactive plot viewing in HTML with the python library *Dash*. Filtering by category and date available.

To further illustrate the findings, we provide examples of text segments that have been rated and labelled. For the cities of Nyon and Rolle, we include samples of both positively and negatively annotated excerpts from the transcripts.

TABLE 3.3: Positive and negative sentiment analysis examples for the city of Nyon, viewed using our interactive app to explore categorised text segments. Due to time constraints and a minimally functional web app, automatic retrieval of the source transcript for each text segment is not currently available. While manual retrieval is possible, automating this process would save time and provide better context for each classified sentence. Additionally, because the text was automatically partitioned from a PDF file, some segments may be cut off mid-sentence or not sectioned properly, although this issue was addressed in the script along with many other potential cases.

	Positive appreciation	**Negative appreciation**
Environment	La première est de travailler sur les énergies renouvelables, comme le Thermorés Ô (si voté plus tard dans la soirée) qui permettra d'avoir une indépendance énergétique beaucoup plus grande et être moins soumis aux variations du marché pour l'énergie, ou encore Novosolis qui installe sur les toitures des panneaux photovoltaïques ou encore grâce à l'augmentation de la nouvelle taxe sur l'efficacité énergétique, dont une partie permettra de financer de manière plus importante la pose de panneaux photovoltaïques sur les toitures privées. RATING: 4.7	D'autre part, puisque ça ne dure pas aussi longtemps que prévu, il ne comprend pas pourquoi on prend de l'argent dans le fonds pour le développement durable. Il n'arrive pas à comprendre ce qui est durable dans ces installations qui sont aussi appelées temporaires. Il n'est pas satisfait de la prestation de cette entreprise et est encore moins d'accord de reprendre les mêmes qui fournissent un travail insatisfaisant sous prétexte qu'ils sont spécialistes. RATING: 1.5
Social	A terme, la Ville de Nyon souhaite être un employeur responsable attractif, respectueux du droit du travail et du droit des personnes. Ils souhaitent promouvoir une politique de la bienveillance 21 et du bien-être au travail, qui permette de faire évoluer la culture organisationnelle vers un environnement plus positif qui valorise les personnes et les compétences au Service d'une Ville inclusive. RATING: 4.8	Il ne va pas revenir sur l'épisode des horaires d'une heure de plus le samedi, accepté par le Conseil, suite au constat qu'il y avait une rupture de dialogue entre les commerçants et le syndicat. Le référendum a été gagné par les représentants du personnel de vente et par les syndicats. A la fin, c'était une perte pour tout le monde puisque la CCT a été dénoncée. Le constat est un échec à tous les niveaux et une perte pour tout le monde. RATING: 1.4
Gouvernance	Hublot est une entreprise nyonnaise et nous 20 sommes fiers qu'elle soit établie dans notre ville. La marque Hublot connaît une visibilité mondiale, grâce à une activité marketing très bien faite. Elle sponsorise, ou est présente, dans de nombreux événements. Une visibilité mondiale donc, mais pas forcément régionale... De ce point de vue, il lui a été suggéré une idée qu'il soumet. RATING: 4.7	Cette exigence de la Municipalité de ne donner l'accès aux documents qu'à une partie restreinte de la COGES est non seulement une ingérence dans l'organisation de la commission, mais a aussi dégradé la capacité de travail et d'analyse de la COGES. La COGES, dans son ensemble, aurait pu mieux se partager la lecture des documents, sachant qu'il s'agissait de 400 pages à lire en trois soirées. RATNG: 1.6

TABLE 3.4: Examples for the city of Rolle

	Positive appreciation	**Negative appreciation**
Environment	Président fait voter l'entrée en matière de cette motion modifiée qui est acceptée à l'unanimité. 7. Motion Founou & Consorts – « Pour plus de jardins potagers et écologiques». M. Founou évoque l'esprit de la motion qui est de vraiment encourager la Municipalité à avoir une réflexion sur la thématique des jardins potagers les plus écologiques possibles. RATING: 4.0	M. Hay se rallie au sous-amendement de la Municipalité mais s'insurge contre le fait que soit qu'en décembre 2023 que l'on découvre que le bâtiment de la commune ne dispose que de vitrages simples a\i rez-de-chaussée. Pour une Cité de l'Energie avec des employés dont le travail porte sur l'étude de l'optimisation de cette énergie, avoir dans le bâtiment communal des vitrages aussi mauvais estvraiment décevant; RATING: 1.5
Social	[...] Municipal conjointe à son activité professionnelle n'est plus compatible avec son état de santé actuel. Cette décision est un crève-cœur pour lui qui a à cœur de travailler pour le bien de sa ville. Il remercie chacun de son engagement pour Rolle, ses collègues avec lesquels il a eu beaucoup de plaisir à partager le début de législature en tant que vraie équipe qui a établi un magnifique programme de législature et il espère que cette dynamique va perdurer pour le bien de Rolle. RATING: 4.5	Safi trouve l'idée plutôt intéressante mais ne voit pas l'intérêt du partenariat avec ENJEU. Elle sait qu'il existe une bourse pour les places d'apprentissage ainsi qu'un site internet, et aimerait déjà savoir comment fonctionne cette bourse et si elle est efficace. Si ce n'est pas le cas, elle ne voit pas comment ça pourrait mieux fonctionner pour des jeunes en recherche de petits jobs. Vollenweider pense que l'idée du Mme postulat est plutôt de créer un site sur Rolle[...] RATING: 2.5
Gouvernance	[...] exceptionnel avec un bénéfice de plus de 4 mios; donc la valeur du point d'impôt 2020 va être prise en considération pour la participation aux investissements d'ENJEIJ en 2022,2023 et 2024. Ainsi qu'expliqué dans le préavis, l'impact cumulé va se situer autour de Fr. 900'000.-, peut-être un peu moins si d'autres communes veulent bien avoir des résultats aussi performants que ceux de Rolle. RATING:4.3	[...]. Elle (Mme. Beck) voit que pour les investissements, entretien du parc immobilier et autres, on fait le minimum. Elle est donc inquiète. M. Hay estime que Mme Beck a sur-interprété les propos de la Syndique qui a juste mentionné que dans d'autres communes le taux d'imposition est revu à la baisse, ce qui est d'ailleurs rapporté dans les journaux. Il n'a pas été question de baisser les impôts et il rappelle que ces dernières années il y a eu quelques votes sur la hausse d'impôt proposée et toujours refusé. RATNG: 1.7

TABLE 3.5: Examples for the city of Vevey

	Positive appreciation	**Negative appreciation**
Environment	Carruzzo Evéquoz indique que la priorité du groupe des Vert.e.s est de trouver un accord budgétaire pour que les services de la Ville, la Municipalité et tous les acteurs concernés puissent avancer efficacement dans les nombreux projets à venir, comme la réalisation des objectifs du Plan climat, l'accélération et la transfor- mation des mobilités, etc. Les projets qui nous attendent sont nombreux et urgents. RATING: 4.7	Concernant la proposition d'un repas carné par semaine, elle se dit extrêmement surprise de voir que les auteurs du postulat – sous le prétexte du climat – souhaitent tuer à petit feu nos agricultrices et agriculteurs. La demande sera si faible que nous ne pourrons plus nous fournir auprès de commerçants locaux proposant de la volaille de la région, des poissons du lac ou de la viande locale. RATING: 1.7
Social	Plus de 200 personnes montent ce festival. Pendant qu'elles travaillent de 3 à 6 mois, elles dorment, consomment, paient leurs impôts à la source à Vevey. Ce sont des dizaines de milliers de personnes qui viennent durant le festival, qui consomment aussi bien à Vevey que dans la région. Les directrices et directeurs des musées de notre ville confirment que, lors de cette manifestation, le nombre d'entrées dans ces institutions explose, y com- pris le Musée historique de Vevey, et c'est réjouissant. RATING: 5	Faire travailler les gens jusqu'à 22h00 pour que pendant les deux dernières heures il n'y ait personne dans les magasins et que la rentabilité soit nulle n'a pas de sens, tout le monde y perd. Le règlement est absolument obsolète, on va attaquer sa révision le plus tôt possible. Il encourage les gens à aller dans les magasins pendant les nocturnes et poser la question aux employés pour voir s'ils sont satisfaits de cette mesure. RATING: 1
Gouvernance	Le groupe PLR, dans sa majorité, est donc satisfait des résultats obtenus pour l'année 2022 dans sa globalité et invite le Conseil à accepter les comptes 2022. Merci à la Municipalité pour ses efforts, mais également à tous les employés pour leur travail, leur contribution tout au long de l'année et leur engagement, qui ne passent pas inaperçus. M. B. RATING:4.7	C'est vrai que Swissmedia Centre a été très rarement rentable, c'est une opération ratée, mais qui n'avait rien à voir avec le logement. La LPPPL prévoit des LUP, on ne peut donc pas faire jouer la LPPPL pour un bâtiment industriel. Ce sont trois mauvais exemples, mais on ne peut pas imaginer qu'une Municipalité fasse toujours tout juste. M. P. Bertschy revient sur la question de la servitude. RATNG: 1.7

3.2 Analysis and Interpretation

3.2.1 Rating Trend with respect to the time - Category wise

In order to analyse the results we have obtained thus far, We firstly present in the figure 3.7 the monthly average ratings for each city, categorized accordingly.

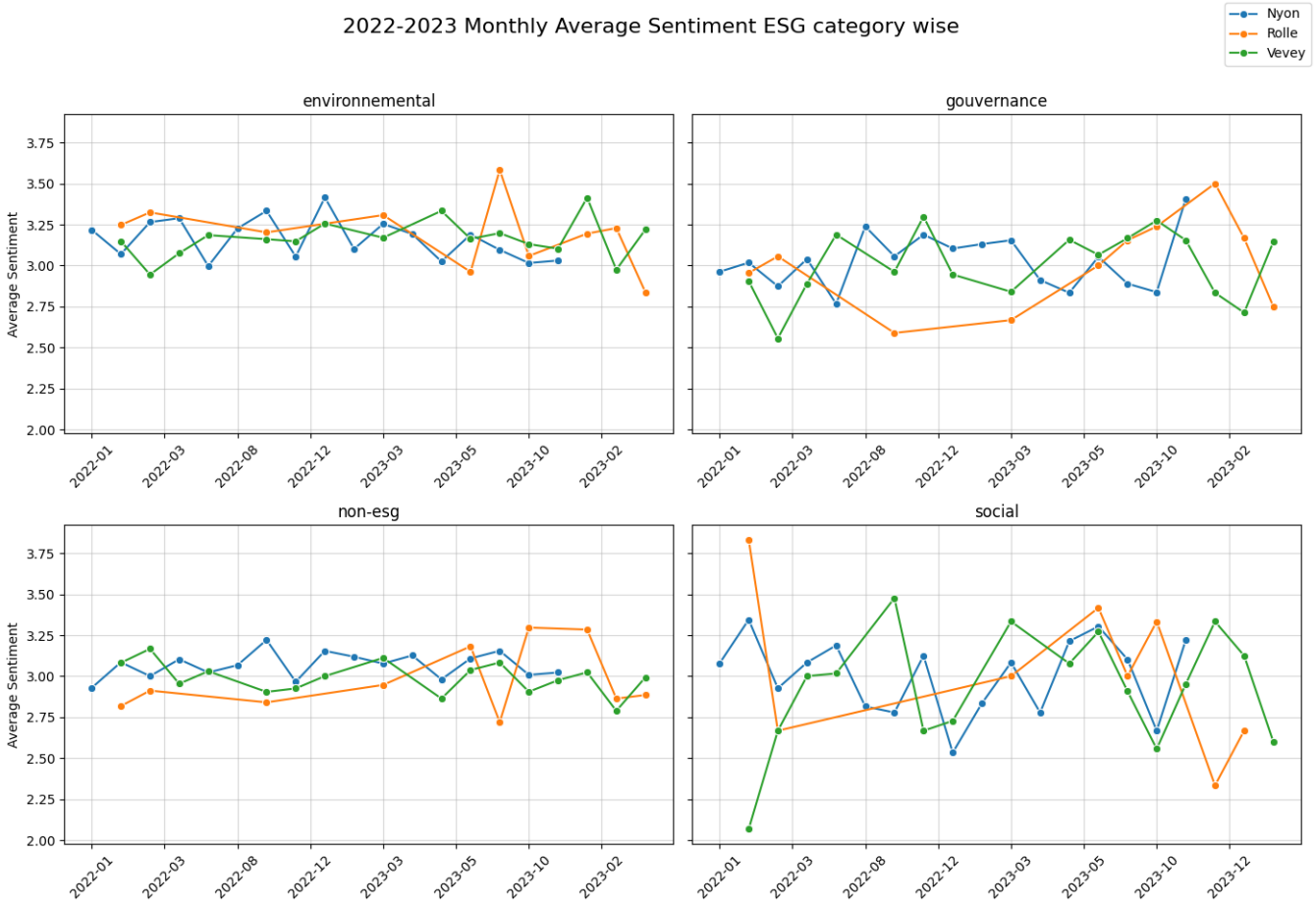


FIGURE 3.7: 2022-2023 average sentiment rating, ESG category wise

Similar to the box plots discussed earlier, the ratings over the two years for the cities of Nyon and Vevey remain quite stable, with some variations observed in the governance and social categories. In contrast, the city of Rolle exhibits more variability across all categories. This increased variability can also be attributed to the lower quantity of data points available for Rolle.

A first trend we can observe is between the period of August and december 2022, where the ratings for Rolle are lower by ~ 0.5 compared to the other two cities which have a neutral rating of ~ 3.0. To understand why it may be lower, we can firstly check from the interactive plot the data from that same period:

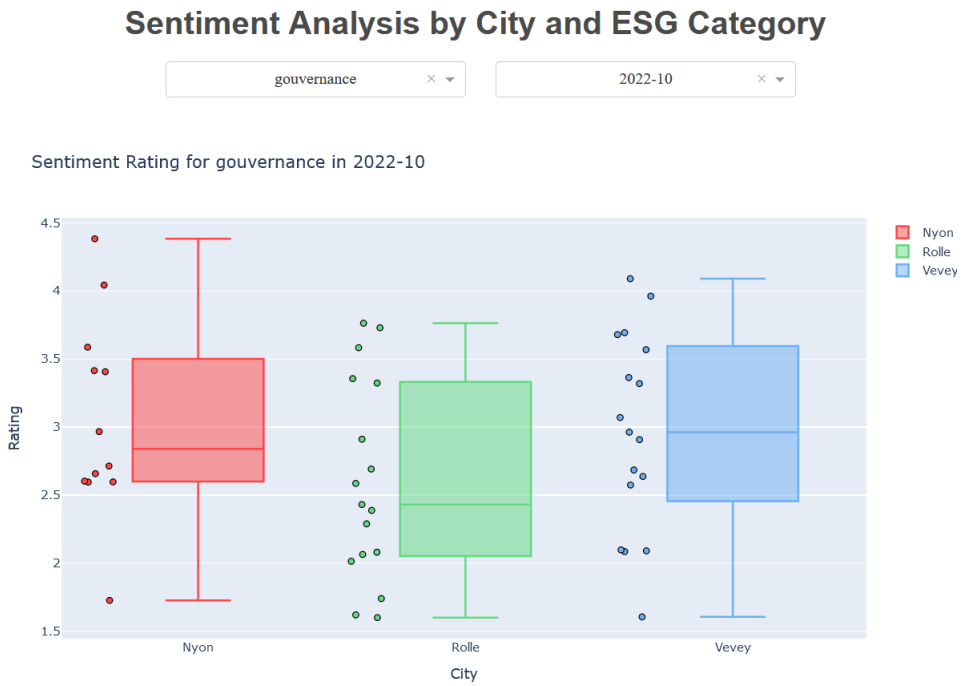


FIGURE 3.8: Ratings' scatter and box-plots for the month of October 2022.

By analysing the text segments for this period with our interactive plot viewer (figure 3.8), we observed that each city discussed tax rates extensively and also continually referred to the state commission of finances (COFIN). Table 3.6 presents specific excerpts from the filtered category.

For the period of October 2022, by viewing the texts classified as governance, it became apparent that each transcript consistently addressed tax rates during their sessions in certain segments, highlighting the challenges cities faced in reducing these tax rates. Although the city of Vevey would not decrease the tax rates, it exhibited a positive outlook by mentioning additional efforts to reduce certain costs (3.6). Consequently, the city of vevey had a better average rating in this period regarding the governance category.

TABLE 3.6: Sample sentences categorised as Governance from October 2022 transcripts

Nyon	**Rolle**	**Vevey**
[3.3 / 5] Un exemple a été donné à la COFIN lors de la séance de septembre qui traitait du taux d'imposition : ils ont emprunté fin 2021 à 0,3\% pour un montant de CHF 10 millions à 7 ans. En juin 2022, pour un montant de CHF 9 millions à 6 ans, le taux était passé à 1,95\%. La prévision de ces nouveaux taux et leurs répercussions financières ont été intégrées dans le budget 2023; la COFIN recevra tous les détails sur la manière dont ils ont calculé ces taux."	[2./ 5] La COFIN prend volontiers l'avis d'une boule de cristal... Mme Beck s'est abstenue lors du vote de la Cofin et ce qui l'inquiète est l'évocation de la Syndique quant à l'éventualité que le taux d'imposition pourrait possiblement être abaissé dans les temps à venir alors que ce qu'elle comprend de la situation actuelle est que Rolle est pénalisée, ...	[3.36/5] Rapport sur arrêté communal d'imposition pour l'année 2023 (2022/P23) Rapport : M. Martino Rizzello Mme S. Marques remarque qu'avec une inflation qui prend l'ascenseur, les prix des matières premières qui grimpent et une incertitude face à l'avenir, le groupe PLR apprécie qu'une hausse du taux d'imposition ne soit pas à envisager aux yeux de la Municipalité. Néanmoins, nous devrions faire mieux pour le pouvoir d'achat de nos concitoyens et concitoyennes.
		[3.67 / 5] Une légère baisse d'impôts aurait été la bienvenue pour alléger un peu les charges des contribuables. Cependant, le PLR se dit aussi conscient que l'administration communale fait beaucoup d'efforts pour diminuer les coûts et il l'en remercie. Dans un contexte économique des plus tendus, avec des projec- tions incertaines des coûts qui seront répercutés sur le contribuable et puisque la Municipalité n'envisage pas de baisse d'impôts, le PLR s'abstiendra sur ce vote.

3.2.2 Interquartile Range (IQR)

To further analyze the trends that can appear in our data, we will use the interquartile range to concentrate on identifying outliers, assessing variability, and comparing these patterns across different cities, time periods and the different ESG categories. We will focus on the Q10th and Q90th percentile data rather than the data close to the median, as this can help detect outliers, highlight unusual data points that might indicate significant events (very highly or poorly rated) or errors.

To further analyze the trends that can appear in our data, we will use the interquartile range (IQR) to concentrate on identifying outliers, assessing variability, and comparing these patterns across different cities, time periods, and different ESG categories. The IQR is a measure of statistical dispersion and is defined as the range between the 25th percentile (Q1) and the 75th percentile (Q3). It is often used to identify outliers and understand the spread of the central portion of the data. However, in our analysis, we focus on the Q10th and Q90th percentile data, which represent the 10th and 90th percentiles. By focusing on these percentiles rather than data close to the median, we can better detect outliers, highlight any unusual data points that might indicate significant events (very highly or poorly rated) or errors, and gain insights into the extremes of the data distribution.

To enhance our analysis, we will also calculate a modified threshold using the formula $Q90 - (3 - Q10)$. This adjustment allows us to more effectively capture significant deviations from typical patterns by focusing on the lower and upper extremes of the data, thus helping to identify any substantial anomalies or trends.

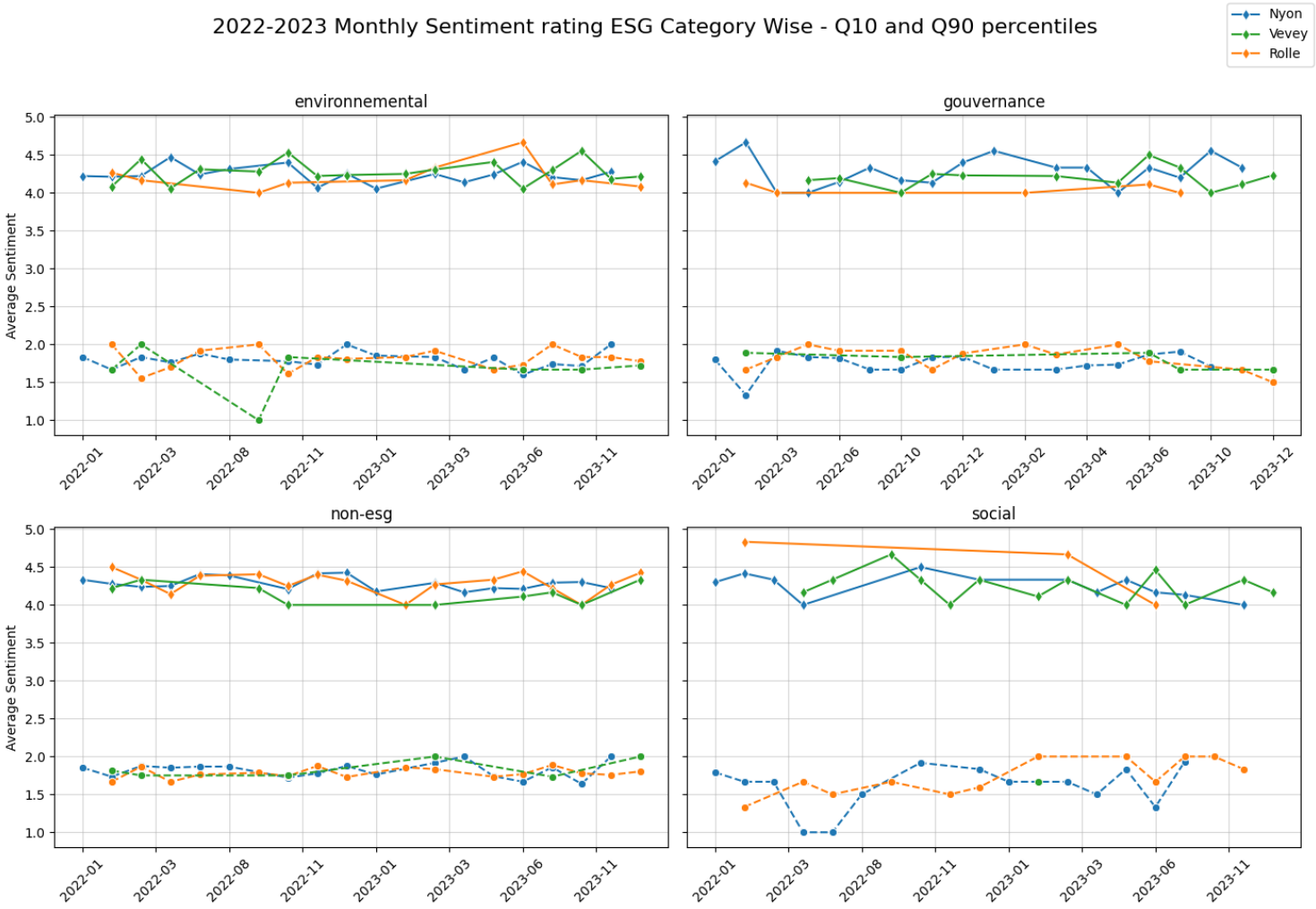


FIGURE 3.9: Rating for each ESG category over the two years period. We take the Q10th and Q90th percentile data.

Overall, as we can see in the above graphs in figure 3.9, our framework tends to avoid assigning extreme ratings, such as 1 or 5, to text excerpts. This likely stems from the use of multiple models to generate ratings, where the individual assessments may counterbalance one another. However, there are instances where the models show strong agreement, as observed in the case of the city of Nyon in the social category during the months of March to August 2022. Below, we provide examples of text passages from Nyon transcripts that received very low ratings (close to 1) and were classified under the social category:

- *Il espère que cette médiatisation de la maladie d'un collaborateur par un Conseiller communal, due probablement à la suggestion positive de la représentante de NRTV, ne nuise pas au retour de ce Chef de service, car cette annonce pourrait malheureusement être mal perçue par celui-ci,mettant ainsi la pression sur un employé qui, au cours de ces sept années passées dans sa*

première période, n'a manqué aucun jour de travail et qui, depuis son retour, n'avait là encore manqué aucun jour, y compris pendant la période de COVID.

- *Tout le monde a dit que GSK allait fermer, des milliers d'emplois allaient disparaître et que ce serait la catastrophe si on ne le fait pas. Mme la Présidente lui rappelle qu'elle s'était élevée en parlant du gaz de Poutine. M. le Municipal Pierre WAHLEN répond immédiatement en remerciant M. le Conseiller d'avoir envoyé l'interpellation assez tôt.*

- *Début 2022, l'ordonnance pénale rendue par le Ministère public ayant fait l'objet d'une opposition de la part du policier concerné, l'affaire a été portée devant le Tribunal d'arrondissement de La Côte. Sa Présidente a rendu, ce jour, son jugement et condamné le policier pour conduite d'un véhicule en état d'ébriété. Le jugement rendu ce matin témoigne d'actes et d'attitudes inappropriés, à fortiori pour un policier.*

Alternatively, we can examine the spider charts below in figures 3.10 and 3.11, which display the $Q10$ and $Q90$ percentile data across each classification category. This can provide a clearer comparison of the lower and upper extremes for each category. Additionally, viewing the three cities together can help us understand how the sentiment can vary between the cities for the same categories, thus highlighting the consistent and polarized areas.

For the Q10 percentile data in 3.10, the city of Rolle exhibits the most variation, with higher ratings in the social and non-ESG categories. As previously mentioned, this can be attributed to the lower number of text segments, which leads to higher skewness and a higher number of negative outliers. Nevertheless, it has better ratings for poorly rated texts compared to the cities of Nyon and Vevey. Despite the slight difference in variance for the social category, Nyon and Vevey exhibit stable sentiment trends, reflecting a more moderate and balanced discussions in these cities.

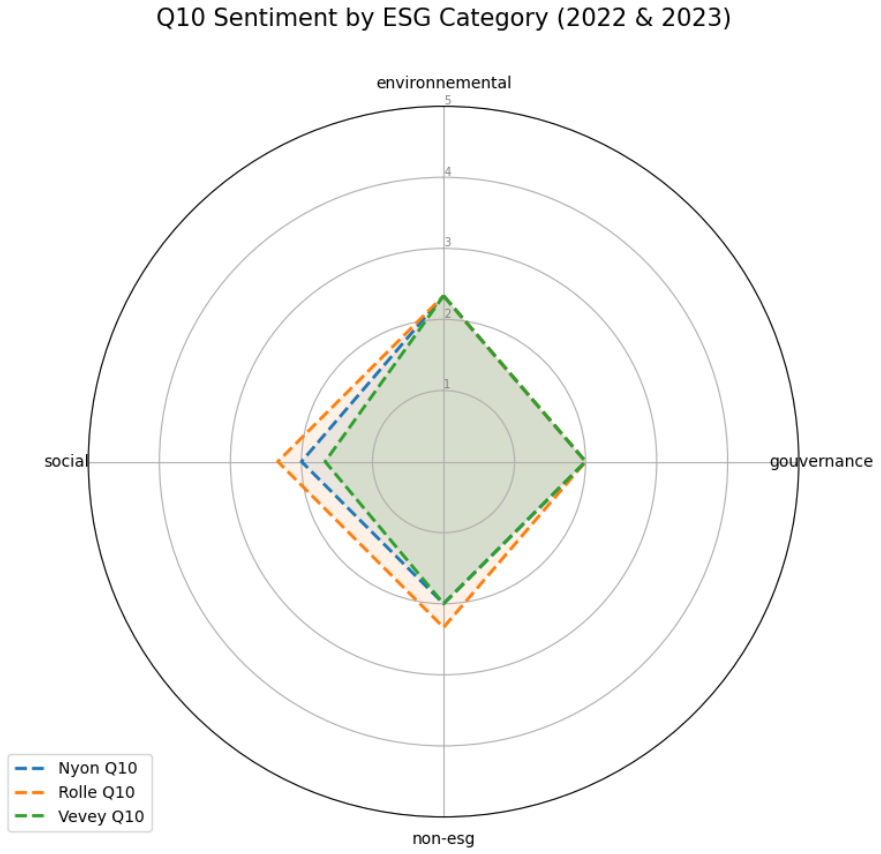


FIGURE 3.10: Q10 percentile data across different ESG categories for the three cities

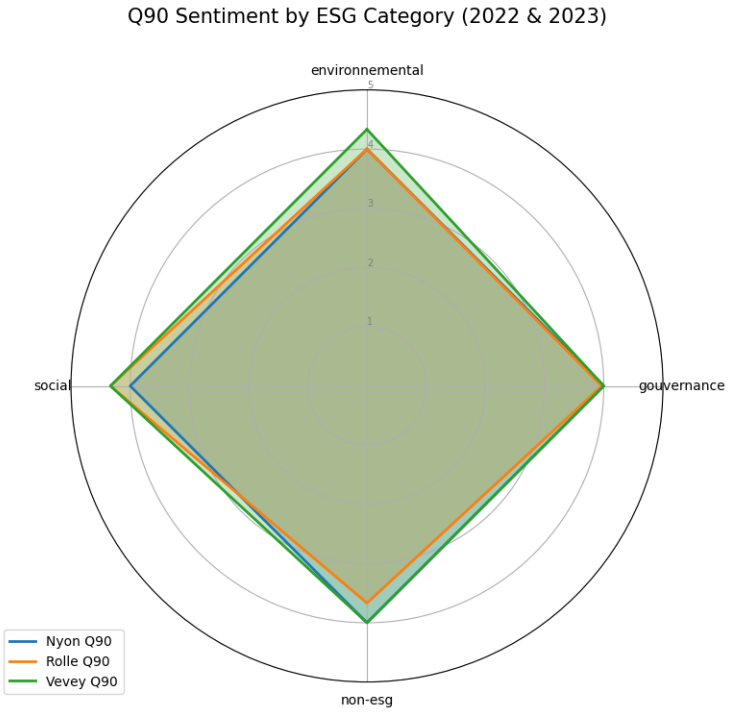


FIGURE 3.11: Q90 percentile data across different ESG categories for the three cities.

On the other hand, With the $Q90$ percentile data (figure 3.11, we can see the cities have a more consistent high sentiment ratings across all the categories. So although Rolle has more variability with lowly rated texts, it shares a comparable level of positive sentiment with the other cities. This duality suggests that Rolle experiences a wider range of sentiment, with both stronger negative and positive reactions, possibly due to the smaller dataset size, leading to higher skewness.

Furthermore, we also combine the two percentiles, $Q10$ and $Q90$, while also subtracting the $Q10$ percentile with the median, thus favouring the higher $Q90$ and the lower $Q10$ values. We get the following the following spyder chart with the three cities together

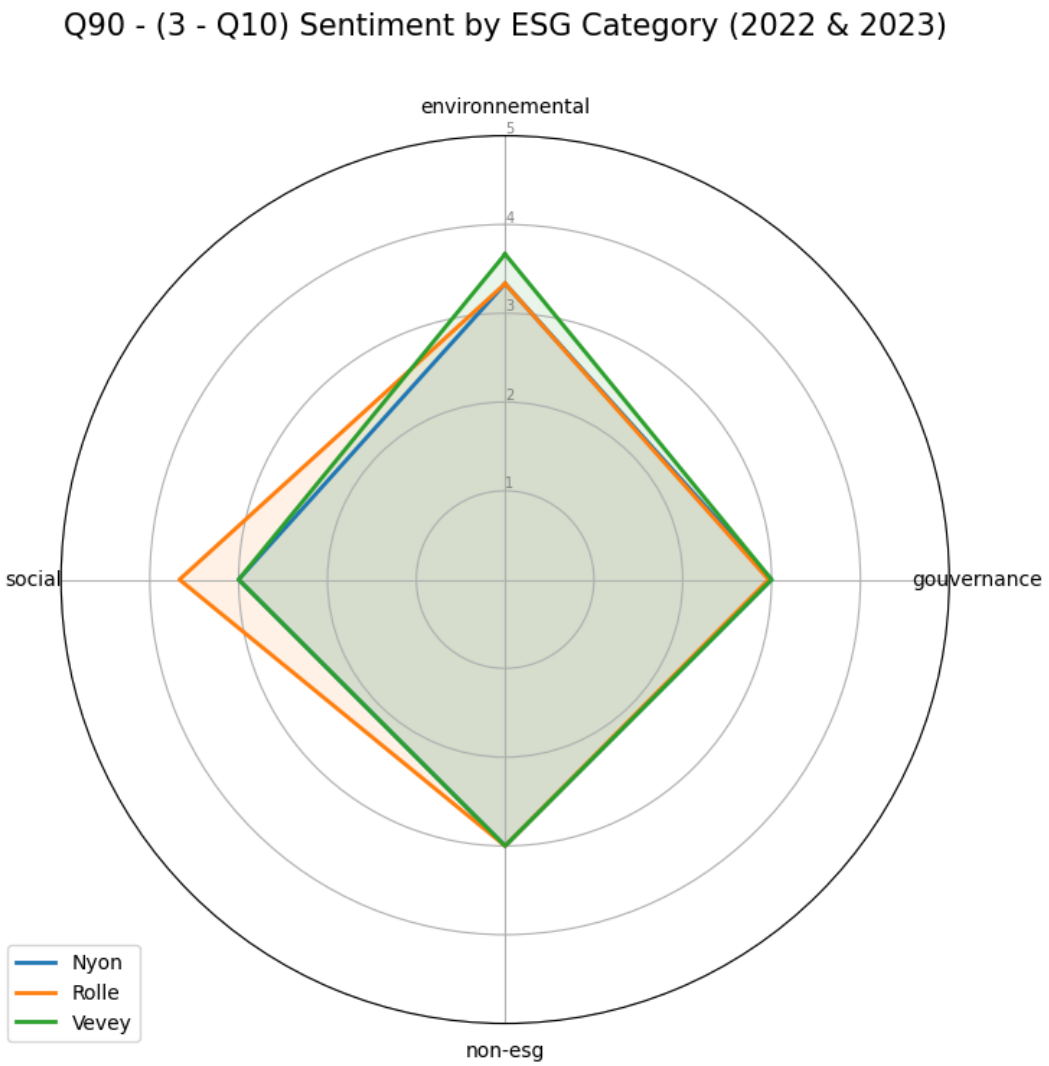


FIGURE 3.12: $Q90 - (3 - Q10)$ percentile data, for all three cities.

Building on the previous analysis, we can observe from the chart 3.12 that the skewness is

significantly higher for the city of Rolle, consistent with our earlier observations. In contrast, the cities of Nyon and Vevey display very similar patterns across all categories, with the exception of the environmental label, where the divergence is slightly evident.

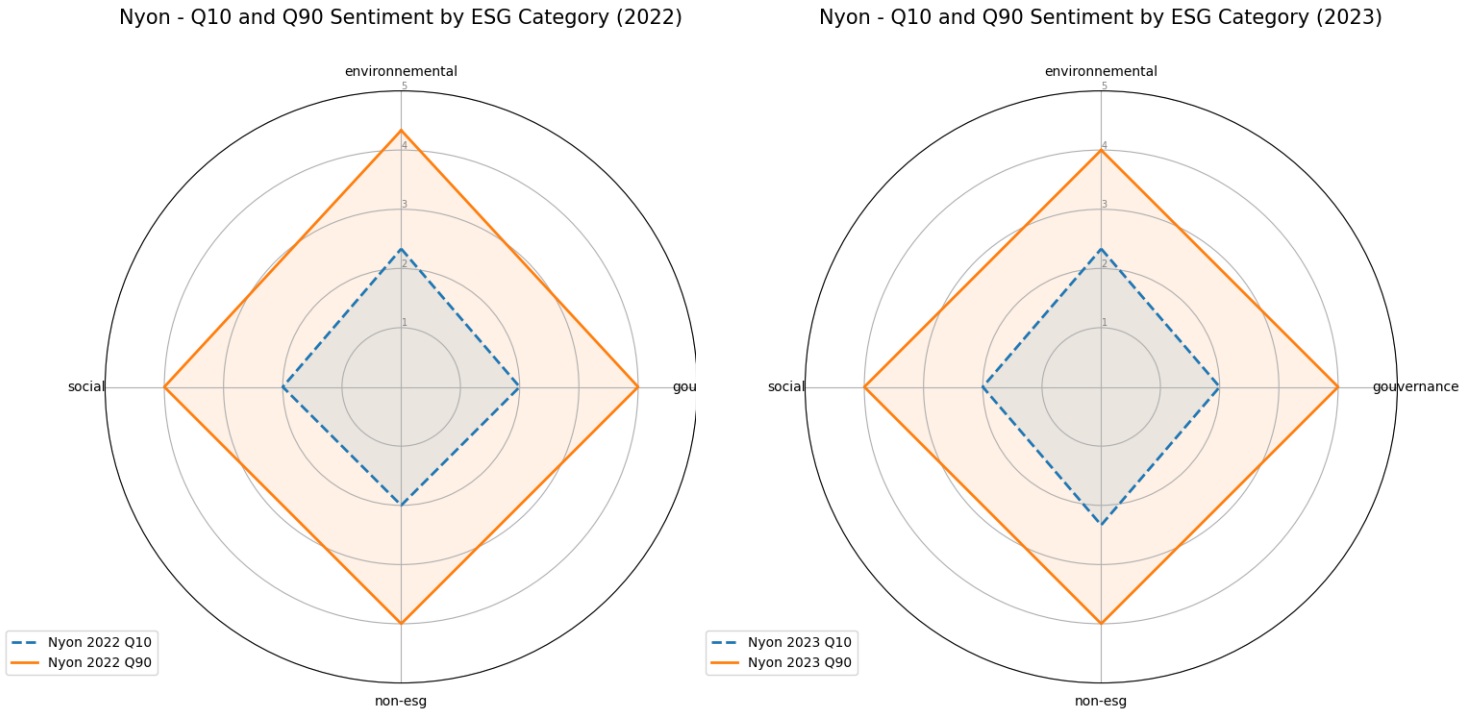
3.2.3 Individual City Breakdown

In the following graphs 3.13,3.14 and 3.15, we can observe the differences between variability in ratings between 2022 and 2023, and if any significant differences can found between each year.

In the chart 3.13 below, the ratings for both percentiles are generally uniform across all categories. However, there are slight differences, such as the variability in the environmental category, where ratings shift from just above 4 to slightly below 4. This can be attributed to the fewer text segments rated above 4.5 in 2023.

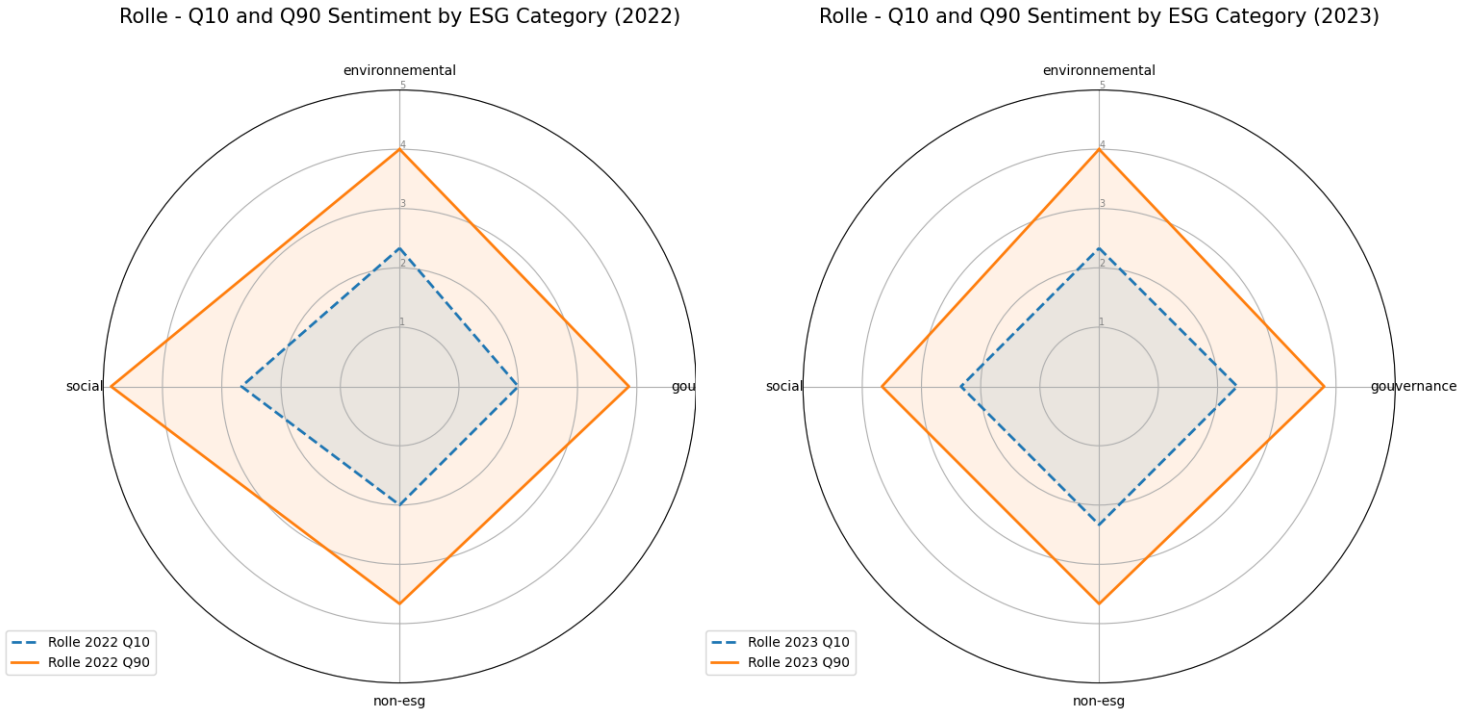
By separating the 2022 and 2023 data for Rolle, the quantity for each chart is further reduced, which increases the skewness, especially given the already limited data. However, despite the lower data quantity for 2023, there is less variability compared to 2022.

Finally, Vevey maintains a certain uniformity across all ESG labels, except for the social label, where the gap between Q10 and Q90 is more pronounced. After Rolle, which has a low amount of data, Vevey shows the highest rating for the upper bound Q90.



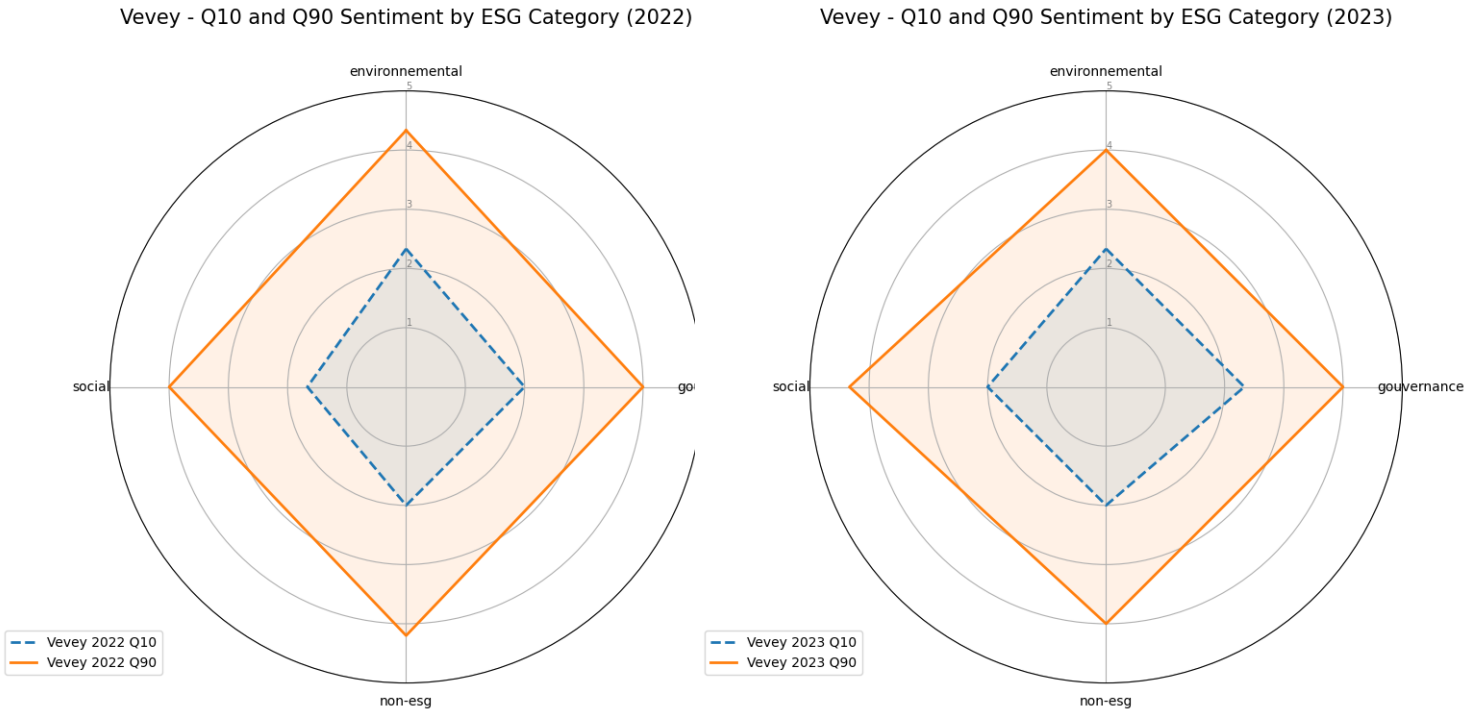
(A) (B)

FIGURE 3.13: Q10 and Q90 percentiles' data for the city of Nyon 2022 and 2023



(A) (B)

FIGURE 3.14: Q10 and Q90 percentiles' data for the city of Rolle 2022 and 2023. The lower amount of data with each having significantly different scores causes the distortions in 2022.



(A) (B)

FIGURE 3.15: Q10 and Q90 percentiles' data for the city of Vevey 2022 and 2023. After Rolle, Vevey in 2023 has the highest rating in the social category.

3.3 Implications and limitations

The automated tools developed for rating transcripts provide several valuable insights for policymakers, researchers, and public administration officials. These tools significantly reduce the time required to manually read, classify, and rate each transcript. By leveraging additional analytical tools, policymakers can assess the issues arising from the transcripts within specific categories and adjust their strategies to meet community needs more effectively. For instance, analysing sentiment trends or the ratio of discussion on the three ESG categories can help citizens better understand the topics covered during joint sessions, hence promoting a greater public engagement.

As the importance of ESG (Environmental, Social, and Governance) factors continues to grow in today's society, these tools can aid different cities in allocating resources more efficiently. This can be achieved through a time-saving analysis of the transcripts, or other pertinent documents. Additionally, this approach for evaluating ESG criteria in government ratings can provide a benchmark for future studies, enabling a comparative analysis over time.

From an investor's perspective, the current state presents both opportunities and challenges. For example, in the city of Rolle, different parts of a single municipal meeting's transcript are hosted on separate URLs, rather than being consolidated into a single document. This fragmentation can make it cumbersome to manually study and analyse the transcripts. Developing a script, as we have done in this study, to automatically extract and combine these transcripts could save significant time. Additionally, many uploaded documents are scanned copies without text recognition, further complicating manual analysis. In this study, we were unable to use scanned documents because free Optical Character Recognition (OCR) tools were insufficient for accurately converting images to text. However, using a paid OCR service in the future could enable us to properly segment these documents and integrate them into our frameworks.

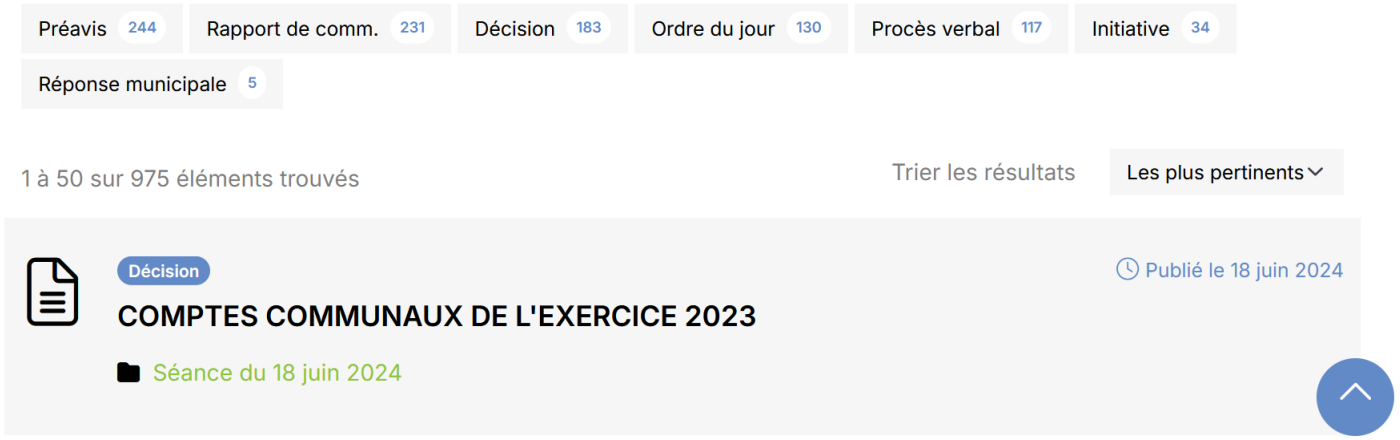


FIGURE 3.16: Rolle's website hosting the transcripts of every meeting, encompassing hundreds of files for every section from municipal meetings. Conducting a manual review to classify and assess the city's sustainability rating with the ESG framework would be both time-consuming and costly. Implementing automation for this task would significantly reduce both time and expenses.

With our system, which automatically classifies and rates yearly transcripts in a matter of seconds, investors can quickly gain an initial understanding of a city's relationship to sustainability. Our system serves as a valuable starting point, providing a general overview of the city's situation. The interactive app enhances this by allowing investors to view positive and negative sentiment at a glance, offering a quick yet informative first impression. Additionally, linking the classified text to specific transcripts for full context would further benefit users, enabling them to explore the material in greater depth as needed.

While ESG categories provide a broad overview, the analysis could benefit from more specificity. Integrating additional ESG key terms could offer a more detailed perspective, and developing an ESG-specific dataset would be necessary to achieve this.

However, certain limitations must be acknowledged. The neutral tone in which the transcripts are written can sometimes lead models to wrongly rating certain text segments as neutral. Additionally, by studying empirically the classified data, the classification process in the beginning may have also mislabelled segments that should be categorised under E/S/G as non-ESG. This issue, combined with lower accuracy rates in the social and governance categories (as noted in Table 2.5), can introduce certain biases in the ratings and subsequent analyses, such as the rating distribution or the sentiment trends that we studied above.

Another smaller constraint of this study was the focus on language. Our research focused on French-speaking municipalities, but the approach can be extended to German- and Italian-speaking cantons of Switzerland. By translating the initial dataset into German and Italian, we can fine-tune large language models, such as [jph23] for German and [gal23] for Italian. For sentiment analysis, pre-trained models in these languages would also need to be used, as there are currently no available datasets specifically designed to rate textual segments in these languages.

Overall, our approach represents a significant step forward in the application of sentiment analysis to public administration. While also offering at the very least a general perspective on the evaluation and improvement of the public administration under the ESG criteria.

Chapter 4

Conclusion

THROUGH the course of this thesis, we developed an automated ESG classification system tailored for municipal joint session transcripts, with the potential to be scaled for use with transcripts from various cities. Our approach utilized fine-tuned CamemBERT models [Mar+20], which were tested on text segments stored in *.csv* files. To fine-tune these models, we leveraged a large corpus of English news headlines classified by ESG criteria. We retrieved the respective articles, translated and summarized them, and created a balanced dataset of up to 20,000 rows of classified French texts.

After training, these models demonstrated a significant improvement in classification, particularly for environmental texts, achieving an average F1-score of up to 0.94. This was a notable enhancement compared to pre-trained models without fine-tuning and models trained for zero-shot classification. Some challenges arose in accurately classifying socially labeled texts, as they were occasionally confused with the non-ESG category. Despite these challenges, by combining multiple trained models, we effectively identified and classified ESG content in transcripts from the municipalities of Nyon, Rolle, and Vevey. Our empirical study of the classified transcripts revealed that some text segments labeled as *non-ESG* could have been categorized under one of the three main ESG categories.

In the final step, we rated these classified transcripts using multiple trained models on the *Huggingface* platform. As no dataset specifically rated texts in this domain, we selected three models, each providing unique insights. We combined the models by standardizing and averaging their output ratings. We found that a large fraction of the transcripts received a neutral rating (of 3), likely due to the neutral tone of speech in the transcripts. Nevertheless, sentiment trends could be observed when focusing on specific periods and categories.

Overall, with tools to interactively study the texts and trends, we can extract general sentiment insights, enabling policymakers to leverage the automated classification system and quickly assess the ESG aspects of municipal discussions, thus facilitating more informed decision-making. This system also contributes to the field of natural language processing by providing a novel application of text classification in the ESG domain in French and the public sector. Additionally, our framework sets a benchmark for future research in automated ESG classification, particularly for non-English datasets.

However, the study is not without limitations. As mentioned, the models occasionally misclassified social content, likely due to the inherent complexities in distinguishing social factors. The dataset used for training, while comprehensive, may not capture all nuances of municipal discussions, potentially affecting classification accuracy. Although this study was concentrated on the french language, translating the datasets and using the relevant datasets can enable to expand the frameworks usage for German and Italian speaking cities in Switzerland.

Future research could explore combining datasets from this study [Sch+23] to improve results globally and reduce potential biases. Enhancing sentiment analysis by creating a new dataset specifically for ESG texts could increase confidence in the predicted ratings. Incorporating manually classified transcripts into the training dataset could further refine the models' accuracies.

In conclusion, this thesis demonstrates the feasibility of applying state-of-the-art deep learning models to a domain that requires further exploration. As ESG considerations continue to gain prominence in both private and public sectors, this framework represents a step towards a more transparent and efficient evaluation of public governance in relation to ESG, aiding local governments in focusing on sustainable development.

Bibliography

[Aba+22] Julien Abadji et al. "Towards a Cleaner Document-Oriented Multilingual Crawled Corpus". In: *arXiv e-prints*, arXiv:2201.06642 (Jan. 2022), arXiv:2201.06642. arXiv: 2201.06642 [cs.CL].

[Bar23] Bards AI. *finance-sentiment-fr-base*. 2023. URL: https://huggingface.co/bardsai/finance-sentiment-fr-base.

[Bla20] Théophile Blard. *Allocine corpus*. 2020. URL: https://huggingface.co/datasets/tblard/allocine.

[cma23] cmarkea. *distilcamembert-base-sentiment*. 2023. URL: https://huggingface.co/cmarkea/distilcamembert-base-sentiment.

[DA22] Cyrile Delestre and Abibatou Amar. "DistilCamemBERT : une distillation du modèle français CamemBERT". In: *CAp (Conférence sur l'Apprentissage automatique)*. Vannes, France, July 2022. URL: https://hal.science/hal-03674695.

[Dai+19] Zihang Dai et al. *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. 2019. arXiv: 1901.02860 [cs.LG].

[Dev+19] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[Fis+23] Jannik Fischbach et al. "Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool". In: *2023 IEEE International Conference on Big Data (BigData)*. 2023, pp. 2823–2830. DOI: 10.1109/BigData59044.2023.10386488.

[gal23] galatolo. *cerbero-7b*. 2023. URL: https://huggingface.co/galatolo/cerbero-7b.

[Hvi17] Søren Hvidkjær. "ESG investing: a literature review". In: *Report prepared for Dansif* (2017).

[Inc16] MSCI Inc. *MSCI ESG Government ratings - Sovereign ratings*. 2016. URL: https://www.smart-und-fair-fonds.de/media/sov_presentation_msci_esg_research_2017-2.pdf (visited on 05/10/2024).

[Inc24] MSCI Inc. *ESG Ratings Key Issue Framework*. 2024. URL: https://www.msci.com/documents/10199/5c0d3545-f303-4397-bdb2-8ddd3b81ca1b (visited on 05/10/2024).

[jph23] jphme. *em_german_leo_mistral*. 2023. URL: https://huggingface.co/jphme/em_german_leo_mistral.

[Keu+20] Phillip Keung et al. *The Multilingual Amazon Reviews Corpus*. 2020. arXiv: 2010.02573 [cs.CL].

[Lau+24] Moritz Laurer et al. *Building Efficient Universal Classifiers with Natural Language Inference*. 2024. arXiv: 2312.17543 [cs.CL].

[Liu+19] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[LLC24] MSCI ESG RESEARCH LLC. *MSCI ESG Government Ratings Methodology*. 2024. URL: https://www.msci.com/documents/1296102/34424357/MSCI+ESG+Government+Ratings+Methodology.pdf.

[Mal+14] Pekka Malo et al. "Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts". In: *Journal of the American Society for Information Science and Technology* (Apr. 2014). DOI: 10.1002/asi.23062.

[Mar+20] Louis Martin et al. "CamemBERT: a Tasty French Language Model". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. DOI: 10.18653/v1/2020.acl-main.645. URL: http://dx.doi.org/10.18653/v1/2020.acl-main.645.

[NLP23] NLP Town. *bert-base-multilingual-uncased-sentiment (Revision edd66ab)*. 2023. DOI: 10.57967/hf/1515. URL: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment.

[PE22] Stefan Pasch and Daniel Ehnes. "NLP for Responsible Finance: Fine-Tuning Transformer-Based Models for ESG". In: *2022 IEEE International Conference on Big Data (Big Data)*. 2022, pp. 3532–3536. DOI: 10.1109/BigData55660.2022.10020755.

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques". In: *Proceedings of EMNLP*. 2002, pp. 79–86.

[Rez21] Mohammad R. Rezaei. *Amazon Product Recommender System*. 2021. arXiv: 2102.04238 [cs.IR].

[San+20] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL].

[Sch+23] Tobias Schimanski et al. "Bridging the Gap in ESG Measurement: Using NLP to Quantify Environmental, Social, and Governance Communication". In: *SSRN Electronic Journal* (Jan. 2023). DOI: 10.2139/ssrn.4622514.

[TK22] Ellia Twinamatsiko and Dinesh Kumar. "Incorporating ESG in Decision Making for Responsible and Sustainable Investments using Machine Learning". In: *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. 2022, pp. 1328–1334. DOI: 10.1109/ICEARS53579.2022.9752343.

[Wen+19] Guillaume Wenzek et al. *CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data*. 2019. arXiv: 1911.00359 [cs.CL].