Overall remark:
- It should be said somewhere that nothing prevent us from doing that in German or in Italien. French is used as an example.
- It is a good work, but you can make it better if you succeed to not make it look like a black box. In the current state, it is very much a black box, give more examples : classifications, rates, rates for Muni, ... I have given hints all along the text.

UNIVERSITY OF GENEVA

MASTER'S IN COMPUTER SCIENCE THESIS

Automated Scoring System for Assessing ESG Sustainability in Swiss Municipalities.

Swiss Municipalities Sustainability, wouldn't that be better?

Author:
Muhammad Azeem ARSHAD

Supervisor:
M. Alexandre DUPUIS

Dr

Important for you, not for me

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)
& in collaboration with Compenswiss

Don't say that, as it is not official.

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'informatique

June 7, 2024

UNIVERSITY OF GENEVA

Abstract

Faculty of Science

Centre Universitaire d'Informatique (CUI)

MSc in Computer Science

Automated Scoring System for Assessing ESG Sustainability in Swiss Municipalities.

by Muhammad Azeem ARSHAD

a research project

This thesis presents an effort to develop and implement an automated scoring system to assess the Environmental, Social, Governance (ESG) factors in certain Swiss municipalities, namely 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 consisted of creating a new balanced dataset in french from a collection of English article headlines and fine-tuning CamemBERT models. Multiple models were produced, and a weighted voting scheme was employed to combine and enhance classification accuracy. The results indicated a strong performance in identifying environmental,governance and non-ESG related content, with a small challenge in distinguishing social aspects. To rate the classified transcripts, we finely selected multiple models trained for general sentiment analysis, as no specific model was available for our specific task. We standardized their outputs and subsequently combined the ratings by averaging them. Finally, we conducted a comprehensive study of the classification and ratings, by examining key findings, limitations, and any emerging trends. This work offers a novel approach for automatic ESG assessment in the public sector in Switzerland and enables for a more informed decision-making.

what about you first say that the framework is for all Muni. Then you say that we apply it to medium size cities in Vaud: Nyon, Rolle and Vevey?

but then what are the key findings, limitations and emerging trends? The summary should be the place to say that.

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor…

Contents

List of Abbreviations

An indentation would help, wouldn't it?

Env	Environmental
Soc	Social
Gov	Governance
ESG	Environmental, Social, Governance
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

For/Dedicated to/To my...

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 refer 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 and analyses various studies performed on corporate esg issues and ratings. Please, here add a paragraph summarising TK22.

Nevertheless, limited research has been conducted on analysing textual resources such as audio or video meeting records, archives, or other forms of documentation to evaluate companies with respect to ESG. what sort of archives? In addition to this limited research, no recent studies have specifically focused on the French language to evaluate its influence within the ESG framework. The study in [PE22] pioneers an approach by fine-tuning transformer-based large language models to simultaneously interpret texts and assign ratings, using companies' annual reports and ESG scores from rating agencies as training data. Does that work? What do they use it for? Any comment on their approach?

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 English I guess? Again how does that compare to our work? I hope Sch+23 keeps on coming in the text below as it could serve as a benchmark

In this thesis, we aim to extend these methodologies by fine-tuning a french transformer-based large language model, classifying publicly available municipal council proceedings from recent years with respect to Environmental (Env), Social (Soc) and Governance (Gov) factors. Subsequently, these classifications will be then scored using sentiment analysis techniques. We will attempt to develop a systematic process to collect council records and assign

ratings by municipality, contributing to a new perspective to the existing body of ESG research.

Shouldn't you already sketch the result? At this stage we have no idea whether you succeeded or not. Do not give the result yet, but a teaser could help embarking the reader.

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. 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.

as one would need a few tens of thousand sentences describing each category.

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.

how does one do that automatically?

The texts were then translated into French using Google's translator, resulting in the following balanced dataset, with Environmental (Env), Social (Soc), Governance (Gov) and Non-ESG categorisation. 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 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.

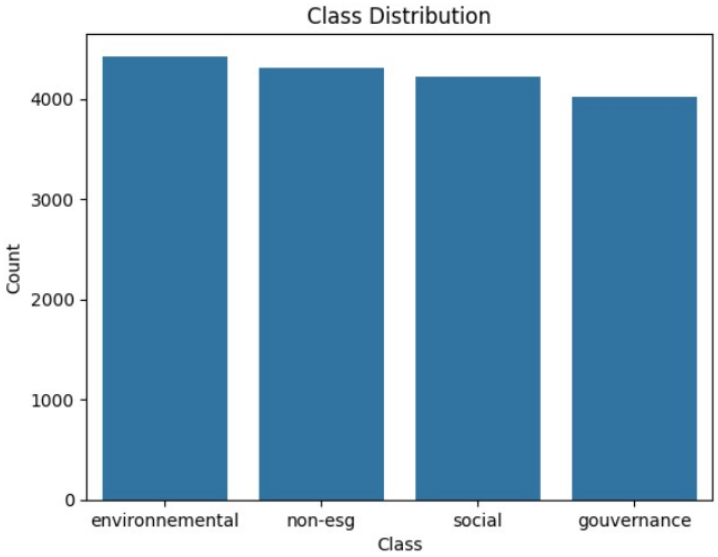


FIGURE 2.1: Dataset class distribution

words? number of characters?

The boxplot in figure 2.2 below displays the average text lengths 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.

in words?

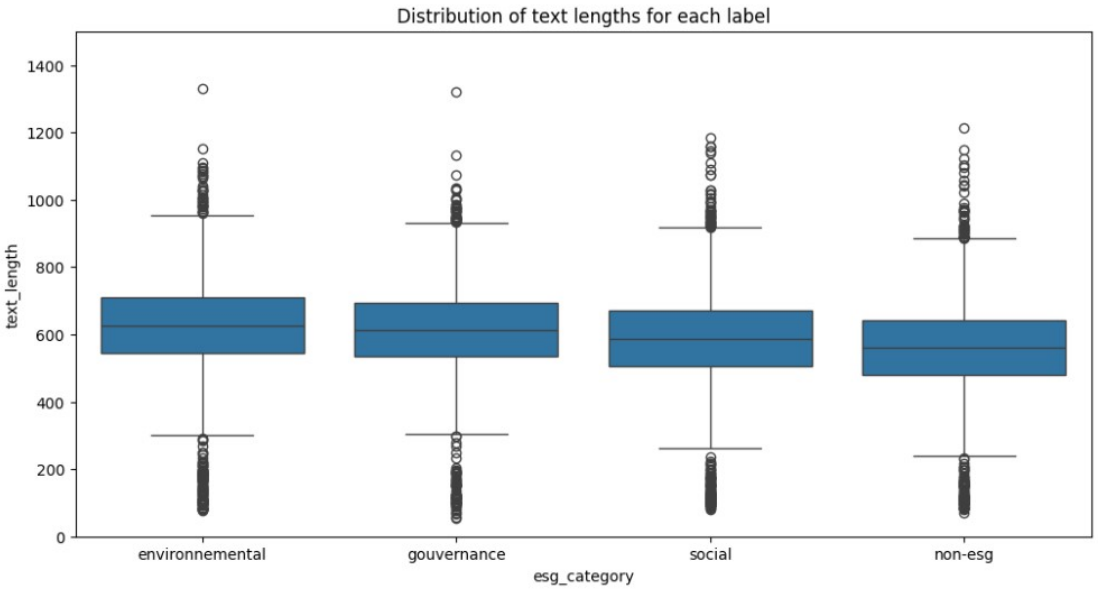


FIGURE 2.2: Dataset input text lengths' boxplot

We use the Term Frequency (TF)-Inverse Document Frequency (IDF) statistics to identify and evaluate the most important terms 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.

why doing that? Please do explain.

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

Interesting table.

TABLE 2.2: Top TF-IDF words for each class

2.2 Methodology and Evaluation Metrics

remind the reader what they are

In this study, we mainly employ macro-averaged metrics for *precision*, *recall*, and *F1-score* to evaluate the performance of our models. Macro-averaging is chosen to ensure that each class contributes equally to the final evaluation, regardless of its frequency in the dataset.

explain more what macro-averaging is. Something short.

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 us with a point of reference to measure the improvement achieved through fine-tuning.

Again you need to say more there. Imagine you explain that to one of your jury who might not be a computer scientist expert. Do not go into the details but summarise what it is

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.

Is CamemBERT only for French? This need to be said.

Please write Camembert in the same way every where.

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.

That's the one, right?

We select Camembert due to its robustness and high performance in various NLP tasks in French, as noted in recent literature. 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.

ref

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

Not so sure I understand the need of having two models, the above clarifications will help me I suppose.

oh that's good

2.4 Fine-tuning

Given their robust performance and advanced components, we selected the Camembert models for fine-tuning [Mar+20]. Bidirectional transformers, as employed in these models,

we did already, right?

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.

say a few words about it

In order to fine-tune, we use the base and large Camembert models. The base model 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.

please restrict yourself later, at this stage all this works for any French speaking muni.

Considering the extensive transcripts from selected Swiss municipalities, 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.

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
```

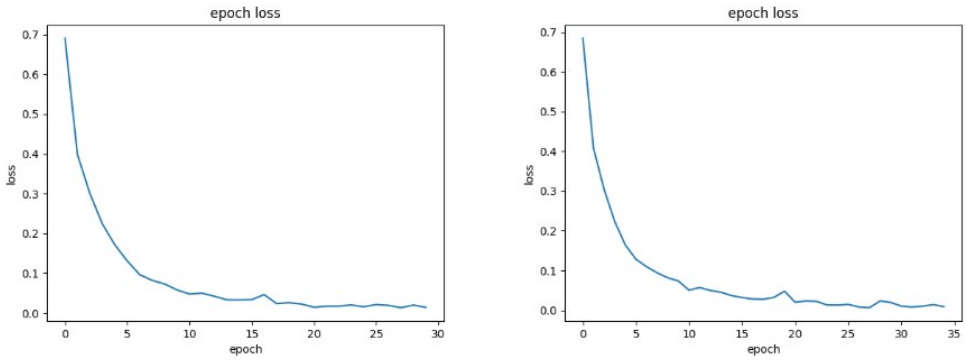
LISTING 2.1: SLURM Job Submission Script

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.

2.4.1 Results

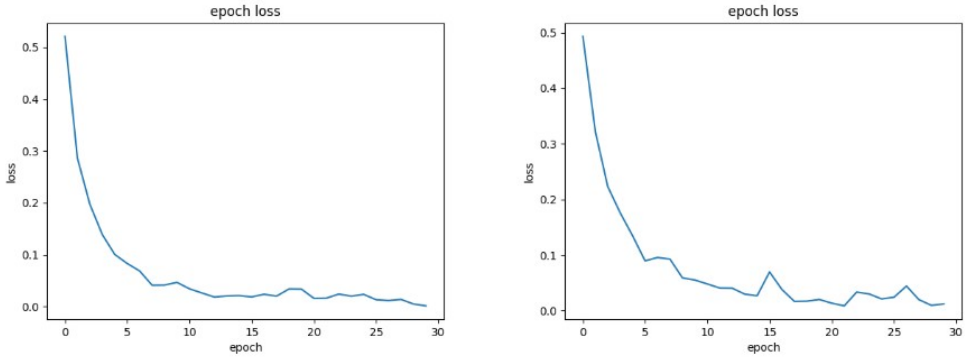
Figures 2.3 and 2.4 illustrate the training losses of the models over a period of 30 epochs, highlighting the stable learning process. It reaches a certain plateau after 20 epochs.

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



(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.

is it ran on a single processor?

TABLE 2.4: Macro-averaged F1-scores and Average Training Times

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. It comprises three separate models, each designed to perform binary classification for one of the ESG categories (Environmental/ Social/ Governance or *None*). Thus, to obtain an accuracy score for the latter, we input the sentences categorised under each ESG label from the 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.

this table caption could be enriched in such a way that one can understand without reading all the text.

TABLE 2.5: Models results comparison with f1-score per label

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

Not I understand. The three first lines are from Sch+23? ANd where doen ESGBert come from?

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. We follow the following steps to generate the weights and label selection: you did construct it by hand? How big it is?

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 $M^{4\times4}$:

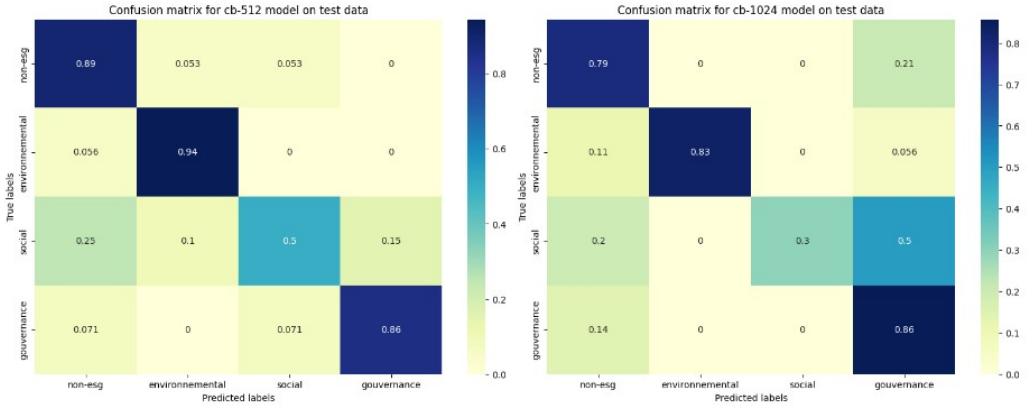
$$\mathbf{M} = \begin{bmatrix} \text{metric}_{1,c1} & \text{metric}_{2,c1} & \text{metric}_{3,c1} & \text{metric}_{4,c1} \\ \text{metric}_{1,c2} & \text{metric}_{2,c2} & \text{metric}_{3,c2} & \text{metric}_{4,c2} \\ \text{metric}_{1,c3} & \text{metric}_{2,c3} & \text{metric}_{3,c3} & \text{metric}_{4,c3}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

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

3. We normalize the matrix w.r.t each column and obtain the final weights matrix $\mathbf{W}$
   how? row?
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 **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 2.5: Confusion Matrices for each fine-tuned model

Overall, in the confusion matrices in the figures 2.5, 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:

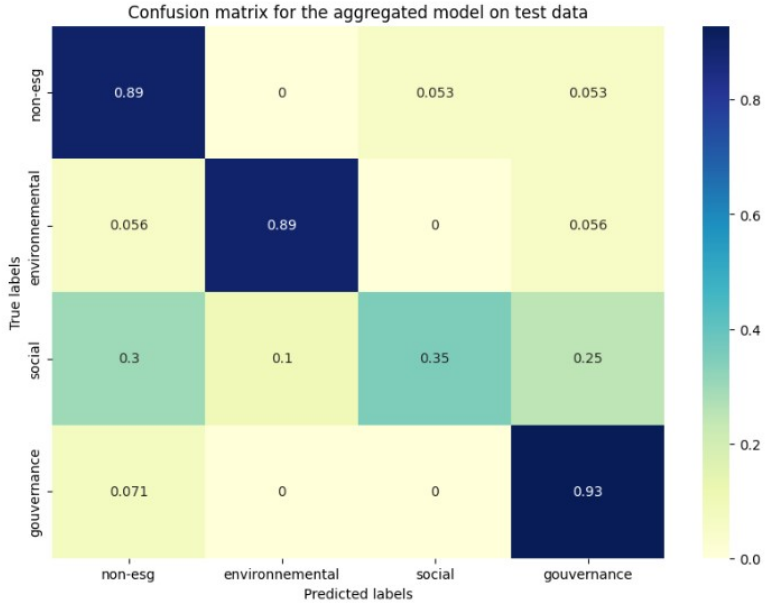


FIGURE 2.6: Confusion matrix with the aggregated model on the smaller, manually annotated dataset

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:

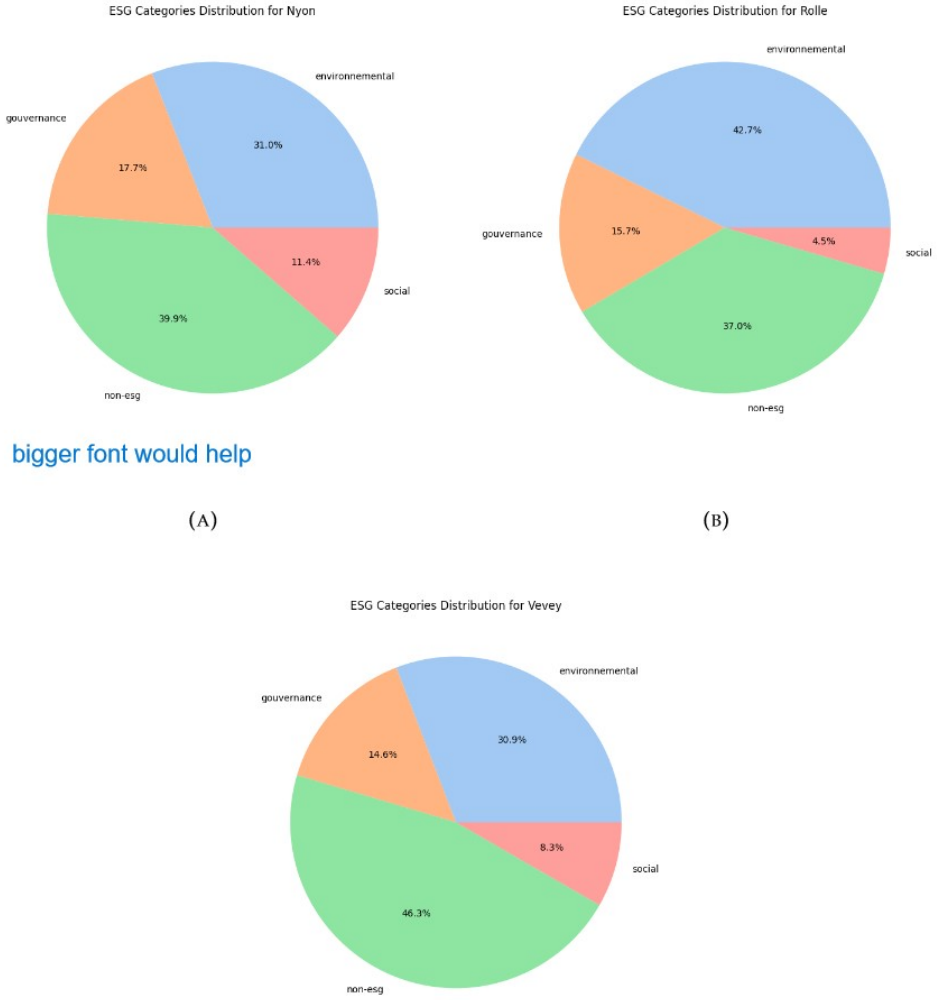


FIGURE 2.7: ESG categories distribution for (A) Nyon, (B) Rolle and (C) Vevey

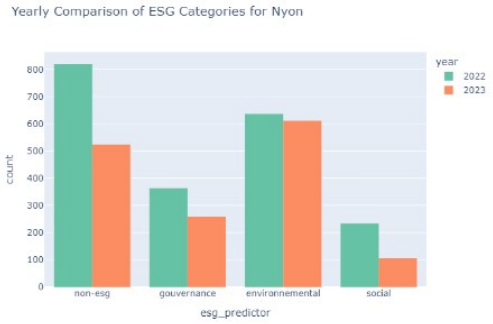
We can first observe that a large proportion of the transcript is classified as *non-ESG*. This classification can be interpreted in two ways. On one hand, it might indicate that certain segments that could potentially belong to one of the three ESG categories were not classified as such. On the other 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

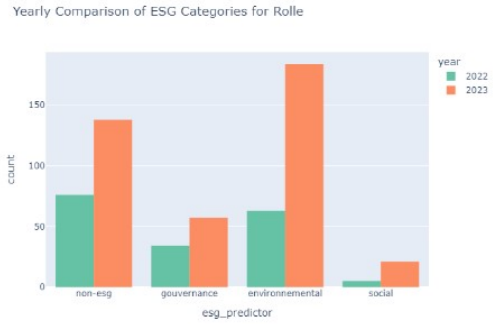
identified as belonging to one of the ESG categories when it does not, while a false negative would be a text that belongs to an ESG category but is not classified as such.

In the following histograms 2.8a, 2.8b and 2.8c, 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.

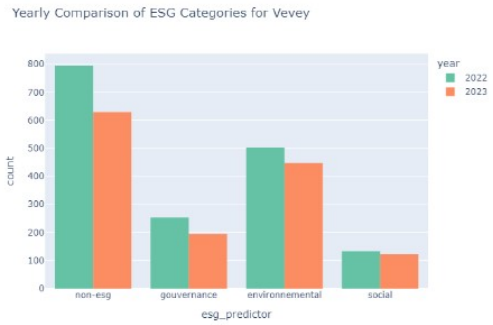
Overall, A possible reason for the social category being small could be that a certain amount of segments labeled as non-ESG should have been categorised under the social label.



(A) Nyon



(B) Rolle



(C) Vevey

FIGURE 2.8: Yearly comparison of ESG categories for Nyon, Rolle and Vevey

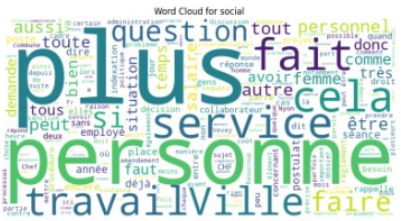
2.6.1 Word Clouds

By observing the word clouds in the figure 2.9 below, we can further visually evaluate the classification results too. 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) nonesg

(B) env



(C) soc

(D) gov

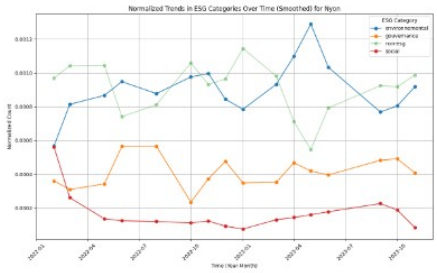
FIGURE 2.9: Word clouds for every classified transcript

Is this for the 3 cities all together?

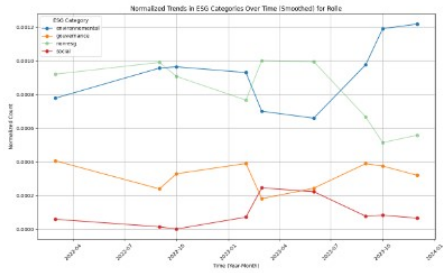
2.6.2 Seasonal trends

Finally, To further analyse the classification results, we can also examine the trends in the frequency of discussions for each category over the past 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.

enlarge fonts

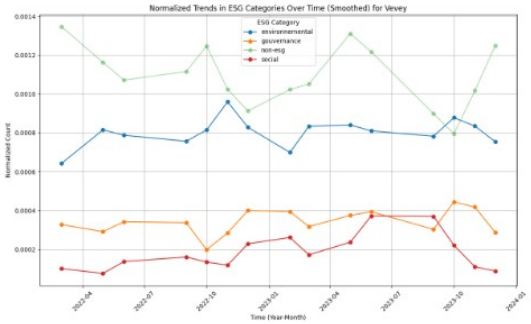


(A) Nyon



(B) Rolle

I don't understand what we do here. What is it that is measured?



(C) Vevey

FIGURE 2.10: 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 trend in figure 2.10 and the histograms in Figure 2.8, it is apparent that the social category is quite underrepresented in the meeting records. A technical 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]: 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 computationally efficient. The authors report that the model achieves an exact accuracy of 61.01% and an off-by-one accuracy of 88.80%. They additionally tested the model above with the same test datasets and obtained accuracies of 54.41% and 82.82% respectively. This supports the reliability of the previous model's performance. Finally, similar to the previous model, this one classifies on a 1 to 5 scale.

61% enough?

- ***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].

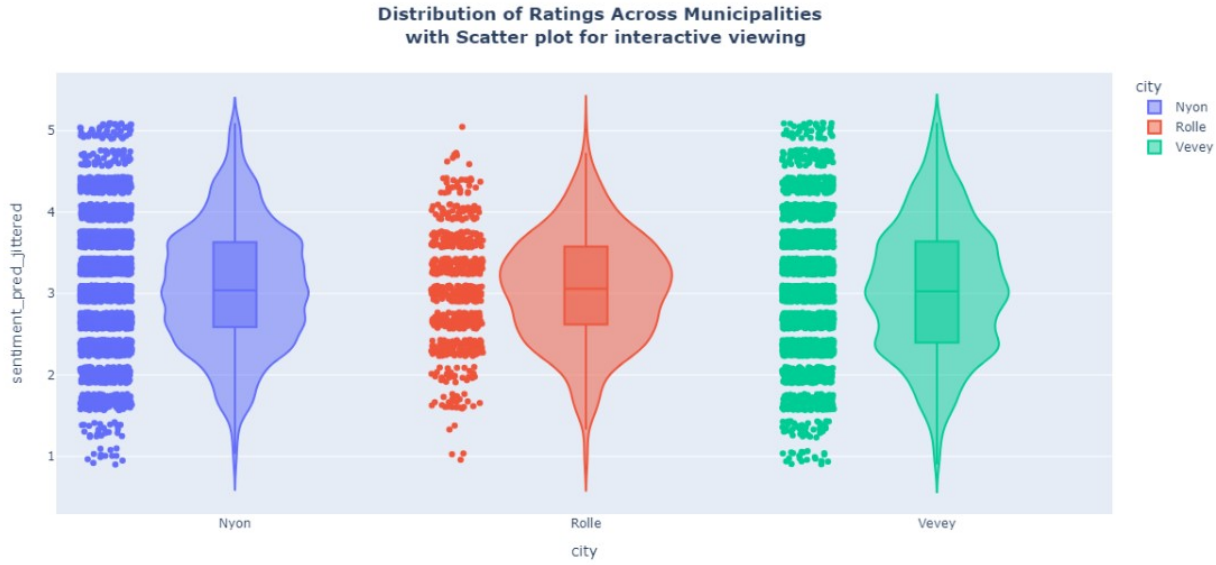


FIGURE 3.1: Distribution of Ratings Across Municipalities with Scatter plot for interactive viewing

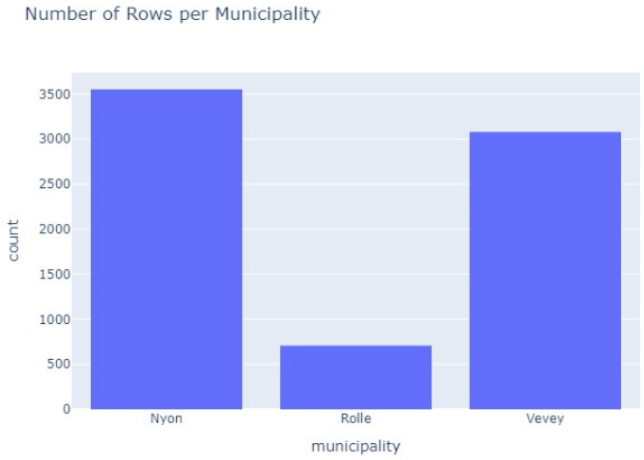


FIGURE 3.2: number of rows per municipality in figure

roughly

However, we observe that the ratings 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).

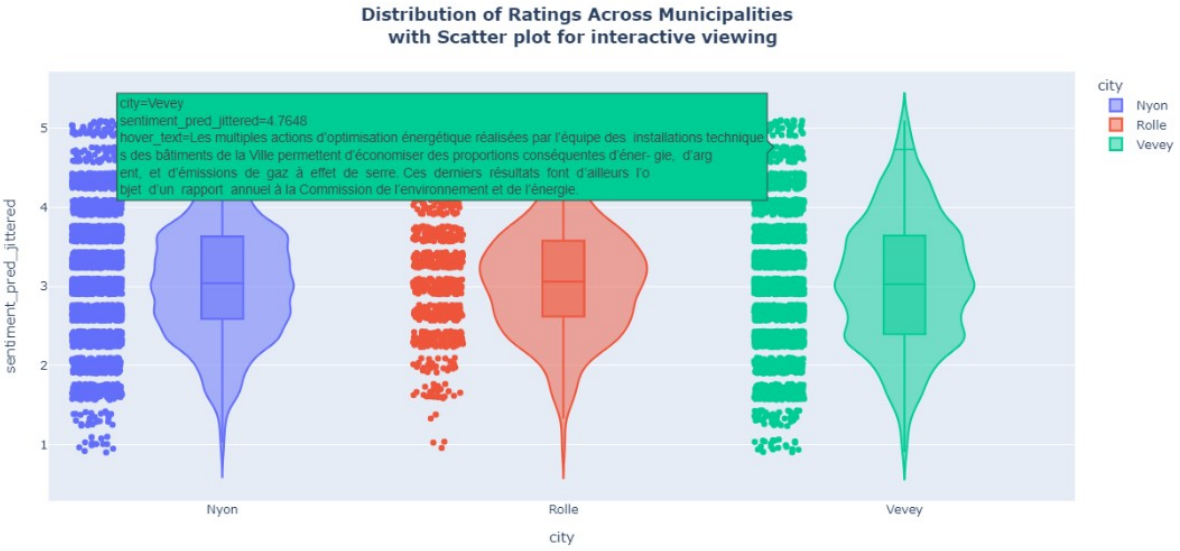


FIGURE 3.3: Interactive view of a text segment's rating in the Distribution of Ratings Across the Municipalities. A slight jitter was added to the scatter plot to ease the view of different samples

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.

isn't that an increase from 3 to 3.4? Why by the way?

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.

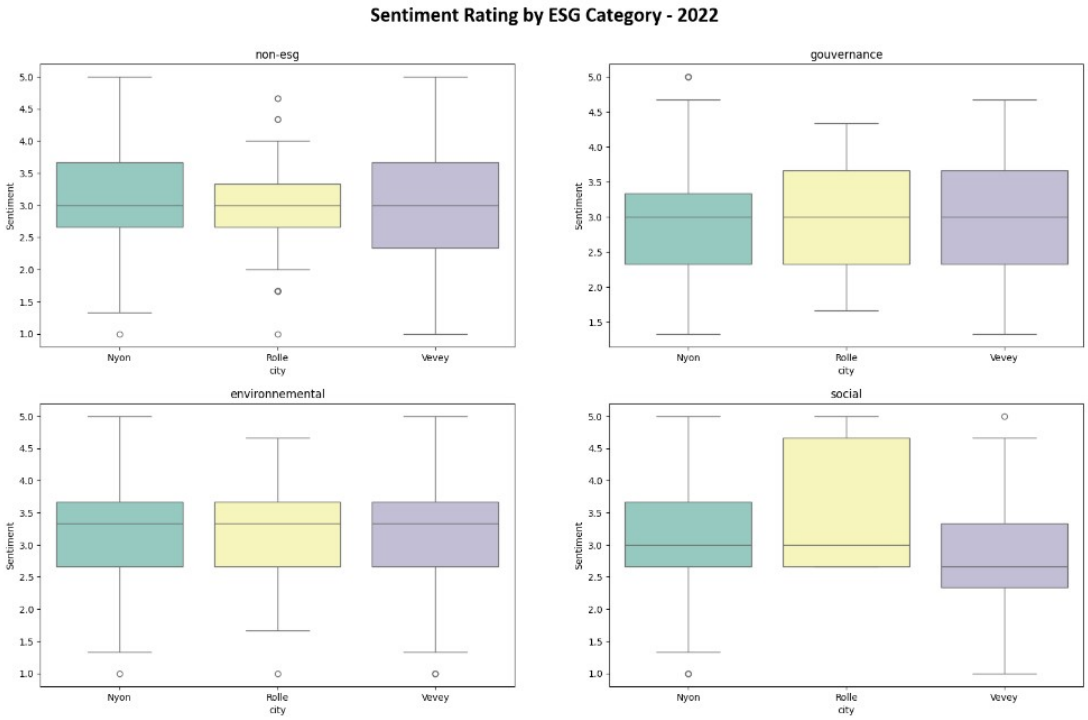


FIGURE 3.4: 2022 Sentiment rating w.r.t the city - Category wise

explain what is being ploted, what are the lines, the dot, the box, ... the caption is the right place to say it.

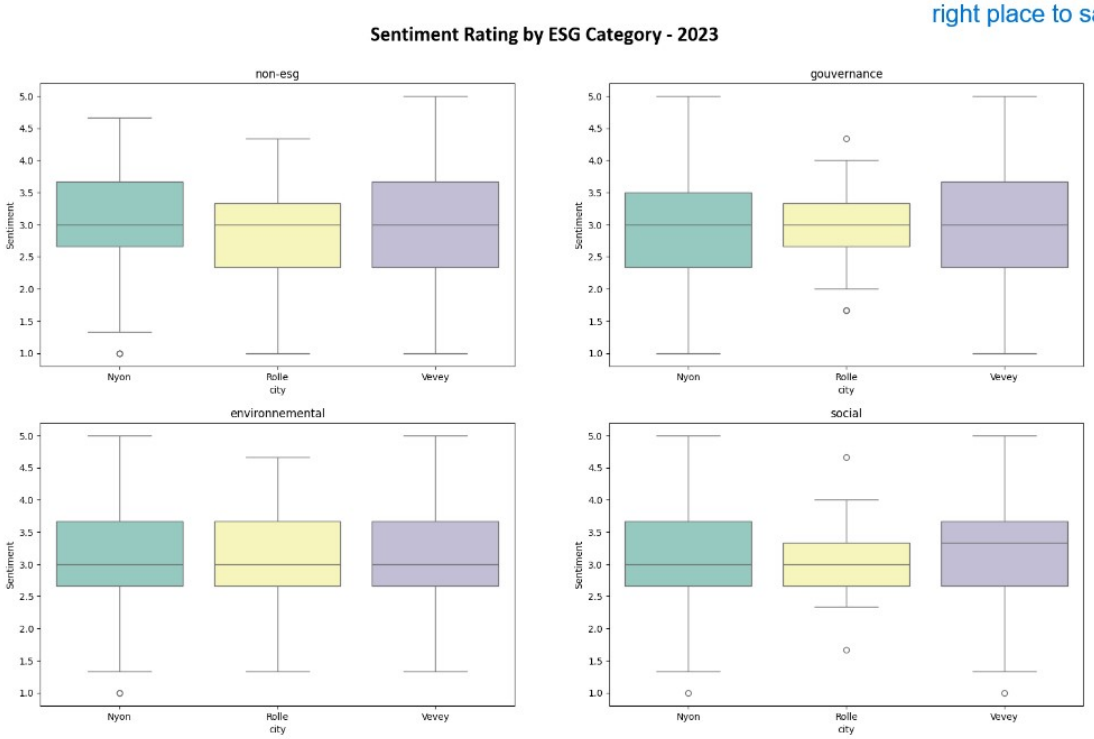


FIGURE 3.5: 2023 Sentiment rating w.r.t the city - Category wise

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.

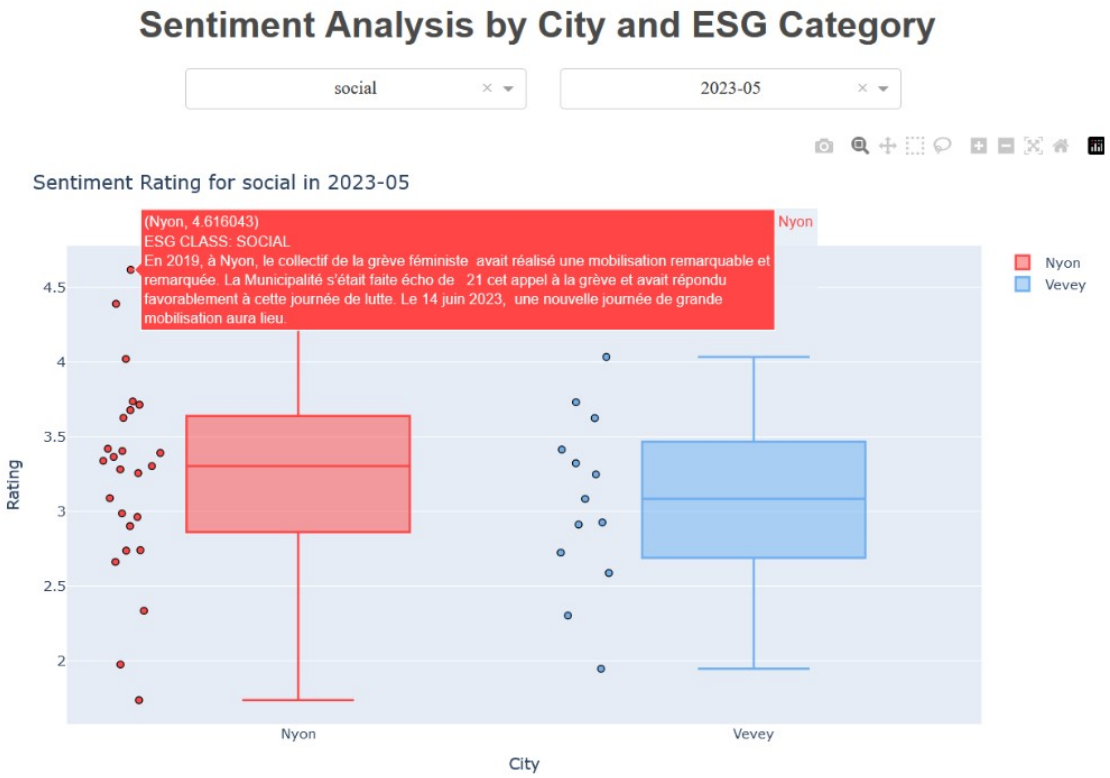


FIGURE 3.6: Interactive plot viewing in HTML with the python library *Dash*. Filtering by category and date available.

I would be interested to see examples where ratings are high and low. For all labels and cities. What about tables citiesxlabel showing example of high and low ratings

3.1.4 Rating Trend with respect to the time - Category wise

We thus present in the figure 3.7 the monthly average ratings for each city, categorized accordingly, providing a detailed comparative analysis of the emerging trends over the two-year period.

is it really the average that one should look at? I would have attempted to look at Quantile 90% for example. Or better, Q90-(3-Q10). This would favour high Q90 and low Q10. What do you think?

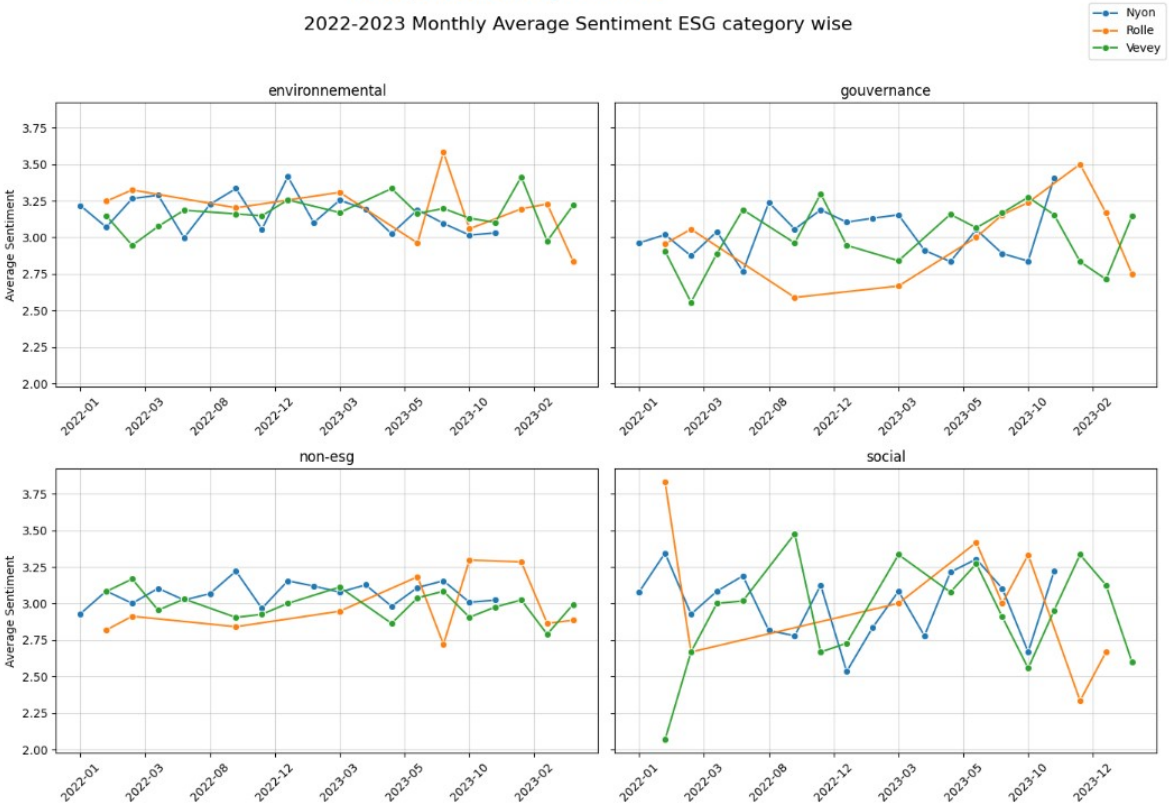


FIGURE 3.7: 2022-2023 average sentiment rating, ESG category wise

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.

We should also attempt to explain the behaviour of the above graphs with examples. Why does it react that way? Otherwise people won't buy it.

What about showing an average over the whole year in a spyder chart? One per city.

3.2 Use-Case Analysis and Interpretation

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 3.2 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 (3.2). Consequently, the city of vevey had a better average rating in this period regarding the governance category.

TABLE 3.2: Sample sentences categorised as Governance from October 2022 transcripts

Nyon	Rolle	Vevey
[3.41 / 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."	[2.08/ 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.

3.2.1 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, analyzing 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.

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.

Wouldn't the logic be the same? I therefor wouldn't call it a limitation.

Another limitation is the language focus. Since the study aimed to rate municipalities in the French-speaking part of Switzerland, the models and rating systems were developed specifically for the French language, potentially limiting their applicability to other 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.

I do miss examples here and there. When texts are classified, and when they are rated. I would suggest to work a little more on the ESG rating itself. The average is not appropriate in my view. Once again, the evolutions of the various ratings need to be explained. It looks too much as a black box at this stage. And people do not like blackbox.

I would also comment on what are the steps to push the idea to the whole Switzerland? What would you suggest one should do?

If I were to decide to each Muni would I lend money, could I in the current state? The spyder chart could help but please do place yourself in the shoes of a decider who has to choose which one to favour. Why would he do that? And what would the Muni would have to improve to rate higher?

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. Furthermore, our focus on French-language transcripts limits the model's generalizability to other languages. The dataset used for training, while comprehensive, may not capture all nuances of municipal discussions, potentially affecting classification accuracy.

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.

[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.

[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).

[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].