Environmental agenda detection: a text classification model employing political data from the Manifesto Corpus

Yerkezhan Abdullayeva
University of Potsdam
Cognitive Systems - Individual Research Module

Abstract

This paper introduces a binary classification model built to determine whether a given sentence is related to environmental protection. The model is trained using hierarchical topic data from Manifesto political corpus, which is a textual collection of political declarations. The study includes experiments to achieve a balanced distribution within the positive category and investigates the dataset's hierarchical nature, comparing the effectiveness of single-step and two-step modeling paradigms. The study evaluates the model's cross-sector applicability by using environmental datasets from the financial industry. Although the results of these evaluations, conducted both within the corpus and on supplementary cross-sector test datasets, did not meet high accuracy standards, the error analysis identified potential areas for advancement and additional research.

1 Introduction

Environmental concerns are also becoming more prevalent in the business sector, according to a previous study on environmental claim detection (Stammbach et al., 2022). The figure 7 from this study demonstrates that since the Paris Agreement in 2015, there has been a steady rise in the number of environmental claims made by businesses and business executives.

Similar pattern can be observed in the Figure 8, which illustrates the frequency of the environmental protection category in the manifestos between 1946 and 2021. A manifesto is a written statement of the goals, objectives, or opinions of the issuer, which might be an individual, an organization, a political party, or even the government (Volkens et al., 2017). A manifesto generally supports a new idea with prescriptive ideas for implementing the changes the author feels should be made, accepts a previously published position or the general agreement, or promotes a new idea.

As the environmental issue becomes more prominent in the political arena too, it is critical to develop an automated method for detecting environmental context in electoral campaigns. This automated method can help identify candidates' positions on environmental issues and hold them accountable for their promises. Additionally, the model trained in the political sector will be evaluated for cross-sector transferability using a financial sector dataset to assess its effectiveness across various sectors.

2 Data

2.1 General description of the corpus

The Manifesto Corpus is a multilingual and annotated collection of electoral programs (Lehmann et al., 2022). The Manifesto Project Dataset was originally created by the Manifesto Research Group (MRG) and later the work was continued under the name Comparative Manifestos Project (CMP). The CMP is an internationally coordinated research project for the analysis of political documents, in particular election and party programs. The project is based on the assumption that the current political positions of the parties are reflected in election programs. The aim of the project is to provide a uniform method for the quantitative content analysis of such documents.

The Manifesto Corpus is the largest annotated electoral program collection available at the moment. The corpus is based on the electoral programs of all parties that won one (Australia, Japan, New Zealand, North America, South Korea, and Western Europe) or two (Central and Eastern Europe) seats in the respective national elections to the lower house. Furthermore, manifestos from parties that were relevant actors in the past but even though did not meet the selection criteria because of significant vote losses have been also added to the corpus.

There are several versions of the Manifesto Project Dataset as it is updated about twice a year and the version 2022a is used for the current analysis. This version covers 1216 political parties from 1945 to 2022. In fact, the total number of documents in the corpus for version 2022a should be 4,850, and 1840 of those documents should be annotated ones. Instead, we get only 4778 documents that we can use for the text classification purpose, 1694 of which have annotation. Overall, most documents of Sweden from 1949 to 1982 and Norway from 1945 to 1989 were lost. This issue was earlier addressed by the authors and it was concluded that this problem is unfortunately inevitable as some manifestos were lost over time.

The data is gathered from manifestos during 779 elections in 56 countries , which were mostly democracies in OECD and Central and Eastern European countries. The source of the corpus is publicly available election statistics and content analysed election programs.

Three different types of information are included in the Manifesto Corpus:

- Digital electoral programs: Machine-readable manifesto of a political party or candidate during an election campaign.
- Meta-data: Each manifesto has the meta information such as party, election date or the language. Overall, there are 16 descriptors of the manifesto. For example, the election date or the party ID.
- Annotation: As was previously mentioned, 1694 documents have content-analyzed annotations at the level of single quasi-sentences, in accordance with the Manifesto Project coding system. Quasi-sentence is a linguistic concept that refers to a sequence of words that may resemble a sentence at first glance, however it expresses a clear idea, even if it lacks the necessary components to form a complete grammatical unit. An annotator assigns a category to each quasi-sentence based on its context. Therefore, the annotation is the label assigned to the quasi-sentence by the specially trained annotator.

Table 1 depicts that 202 of the 1694 annotated manifestos are in English. To unify disparate data into a single master data set, it was decided to translate the rest of the documents into English with the help of Google Translate.

Languages	Documents	Sentences
English	202	174489
German	137	175778
Spanish	124	168245
Dutch	101	177958
Hebrew	63	30478
Greek	62	42388
Ukrainian	56	6670
Danish	55	18373
Croatian	54	30764
French	52	103288
Macedonian	43	74072
Slovenian	41	39392
Slovak	40	29735
Romanian	39	15371
Montenegrin	38	15547
Icelandic	34	8012
Bosnian	34	19287
Lithuanian	34	35141
Swedish	33	17485
Italian	33	21684
Polish	32	28048
Norwegian	31	83258
Finnish	31	21924
Czech	31	25812
Latvian	30	2031
Portuguese	28	46953
Turkish	28	56716
Hungarian	28	46857
Russian	28	5838
Estonian	24	16322
Georgian	23	12585
Armenian	22	7466
Serbian-cyrillic	20	10377
Bulgarian	18	12131
Catalan	13	22823
Serbian-latin	11	3415
Japanese	9	3632
Galician	6	5885
Korean	5	6622
Bosnian-cyrillic	1	728
Total	1694	1623580

Table 1: Language-specific numbers of annotated documents and sentences.

However, it is important to note that the translation of the sentences to English was not completed by the corpus's authors, and the accuracy of the translations was never officially verified. Therefore, there may be possible errors or inconsistencies in the translated text.

2.2 Data extraction

The Manifesto Corpus is stored as an online database and in order to access the corpus it is mandatory have Manifesto Project Database API Key, consequently to register on the main website : `https://manifesto-project.wzb.eu/`

There are 4 ways of extracting the coded election programs from the corpus:

1. Manifesto Project Data Dashboard: There is a dashboard website where you can download either the manifesto files with .csv extension individually or make customized datasets by restricting the filters on the website. This method is more time-consuming compared to other methods, but its advantage is that it does not require programming skills therefore makes the corpus accessible for everyone.

2. Access using ManifestoR: There is a free package for the statistical software R called ManifestoR, which transforms downloaded documents into corpus format and also has specific functions for scaling of coded political texts.

3. Access using Manifestata: Similar to ManifestoR, Manifestata is a free add-on for the statistical software Stata. However, this method is outdated and some errors stayed unresolved.

4. Access via API: This method has an advantage as it lets the user to have direct access to the database and returns data in a standardised JSON format, therefore it was applied for the text classification purpose.

2.3 Category classification

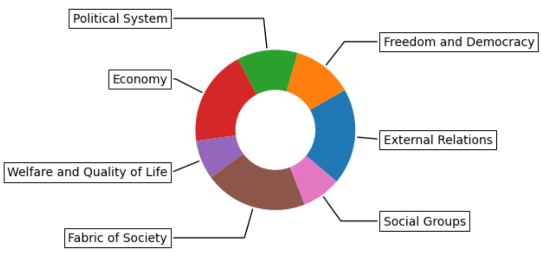

Figure 1: Distribution of the categories among seven policy area of the Manifesto Corpus.

The CMP classification scheme officially provides 139 categories , which are assigned to seven policy areas in the 2022a version of the Manifesto Corpus. Figure 1 represents the distribution of the categories among these seven domains. However, the real number of actually available categories is 132 as some documents were lost over time.

Seven domains of the Manifesto Corpus:

- "Welfare and Quality of Life" domain includes 11 categories. The target category "Environmental Protection" is category of this domain.

- "External Relations" domain includes 23 categories.

- "Freedom and Democracy" domain includes 15 categories.

- "Political System" domain includes 17 categories

- "Economy" domain includes 27 categories

- "Fabric of Society" domain includes 28 categories

- "Social Groups" domain includes 11 categories

Table 13 depicts these 7 domains with overall 139 categories assigned to them. The table shows that the corpus is hierarchically arranged and has sub-classes, indicating that domains are general classes with sub-classes as categories inside these classes.

2.4 "Environmental Protection" category

In the codebook of Manifesto Project Dataset (Volkens, 2020) the "Environmental Protection" category is described as general policies in favour of protecting the environment, fighting climate change, and other “green” policies. The category may include a wide range of regulations with the unified objective of environmental protection, such as:

- General preservation of natural resources;

- Preservation of countryside, forests, etc.;

- Protection of national parks;

- Animal rights. May include a great variance of policies that have the unified goal of environmental protection.

Figure 8 the frequency of the occurrence of "Environmental Protection" category in the documents. The figure demonstrates that over time, there has been a significant increase in the number of times the environmental protection category has been mentioned in political campaigns.

Environmental preservation and nature conservation emerged as one of the most important subjects

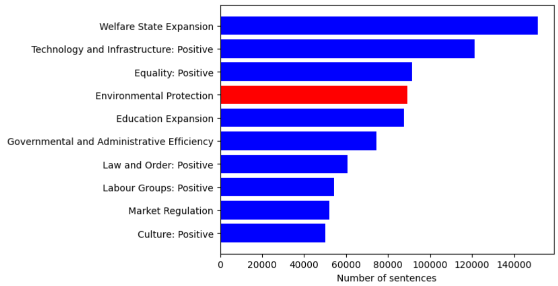

Figure 2: The ten most frequent categories in the Manifesto corpus.

in the entire dataset after thorough study and corpus analysis. Figure 2 shows that environmental protection is the fourth most common category in the dataset.

3 Task description

The objective of this study is to develop and evaluate a binary classification model for assessing if a statement is relevant to environmental protection. It investigates several strategies for balancing classes and raising effectiveness. To find possible areas for development and to make sure the model works across other sectors, its applicability is evaluated, with a focus on financial datasets.

For the sake of brevity, this paper uses the abbreviation WQL for "Welfare and Quality of Life," Non-WQL for domains excluding it, and EP for the "Environmental Protection" category. The abbreviation Non-EP encompasses all categories except the "Environmental Protection" category.

Two more datasets are used to evaluate the classifier's transferability:

- Dataset for Detecting Real-World Environmental Claims (Stammbach et al., 2022) : an expert-annotated dataset for detecting real-world environmental claims. The dataset supports a binary classification task of whether a given sentence is an environmental claim or not.

- The 10K-files document (Velez-Calle and Robledo-Ardila, 2020) is a US-listed company's annual regulatory filings, labeled to indicate if the sentence contains climate-related risks that are material to their business.

The research aims to achieve the highest possible result in a binary text classification task focusing on the "Environmental Protection" category through various experiments.

Due to the hierarchical nature of the corpus, two methods were applied:

1. The one-step classification method is an approach that focuses solely on the category, ignoring the domain label. It involves preprocessing data before delivering it to a category identification model, which predicts whether a sentence belongs to the EP or non-EP category.

2. The two-step classification method involves the use of both domain and category labels. Figure 3 shows a model architecture utilizing the two-step classification method. The data is first pre-processed before being sent to the domain identification model, which is a binary model that predicts whether the domain of the sentence is WQL or non-WQL. The sentences predicted to have a WQL domain are filtered and sent to a category identification model, which determines if the sentence belongs to the positive EP or negative non-EP category.

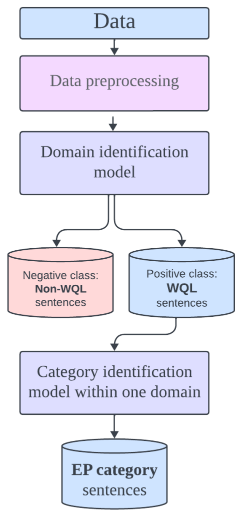

Figure 3: Model architecture utilizing the two-step classification method.

4 Experiments

The text classification model for identifying environmental protection contexts is developed through four steps:

1. Compare various balance combinations of the positive class by applying machine learning models to both methods. There is a need to

investigate how the same model would behave with different balance combinations of the positive class.

2. Select the top 3 balance combinations and apply a large pre-trained language model, XLM-RoBERTa (Conneau et al., 2020), to those 3 balance combinations for both methods. Select a model with the best performance considering both methods.

3. The process involves three optimization attempts of the best performing model: comparing pre-trained models, extracting confusing categories, and adjusting hyperparameters to improve performance.

4. The next step involves evaluating the cross-sector transferability of the best-performing model. The motivation is to test a model trained on political data on financial and environmental datasets.

4.1 Test set

It is important to note that the test set was consistent across all experiments. A fair test is produced when several models are contrasted on the same test set. This enables us to evaluate the performance of different models and make meaningful comparisons between them.

Therefore, the test set is created with adjusted sample percentages for the specific EP category while maintaining the overall balance of the categories as in the Manifesto Corpus itself. The test set contains 84860 sentences. Since environmental protection is the focus of the research, the test set was specially crafted to ensure it was well represented. Firstly, the number of annotated sentences for each category in the corpus is identified. Later, 0.08% of each category is added to a test set, with the exception of the EP category; the percentage for the target EP category is 0.15.

4.2 The first Step

Even though, with the introduction of large language models, traditional machine learning algorithms such as SVC are rarely used for text classification tasks nowadays, training and fine-tuning pre-trained deep learning models may be time-consuming.

The Linear Support Vector Classification (SVC) model, which performed the best of the 36 machine learning models that were initially tested, was used to compare balance combinations of a WQL domain or EP category.

№	EP %	Accuracy	F1 Score(Weighted)	Macro Average	Non-EP F1 score	EP F1 score
1	10	0.93	0.93	0.77	0.96	0.58
2	15	0.93	0.93	0.78	0.96	0.60
3	20	0.92	0.93	0.78	0.96	0.61
4	30	0.9	0.91	0.76	0.94	0.58
5	40	0.88	0.9	0.74	0.93	0.55
6	50	0.85	0.87	0.70	0.91	0.50
7	60	0.8	0.84	0.66	0.88	0.44
8	70	0.73	0.79	0.60	0.83	0.37

Table 2: "Environmental Protection" (EP) category balance experiments with one-step classification method conducted with Linear Support Vector Classification (SVC) model.

Preparing the text for analysis by removing noise, unimportant components, and inconsistent formatting is generally a crucial stage in the text classification process. Therefore, the first step in all experiments was data pre-processing: text input is converted to lowercase, non-alphanumeric characters are stripped except for a few punctuation marks, URLs and HTML tags are removed, various punctuation marks and special characters are stripped, stop-words are filtered out, and emojis are lastly removed.

4.2.1 One-step classification with SVC model

Table 2 depicts eight EP category balance experiments with a one-step classification method trained on a linear support vector classification (SVC) model. The table clearly illustrates that when the percentage of EP categories increases, the outcome decreases.

The dataset with the best-performing SVC model contains 15% sentences from the EP category, whereas the dataset with the worst-performing SVC model contains 70% sentences from the EP category. The F1 score columns for EP categories and non-EP categories reveal another clear trend on the table: all models perform better at identifying negative non-EP categories than positive EP categories, as the positive class has a significantly lower F1 score than the negative class.

4.2.2 Two-step classification with SVC model

The outcome of WQL domain balance experiments using the two-step classification method and the SVC model are displayed in Table 3. According to the table, the domain classification model also demonstrates a trend of the negative class having a higher F1 score. Furthermore, the table indicates that the best results are obtained when the percent-

№	WQL %	Accuracy	F1 Score(Weighted)	Macro Average	Non-WQL F1 score	WQL F1 score
2	10	0.70	0.63	0.55	0.81	0.30
2	15	0.73	0.69	0.63	0.82	0.43
3	20	0.75	0.72	0.67	0.83	0.52
4	30	0.77	0.76	0.72	0.83	0.61
5	40	0.77	0.77	0.74	0.83	0.66
6	50	0.76	0.76	0.74	0.80	0.67
7	60	0.72	0.73	0.72	0.76	0.67
8	70	0.67	0.67	0.67	0.69	0.65

Table 3: WQL domain balance experiments with two-step classification method conducted with SVC model.

№	EP %	Accuracy	F1 Score(Weighted)	Macro Average	Non-EP F1 score	EP F1 score
1	10	0.92	0.91	0.88	0.95	0.81
2	15	0.93	0.93	0.90	0.95	0.84
3	20	0.93	0.93	0.91	0.96	0.86
4	30	0.94	0.93	0.91	0.96	0.87
5	40	0.93	0.93	0.91	0.95	0.86
6	50	0.92	0.92	0.90	0.95	0.85
7	60	0.90	0.91	0.88	0.93	0.83
8	70	0.87	0.88	0.85	0.91	0.79

Table 4: EP category balance experiments with two-step classification method within the WQL domain conducted with SVC model.

age of the positive WQL domain class is between 30 and 50.

Table 4 shows the results of the second part of the two-step classification method, to be precise, the EP category balance experiments within the WQL domain with the two-step classification method were conducted with also the same SVC model. The table shows that the F1 scores are quite high for both the negative and positive classes, and that there is also little evidence of a trade in the balance problem between the two classes. The positive EP category receives an F1 score of 87%, compared to a maximum F1 score of 60% with one-step classification.

In a two-step classification method, the first step's accuracy is crucial as it narrows down the desired domain. Contrarily, as demonstrated in Tables 3 and 4, the accuracy of the second step is around 20% higher, but the accuracy of the first step is still critical in determining the efficacy of the two-step classification model as a whole. This is due to the fact that if the first step's model misclassifies positive WQL domain sentences and subsequently potentially EP category sentences, the second step's model will never receive those misclassified sentences as input because they were automatically classified as negative and filtered out by the first step's model.

№	EP %	Accuracy	F1 Score(Weighted)	Macro Average	Non-EP F1 score	EP F1 score
1	10	0.93	0.94	0.81	0.96	0.66
2	15	0.93	0.93	0.81	0.96	0.65
3	20	0.92	0.92	0.79	0.95	0.63

Table 5: The result of Environmental Protection (EP) category balance experiments with a one-step classification method conducted with XLM-RoBERTa model.

4.3 The second Step

Traditional machine learning models such as SVC are typically not as effective for text categorization problems as language models that have been pre-trained on a large corpus. Generally, pre-trained models are more preferable for natural language classification tasks as they have the capacity to understand intricate patterns and semantic representations from vast amounts of text input.

For the experiments of the second step, the pre-trained XLM-RoBERTa model (Conneau et al., 2020) was applied. XLM-RoBERTa is a cross-lingual pre-trained language model based on the BERT (Devlin et al., 2019) architecture, designed to understand and generate text in multiple languages.

Initially, 3 best performing balance combinations from the experiments of the first step were sorted out and the XLM-RoBERTa model was fine-tuned for those data collections. The main goal remains as it was in the first step: to test the balance combination of the target class, therefore the hyperparameters of the model did not change and stayed the same. The parameters of the model:

- Maximum sequence length: 151 or 135. There is an option because the data for tokenization changed depending on the balance combination.
- Training batch size: 16
- Learning rate: 2e-5.
- Optimizer: Adam optimization algorithm

4.3.1 One-step classification with XLM-RoBERTa model.

The previous one-step classification method experiment with the linear SVC model revealed that the model performed best when the percentages of EP categories were 10, 15, and 20. In order to enhance the results, these three balance combinations were compared using a XLM-RoBERTa pre-trained model.

№	WQL %	Accuracy	F1 Score(Weighted)	Macro Average	Non-WQL F1 score	WQL F1 score
1	30	0.80	0.80	0.77	0.85	0.69
2	40	0.80	0.80	0.78	0.84	0.71
3	50	0.78	0.79	0.78	0.83	0.72

Table 6: Results of WQL domain balance experiments using domain-specific XLM-RoBERTa model, stage one of the two-stage classification model.

№	EP %	Accuracy	F1 Score(Weighted)	Macro Average	Non-EP F1 score	EP F1 score
1	20	0.94	0.94	0.92	0.96	0.89
2	30	0.94	0.94	0.93	0.96	0.89
3	40	0.94	0.94	0.92	0.96	0.89

Table 7: Results of EP category balance experiments within one domain using category-specific XLM-RoBERTa model, stage two of the two-step classification model.

Table 5 illustrates positive EP category balance experiments with one-step classification method conducted with XLM-RoBERTa model. When the results of the one-step classification method using the linear SVC model and the RoBERTa model are compared, it is revealed that the RoBERTa model exceeds at classifing posive EP class.

4.3.2 Two-step classification with XLM-RoBERTa model.

Table 6 shows the outcomes of WQL domain balance experiments using the domain-specific XLM-RoBERTa model, which is the first part of the two-stage classification model. Table 7 shows the results of EP category balance experiments within one domain using the category-specific XLM-RoBERTa model.

The two-step classification model was not previously combined into a single model. Therefore, rhe best-performing XLM-RoBERTa models for domain and category classification are chosen in order to combine two steps into one model. According to the experiments, the best category-specific XLM-RoBERTa model is obtained when the percentage of the EP category is 30, and the best domain-specific XLM-RoBERTa model is obtained when the percentage of the WQL domain is 40. The table 8 shows that, in contrast to the one-step model, the combined two-step classification model performs significantly worse.

4.4 The third step: optimization

When all of the experiment results from the second step were compared, it was found that the XLM-RoBERTa model with the one-step classification method and 10% of the total EP category sentences as input was the most effective model.

WQL %	EP %	Accuracy	F1 Score(Weighted)	Macro Average	Non-EP F1 score	EP F1 score
40	30	0.69	0.61	0.53	0.80	0.26

Table 8: The outcome of two combined models: the WQL domain classification model and the EP category classification model.

Model	Accuracy	F1 Score(Weighted)	Macro Average	Non-EP F1 score	EP F1 score
XLM-RoBERTa base	0.93	0.94	0.81	0.96	0.66
RoBERTa base	0.93	0.94	0.81	0.96	0.65
BERT base	0.93	0.94	0.81	0.96	0.65

Table 9: The outcome of comparing pre-trained language experiments.

4.4.1 Comparing pre-trained language

For the balance experiments of the second step, only the XLM-RoBERTa model was applied. Thus, to improve the results of the best-performing model, three pre-trained language models were compared:

- BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) base model introduced the idea of bidirectional pre-training, where some input tokens were hidden and their predictions were made using context from both directions. With 110 million parameters, it has 12 layers of transformer encoders.

- RoBERTa (A Robustly Optimized BERT Pre-training Approach) base (Liu et al., 2019) base model is an improved of BERT which makes use of more training data and tweaks several hyperparameters. Additionally, dynamic masking is used. RoBERTa base is slightly bigger than BERT base model with 12 transformer layers and 125 million parameters.

- XLM-RoBERTa (Cross-lingual Language Model RoBERTa) (Conneau et al., 2020) base model is similar to RoBERTa in architecture but is trained on a variety of languages, making it effective for understanding text in multiple languages and it was earlier applied for the balance experiments of the second step.

The training dataset and hyperparameters were identical for all three of these language models in order to conduct a fair comparison. Moreover, only the base versions of these models were applied for the comparison. Table 9 shows the result of the comparative analysis of these three lanuage models. The results of these three language

Experiment	Accuracy	F1 Score(Weighted)	Macro Average	Non-EP F1 score	EP F1 score
Baseline	0.93	0.94	0.81	0.96	0.66
Sustainability removed	0.93	0.94	0.81	0.96	0.65

Table 10: The result of the experiment with Sustainability:Positive category derived.

models were evidently nearly identical, with the XLM-RoBERTa model's F1 score in the positive class being only 1% higher than the results of the other two models, as shown in the table. The XLM-RoBERTa model was subsequently used for the following optimization experiments and stayed as a baseline model.

4.4.2 Sustainability:Positive category

The Sustainability: Positive category in the codebook is defined as an appeal for sustainable economic growth and opposition to growth that harms society or the environment. This category is similar to our target category, environmental protection, based solely on description.

An experiment connected with Sustainability category was conducted. The goal of the experiment is to identify whether adding Sustainability: Positive as a non-EP class can be detrimental for the model as the similarity of the content can confuse the model. Table 10 shows the result of the experiment with Sustainability:Positive category derived. As can be seen from the table, generally, deriving the Sustainability:Positive category from the training dataset did not affect the result. The only difference is that the F1 score in the EP category has dropped by 1%.

To summarize, even though the descriptions of the two categories are similar, Sustainability:Positive category had no negative impact on the training process, and thus no changes to the baseline model will be made.

4.4.3 Hyperparameters optimization

As prior main focus of the experiments was binary data balance, it can be helpful to experiment with hyperparameters optimization of the best performing(baseline) model.

Table 11 shows the result of hyperparameters optimization experiments. The table clearly shows that various combinations with batch sizes ranging from 8 to 32, learning rates ranging from 1 e-5 to 3 e-5, and epoch numbers ranging from 2 to 4 were evaluated and compared for their performance.The baseline model's initial hyperparameters, however, continued to have the best performance. Therefore, it was decided not to change the baseline model and apply it for the fourth step.

4.5 The fourth step

In the fourth stage, two environmental datasets from the financial sector are used to evaluate the cross-sector transferability of a model trained on political data of the Manifesto corpus. To be more precise, the major objective is to determine whether the model can correctly categorize environmental statements in another sector.

4.5.1 Environmental claim detection dataset

The environmental claim detection dataset (Stammbach et al., 2022) is an expert-annotated dataset for detecting real-world environmental claims made by companies. The dataset supports a binary classification task that determines whether a given sentence is an environmental claim or not.

The dataset was created primarily using publicly available text data that was made available by companies and public databases. For example, a text from corporate annual reports, sustainability reports, and transcripts of earnings calls. Environmental claims aim to create a significant environmental benefit that is acceptable to the target audience, including consumers or stakeholders. Typically, an environmental claim is made to improve the environmental reputation of a business or a product.

The environmental claim detection dataset was used to evaluate the baseline model's realm portability. The environmental claim detection dataset's (Stammbach et al., 2022) authors divided it into train, test, and validation sets; however, for our needs, this division is unnecessary, so all three sets are combined into a single dataset with 2647 sentences. Example sentences of the environmental claim detection dataset:

- Positive sentence: "The project will make a significant contribution to the German and European hydrogen strategy and hence to achievement of the climate targets."
- Negative sentence: "TOTAL performs sensitivity tests to assess the ability of its asset portfolio to withstand an increase in the price per ton of CO_2."

4.5.2 The U.S. Securities and Exchange Commission (SEC) 10-K files dataset

The U.S. Securities and Exchange Commission

Experiment	Batch size	Epochs	Learning Rate	Accuracy	F1 Score(Weighted)	Macro Average	Non-EP sentence, F1	EP-sentence, F1
Baseline	16	3	2e-5	0.93	0.94	0.81	0.96	0.66
1	8	3	2e-5	0.93	0.94	0.80	0.96	0.64
2	32	3	2e-5	0.93	0.94	0.81	0.96	0.66
3	16	3	1e-5	0.93	0.94	0.81	0.96	0.65
4	16	3	1e-3	0.91	0.87	0.48	0.95	0.00
5	16	3	2e-3	0.91	0.87	0.48	0.95	0.00
6	16	3	3e-3	0.91	0.87	0.48	0.95	0.00
7	16	3	3e-5	0.93	0.93	0.80	0.96	0.64
8	16	2	2e-5	0.93	0.94	0.81	0.96	0.65
9	16	4	2e-5	0.93	0.94	0.81	0.96	0.65

Table 11: Hyperparameters optimization experiments.

Model №	Train dataset	Test dataset	Accuracy	F1 score	F1 score (weighted)	Macro Average	Precision	Recall	EP, F1 score	Non-EP, F1 score	EP, support	Non-EP, support
1	Manifesto	Manifesto	0.93	0.61	0.94	0.81	0.61	0.71	0.66	0.96	7475	77385
2	Manifesto	Environmental claims	0.77	0.62	0.78	0.73	0.52	0.77	0.62	0.83	665	1982
3	Environmental claims	Manifesto	0.91	0.43	0.91	0.69	0.49	0.39	0.43	0.95	7475	77385
4	Manifesto	10-K files	0.93	0.43	0.94	0.70	0.3	0.74	0.43	0.96	125	3175
5	10-K files	Manifesto	0.91	0	0.87	0.48	0	0	0.00	0.95	7475	77385

Table 12: The outcome of the cross-sector transferability testing on environmental claims and 10-K file datasets of the baseline model trained on the Manifesto dataset. Inversely trained models are also included.

(SEC) 10-K files dataset is annual regulatory filings in which listed companies in the US are required to self-identify climate-related risks that are material to their business.

SEC regulates the securities industry, including exchanges, brokers, dealers, investment advisors, and mutual funds (Velez-Calle and Robledo-Ardila, 2020). Its key regulatory requirement for publicly traded companies is the filing of financial reports, including the annual Form 10-K. Form 10-K is an annual report submitted by publicly traded companies to the SEC that details financial performance, business operations, risk factors, and Management discussion and analysis, which is critical for investors and analysts. Therefore, the dataset can be helpful for financial analysis, investment research, and other uses like identifying market trends and patterns. The 10-K file dataset, like the environmental claim detection dataset, was chosen to test the cross-sector transferability of the baseline model.

However, the Form 10-K is not an annotated document; however, the authors of the ClimaText dataset (Varini et al., 2021) annotated 3000 sentences from the relevant sections of the 10-K file for their Active Learning algorithm. For example, Item 1A of the 10-K files from 2014, as this is the relevant section in which climate risk must be reported, was annotated manually. Additionally, the authors created a test set using the 10-K files by randomly selecting 150 examples within the positive and negative predictions. Both sets were binary, with positive and negative labels determining whether or not the sentence's content was about climate change.

Finally, these two annotated sets of Form 10-K were combined into a single set of 3300 sentences to test the transferability of our baseline model. Example sentences of the 10-K files dataset:

- Positive sentence: "Increases in transportation or shipping costs, climate change regulation and other factors may negatively impact our results of operations."
- Negative sentence: "Price fluctuations in our common stock could result from general market and economic conditions and a variety of other factors."

4.5.3 Evaluation of cross-sector transferability

Table 12 displays the results of cross-sector transferability testing on environmental claims and 10-K file datasets of the baseline model trained on the Manifesto dataset. The table also includes the outcomes of inverse training. For instance, after the baseline model was evaluated using the Environmental Claims dataset, the model was subsequently trained using the Environmental Claims dataset, and the Manifesto dataset test set was utilized for evaluation. This inverse training method is employed for the 10-K file dataset too.

The significant capacity of the models to precisely detect and classify non-EP phrases or negative situations may be explained by the tendency of high performance on negative examples across all models and dataset scenarios. Accordingly, regardless of the dataset source or sector, there is a constant ability to decrease false positive predictions for negative cases, as seen by consistently high F1 scores and precision values over 0.9. This pattern highlights the models' accuracy in identifying and accurately categorizing non-EP phrases, which is an important competency in a variety of natural language processing tasks.

Models exhibit a consistent trade-off between precision and recall when tested on different datasets from their training data. As models improve recall, their precision decreases, while vice versa. This trade-off reflects the challenge of adapting models to different data sources or sectors, as they must balance minimizing false positives and capturing as many true positives as possible. This highlights the need for careful consideration when testing models on different datasets.

Moreover, it is important to note that the model trained on the 10-K file dataset had a recall and precision of 0 for positive examples, indicating it couldn't accurately identify or classify any positive instances in the "Manifesto" dataset. However, it excelled in identifying negative examples with a strong F1 score of 0.95.

5 Error analysis

It is essential to conduct an error analysis following a text categorization task that produced an unsatisfactory result. It aids in determining and comprehending the causes of the poor performance and offers suggestions for enhancing upcoming models. Therefore, the error analysis of the manifesto claims dataset has been conducted.

The 10-K files dataset was ignored as the model trained on this dataset could not predict any of the sentences in the positive EP class, thus the F1 score was 0. This suggests that the 10-K files dataset is not suitable for enhancing upcoming models.

5.1 Manifesto dataset

5.1.1 Repeated sentences with different labels

It was discovered that there were 209 repeating sentences in the Manifesto test set. 206 of these sentences had two repetitions, and the remaining 3 sentences had three repetitions.

The 159 repeated sentences had no direct negative effects on the testing procedure, except the inflated model performance and reduced diversity, because they were exact duplicates with the same domain and category labels. The rest of duplicated 50 sentences of the test set can be divided into two categories:

1. **Sentences with the same domain but different categories**: there are 15 sentences belonging to the same domain but having different categories.

   Most of them are very short sentences with just one or two general words. For example, "Justice" is a one-word sentence that is repeated twice in the test set. In both cases, it has the same 'Welfare and Quality of Life' domain; however, the categories differ, as in one case it is the 'Equality: Positive' category and in the other case the category is our target category, 'Environmental Protection'.

   Nevertheless, there are five repeated sentences that are more detailed and longer than three words. For instance,the sentence 'Determined support for local production is an important part of the solution to the current global crisis, due to the innumerable knock-on effects it exerts on other sectors, its added value, the quality employment it generates and the deep roots that certain productive activities keep with the Canary Islands.' was duplicated twice and was presented in two different categories: 'Economic Growth: Positive' and 'Protectionism: Positive'.

2. **Sentences with different domains and categories**: there are repeated 35 sentences with distinct domains and categories.

   Once more, the majority of the repeated sentences were fairly brief. For instance, the phrase 'Financing' was offered in two different labeling variants: the first form had the category 'Agriculture and Farmers: Positive' of the 'Social Groups' domain, while the second variant had the category 'Economic Planning' of the 'Economy' domain.

   Additionally, there were lengthy sentences like, "Basic moral regulations will be made to cover public administration, civil society, the media, and private enterprise in order to prevent corruption." presented in two versions, the first of which has the category "Political Corruption" of the "Political System" domain, and the second of which has the category "Traditional Morality: Positive" of the "Fabric of Society" domain.

The training set, however, did not experience the duplication error. Supposedly, as a result of the set's prior processing.

The reason

The duplication problem may have existed in the corpus from the start, or it may have appeared as a result of the shift in the Excel sheets being used for the translation process.

In total, the corpus contains 1623580 annotated sentences, and 63821 of those sentences were repeated. Investigation revealed that 50507 of these repeated sentences are exact duplicates with same annotation, indicating that the remaining 13314 repetitions fall under separate categories.

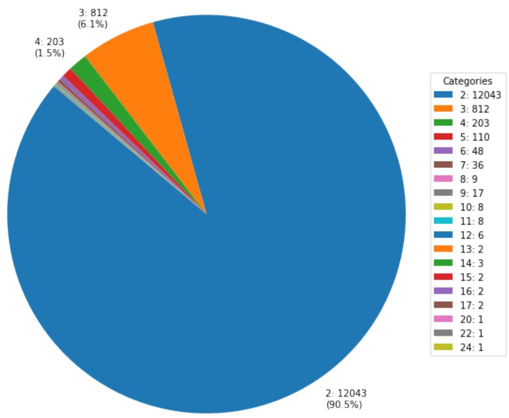

Figure 4: The distribution of duplicates in the Manifesto corpus based on how many times they are repeated.

The Figure 4 shows the distribution of duplication frequencies. It is clear from the graph that most of the differently annotated repetitions are binary duplication. Moreover, the graph shows that there are sentences that are annotated differently more than 20 times. For example, the sentence that was assigned 22 different labels is the one-word sentence "Transport". Despite the fact that the sentence can be used in a variety of contexts, combining sentences from the corpus could lead to model confusion.

The longest duplicate in the entire corpus contained the most intriguing repeat with various labels. Its 128 words were originally written in Spanish. Table 14 contains the original, the translation to English that was completed with Google Translate, labels, and supplementary data. The interesting fact is that it was described by two binary categories: 'Constitutionalism: Positive' and 'Constitutionalism: Negative'. Generally, the sentence was repeated three times by the same party for three different elections. The difference in the label could be the result of changes in strategic messaging by the party. Also, it could be the result of a wrong annotation. To investigate this issue, the whole manifesto should be examined.

5.1.2 Errors of baseline model, trained and tested on Manifesto dataset

When the model was tested on the 84860-sentence test set of the manifesto corpus after being trained on the Manifesto dataset, it predicted incorrectly 5586 sentences. As previously stated, the Manifesto test set contained 50 sentences that were labelled differently at least twice, and 12 of the 5586 incorrectly predicted sentences were among these 50 sentences.

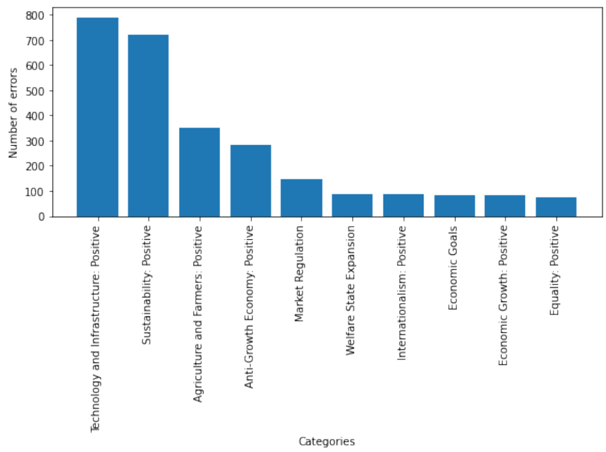

Figure 5: The Top 10 categories where the model mispredicted sentences as positive EP category (Manifesto test dataset).

False positive instances accounted for 3432 of the 5586 incorrectly predicted sentences, or 61 percent of the total. Figure 5 displays the most frequent 10 categories in which the model mispredicted sentences to belong to the positive EP category. The figure clearly shows that the "Technology and Infrastructure: Positive" and "Sustainability: Positive" categories account for the vast majority of false positive cases, when the model predicts sentences from these categories as positive EP. However, the experiment with removing the "Sustainability: Positive" category from the training set did not improve the model's overall outcome. This could be due to the fact that the "Sustainability: Positive" category was not also excluded from the test set.

Some examples of false positive instances:

1. Sentence: 'Production of energy-saving and environmentally friendly technology, biotechnology and more and more'.

   True label: "Technology and Infrastructure: Positive".

   Assumption: The creation of environmentally friendly technology and biotechnology, which are closely tied to environmental protection, are included in this clause. It's possible that the model misclassified it as a technological statement despite the fact that it also raises environmental issues.

2. Sentence: 'Our energy consumption must be reduced'.

   True label: 'Sustainability: Positive'.

   Assumption: The idea of reducing energy consumption frequently arises in both the sustainability and environmental protection categories, and these two categories are very similar and overlap significantly. The model may have placed more emphasis on the widely held idea of reducing energy consumption, which resulted in a false-positive prediction in the environmental protection category.

3. Sentence: " Let's stop climate change, protect our common environment".

   True label: 'Economic Planning'.

   Assumption: Both "stop climate change" and "protect the environment," which are important goals of environmental protection activities, are directly urged in the sentence. These significant elements may have emphasized these strong environmental cues.

4. Sentence: 'In parliament, only the PVDA spoke out against the introduction of the carbon tax'.

   True label: 'Political Authority: Party Competence'.

   Assumption: The term "carbon tax" is frequently used in connection with environmental policy because its traditional objectives include lowering carbon emissions and preserving the environment. It's possible that the model became biased with this environmental word and incorrectly labeled it as positive EP category.

5. Sentence: 'Reduce CO2 emissions from agriculture'.

   True label:'Agriculture and Farmers: Positive'

   Assumption: Reducing CO2 emissions, a crucial component of environmental conservation, is specifically mentioned in the sentence. This environmental problem may have received more priority in the model's prediction while ignoring "agriculture" in the sentence.

6. Sentence: 'We will replace old polluting vehicles by creating a bonus of 1000 euros to buy a new or less polluting second-hand vehicle'.

   True label: 'Sustainability: Positive'

   Assumption: In order to achieve sustainability and environmental goals, the statement discusses the replacement of polluting cars with alternatives that are less polluting. The model might have established a strong link between the EP category and an environmental objective.

Although not all false positive occurrences have been investigated, certain inferences may be drawn from the examples of false positive instances. All of these false positive examples may be assigned to the model's predictions due to its sensitivity to environmental keywords and ideas within the sentences, even though broader context may also include other topics.

False negative instances made up 2154 of the 5586 incorrectly predicted sentences, or 39 percent of the total.

Some examples of false negative error instances:

1. Sentence: 'Rural and agricultural policy'

   Assumption: Despite the clear link between rural and agricultural policies and environmental protection, the model may have missed the indirect reference to environmental concerns in this sentence because it lacked environmental keywords.

2. Sentence: 'What problems do we consider to be key?'

Assumption: This sentence appears to be a general question and does not contain explicit environmental terms.

3. Sentence: 'The financing will be realized partly with budget funds and partly with foreign donations.'

   Assumption: While this sentence discusses financing, it lacks explicit environmental terminology. The model, on the other hand, may have expected some mention of environmental funding and conservation efforts, resulting in a false negative prediction.

4. Sentence: 'The main sectoral proposals can be summarized as follows:'

   Assumption: Without specifically referencing environmental concerns, this statement looks to be the beginning of a description of sectoral ideas. Given that it is questionable whether this sentence belongs in the EP category, the model's error in labeling it as such may be the cause.

In these false negative instances, the lack of clear environmental phrases or context in the sentences may be to blame for the model's false negative mistakes. The sentences seem to cover broader topics or lack particular environmental signals, which causes the positive EP class sentences to be misclassified as negative class.

5.2 Environmental claim detection dataset

The environmental claim detection dataset did not have a duplication issue. When the baseline model was applied to the claim detection dataset, 620 of the 2647 sentences were incorrectly predicted.

False positive instances accounted for 155 of the 620 incorrectly predicted sentences, or 25 percent of the total. Some examples of false positive instances:

1. Sentence: 'In this way, it offers all occupants a more comfortable and less tiring ride and enables them to respond quickly to environmental changes.'

   Assumption: Due to phrases like "respond quickly to environmental changes," which can be viewed as a positive aspect related to adaptability to environmental conditions, the sentence may have been interpreted positively by the model. Nevertheless, the sentence's broader context places more emphasis on user comfort and adaptability than it does on an explicit environmental argument.

2. Sentence: "So we've had a primary focus on helping the OEMs meet their fuel economy/CO2 emissions requirements."

   Assumption: The model might have been misled into assuming that the line is addressing effective measures taken to address environmental issues because it included "helping original equipment manufacturers (OEMs) meet their fuel economy and CO2 emissions requirements". However, rather than presenting a clear environmental argument, the main emphasis here is on economic goals and assistance for OEMs.

3. Sentence: "NIB reports on the impact of the projects that are financed by the proceeds of NIB Environmental Bonds on an annual basis."

   Assumption: The model may have classified "NIB Environmental Bonds" expression positively, but the sentence mainly pertains to financial transaction reporting, with the environmental aspect being part of the reporting process.

4. Sentence: "The aim of the acquisition is to expand the range of polycarbonate materials for major industries to include composites made from continuous fiber-reinforced thermoplastics."

   Assumption: The sentence may be seen as positive as it discusses expanding materials for major industries, potentially promoting industrial growth, but it primarily focuses on industrial material expansion.

5. Sentence: "Companies' climate strategy such as ambitious decarbonisation targets are also outside the scope of the model."

   Assumption: The sentence suggests ambitious decarbonization targets, potentially promoting climate action and environmental goals, but it clarifies that these topics are not part of the model's analysis.

6. Sentence: "Planning for these works includes ensuring that the cooling infrastructure meets

potential future needs in a climate change impacted future."

Assumption: The model could classify the sentence as positive due to its consideration of future climate-related needs, but it primarily discusses infrastructure planning in the context of climate change, without explicitly claiming environmental impact.

According to these false positive examples, the model's incorrect classification as positive may be due to its sensitivity to environmental keywords and concepts, despite the sentences' broader context or intent not necessarily aligning with explicit environmental claims.

False negative instances made up 465 of the 620 incorrectly predicted sentences, or 75% percent of the total. Some examples of false negative error instances:

1. Sentence: "Environmentally and socially responsible supply chains – Sustainability To meet these challenges, Kering has established a dedicated organization.".

   Assumption: The sentence mentions environmentally and socially responsible supply chains and sustainability initiatives.

2. Sentence: 'The CO2 savings achieved in recent years have proved to be much larger than expected.'

   Assumption: The sentence discusses CO2 savings, which are linked to environmental claims about reducing carbon emissions.

3. Sentence: 'We are developing new solutions that meet increasing customer demand for 24/7 renewable power and greater energy efficiency.'

   Assumption: The sentence discusses developing renewable power and energy efficiency solutions to meet customer demand, addressing environmental claims and sustainability goals.

4. Sentence: 'The three-year project aims to improve sustainability in the mint supply chain and help farmers optimize their farm management in a sustainable way.'

   Assumption: The sentence emphasizes improving sustainability in mint supply chains and optimizing farm management.

The model's false negative errors are likely due to its inability to recognize the significant environmental cues in sentences that explicitly mention sustainability.

5.3 Intersection of errors

Two models are evaluated on the Manifesto test set, one trained on Manifesto and the other on environmental claims datasets. Given that the test set is the same, analysis may be done on the common mistakes that exist in both sets.

When the model was trained on the Environmental Claims dataset and tested on the Manifesto test set, 7647 out of the total 84860 sentences were wrongly predicted. Later, these 7647 misclassified sentences were compared with 5586 errors in the model trained on the Manifesto dataset. It was revealed that 3309 of these misclassified sentences are common. There are 1414 false-positive cases among the total of 3309 misclassified sentences.

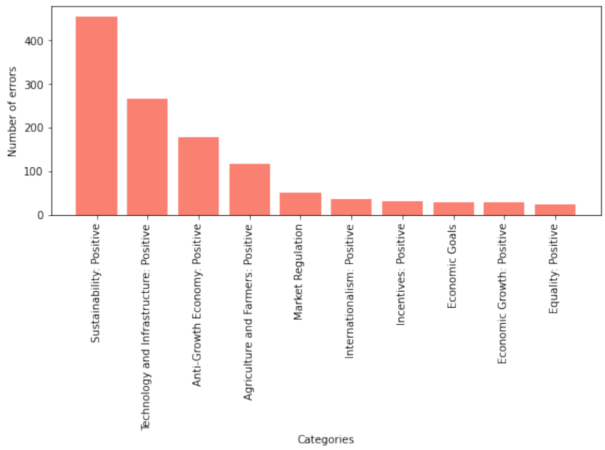

Figure 6: The Top 10 categories where the model mispredicted sentences as positive EP category(Intersection of errors).

Figure 6 shows the ten categories that were incorrectly predicted as the positive EP category the most frequently when only the error intersection of two models was used. Figures 6 and 5 appear to be very similar, highlighting the need to investigate the categories that these two models frequently misinterpret as the positive EP category.

6 Discussion

The baseline model was developed through a series of tests designed to improve the model's output.The Manifesto corpus evaluation of the baseline model did not yield particularly impressive results, and when the baseline model was tested for cross-sector transferability on additional datasets, the results drastically decreased.

There are numerous ways to possibly enhance the model's output:

1. Short sentences should not be used because they are too general, lack specificity, and can be used for any category, which could lead to confusion or misunderstanding. Particularly one-word sentences can be very perplexing and context-free.

2. Duplication errors should be considered before creating the datasets.

3. The error analysis also revealed that there are certain categories that make up the majority of the false error cases. It would be interesting to investigate these categories further to understand the underlying causes of the false error cases.

4. The specific annotation method for the Manifesto corpus should be considered, as sentences are sometimes labeled incorrectly solely because of neighboring sentences.

7 Conclusion

The hierarchical data of the Manifesto corpus was investigated in this paper, as well as the development of an environmental agenda detection model trained on this corpus. Several experiments were carried out to determine the best-performing model. The study investigated the proportion of positive classes, employed one-step and two-step classification approaches, and conducted optimization experiments. Despite numerous experiments, even the top-performing model had some efficiency limitations, with the F1 score reaching only 61 percent.

The study reveals a consistent trend of high performance on negative examples across various model and dataset scenarios, emphasizing the need to address model biases and ensure robust performance.

In conclusion, the evaluations on corpus and cross-sector datasets lacked high accuracy standards, but error analysis revealed possible areas for improvement and additional study.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pola Lehmann, Tobias Burst, Theres MatthieÃŸ, Sven Regel, Andrea Volkens, Bernhard WeÃŸels, and Lisa Zehnter. 2022. The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2022a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2022. A dataset for detecting real-world environmental claims.

Francesco S. Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2021. Climatext: A dataset for climate change topic detection.

Andres Velez-Calle and Cristina Robledo-Ardila. 2020. Exploring the u.s. securities and exchange commission's edgar database by sampling joint venture contracts. *International Journal of Disclosure and Governance*, 17.

Andrea Volkens, Pola Lehmann, Theres MatthieÃŸ, Nicolas Merz, Sven Regel, and Bernhard WeÃŸels. 2017. The manifesto data collection. manifesto project (mrg/cmp/marpor). version 2017a.

Tobias / Krause Werner / Lehmann Pola / Matthieß Theres / Merz Nicolas / Regel Sven / Weßels Bernhard / Zehnter Lisa Volkens, Andrea / Burst. 2020. The manifesto project dataset - codebook. manifesto project (mrg / cmp / marpor). version 2020b.

Appendix

A Reflections

Due to the lack of a baseline to follow, the studies were carried out in an exploratory manner. Therefore, after all experiences have been taken into account, there are various ways to improve the outcome:

1. Unfortunately, the most recent version of the Manifesto Project Dataset (version 2023a) was released after most of the experiments had been completed. In the new version of the corpus, the number of manifestos rose to 5089, and additionally, the Codebook for Political Parties was also published. With more annotated documents, the model's accuracy may improve.

2. Modern Python hyperparameter optimization libraries, like Ray Tune or Optuna, enable a variety of search techniques and parallelisms and are built for performance and versatility. However, it was not possible to use those libraries on the complete dataset due to the GPU's limited resources. Manual hyperparameter optimization was the sole choice because the dataset's balance played a significant role in the training process. Compared to using those libraries, this method may be more time-consuming and ineffective.

3. Unfortunately, due of restricted resources, I was unable to select up-to-date huge generative models. The first reason is that the GPU available at the institution only has 12 GB of random-access memory, making it incapable of running big models trained on billions of pieces of data, such as GPT or LLaMA. It would be very fascinating to test this balance combination with such powerful models.

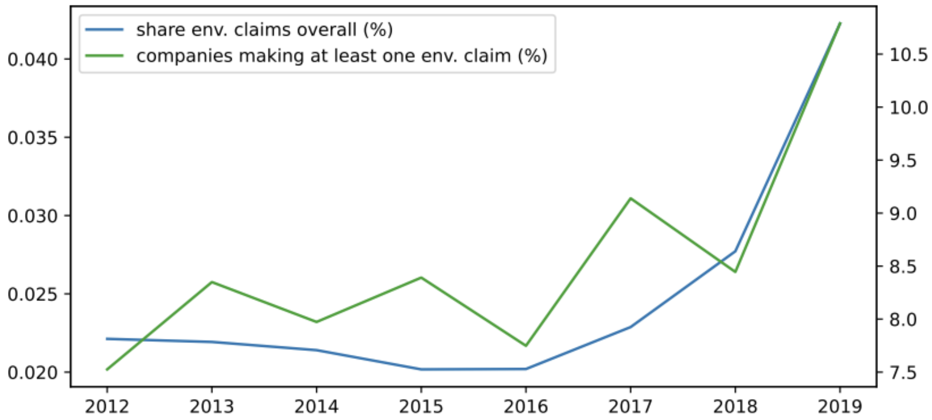

Figure 7: Amount of environmental claims (in %) made in earning calls answer sections. The blue line (y-axis on the left) shows the share of environmental claims made each year. The green line shows the share of companies making at least one environmental claim in a given year (y-axis on the right).

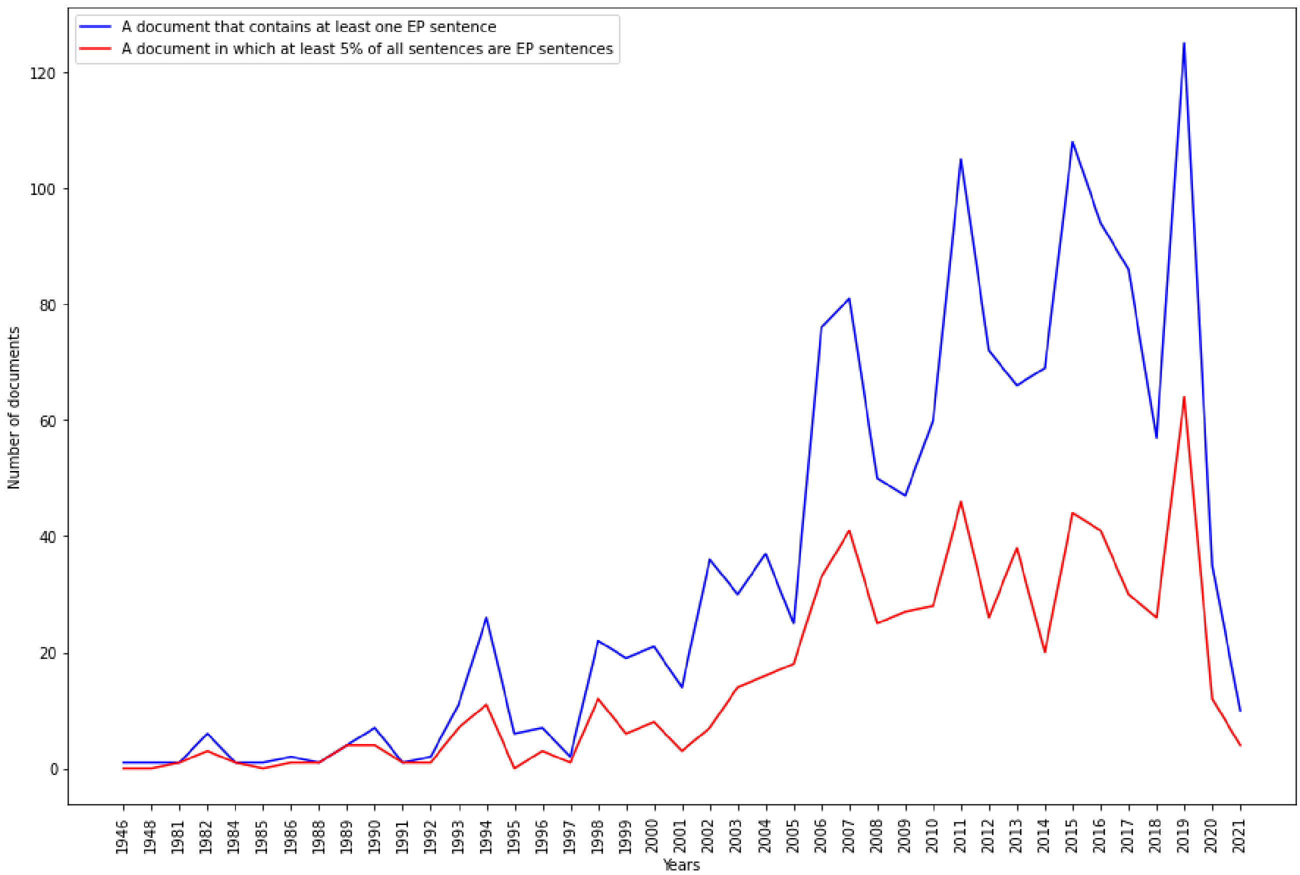

Figure 8: The frequency with which the environmental protection category appears in electoral campaigns. The blue line (y-axis) shows the number of documents that contain at least one EP category sentence. The red line depicts the documents in which at least 5% of all sentences are EP category sentences.

Domain 1. Welfare and Quality of Life:	**Domain 4. Economy:**	**Domain 7. Fabric of Society:**
Environmental Protection	Free Market Economy	National Way of Life: Positive
Culture: Positive	Incentives: Positive	National Way of Life: Negative
Equality: Positive	Market Regulation	Traditional Morality: Positive
Welfare State Expansion	Economic Planning	Traditional Morality: Negative
Welfare State Limitation	Corporatism/Mixed Economy	Law and Order: Positive
Education Expansion	Protectionism: Positive	Civic Mindedness: Positive
Education Limitation	Protectionism: Negative	Multiculturalism: Positive
Private-Public Mix in Culture: Positive	Economic Goals	Multiculturalism: Negative
Private-Public Mix in Social Justice: Positive	Keynesian Demand Management	The Karabakh Issue: Positive
Private-Public Mix in Welfare: Positive	Economic Growth: Positive	Rebuilding the USSR: Positive
Private-Public Mix in Education: Positive	Technology and Infrastructure: Positive	National Security: Positive
Domain 2. External Relations:	Controlled Economy	Cyprus Issue
Foreign Special Relationships: Positive	Nationalisation	General Crisis
Foreign Special Relationships: Negative	Economic Orthodoxy	Cultural Autonomy: Positive
Anti-Imperialism	Marxist Analysis	Multiculturalism pro Roma: Positive
Military: Positive	Anti-Growth Economy: Positive	Multiculturalism pro Roma: Negative
Military: Negative	Privatisation: Positive	National Way of Life General: Positive
Peace	Control of Economy: Negative	National Way of Life: Immigration: Negative
Internationalism: Positive	Property-Restitution: Positive	National Way of Life General: Negative
European Community/Union: Positive	Privatisation Vouchers: Positive	National Way of Life: Immigration: Positive
Internationalism: Negative	Social Ownership: Positive	Law and Order: Negative
European Community/Union: Negative	Mixed Economy: Positive	Civic Mindedness General: Positive
Russia/USSR/CIS: Positive	Publicly-Owned Industry: Positive	Civic Mindedness: Bottom-Up Activism
Western States: Positive	Socialist Property: Positive	Multiculturalism General: Positive
Eastern European Countries: Positive	Property-Restitution: Negative	Multiculturalism: Immigrants Diversity
Baltic States: Positive	Privatisation: Negative	Multiculturalism: Indigenous rights: Positive
Nordic Council: Positive	Sustainability: Positive	Multiculturalism General: Negative
SFR Yugoslavia: Positive	**Domain 5. Social Groups:**	Multiculturalism: Immigrants Assimilation
Russia/USSR/CIS: Negative	Labour Groups: Positive	Multiculturalism: Indigenous rights: Negative
Western States: Negative	Labour Groups: Negative	**NA: No other category applies**
East European Countries: Negative	Agriculture and Farmers: Positive	
Baltic States: Negative	Middle Class and Professional Groups	
Nordic Council: Negative	Underprivileged Minority Groups	
SFR Yugoslavia: Negative	Non-economic Demographic Groups	
Russian Army: Negative	Minorities Inland:Positive	
Independence: Positive	Minorities Abroad: Positive	
Rights of Nations: Positive	War Participants: Positive	
Anti-Imperialism: State Centred Anti-Imperialism	Refugees: Positive	
Anti-Imperialism: Foreign Financial Influence	Agriculture and Farmers: Negative	
Domain 3. Political System:	**Domain 6. Freedom and Democracy:**	
Political System	Freedom and Human Rights	
Decentralization	Democracy	
Centralisation	Constitutionalism: Positive	
Governmental and Administrative Efficiency	Constitutionalism: Negative	
Political Corruption	Transition to Democracy	
Political Authority	Restrictive Citizenship: Positive	
Republican Powers: Positive	Lax Citizenship: Positive	
Public Situation: Negative	Presidential Regime: Positive	
Communist: Negative	Republic: Positive	
Rehabilitation and Compensation: Positive	Checks and Balances: Positive	
Political Coalitions: Positive	Monarchy: Positive	
Political Authority: Party Competence	Freedom	
Political Authority: Personal Competence	Human Rights	
Political Authority: Strong government	Democracy General: Positive	
Transition: Pre-Democratic Elites: Positive	Democracy General: Negative	
Transition: Pre-Democratic Elites: Negative	Representative Democracy: Positive	
Transition: Rehabilitation and Compensation	Direct Democracy: Positive	

Table 13: Coding scheme of the Comparative Manifestos Project: seven domains and categories within these domains.

Original sentence (Spanish)	La Constitución del 78 recoge derechos como: el derecho a un trabajo suficientemente remunerado (art. 35), el derecho al acceso a la cultura (art. 44), derecho a una vivienda digna y adecuada (art. 47) y que los poderes públicos “establecerán los medios que faciliten el acceso de los trabajadores a la propiedad de los medios de producción” (art. 129) y que “el Estado, mediante Ley, podrá planificar la actividad económica general para atender a las necesidades colectivas, equilibrar y armonizar el desarrollo regional y sectorial y estimular el crecimiento de la renta y de la riqueza y su más justa distribución” (art. 131.1).
Translation	The Constitution of ’78 includes rights such as: the right to a sufficiently remunerated job (art. 35), the right to access to culture (art. 44), the right to decent and adequate housing (art. 47) and that the powers public “shall establish the means that facilitate workers’ access to ownership of the means of production” (art. 129) and that “the State, through Law, may plan general economic activity to meet collective needs, balance and harmonize regional and sectoral development and stimulate the growth of income and wealth and their fairer distribution” (art. 131.1).
Labels	’Constitutionalism: Positive’ and ’Constitutionalism: Negative’
Election dates	Novermber 2011; April 2009; November 2019
Party ID	33220

Table 14: The longest duplicate in the corpus with two different labels.