Analysis of the National Energy and Climate Plans of EU member states using Natural Language Processing (NLP)

MSc Bioinformatics and Biostatistics (UOC – UB)
Statistical Bioinformatics and Machine Learning

Adrian Carrascosa Lopez

Iván Contreras
January 2024

FINAL WORK SHEET

Work title	Analysis of the National Energy and Climate Plans of EU member states using Natural Language Processing (NLP)
Author's name	Adrian Carrascosa Lopez
Consultor's name:	Iván Contreras
PRA's name:	Carles Ventura
Tutors at BSC:	Eulàlia Baulenas, Paula Checchia, Mercè Crosas
Delivery Date:	01/2024
Program:	MSc Bioinformatics and Biostatistics
Work Area:	Statistical Bioinformatics and Machine Learning
Work language:	English
Keywords:	Natural Language Processing, Machine Learning, National Energy and Climate Plans

Resum del Treball

Durant els últims anys, el Processament del Llenguatge Natural (NLP) ha experimentat grans avenços gràcies al ràpid desenvolupament de models de *Deep-Learning* i tècniques de processament de text. Aquests avenços han revolucionat la forma en la que interactuem els humans amb les màquines, permetent una comprensió més profunda i contextual.

En aquest context, el present treball es centra en utilitzar aquesta tecnologia de vanguardia per analitzar els Plans Nacionals d'Energia i Clima, uns documents on cada país de la Unió Europea presenta com a full de ruta per tal de complir una sèrie d'objectius i mesures amb data 2030.

L'anàlisi consta, per una banda, d'un primer anàlisi de tipus *bottom-up* on s'identifiquen els principals tòpics que tracten els plans europeus, i per altra banda, un anàlisi *top-down* on classifica i qualifica cada país segons discursos preestablerts.

Un cop fet l'anàlisi, s'identifiquen les estratègies que opta cada país per afrontar els objectius europeus i, mitjançant una feina inductiva, s'avalua la coherència i exactitud del propi algoritme creat.

Abstract

During recent years, Natural Language Processing (NLP) has experienced significant advancements thanks to the fast development of *Deep-Learning* models and text processing techniques. These advancements have revolutionized the way humans interact with machines, enabling a deeper and more contextual understanding.

In this context, the present work focuses on leveraging this cutting-edge technology to analyse de National Energy and Climate Plans (NECP), documents where each European Union country outlines its roadmap to achieve a serie of objectives and measures by the year 2030.

The analysis comprises, on the one hand, a *bottom-up* analysis where the main topics addressed in the European plans are identified, and on the other hand, a *top-down* analysis that classifies and assesses each country according to predefined discourses typologies.

Once the analysis is completed, the strategies adopted by each country to meet European objectives are identified, and through and inductive approach, the coherence and accuracy of the algorithm developed are evaluated.

Table of content

I. List of figures

II. List of tables

III. Glossary

NLP – Natural Language Processing

ML – Machine Learning

AI – Artificial Intelligence

NECP – National Energy and Climate Plan

EU – European Union

LDA – Latent Dirichlet Allocation

TF-IDF – Term Frequency – Inverse Document Frequency

BoW – Bag of Words

IR – Information Retrieval

EDDI – Environmental Discourse Dataset Index

ADM_RAT – Administrative Rationalism

DEM_PRA – Democratic Pragmatism

ECO_RAT – Ecological Rationalism

GRE_RAD – Green Radicalism

SURV – Survivalism

SUST – Sustainability

LLM – Large Language Model

DL – Deep Learning

1- Introduction

1.1 – Context and justification

During the last decades, the climate change situation has led many countries around the world to question their own economic model, increasingly opting for renewable energies of more efficient energy management plans. Extreme phenomena such as droughts, floods, or the rising of the sea level are direct consequences of this global warming, primarily caused by human industrial actions.

Within the European framework, alignment of action among its member states is crucial for the proper and optimal development of actions and measures. In order to achieve an ecological transition, the European Commission launched the National Energy and Climate Plans (NECP), plans in which member states were required to outline their objectives and action measures to achieve climate neutrality by 2050.

Although these plans have a common structure and policy, each member country is free to implement the transition measures in a freely according to its needs and possibilities, as long as they are plausible. The discursive method and the terms used in NECPs also provide us with indirect information about what perspective takes each state in a specific theme.

Social sciences are responsible for studying the perspectives embraced by each climate discourse, and in this thesis, the analytical approach of discourse is linked to language processing techniques such as Natural Language Processing (NLP).

Whilst discourse analysis has a long presence in social science research and also in the study of energy policy, its implementation in research has been greatly explored through single case studies or small-N comparative studies, which make difficult the cumulation of knowledge with regards discourses, practices and their influence in policy-making and change (Hajer & Versteeg, 2005) and (Berrang-Ford et al., 2021).

One of the methods used for discourse analysis is content analysis, nonetheless, which could be a tool used with a large number of observations systematically and to explore open hypotheses about the effects of discourse on social and political systems. Therefore, one of the first research gaps this thesis aims to address is the lack of comparative studies using discourse analysis.

In addition, few studies have merged the discourse analysis with NLP, which emerges as a cutting-edge technology in many of the daily-use applications that are indispensable for a large part of the population. This technology, defined as a branch of artificial intelligence focused on machines understanding and generation of human language, brings significant benefits to discourse analysis, turning it a quick and effective tool for analysing large amounts of texts.

1.2 – Objectives

General objective

The main objective of this work is to develop an automated tool for the systematic analysis of Energy and Climate Plans across European countries using Natural Language Processing techniques.

Specific objectives

- Conduct a comprehensive bottom-up analysis of individual sections identifying thematic categories and strategies outlined in the NECPs.
- Implement a top-down analysis of NECPs using predefined linguistic patterns of relevant environmental discourses.

1.3 – Impact on sustainability, ethical-social and diversity

The present master's thesis is framed within the values mentioned in the UOC's cross-cutting guide on the transversal competency "Ethical and Global Commitment". Its starting point is to act in an honest, ethical, sustainable, socially responsible, and respectful manner towards human rights and diversity, aiming to design solutions for the improvement of these practices.

Secondly, the context in which this work is situated, the European plans for energy and climate, makes the research encompass sustainable objectives such as sustainability (affordable and clean energy, sustainable cities and communities, or climate action), or the ethical behaviour and social responsibility (no poverty, decent work and economic growth, or clean water and sanitation). In addition, thanks to this automated tool, could help the team to understand what the NECPs are talking about, making comparisons between them and also classify them into some predefined categories.

Finally, the analysis of these lines of work in terms of climate and energy action can have a direct impact on identifying the strengths and weaknesses of European objectives and measures, regarding the climate issue while finding opportunities for improvement.

1.4 – Approach and methodology pursued

The methodology followed in this work started by identifying the target documents to analyse, and defining the pipeline that the team should follow.

In a first instance, the pipeline had four basic steps: 1) text pre-processing, 2) vectorization, 3) clustering and finally 4) visualization. Once these steps were achieved, new scopes needed to be met, so an identification of the environmental discourses was done while building each discourse dictionary dataset, which will be explained with more detail in further chapters of this work.

Finally, once the previous steps were completed, visualization of the most representative results were done.

1.5 – Work planning

The work planning carried out during this master thesis is represented in the following calendar:

Thesis calendar	Start	End	Delivery	Duration (days)
Scope definition				
Identify project objectives	01/10/2023	10/10/2023		10
Conduct requirements analysis	01/10/2023	20/10/2023		20
Define deliverables	01/10/2023	20/10/2023		20
Research and data collection				
Collect NECPs	10/10/2023	15/10/2023		6
Search best practices in NLP	10/10/2023	01/11/2023		23
Data preprocessing				
Search preprocessing techniques	15/10/2023	01/11/2023		18
Coding (python)	15/10/2023	07/11/2023		24
Vectorization				
Find most suitable word embeddings	25/10/2023	10/11/2023		17
Coding	25/10/2023	15/11/2023		22
Visualization				
Dimensionality reduction techniques	25/10/2023	15/11/2023		22
Clustering techniques	25/10/2023	25/11/2023		32
Modelling				
Collect most suitable NLP model	01/11/2023	25/11/2023		25
Model training	07/11/2023	01/12/2023		25
Adjustments and testing	15/11/2023	07/12/2023		23
Analysis				
Bottom-up analysis	01/12/2023	20/12/2023		20
Develop discourse dictionary	15/12/2023	01/01/2024		18
Top-down analysis	15/12/2023	07/01/2024		24
Other				
Literature research	01/10/2023	31/12/2023		92
Memory writing	01/12/2023	14/01/2024		45
Prepare presentation	14/01/2024	02/02/2024		20
Deliverables				
PAC1 (Definition and working plan)			16/10/2023	
PAC2 (Work development)			20/11/2023	
PAC3 (Work development)			23/12/2023	
PAC4 (Memory closing)			14/01/2024	

***Table 1**: work planning calendar*

2- State of art

2.1 – National Energy and Climate Plans (NECPs)

During the last years, the European Union has been implementing a significant transformation in its energy policies and environmental impacts. This evolution culminated in European Commission's decision in 2020 to introduce a comprehensive strategy involving all member states of the Union. The National Energy and Climate Plans (NECPs) represent a big step toward the realization of a so-called "zero-carbon" economy. (Maris & Flouros, 2021)

Each European plan is elaborated with a set of key objectives in mind: firstly, we find the *Decarbonisation Dimension*, which explains how the nation will become a carbon-neutral country by 2050. Many countries include the reduction of Green House Gases (GHG) emissions and the impulse of the renewable energy as the main ways to achieve this objective.

Within the *Energy Efficiency Dimension,* nations describe how energy will be implemented more efficiently in primary sectors and particularly in transport and industry. Moreover, in the *Energy Security Dimension*, each country ensures the diversification of the energy and guarantee security on the supply of itself.

Regarding the Internal *Energy Market Dimension*, it promotes a more competitive, flexible, and transparent energy market to promote the cross-border trade.

Finally, *Research & Innovation Dimension* focuses on how to reach these energy and climate targets, describing national budgets to research and innovation and how this support will impact in clean energy technologies.

Having these dimensions of the plans correctly identified, a comprehensive analysis will be conducted to find which are the most relevant characteristics to consider and to find out which is the already existing discourse style that aligns most effectively with the text.

2.2 – Types of discourses

Since the Industrial Revolution, politics in Earth has featured a wide range of issues and changes regarding concerns with natural resources, pollution, and population growth. Over the last decades, other concerns like climate change, energy supply, environmental justice or species extinction have joined the previous issues, raising numerous questions about the ways in which humans must interact with the planet.

As John S. Dryzek writes in his book "The politics on Earth – Environmental discourses", (John S. Dryzek, 2022) environmental issues do not present themselves in a well-defined and labelled box. Instead, most of them are interconnected and multidimensional. In addition, the more complex is a situation, the greater number of perspectives and plausible opinions. Here is where are born environmental discourses.

In his book, Dryzek defines two different dimensions in order to finally obtain four main environmental discourses. The first dimension is to differentiate if a discourse is *prosaic* or *imaginative*. Prosaic departure is seen mainly in terms of troubles encountered by the established industrial political economy. On the other hand, imaginative discourse seeks to redefine the chessboard, seen more opportunities than troubles and not harmonizing the environmental concerns with the economic ones.

The second dimension defined by Dryzek was to differentiate between *reformist* and *radical*. While the reformist perspective seeks to make changes within the existing system and making gradual reforms and pragmatic policies to address environmental issues, radical discourse focus on transformative changes in the system to achieve rapid and significant changes.

Combining these two dimensions, we obtain the following four environmental discourses with their respective definitions:

	Reformist	**Radical**
Prosaic	Problem solving	Survivalism
Imaginative	Sustainability	Green radicalism

***Table 2**: Dryzek's four types of environmental discourses*

Environmental problem solving

It is defined by taking the political economic status to cope with environmental problems, specially via public policies. Such adjustments have three focus areas: administrative rationalism, democratic pragmatism, and economic rationalism. These three approaches of environmental problem solving will be considered separately in the experiments conducted in this work.

Administrative rationalism

This discourse consists of institutionalizing environmental expertise in its operational procedures and emphasizing the role or the expert rather than the citizen or the producer (Takahashi, 2020).

A politically neutral and professionalised bureaucracy must produce accurate public policy decisions to solve environmental problems, which are considered as technical in nature: problems and solutions must be segmented into a "reactive, tactical, piecemeal and end-of-pipe" approach (John S. Dryzek, 2022).

Democratic pragmatism

Democratic pragmatism takes profit of problem-solving capacities of liberal democratic governments to facilitate environmental actions. The word "pragmatism" gets two different connotations within this environmental discourse: the first way is about solving problems in a world full of uncertainty, and the second one is used as a realistic orientation of the world. Problem solving must be a flexible process involving many voices and cooperation across a plurality of perspectives (John S. Dryzek, 2022).

Economic rationalism

Its main idea is to put price tags on environmental harms and benefits, while having policy measures including emissions trading, environmental taxes, and households waste fees and deposit systems in what waste management is concerned (Takahashi, 2020). This tendency expects to reduce waste generation increasing recyclable products and resource efficiency by extending the producer responsibility, requiring the coverage of the production and recycling costs.

Survivalism

Its basic idea is that unmeasured population growth and economic expansion will eventually exceed the Earth's natural resources and the capacity of ecosystems to support human industrial and agricultural activities. It is radical because its idea of power redistribution of industrial political economy, and prosaic because it sees plausible solutions. Furthermore, there is a raise tendency that seems to be motivated by the idea of that collapse is inevitable and thus preparing for it is preferable to mitigating or avoiding it.

Sustainability

Dryzek defines sustainability as a way to dissolve the conflicts between environmental and economic values. Sustainable development and ecological modernization should go hand in hand.

Sustainable development stands for a process of change in which the exploitation of resources, the direction of investments, the orientation of technological development, and institutional change are all in harmony and enhance both current and future potential to meet human needs and aspirations.

On the other hand, the main ideas of ecological modernization can be list as follows: 1) positive sum outcome between economic and environmental objectives, 2) economic development and ecological protection are both desirable objectives for the planet and future generations welfare, 3) the “polluter pays” principle, among others (Toke, 2011).

Green radicalism

Green radicalism rejects the basic structure of industrial society and the way the environment is conceptualized, preferring (within the most radical version) a landscape without humans (John S. Dryzek, 2022). This direction argues that solutions for the environmental challenges can not be found within the today’s economic system, and addressing climate change requires a fundamental reorientation of economic behaviour (Stevenson, 2014). Concerns related to human rights, justice, and equality, tend to be common trending topics, most of which in a short-term perspective.

2.3 – Natural Language Processing

Natural Language Processing (NLP) is a subfield of Artificial Intelligence (AI) and linguistics that focuses on the interaction between the statements or words written by humans and the understanding of the computers. The goal of NLP is "*to accomplish human-like language processing*" and being able to perform such different tasks as paraphrase text and summarise it, translation, or question-answering (Chopra et al., 2013).

The history of NLP usually starts in 1950s, since the interest of the translation of "syntactic structures" by Chomsky. In the same year Alan Turing proposed the famous Turing test to determine if a computer can think like a human or not (Liddy, 2001).

The recent availability of large amounts of data has led to the emergence of models such Large Language Models (LLM), Transformers or the chatbot 'Chat GPT' itself, which are now present in multiple aspects of our daily lives from the 2010s to the present.

Core tasks and popular applications

Being a multidisciplinary area of computer science and linguistics, NLP comprises a wide range of tasks and applications. Some of the fundamental tasks are described below: (Sowmya V. B., 2020)

- *Text classification*: assigning predefined categories or labels to the text based on its content. This technique is present in a wide variety of applications, such as email spam recognition or sentiment analysis.
- *Information extraction*: identifying structured information from unstructured text, such extracting calendar events or names of people mentioned in a post.
- *Question answering:* developing systems that can answer questions in a natural language.
- *Information retrieval:* finding relevant documents from a large collection with a single query. One example would be Google Search.
- *Conversational agent: building a* conversation between machines and humans (Alexa, Siri, etc).

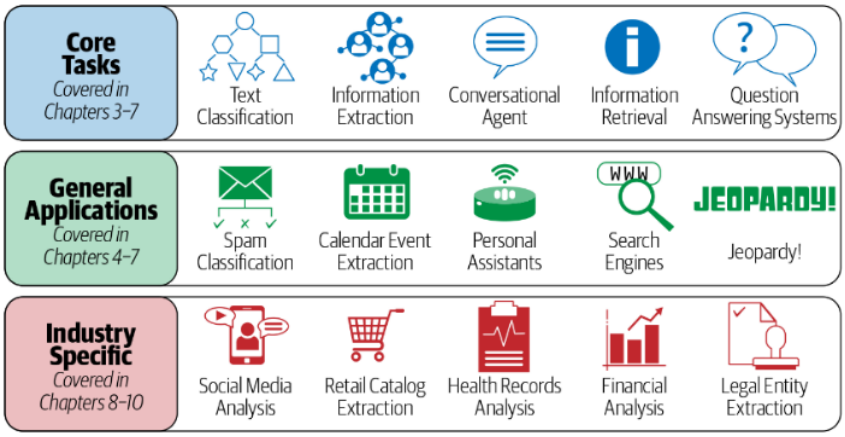

Fig 1: *Core tasks of NLP. Source:* (Sowmya V. B., 2020)

3- Materials and Methods

3.1 – Data collection

For the comprehensive analysis taken, we sourced data from 27 NECPs directly downloaded from the European Commission's website in a pdf format.

After an inductive analysis, three distinct approaches were employed in order to organize and consolidate the most relevant information:

- **Full text:** this involved compiling raw data containing the entirety of the plans, including summaries, objectives, policies, impacts and annexes.
- **Dimensions separately by objectives and measures:** the texts were split based on sections.
- **Dimensions jointly by objectives and measures:** this approach involved an integration of data, considering both objectives and measures simultaneously.

This comprehensive methodology facilitated the identification of variations between countries across all discussed levels.

3.2 – NLP pre-processing

Cleaning the unstructured source text is crucial during any process of training a model or even in doing some Information Retrieval (IR). There is many information that is irrelevant for our research, as for example unwanted words, punctuation or characters that are not useful for our analysis.

The pre-processing procedure simply means to convert the document into a understandable, predictable and analysable format for the machine (Tabassum & Patil, 2020). Some of the most common and widely used pre-processing techniques are the following ones:

1- Sentence segmentation

By segmenting the text into sentences, we can obtain word boundaries so further analysis can be carried out on each sentence (Tabassum & Patil, 2020). Although the simple way would be to split the full corpus by dots, there are many exceptions that could break our rule (e.g., Dr., Mr., etc). Thankfully, Python library NLTK (Natural Language Tool Kit) counts with some useful functions like *'sent_tokenize'* or *'word_tokenize'.*

Given that three out of four tonnes of greenhouse gases originate in the energy system, its decarbonisation is the cornerstone on which the energy transition and decarbonisation of the economy are based. However, the INECP also devotes a great deal of attention to measures to reduce greenhouse gas emissions in other sectors.

[Given that three out of four tonnes of greenhouse gases originate in the energy system, its decarbonisation is the cornerstone on which the energy transition and decarbonisation of the economy are based.,

However, the INECP also devotes a great deal of attention to measures to reduce greenhouse gas emissions in other sectors.]

***Fig 2**: example sentence segmentation*

2- Word tokenization

Similar to text tokenization, text tokenization refers to splitting sentences into words, characters or punctuations, all of which are called as *tokens*. This step will be fundamental when filtering unwanted words in further steps

3- Stop words removal

Many words like "*the", "and", "or"* or "*is*" do not have a significance in Natural Language Processing excepting some specific use cases (Tabassum & Patil, 2020). The reason why these words should be removed is that they make the text look heavier and less important for analysis so are meaningless. By removing these 'stop words' (which are usually prepositions, articles, and pro-nouns) we can reduce the dimensionality of the term space (Mohan, 2015)

4- Stemming and Lemmatization

Both stemming and lemmatization are shortening techniques that focuses on getting the root of a word. While stemming removes the suffix of a word and reduces it to some base form (Sowmya V. B., 2020), lemmatization tries to find the *lemma* of a word, the base form in a more grammatically correct form.

Although both techniques seem similar, lemmatization requires more linguistic knowledge and is more computationally expensive, since it will always find a meaningful word (unlike a stemmed word).

Stemming	Lemmatization
Studying → Study	Caring → Care
Coding → Cod	Better → Good

***Table 3**: differences between Stemming and Lemmatization*

5- Other pre-processing steps

Other pre-processing steps like removal of punctuation (!?, etc) and lowercasing are crucial steps for the data cleaning and normalization, so we are removing more noise that the machine could not understand. After all the explained steps, and following with the example sentence, we would get the following result:

['However', ',', 'the', 'INECP', 'also', 'devotes', 'a', 'great', 'deal', 'of', 'attention', 'to', 'measures', 'to', 'reduce', 'greenhouse', 'gas', 'emissions', 'in', 'other', 'sectors', '.']

['however', 'inecp', 'also', 'devote', 'great', 'deal', 'attention', 'measure', 'reduce', 'greenhouse', 'gas', 'emission', 'sector']

***Fig 3**: example pre-processed to processed sentence*

3.3 – Feature Engineering

Once we have cleaned properly our text, we still need a method that converts those sentences into a suitable language for a ML task. Such method is called 'Feature Engineering' and is one of the most important processes during the NLP pipeline.

The aim of feature engineering is to capture the characteristics of the text (in our case each word of each sentence) and convert it into a numerical vector understandable for the machine. This is known as "*word embedding*". Furthermore, this method can be summarized into two crucial steps: *feature extraction* and *feature selection*. (Hladka & Holub, 2015)

Whereas the purpose of *feature extraction* is to collect all the information potentially useful for the analysis, *feature selection* focuses on selecting a subset of the most relevant features in order to reduce the dimensionality of the corpus and get rid of irrelevant parts. By implementing these two steps we will reduce the model complexity and the computational cost, while improving the prediction performance and the model interpretability. (Hladka & Holub, 2015)

Some different strategies can be followed up to represent texts quantitatively. Two different approaches are discussed: classical and Deep-Learning (DL) based NLP approach (Sowmya V. B., 2020)

3.3.1 – Classical NLP approach

The traditional process of feature engineering counts with some count-based strategies (such bag of words, term frequency, TF-IDF, bigrams, etc) and are build using statistical and mathematical methodologies. (Sarkar, 2019). Although they are effective methods to extract features from text, we lose information such context and semantics and some handcrafted tasks must be implemented. We look now at some of these models with some examples:

- **Bag of words**

The key idea behind this technique is to represent the text as a collection of words ignoring context and order and only counting the times each word appears in every text. Its basic intuition is that if two corpus have a similar distribution of word counts, they both belong to the same class. Let's see an example:

	A	behind	cast	flows	mountains	peacefully	river	sets	shadow	sun	the	through	valley
The sun sets behind the mountains.	0	1	0	0	1	0	0	1	0	1	1	0	0
The mountains cast a shadow as the sun sets.	0	1	1	0	1	0	0	1	1	1	1	0	0
A river flows peacefully through the valley.	1	0	0	1	0	1	1	0	0	0	1	1	1

***Table 4**: bag of words representation*

With this kind of representation, corpus having similar word counts will have their vector representation closer between them (many distance formulas like Euclidean distance will be discussed in the next chapters). However, the model complexity increases while

increasing the vocabulary. In addition, BoW does not allow the capture neither the context nor different words with the same meaning. (Sowmya V. B., 2020)

- **TF-IDF**

TF-IDF stands for Term Frequency – Inverse Document Frequency. Its main objective is to consider the importance of a word in a particular corpus. Specifically, TF (Term Frequency) denotes the frequency of a word in text and IDF (Inverse Document Frequency) represents the general importance of a word in text. Let's see its mathematical representation considering as an example word '*tree*' in text *j:*

$$TF_{tree,j} = \frac{frequency\ 'tree'\ in\ text}{total\ word\ counts\ in\ text} \quad (1)$$

$$IDF_{tree} = \log\left(\frac{Total\ number\ of\ texts}{Number\ of\ texts\ containing\ 'tree' + 1}\right) \quad (2)$$

$$TF\ IDF_{tree,j} = TF_{tree,j} \times IDF_{tree} \quad (3)$$

Summarising, the intuition behind this metric will be: if we get a high value for word *w* in document *i* and lower value for document *j*, then *w* will be of great importance in document *d*. in addition to BoW, TF-IDF is widely used in NLP tasks such as text classification and information retrieval (Sowmya V. B., 2020).

3.3.2 – Deep-Learning NLP approach

As we have seen in the previous chapter, the main limitations of classical approach in text vectorization are, on the one hand, its discrete representation of texts (unable to capture relationships between words), and in the other hand, the fact that the vector dimensionality increases with the size of the vocabulary used, increasing the complexity of the model. (Sowmya V. B., 2020)

To overcome these limitations, many solutions have emerged. One is to represent each word in a semantic vector space by *learning* a low-dimensional vector representation of each word, known as *word embedding*. The main idea is to map each word in a vector space, in a way that similar words are closer. (Yang Li & Tao Yang, 2018)

Vocabulary Dictionary	
Word	Representation
Decarbonisation	[1.21, 1.52, 1.52, 0.14 ...]
Policy	[0.31, 0.46, 0.45, 0.53 ...]
Measures	[0.42, 0.77, 1.13, 0.58 ...]
Are	[1.01, 0.51, 0.88, -0.14 ...]
Summarised	[0.59, -0.01, -0.24, 0.19 ...]
...	...

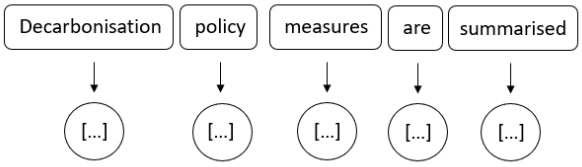

***Fig 4**: example textual data represented as a word embedding*

This revolutionary scope was proposed recently in 2013 by Mikolov (Mikolov et al., 2013) and the computational tool for representing continuous distribution of words was called

Word2Vec. Such model was based on *distributional similarity*, meaning that a word can be understood by its context (connotation) being able to capture word analogy relationships like:

vector("King") – vector("Man") + vector("Woman") = vector("Queen)

Unlike most text classification techniques, Word2Vec can be considered both unsupervised and supervised technique(Lilleberg et al., 2015). It is supervised because the model derives a supervised learning task from the corpus, but it is also unsupervised because you can provide any corpus for your model.

The original Word2Vec approach proposed two different architectures: Continuous Bag of Words (CBOW) and SkipGram.

- **CBOW**

The goal of Continuous Bag of Words is to predict a target word based on its context. Its name comes from the fact that the order of the words in context does not influence the final result (Mikolov et al., 2013).

Fig 5: *CBOW target word prediction scheme*

CBOW's architecture is composed by:

- **Input layer:** it consists of one neurone for each word in context window. It is usually represented by a one-hot-encoding vector where each element corresponds to one context word.
- **Hidden layer:** one-hot encoding vectors are multiplied by a projection matrix and word embeddings are created in a lower dimensionality space.
- **Output layer:** after the summation or average of vectors from the hidden layer, the output layer is a soft-max layer that generates a probability distribution over the full vocabulary. The target word will be the one with the highest probability.

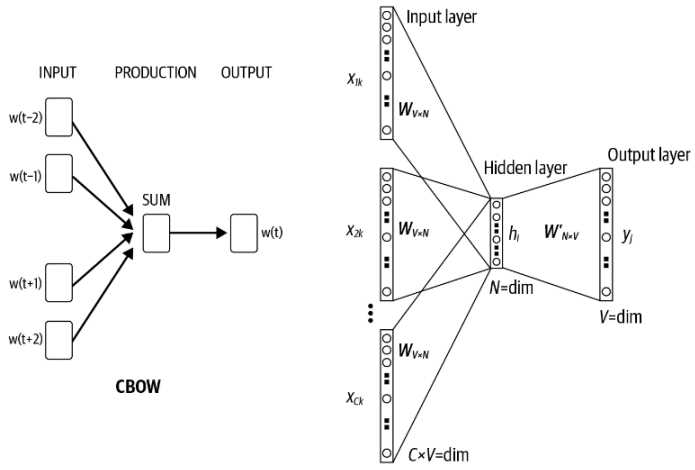

Fig 6*: CBOW's architecture scheme. Source:* (Mikolov et al., 2013)

- **SkipGram**

The second structure of Word2Vec, instead of predicting a word based on its context, tries to find the context words based on the centre word (Sowmya V. B., 2020). The model works considering each word as the target word, their context the words that occur in a vicinity within a specific window size. In the following example, such window would be ±3:

Fig 7: *skip-gram context prediction scheme*

The neural architecture for SkipGram (**Fig 8**) is very similar to the CBOW one, with a single hidden layer:

- **Input layer**: it consists in a one-hot encoding vector representing the target word
- **Hidden layer:** the one hot encoding vector is multiplied by a weight matrix and create the word embedding for the target word.
- **Output layer**: the hidden layer is again multiplied by another weight matrix producing the output layer, which uses a soft-max activation function to convert the output into probabilities among the entire vocabulary.

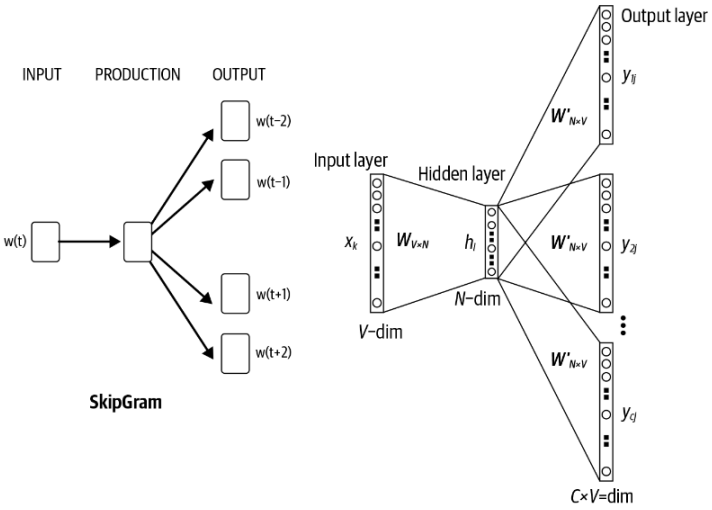

Fig 8: *skip-gram architecture scheme. Source:* (Mikolov et al., 2013)

3.4 – Modelling and evaluation

3.4.1 – Topic modelling

For topic modelling task, it was applied a Latent Dirichlet Allocation (LDA) model. LDA is a generative probabilistic model of a corpus widely used in NLP tasks, which ‘‘involves the use of machine learning techniques to perform semantic analysis of a corpus by building structures that approximate concepts from a large set of documents’’ (Gross & Murthy, 2014).

Basically, LDA models the relationship between words, documents and topics inside a corpus using such generative probabilistic model. Within the model, the documents are represented as random mixtures of topics, and each topic is modelled as a unique distribution of the entire vocabulary of the corpus (Blei et al., 2003).

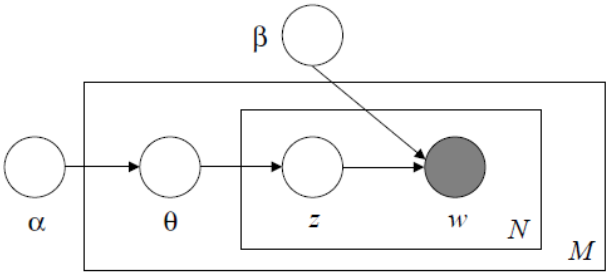

Fig 9: *LDA scheme. Source:* (Blei et al., 2003)

The process represented in the last scheme can be understood as a graphical representation with ‘plate’ notation. The N plate is the number of words in each document, and M represents the number of documents. In **Fig 9**, words *w* are the only observable variable. Z is the topic assignment for a given word in a document and alfa, q and beta are hyperparameters set manually in the model, as in the same way that K, the number of topics (Gross & Murthy, 2014).

3.4.2 – Discourse dictionary

After meticulous inductive research of the NECPs structure and being identified the most relevant topics to analyse, the types of discourses presented in chapter 2.2 were selected. On the one hand, environmental problem solving was processed as three separated discourses: administrative rationalism, democratic pragmatism, and economic rationalism. On the other hand, the two sub discourses of sustainability were processed together, having sustainability as a summary of both. Survivalism and green radicalism were also processed individually.

Once the six discourses were identified, the next step was to identify 10 words that most represented each of the discourses. The set of words was called “seed words”. After deciding the 10 most suitable seed words, 10 synonyms were searched in NECPs to complete each of the environmental discourse dictionaries, having a total of 660 words (10 seed words for each discourse (60) plus 10 for each seed word (600)).

3.5 – General workflow

Finally, once all methodologies have been explained, this chapter provides a comprehensive explanation of how code was structured for both analytical approaches: bottom-up and top-down analysis. Both analyses were developed in Python using Jupyter Notebook, a user-friendly tool that helps the user to execute the program step by step.

In order to make the code reproducible for future analysis and create and automated tool, the process is helped by a spreadsheet where the user can specify the file to read and also the pages at the plans which from and to the program should read.

Bottom-up analysis

The very first step of the bottom-up analysis (as explained before) is to take the raw NECPs documents and place them all together in the same folder. Then the Python script can start by applying the pre-processing methods explained in chapter 3.2. After the sentences cleaning, clean documents are stored as .txt file in another folder, ready for their analysis. Their segmentation will depend on the analysis we want to take.

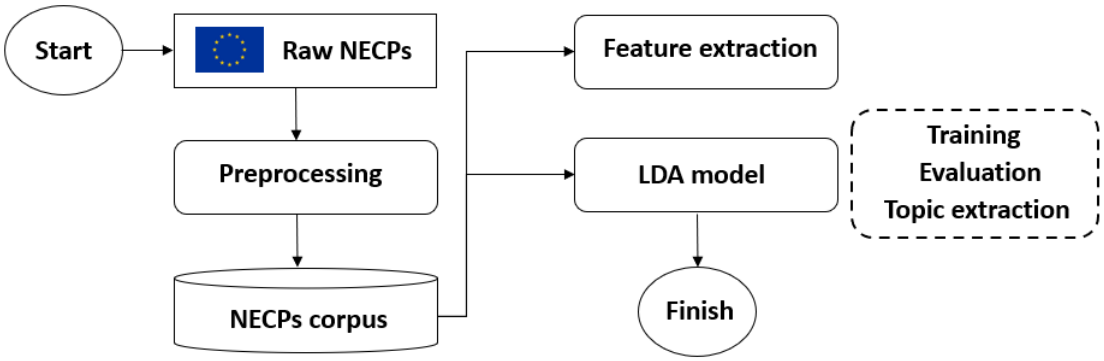

***Fig 10**: Bottom-up workflow scheme*

Feature extraction is done as a previous step to modelling in order to get the first results in TF-IDF or BoW matrices. Then the LDA model is trained and evaluated to find the best coherence to get the most suitable number of topics. As a result, we get the n number of topics that would explain our input texts, as well as the score for each topic that obtain each country.

Top-down analysis

As well as in bottom-up analysis, the first step of top-down analysis is to store the NECPs as .txt files (not necessarily if we have executed the other procedure previously). A Word2Vec model is created with the full extension of the plans.

The first step of the Word2Vec model development is to train it with the full length of texts. After training, it is important to find the bigrams that appeared the most. In other words, which pairs of words are commonly repeated that we can consider them as a single word. The requirement selected was that the pair of words must appear together at least five times.

Finally, the model is saved as a *.model* file. Thanks to this, it is not necessary to build the model every time we want to execute the file. The algorithm only needs to check if the file exists, and if not, create the model.

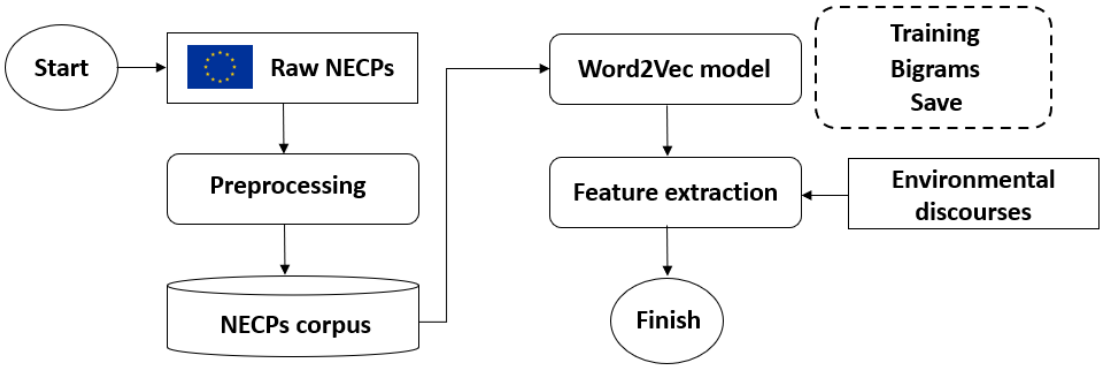

***Fig 11**: top-down workflow scheme*

At the same time, environmental discourses are constructed, with their respective dictionaries containing the seed-words that most explain each discourse. Having a well-constructed dictionary will be a crucial part of the analysis, because of its direct influence in the results.

Combining the feature extraction with the seed-words selected, we get as a result an index for each country explaining how much their narratives are related with a from environmental discourses.

4- Results and discussion

4.1 – Descriptive results

As introduced in previous chapter, the 27 National Energy and Climate Plans were divided into its dimensions, which are decarbonisation, energy, efficiency, energy security, energy market and research and development. Each dimension is also divided into two sections: “national objectives and targets” and “policies and measures”.

In order to find out how much importance have each country spent in each section, a preliminary descriptive analysis was carried out. The following plot shows the overall result on how many words have the countries written in each section, considering the number of words as a measure of importance and dedication:

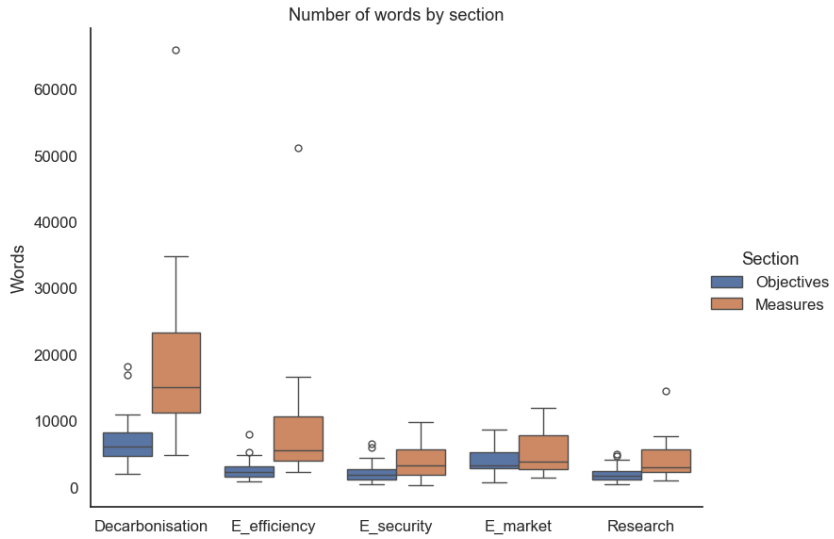

Fig 12: *boxplot of word numbers by section*

Being decarbonisation measures the section in which European countries have focused on the most, it can be observed that the plans are more extensive in the measures than in the set objectives. Decarbonization is clearly the most detailed section, followed by energy efficiency and the energy market.

Furthermore, there were certain countries that expanded much more than the average in some sections. Such countries were Belgium and Italy for decarbonisation objectives and Belgium again for measures, Belgium and Cyprus in energy efficiency objectives and Belgium again for measures, Italy and Slovakia for energy security objectives, Italy, and Spain for research objectives and finally Belgium in research measures (any outlier was found under any box diagram).

Same analyses were done in regard with sentences count obtaining similar results. The overall results, including the cleaning percentage and word per sentence count can be found in the GitLab repository.

The pre-processing procedure helped to clean stop words from text and reduce noise in it. The average removal of stop words reached approximately the 50% across most of the sections, contributing to the refinement of sentences. The resulting sentences were well suited for an accurate examination, providing a solid foundation for initial exploration, such as creating a word cloud.

The word cloud represents the most frequently occurring words directly proportional to their size in plot, offering a very first snapshot of the input's key themes:

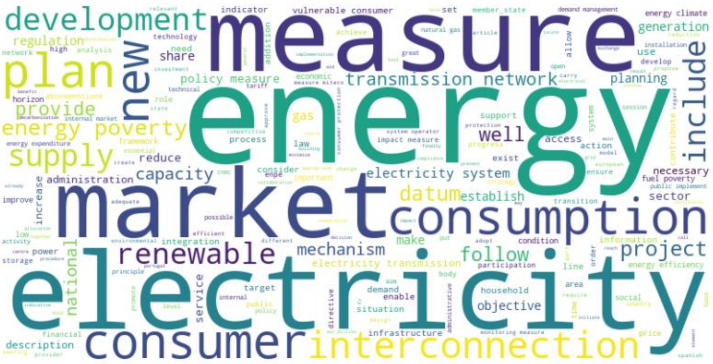

Fig 13: *Word-cloud example (energy market measures of Spain)*

Furthermore, it is interesting to find how these frequent words interact between each other. In other words, finding how many bigrams do we found in texts and how they interact can give us a fast screenshot of how the text is organised. For example, looking at the following bigram network of decarbonisation objectives, bigrams like energy-consumption, climate-emission or emission-reduction give us a primary idea of the text's content.

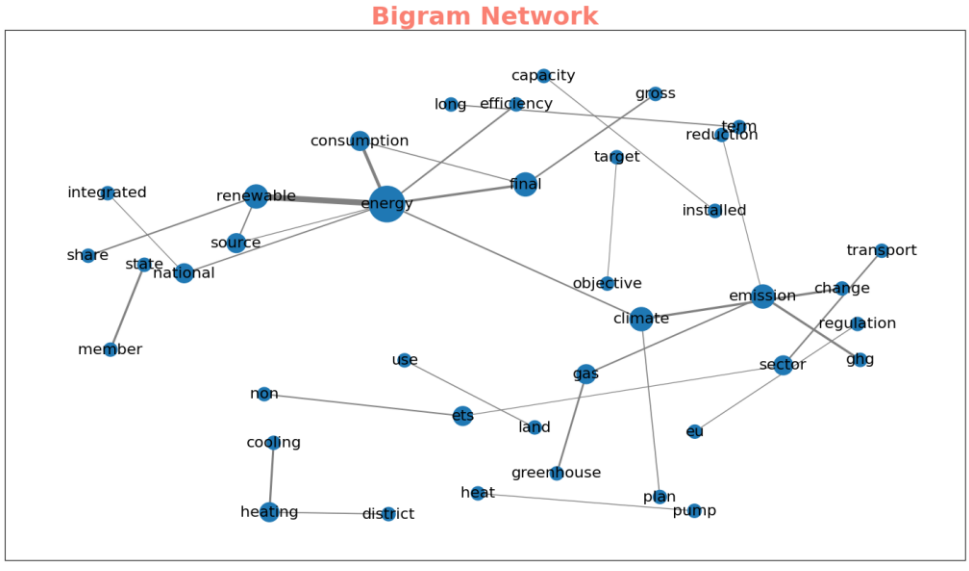

Fig 14: *Bigram network – decarbonisation objectives*

After this initial and preliminary analysis, a subsequent examination was conducted in order to explore how the most frequently used topics interrelate. This process is known (as explained before) as topic modelling.

4.2 – Topic modelling

4.2.1 – Finding the optimal number of topics

Having cleaned the sentences, the data is ready for themes identification and extraction. As explained in previous chapters, this procedure was carried out using LDA analysis.

The first step involved identifying the optimal number of topics that best suited each section of the plans. The metric employed to determine these values was the coherence function provided by the genism library itself. This function uses a general corpus associated with several coherence models (in our case was model C_V) measuring interpretability and consistence of themes.

Each section was addressed separately with an independent model for each one and applied a model coherence examination from 2 to 9 topics. Although in most cases the optimal number of topics was 2, it was preferred to select a lower coherence model in order to identify more distinct topics inside the text, so a little bit of coherence was sacrificed to obtain a better and complete representation of the plans.

In the following image is shown how the model coherence varies while increasing the number of topics:

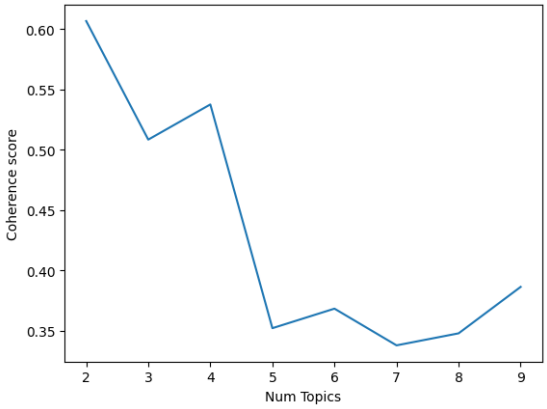

***Fig 15**: Decarbonisation coherence example. Optimal number of topics = 4*

Once we have obtained the most optimal value for the number of topics we want to analyse, the next step is to build a LDA model and stablish its characteristics. The chosen model was the one provided by the *gensim* library and looks like this:

```
lda_model = gensim.models.ldamodel.LdaModel(corpus = corpus,
                                            id2word = id2word,
                                            num_topics = 5,
                                            random_state = 100,
                                            update_every = 1,
                                            chunksize = 10,
                                            passes = 10,
                                            alpha = 'symmetric',
                                            iterations = 100,
                                            per_word_topics = True)
```

***Fig 16:** Python code example of LDA model*

The parameters values were chosen using 1) literature research of recommended values and 2) choosing the one which suited the most with the model performance. Its parameters are: **corpus** for the document vectors to analyse, **id2word** for the corpora dictionary, **num_topics** for the number of topics (calculated previously), **random_state** for reproducibility, **update**_every for iterative learning, **chunksize** for the number of sentences to be used in each training chunk, **passes** for the number of passes during the training, **alpha** for symmetric topic distribution within each document, **iterations** for number of iterations during the iteration and finally **per_word_topics** for the generation of a list with the most likely topics for each word.

4.2.2 – LDA performance

Now that the model has been build, we can extract the topics obtained from the LDA model using the function *print_topics()*. Taking as an example the dimension of energy market measures, each obtained topic is a combination of keywords contributing each one with a certain weightage to the topic:

Topic 1: 0.061*"electricity" + 0.044*"market" + 0.022*"consumer" + 0.021*"gas" + 0.020*"network" + 0.020*"price" + 0.018*"development" + 0.018*"system" + 0.015*"transmission" + 0.013*"new"

Topic 2: 0.039*"measure" + 0.039*"energy" + 0.019*"plan" + 0.018*"poverty" + 0.016*"project" + 0.016*"national" + 0.013*"policy" + 0.012*"include" + 0.011*"support" + 0.011*"consumer"

Topic 3: 0.048*"energy" + 0.019*"consumption" + 0.016*"renewable" + 0.014*"generation" + 0.013*"household" + 0.012*"use" + 0.012*"increase" + 0.011*"system" + 0.010*"cost"+ 0.010*"demand_response"

In this particular example, the different topics were evaluated, resulting in three different themes that countries cover in their respective national plan.

Within the first topic example, the topic seems to revolve around the electricity market, including elements such as consumer behaviour, gas-related aspects, network infrastructure and introduction of new elements. However, at this precise moment we don't know which words are related with 'price' or 'gas', because neither the word 'increase' or 'decrease' have appeared.

The second topic appear to be related to energy measures, focusing on addressing energy poverty. It mentions national plans, as well as projects and policies to support customers issues regarding poverty.

Finally, the third topic seems to centre around renewable energy, particularly in its consumption and generation. It includes terms related to household energy use, the renewable energy increase and other considerations like cost and demand response.

This procedure was applied to each of the commented subsections, and the obtained results are explained in the following table:

Section		N topics	Coherence	Keywords
Objectives	Decarbonisation	4	0.53	1: energy, renewable, increase, consumption, heating... 2: sector, electricity, transport, biofuel, biomass... 3: climate, change, measure, adaptation, impact... 4: target, emission, reduction, regulation, land…
	E. efficiency	4	0.45	1: energy, target, efficiency, directive, saving... 2: national, strategy, transport, sector, supply... 3: building, reduce, residential, public, requirement... 4: consumption, final, primary, renewable, reduction...
	E. security	3	0.39	1: electricity, capacity, system, power, storage... 2: energy, supply, gas, source, security... 3: oil, regulation, measure, emergency, stock…
	E. market	3	0.44	1: capacity, price, interconnection, power, indicator... 2: energy, poverty, measure, customer, support... 3: market, system, electricity, gas, objective…
	Research	3	0.46	1: research, innovation, energy, climate, policy... 2: development, new, economy, cost, solution, knowledge... 3: renewable, system, transition, change, industrial…
Measures	Decarbonisation	3	0.51	1: tax, reduce, transport, waste, carbon, emission… 2: support, system, work, public, information… 3: energy, climate, measure, renewable, government…
	E. efficiency	3	0.48	1: energy, efficiency, policy, promote, saving... 2: tax, network, electricity, fuel, consumption... 3: support, regional, fund, programme, initiative…
	E. security	3	0.41	1: measure, security, regulation, plan, supply... 2: gas, stock, natural, oil, storage, transmission... 3: energy, electricity, system, market, security...
	E. market	3	0.45	1: electricity, market, consumer, gas, price... 2: measure, energy, poverty, support, consumer... 3: energy, consumption, renewable, generation…
	Research	3	0.42	1: energy, system, new, solution, renewable, development... 2: climate, society, transition, knowledge, challenge... 3: research, energy, innovation, plan, government...

Table 5: *LDA summary for each NECP's dimensions and sections*

As shown in the previous table, coherence values went from 0.41 to 0.53, obtaining different number of topics, always between 3 and 4. Depending on the section, the keywords of each topic may have more or less relevancy and will explain more or less variability between them.

4.2.3 – Labelling countries discourses

Once the LDA model has been trained and modelled, we can evaluate each of the NECPs sentences and classify them into one of the topics. As a result, we get a topic distribution for all the countries and for all the sections describing the percentage of sentences assigned for each topic.

Taking as an example the dimension of Decarbonisation objectives, and labelling all the sentences into one of the four topics found (see **Table 5** for more information), the following results are obtained:

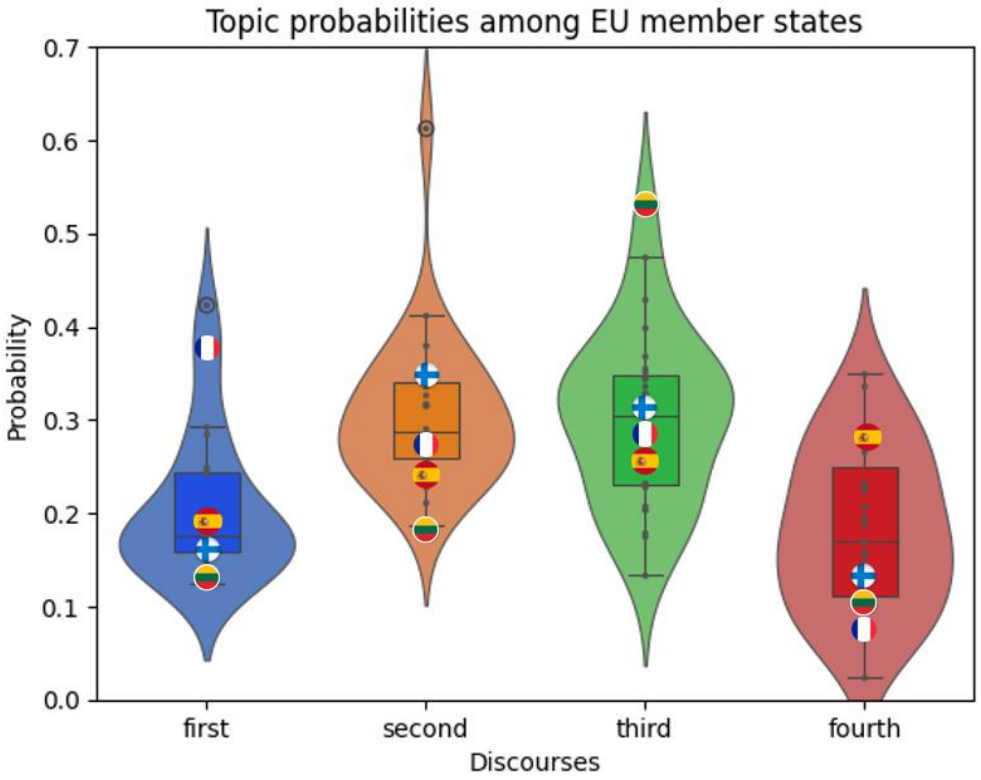

***Fig 17**: France, Finland, Spain, and Lithuania topic distribution in the Decarbonisation objectives dimension*

As it is shown in **Table 5**, four topics were obtained applying a LDA model to Decarbonisation section. The first one was related to renewable energy, the second one to transport and energy sources, the third one to climate change and adaptation, and finally the fourth one to emissions reduction.

Considereing all the sentences of all NECPs and classifying the into one of these 4 topics, as it is shown in **Fig 17,** there were only 4 countries classified as outliers that talked more that the rest of an specific topic. These countries were France and Latvia for the first discourse, Romania for the second one, and Lithuania for the third one.

Taking as an example Spain, France, Finland, and Lithuania, all of them have different topics distributions. However, Lithuania is the only one that devotes more than 50% of its sentences to talk about the third discourse. On the other hand, Spain counts with one of the most balanced distributions among european countries.

Being Europe such a diverse continent (both geographical and political), we can go a step further and draw a geographical representation in order to find similarities and differences between countries located at the same latitude. For example, looking at the energy market measures dimensions among european countries, we obtain the following distribution:

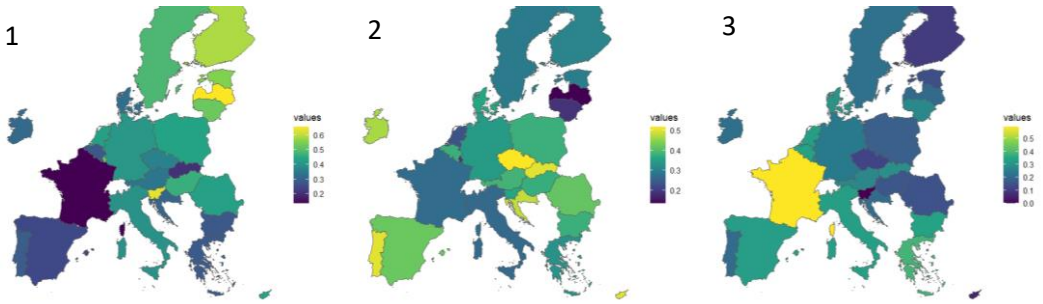

Fig 18: *energy market measures – topics 1, 2 and 3*

As shown in **Table 5**, three topics were obtained in energy market measures, which we can name as "**Electricity market evolution"** *(electricity*, *market*, *consumer*, *gas*, *price…),* "**Energy prosperity**" *(measure*, *energy*, *poverty*, *support*, *consumer…)* and **"Renewable future"** *(energy*, *consumption*, *renewable*, *generation…)*.

Fig 18 shows the geographical distribution of topics across the European country. Although norther countries focus more on the first topic, southern ones do it on the second topic, while and homogenous distribution is present in the third one.

As a result, the variation in topic distribution underscores the evidence that the climate and the intrinsic conditions of each country have a direct impact on the energy and climate plans.

4.3 – Discourses in NECPs

4.3.1 – Environmental dictionary

Once the bottom-up analysis was done, top-down analysis was carried out considering the discourses presented in chapter 2.2. An inductive literature research was done in order to find the top 110 words/bigrams that most suited each discourse and create a discourse dictionary. These 110 words/bigrams will be known as seed words.

To ensure that those seed words are relatively authoritative and convincing (Tian et al., 2023), the following procedure was followed: first of all, an extensive analysis through environmental discourses was taken with experts at BSC. As a result, the sustainability dimension included the subdimensions of sustainable development and ecological modernisation, the environmental problem-solving dimension included administrative rationalism, democratic pragmatism and economic rationalism, the green radicalism dimension included green consciousness and green politics, and finally survivalism including survivalism and promethean.

Secondly, the discourse dictionary seed words followed the following criteria: (1) the selected seed words must appear in NECPs, and (2) after training, the synonyms must complement the meaning of seed words, following the extension of *Word2Vec* model. Finally, these 60 seed words were obtained.

In addition to these 60 seed words, 10 synonyms for each word were obtained using the *cosine similarity* method. This method uses the property of Word2Vec of reducing the dimensions of each word into a vector in such a way that it can summarise the meaning of the word in a comprehensive way without being redundant (Tian et al., 2023). Then the similarity between two words using the cosine of the angle between the two-word vectors, following the next formula, using A and B as vectors of words *a* and *b*:

$$sim(A,B) = \frac{A * B}{||A|| * ||B||} \quad (4)$$

To ensure enough synonyms, we firstly selected 50 most suitable synonyms following the cosine similarity methodology, and then retained the most suitable ones inductively. Ultimately, an environmental discourse dictionary was built with a total of 660 words (60 seed words + 60 seed words * 10 synonyms).

Finally, the selected seed words are described in the following table (their respective synonyms can be found in **Annex I**):

Green radicalism	Administrative rationalism	Democratic pragmatism
idea	expert	participatory
equality	administration	flexibility
structural_change	control	interactive
nature	advisory	legitimacy
reflection	evaluation	consultation
interaction	ecosystem_service	mediation
progress	impact_assessment	voluntary_agreement
intuitive	evidence-based	community
…	…	…
Economic rationalism	**Sustainability**	**Survivalism**
market	integration	limit
property	sustainable	collapse
price	socially_just	scarcity
tax	environmental_protection	depletion
cost	cooperation	strain
auction	innovation	aggregate
trading	recycling	pressure
permit	conservation	nuclear_power
…	…	…

***Table 6**: example seed words by environmental discourse*

4.3.2 – Construction of environmental discourse dataset

The procedure of construction of each environmental discourse dataset can be divided into two crucial steps: the weighting scheme and the calculation of the environmental discourse dataset index (EDDI).

In the very first step, the TF-IDF methodology was used to calculate the weight of each word in the energetic plans. This measure, as explained in previous chapters, takes into account the importance of each seed word in a particular NECP and the entire NECP corpus of all countries, giving as a result a very well-suited metric for our case.

Within the second step, the EDDI was calculated from the weighted sum of all word's frequency in the corresponding environmental dictionary, applying such methodology to each of six discourses: administrative rationalism (**EDDI_ADM_RAT**), democratic pragmatism (**EDDI_DEM_PRA**), ecological rationalism (**EDDI_ECO_RAT**), green radicalism (**EDDI_GRE_RAD**), survivalism (**EDDI_SURV**) and finally sustainability (**EDDI_SUST**).

Each of these datasets will have 27 rows, corresponding to each country, and 110 columns, corresponding to the value of the TF-IDF value of each seed-word.

4.3.3 – Analysis of the environmental discourse dataset

Once the environmental discourse datasets have been constructed, the TF-IDF values of each seed-word from each discourse were grouped and the mean value were calculated, obtaining the following results:

	EDDI ADM_RAT	EDDI DEM_PRA	EDDI ECO_RAT	EDDI GRE_RAD	EDDI SURV	EDDI SUST
Austria	3.87	4.36	8.73	1.71	0.76	7.77
Belgium	4.41	4.91	7.27	1.80	1.19	5.00
Bulgaria	3.96	2.64	10.00	1.35	1.86	5.54
Croatia	5.14	3.55	10.00	1.26	1.27	7.50
Cyprus	4.23	3.73	9.45	1.35	1.10	4.73
Czechia	4.86	2.55	8.36	1.44	1.78	5.36
Denmark	4.05	4.00	7.73	1.98	0.93	6.07
Estonia	3.69	3.73	8.00	1.44	1.86	5.89
Finland	4.41	3.36	9.55	2.34	1.44	4.82
France	2.61	4.36	8.18	1.80	1.61	4.91
Germany	5.23	3.27	9.09	1.89	1.44	6.70
Greece	3.33	3.82	8.55	1.71	1.53	7.68
Hungary	2.97	3.09	11.09	1.08	1.69	5.36
Ireland	3.24	3.91	8.73	2.43	0.76	5.00
Italy	3.42	4.45	10.27	1.35	1.61	6.16
Latvia	3.24	3.00	10.45	1.35	1.44	6.07
Lithuania	3.24	3.09	8.82	1.71	1.53	6.61
Luxemburg	4.05	4.45	7.55	2.07	1.19	7.14
Malta	4.86	3.27	7.36	1.44	0.93	4.82
Netherlands	3.42	3.00	9.09	2.07	1.02	5.00
Poland	4.05	2.45	9.27	1.62	1.95	5.98
Portugal	4.23	4.55	7.82	2.16	1.53	6.16
Romania	3.96	4.18	10.73	1.89	2.20	7.59
Slovakia	4.50	3.27	9.18	1.26	2.20	4.46
Slovenia	3.51	3.73	9.82	1.35	2.20	6.34
Spain	4.14	4.27	7.91	2.16	2.03	6.52
Sweden	4.05	3.27	10.45	2.97	1.86	4.73

***Table 7**: EDDI for each European country NECP*

Moreover, the summary statistics for each EDDI is:

	Mean	Max	Min	St. Dev.
EDDI_ADM_RAT	3.951	5.230	2.610	0.655
EDDI_DEM_PRA	3.639	4.910	2.450	0.662
EDDI_ECO_RAT	9.017	11.090	7.270	1.090
EDDI_GRE_RAD	1.740	2.970	1.080	0.437
EDDI_SURV	1.515	2.200	0.760	0.433
EDDI_SUST	5.923	7.770	4.460	1.015

***Table 8:** descriptive statistics of environmental discourses*

**EDDI values were multiplied by 10000 for a better visualization*

As it is shown in the previous table, the mean value of economic rationalism dimension (**EDDI_ECO_RAT**) is the largest, suggesting that markets, investors, and companies play a very important role, as well as the interaction between them too. In addition, sustainability dimension (**EDDI_SUST**) holds the second most used discourse, meaning that countries also focus on the sustainable interactions of the last elements.

Moreover, survivalism (**EDDI_SURV**) and green radicalism (**EDDI_GRE_RAD**) obtain the lowest mean values, suggesting that topics like preparing for emergency situations or radical positions regarding environmental protection are not considered so much in the NECPs.

No outlier was found calculating the mean value for each EDDI. The country with the highest value of **EDDI_ECO_RAT** was Hungary with a 11,09 score, while the lowest one was Belgium, with a score of 7,27. Furthermore, the highest **EDDI_SURV** value was Austria while the lowest were Romania, Slovakia, and Slovenia.

Merging all six TF_IDF matrices, and without considering from which discourse was each word, an unsupervised classification was done by using a dendrogram cluster and a 'ward' methodology, which uses the Ward variance minimization algorithm. This distance formula calculates the new entry distance between cluster *u* and *v* as follows:

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2} \qquad (5)$$

Where |u| and |v| are the size of cluster u and v respectively, and T is the total size of all points of the cluster joined. As a result, the following hierarchical plot is obtained:

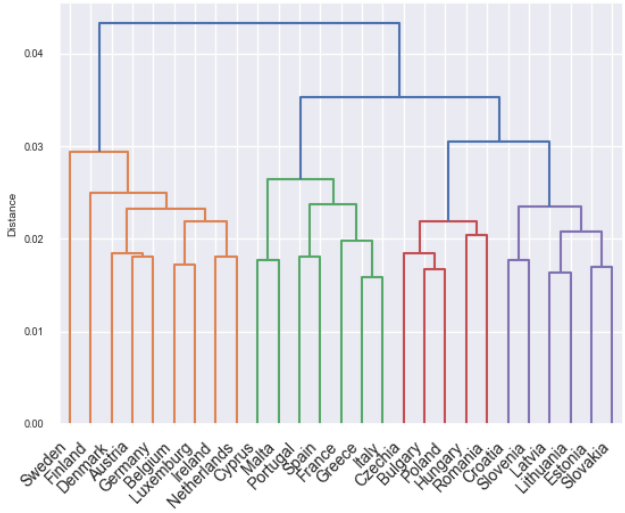

Fig 19: *hierarchical cluster European discourses*

Four clear clusters can be easily distinguished from the last figure. We can get a more visual representation by painting the European countries geographically:

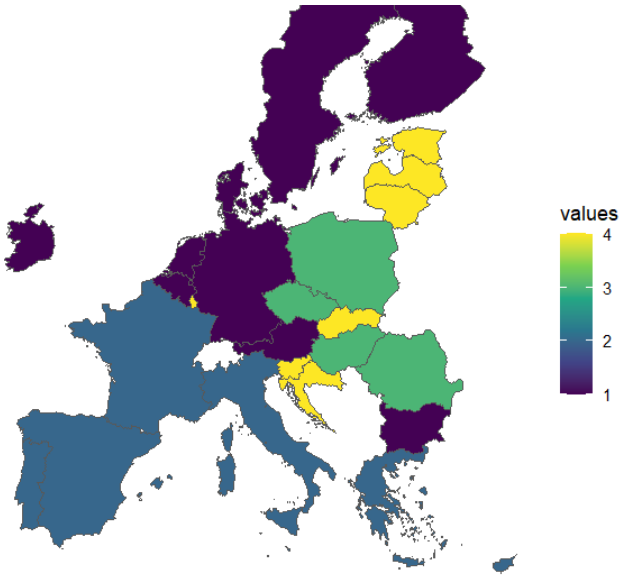

***Fig 20**: clustering classification of European countries using environmental topics*

As **Fig 20** shows, several conclusions can be drawn. The first observation is the evident geographical distribution of each cluster: for instance, cluster number 2 prevails in southern countries like Portugal, Spain, France, Italy, or Greece. Northern countries like Sweden, Finland, Denmark, and also Germany, Ireland or the Netherlands all fill in the same cluster 1.

Furthermore, Baltic countries fall into the same cluster 4, along with Croatia or Slovakia. Finally, eastern countries like Poland, Czechia, Hungary, and Romania are all present in cluster number 3.

Europe is a very diverse continent, not only culturally, linguistically, or historically, but also geographically and climatically. Such conditions make the European countries to have very wide range of opinions, priorities, and objectives in terms of economics and environmental politics.

The availability of energetic resources can be another fundamental pillar within a climatic context. Southern countries like Spain or Italy may have different energetic sources than Sweden of Finland: southern countries will have more availability to solar or thermic ones, while northern will have hydroelectric or wind power.

In summary, all of these opinions, priorities and resources available have an impact on the development and implementation of environmental and climatic plans in each European country. This variety approaches highlights the complexity of dealing with climate issues in Europe while emphasizing the need for countries to work together on shared solutions for the current environmental challenges.

5- Conclusion

During the course of this master's thesis, the development of an automated tool has been successfully carried out, enabling the analysis of texts through NLP technologies from two different perspectives: bottom-up and top-down.

Within the framework of the first perspective, an LDA model has been employed to extract both the optimal number of topics present in a text and the keywords that describe these topics. This has resulted in a useful, effective, and above all, reproducible tool applicable to any type of text.

In addition, some literature research has been carried out, identifying similar patterns and methodologies by applying different approaches to the NECPs. High similarity between similar countries like Spain, Portugal and Italy were also found in (Zółkowski et al., 2022) regarding the leveraging topic modelling and clustering of the comparative analysis done in the bottom-up analysis.

Regarding the second approach, in collaboration with BSC experts, an inductive search of the main environmental discourses has been conducted. This process involved identifying the words that best described those discourses, striving to make them as representative as possible, and having a final environmental discourse dictionary of 110 seed-word per discourse.

Once the search was completed, we analysed the best techniques to identify the impact that words can have on a text, ultimately selecting the TF-IDF methodology.

Alongside the environmental discourses, each EDDI was extracted, resulting in a European majority preference for discourses on economic rationalism and sustainability. Furthermore, relevant geographical relationships between discourses were obtained, which can explain how the climate and internal factors of each country may influence the drafting of a national energy and climate plan.

Finally, as a future research line, it is proposed to combine the NECPs sections to conduct more homogeneous and cross-sectional analyses. This approach should not overlook the investigation of other factors that may explain the distribution of topics among countries. These factors could include, in addition to location, the ruling political party, per capita GDP, or the percentage of renewable energy used by each country.

6- Bibliography

Berrang-Ford, L., Siders, A. R., Lesnikowski, A., Fischer, A. P., Callaghan, M. W., Haddaway, N. R., Mach, K. J., Araos, M., Shah, M. A. R., Wannewitz, M., Doshi, D., Leiter, T., Matavel, C., Musah-Surugu, J. I., Wong-Parodi, G., Antwi-Agyei, P., Ajibade, I., Chauhan, N., Kakenmaster, W., … Abu, T. Z. (2021). A systematic global stocktake of evidence on human adaptation to climate change. *Nature Climate Change 2021 11:11*, *11*(11), 989–1000. https://doi.org/10.1038/s41558-021-01170-y

Blei, D. M., Ng, A. Y., & Edu, J. B. (2003). Latent Dirichlet Allocation Michael I. Jordan. In *Journal of Machine Learning Research* (Vol. 3).

Chopra, A., Prashar, A., & Sain, C. (2013). Natural Language Processing. *INTERNATIONAL JOURNAL OF TECHNOLOGY ENHANCEMENTS AND EMERGING ENGINEERING RESEARCH*, *1*(4). DOI: 10.1109/INMIC.2004.1492945

Gross, A., & Murthy, D. (2014). Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing. *Neural Networks*, *58*, 38–49. https://doi.org/10.1016/j.neunet.2014.05.008

Hajer, M., & Versteeg, W. (2005). A decade of discourse analysis of environmental politics: Achievements, challenges, perspectives. *Journal of Environmental Policy & Planning*, *7*(3), 175–184. https://doi.org/10.1080/15239080500339646

Hladka, B., & Holub, M. (2015). A gentle introduction to machine learning for natural language processing: How to start in 16 practical steps. *Language and Linguistics Compass*, *9*(2), 55–76. https://doi.org/10.1111/lnc3.12123

John S. Dryzek. (2022). *POLITICS OF THE EARTH: ENVIRONMENTAL DISCOURSES*. https://books.google.com/books/about/The_Politics_of_the_Earth.html?hl=es&id=sjVKEAAAQBAJ

Liddy, E. D. (2001). *Natural Language Processing*. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and Word2vec for text classification with semantic features. *IEEE International Conference on Cognitive Informatics and Cognitive Computing*, 136–140. https://doi.org/10.1109/ICCI-CC.2015.7259377

Maris, G., & Flouros, F. (2021). The green deal, national energy and climate plans in Europe: Member states' compliance and strategies. *Administrative Sciences*, *11*(3). https://doi.org/10.3390/admsci11030075

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://arxiv.org/abs/1301.3781

Mohan, V. (2015). *Preprocessing Techniques for Text Mining-An Overview*. https://www.researchgate.net/publication/339529230

Sarkar, D. (2019). Text Analytics with Python. In *Text Analytics with Python*. Apress. https://doi.org/10.1007/978-1-4842-4354-1

Sowmya V. B. (2020). *Practical natural language processing : a comprehensive guide to building real-world NLP systems* (B. Majumder, A. Gupta, & H. Surana, Eds.; First edition.) [Book]. O'Reilly Media.

Stevenson, H. (2014). Representing Green Radicalism: The limits of state-based representation in global climate governance. *Review of International Studies*, *40*(1), 177–201. https://doi.org/10.1017/S0260210513000077

Tabassum, A., & Patil, R. R. (2020). A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. *International Research Journal of Engineering and Technology*. www.irjet.net

Takahashi, W. (2020). Economic rationalism or administrative rationalism? Curbside collection systems in Sweden and Japan. *Journal of Cleaner Production*, *242*, 118288. https://doi.org/10.1016/J.JCLEPRO.2019.118288

Tian, J., Cheng, Q., Xue, R., Han, Y., & Shan, Y. (2023). A dataset on corporate sustainability disclosure. *Scientific Data*, *10*(1). https://doi.org/10.1038/s41597-023-02093-3

Toke, D. (2011). Revising Ecological Modernisation Theory. *Ecological Modernisation and Renewable Energy*, 7–40. https://doi.org/10.1057/9780230302167_2

Yang Li, & Tao Yang. (2018). *Guide to Big Data Applications* (S. Srinivasan, Ed.; Vol. 26). Springer International Publishing. https://doi.org/10.1007/978-3-319-53817-4

Zółkowski, A. ˙, Krzyzí, M., Wilczy´nski, P., Wilczy´nski, W., Gizí, S., Wi´snios, E. W., Pielí, B., Sienkiewicz, J., & Biecek, P. (2022). *Climate Policy Tracker: Pipeline for automated analysis of public climate policies*. https://arxiv.org/abs/2211.05852v1

7- Annexes

Annex I – Environmental discourses seed words

- **Green radicalism**

idea, equality, structural_change, nature, reflection, interaction, progress, intuitive, consciousness, academic_institution, gender, phase_out, more_resilient, environmentally_responsible, main_determinant, effort, soft, perceives, start_ups, human_right, retraining, natural_heritage, guidance, main_vector, scoreboard, traffic_management, agency, incubation, actively_involved, gradual_reduction, preservation, emotion, trans_sectoral, significant_contribution, car_pooling, existential, entrepreneurial, inclusive_transition, just_transition, habitat, independent_expert, interaction_between, indicator, carpooling, aspiring, knowledge_institution, peace, job_creation, ecosystem, interdepartmental, main_exogenous, commitment, bus_lane, irreversible, laboratory, gender_equality, fair_transition, biosphere, suggestion, interdependence, substantial_contribution, pedestrian, ourselves, pioneer, horizontally, restructuring, nature_conservation, formulating, especially_vulnerable, considerable_effort, naturalist, learning, social_inclusion, economic_crisis, biodiversity, ethical, factor_influencing, chievements, understand, justice, labour_productivity, terrestrial_ecosystem, coherence, correlating, climate_neutrality, cutting_edge, woman, gdp_growth, landscape_protection, argument, potential_overlap, decisive_contribution, knowledge, girl, employment, transversal, synergy, progress_report, creative, labour_market, external_factor, alliance, implication, interlinkages_between, practical, coal_mining, relationship_between, cultural, regionally, value_added

- **Administrative rationalism**

expert, administration, control, advisory, evaluation, ecosystem_service, impact_assessment, evidence_based, technical, consultant, state_administration, supervisory, advisory_service, multivariate, restore, macroeconomic_impact, adaptation_strategy, organisational, relevant_stakeholder, automation, reactive, upskilling, ipcc_guideline, environmental_status, planned_policy, observatory, administrative, close_collaboration, social_insurance, diagnostics, assisting, adequacy_forecast, protected_area, reference_scenario, enquiry, diagnosis, working_group, competent, frequency, lifelong, mathematical, macroeconomics, risk_arising, expert_group, authority, detection, accreditation, coverage_analysis, sensitivity, legal, steering_group, contracting, observability, educational_establishment, upscale, analysis, licensing, intergovernmental, local_government, inspection, educational_institution, econometrically, extent_feasible, suitability, adviser, facilitator, supervision, specialist, tolerability, mandate, plenary_meeting, responsibility, monitoring, practitioner, modelling, qualification, technical_constraint, directorate, instrumentation, qualified, analytic, testing, representative, public_administration, measurement, projection, standardisation, appointed, municipal, sensor, assessment, verifying, ministerial_department, remote_reading, sensitivity_analysis, manager, contracting_authority, government, parliamentary, policy_statement, advisory_body, task_force, declaration, resolution, pact, intention, federal_government, climate_action, authorizes, sustainable_finance

- **Democratic pragmatism**

participatory, flexibility, interactive, legitimacy, consultation, mediation, voluntary_agreement, community, visibility, citizen, public, engagement, distributed_generation, easily_accessible, independent_monitoring, respondent, voice, voluntary, association, critical_mass, people, public_body, eligible, transparency, local_community, deploying_domestic, facilitateur, inclusivity, public_debate, consultation_process, voluntary_commitment, own_consumption, mobilising, empower, proactive_responsibility, cooperative, flexibility_option, digital_platform, social_acceptance, opinion, charitable_organisation, behavioural, self_consumption, leveraging, aware, public_authority, active_participation, greater_integration, bring_together, misinformation, consultation_session, enabling_framework, capability, informed, renovation, active_involvement, cycling_infrastructure, awareness_campaign, multiplicative, participatory_process, proactive_role, strength, vulnerable_group, local_authority, public_service, social, improve_knowledge, adjustable, knowledge_transfer, discussion, active_customer, human_resource, society, central_government, citizen_participation, decentralised_production, training_course, social_partner, collective, dissemination, criticism, car_sharing, mobilisation, storage, civil_society, self_generation, broadening, choice, empowering, decentralisation, round_table, intermediary, collaboration_between, most_vulnerable, public_consultation, crowdfunding, battery_storage, hearing, prosumption, campaign, publicly_owned, regular_exchange, prosumers, inequality, information_campaign, autoproduction, easy_access, leverage, territorial_cohesion, disseminate_information

- **Economic rationalism**

market, property, price, tax, cost, auction, trading, permit, operator, ownership, commodity, registration_tax, operating_cost, auctioned, traded, authorisation, competition, solvency, price_volatility, tax_rate, maintenance_cost, green_bonus, intraday, construction_permit, price_formation, market_liberalisation, competition_between, balancing_market, market_integration, market_player, competitive, well_functioning, internal_market, certified_installers, pricing, new_entrant, retail, fully_liberalised, hourly_price, permit_granting, predictable, contractor, freely, bidding_procedure, taxation, fixed_cost, price_premium, volume, issuing, personal_income, expense, tendering, liquidity, planning_permission, refund, investment, competitive_tendering, wholesale, permission, tax_exemption, benefit, green_certificate, transaction, licence, regulated, tax_reform, discount_rate, emission, broker, issued, bid, wholesale_price, cost_incurred, tax_revenue, deficit, tender, wholesale_market, authorisation_procedure, carbon_tax, deduction, quota, capacity_allocation, procedure, reduced_rate, profitability, revenue, market_coupling, simplification, co_pricing, minimising, certificate, market_participant, environmental_permit, marginal, remuneration, entitlement, investing, competitive_bidding, emission_allowance, legal_certainty, benefit_analysis, remuneration_scheme, balancing, concession, investor, renegotiating, traded_volume, holder, competitiveness, tariff

- **Sustainability**

integration, sustainable, socially_just, environmental_protection, cooperation, innovation, recycling, conservation, modernisation, taking_advantage, smart_city, ultimate_goal, radioactive_waste, partnership, experimental_development, plastic, natura_site, rehabilitation, automation, efficient, living_environment, rural_development, regional_cooperation, higher_education, landfilling, organic_soil, expansion, full_integration, environmentally, equitable, air_protection, multilateral, scientific_research, recycled_material, carbon_sequestration, upgrading, smartly, bioeconomy, lifestyle, soil_degradation, exchanging_information, technological_development, recovery, regeneration, renewal, digitalisation, climate_friendly, cohesive, water_management, collaboration, mission_oriented, reuse, wellbeing, reconstruction, innovative_solution, resilient, climate_resilient, waste_management, international_cooperation, specialisation, biodegradable_waste, aquatic_ecosystem, mechanisation, interoperability, ecology, fair, risk_prevention, participating_country, innovation_agenda, waste_treatment, forest_ecosystem, optimisation, high_quality, greener, dialogue, entrepreneurship, recyclable_material, protected_specie, infrastructure, low_carbon, ecologically, intergovernmental_agreement, smart_specialisation, circular_economy, development, territorially, prosperity, closer_cooperation, talent, organic_fraction, technology, electromobility, prosperous, mutual, science_technology, bio_based, protecting_vulnerable, partner, societal_challenge, create_synergy, excellence, tracking, credibility, holistic_approach, cleantech, complexity, sustained, enhancement, forum, future_oriented

- **Survivalism**

limit, collapse, scarcity, depletion, strain, aggregate, pressure, nuclear_power, prescribe, shortage, mitigate, safety_standard, natural_resource, stringent, combusted, maximum, pose_challenge, deterioration, interference, severe, conservative, high_pressure, nuclear, limit_value, extreme_event, casualty, failure, harmful, decommissioning, threshold, suffering, slope, malicious, gas_pipeline, fresh_nuclear, ceiling, excessive, containment, endanger, susceptibility, nuclear_reactor, restriction, fragmentation, turbidity, extreme_weather, exposure, shutting_down, reduce, erosion, particulate_marginally, combusted, adverse_impact, power_station, exceed, worsening, eutrophication, instability, margin, unpredictable, consequential, hazard, undesirable, radioactive_waste, minimum, sensitive, marginally_reduces, overload, bursting, dismantling, limitation, disturbance, unlawful, unavailability, interruption, sensitivity_scenario, contingency, damage_caused, overgrowth, contamination, eliminate, adverse_weather, danger, global_temperature, disrupted, tension, stipulated, susceptible, warning_system, accidental, protection_against, geopolitical, imposes, calamity, melting, strict, malicious, catastrophic, boundary, exposed, disease, penalty, heatwaves, soil_erosion, unrealistically, undesirable, terrestrial_ecosystem, interventional, ancient, blackout, attack