Literature Review on How to Measure Illiberalism Using Text Data

D3.1 Deliverable for the AUTHLIB Project

FIRST DRAFT

31 August 2023

SciencesPo

This report is a work in progress.
Do not cite without the permission of the authors.

Please, contact Jan Rovny: jan.rovny@sciencespo.fr

Published in the framework of the project "AUTHLIB – Neo-Authoritarianisms in Europe and the Liberal Democratic Response". Funded by the European Union and the UK Research and Innovation. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or UK Research and Innovation. Neither the European Union nor the UK Research and Innovation can be held responsible for them.

1. Introduction

The goal of this paper is to review and assess the most innovative methodologies available to conceptualise illiberalism using text data. To do so, the paper is structured in four parts: a summary of the main underlying concepts, a review of methodologies linked to Language Models, a review of data sources and data-related problems, and a roadmap of the models we are planning to build.

The first section defines what we are trying to measure, illiberalism, and the most innovative way to measure a concept using text data, which are Language Models. We define illiberalism in relation to the concepts of authoritarianism and populism. We do so to avoid the measurement of overlapping concepts and to avoid confusion in our models. For this reason, we define illiberalism in terms of minimum characteristics according to the literature and the work of AUTHLIB's Work Package 2. In this section, we also reflect on two possible approaches to conceptualization: an inductive one and a deductive one, and on which one would be better for a Language Model.

In the second part of the first section, we briefly introduce the definitions of Natural Language Processing (NLP), Language Models (LM), and Large Language Models (LLM). We do so for two reasons: to give a simple overview of why the methodology we will use is the most innovative one and to explain why in the second section we go through a literature that builds from basic machine learning concepts up to BERT models.

In the third section, we review all the concepts that make us able to understand the family of Language Models that we will use, BERT models, starting from the basics. We start from simple concepts such as the different types of machine learning, up to deep learning, transformer architecture and attention mechanisms. Following that, we review available data sources and the most common methodological challenge for an LM focused on classification: the choice and creation of training data. We also briefly review our options for dealing with the multiple languages like in our case.

Last, we build a roadmap of the models that we are planning to use that includes the choice of BERT models, the preferred training data, fine-tuning considerations, and how we go from our models' outputs to a map of 'illiberalisms' in Europe.

2. Problem Definition and Underlying Concepts

Illiberalism

Our goal as a Work Package is to conceptualise illiberalism using large text data. Before delving into the methodological part, we try to conceptualise illiberalism both for conceptual clarity and for measurement purposes. There is a growing consensus in the literature that, since the early 2000s, we assisted in a decline in the quality of democracy together with a growing sense of normative criticism towards liberal democratic ideas, institutions, and practices. This phenomenon is especially relevant in Europe, where there is a tendency to challenge liberal democratic ideas with illiberal ones via the use of populist rhetoric (Mudde, 2010).

The literature on this topic usually focuses on three main concepts: populism, authoritarianism, and illiberalism. All of them are, in principle, applicable to regimes, ideas, ideologies, attitudes and behaviour. However, the three concepts have fundamentally different ways of being understood. Populism can mainly be understood in a discursive way, while authoritarianism tends to be understood either as a psychological phenomenon or as a largely non-democratic regime type (Hawkins et al., 2021; Laruelle, 2022).

Illiberalism is, on the other hand, a complex phenomenon that has both ideological and regime-type elements. The ideological elements of illiberalism focus on direct or indirect attacks against the values of political liberalism such as: "human rights, justice, equality and the rule of law, its commitment to multiculturalism and tolerance, ideas of Isaiah Berlin's 'negative liberty', Karl Popper's 'open society', John Rawls' 'overlapping consensus', or Ronald Dworkin's equality as the 'sovereign virtue'" (Halmai, 2021: 813).

As a regime type, illiberalism focuses on ideas undermining constitutionalism and the rule of law. This is why illiberalism is often defined using both its ideological and regime-type elements together. The Illiberalism Studies Program at George Washington University defines illiberalism as

"a strain of political culture, a set of institutional reforms (such as assaults on an independent judiciary) and broader societal processes (such as declining trust in liberal democratic institutions) that, over the past two decades, has emerged in response to liberalism as experienced by various countries" (illiberalism.org, 2021). Similarly, The Routledge Handbook of Illiberalism offers the following comprehensive definition: "Illiberalism refers to a set of social, political, cultural, legal, and mental phenomena associated with the waning of individual liberty (personal freedom) as an everyday experience" (Sajó, Uitz and Holmes, 2021: xxi). According to these authors, "illiberalism is not an ideology or a regime type"; it is a broader phenomenon that undermines freedom.

Because of these elements, illiberalism is prone to be measured with text data. As a consequence, this work will focus on illiberalism rather than populism or authoritarianism as AUTHLIB mainly focuses on ideas and ideological configurations. To avoid confusion or overlapping results, we focus on the measurement of the cumulative concept of illiberalism (Sartori, 1970). Some ideas, as highlighted by Work Package 2, overlap with authoritarianism and populism and should therefore be discarded. Distaste for representative institutions is also part of the definition of populism, opposition to procedural democratic norms is also part of authoritarianism, and anti-pluralism is both part of populism and authoritarianism. The following characteristics are the only ones unique to illiberalism and consequently relevant to our analysis.

- erosion of freedom
- censure of individualism
- disrespect for minority rights (e.g., LGBTQ)
- politics of exclusion
- defence of cultural homogeneity
- defence of national sovereignty (Euroscepticism)
- economic protectionism at the nation-state level

Classifying Concepts Using Text Data

Our problem concerns classification using text data. The standard methodology to go about it would be to use a methodology such as a cluster analysis or any combination of descriptive statistics methods to achieve this goal. However, this approach is neither the most innovative nor the most appropriate to process a large amount of text data.

To understand why classical statistical approaches are not the most innovative ones, we can just look at the evolution of Natural Language Processing P). NLP is a field whose main goal is to use computers to understand natural language. Natural language is any form of written, spoken or sign human language naturally created by humans to be able to communicate. NLP started by using statistical approaches to understand human language in the 1990s. As soon as machine learning and neural networks became available, the field adopted them.

Language Models (LM) refer to models that use neural networks or statistical methods to generate probabilities of a series of words occurring. Neural networks are overall more innovative than statistical approaches as they require less feature engineering and can capture semantic properties of words through the transformer, as explored later. Large Language Models (LLM) are a further methodological innovation in the field as they can process a large amount of data to capture semantic meaning.

3. Review of Methodologies Linked to Language Models

In this section we review concepts related to machine learning from the generic ones down to specific ones to help us understand the construction of the most appropriate LLM for our purpose. Even if some of the concepts in this section might seem out of context, they all come to be used to understand the model construction section.

We start by exploring some basic concepts related to machine learning. Machine learning is a multi-disciplinary field focused on helping computers find their own rules to solve a specific problem. NLP uses different machine learning types to perform tasks such as text classification,

summarization, and generation, among others. There are three main families of machine learning: supervised learning, unsupervised, and deep learning. All of them only have one common characteristic to keep in mind: they use a small set of data, also called training data, to train the algorithm on a larger set of data, also called test data.

In the following subsections, we briefly explain how these three families differ and we explain how they could be used to classify illiberalism. The final goal is to understand why the deep learning family is the most innovative solution for our purpose.

Option 1: Supervised Machine Learning

Supervised machine learning uses labelled training data to run a model on test data. Labelled data is a dataset where observations are labelled according to the goal of the model. The control over this data is what the 'supervision' part of the name refers to. In our case, the labelled data would be for example a set of sentences that an expert classified as illiberal or non-illiberal.

Classical supervised learning uses algorithms both for prediction and classification tasks, even if a prediction is only possible using continuous data, and it is therefore out of the NLP realm (Nasteski, 2017). Conversely, classification problems in classical supervised learning use categorical data. The most common classifier options for supervised machine learning, mentioned here for reference, are logistic regression, Naïve-Bayes classifiers, decision trees, random forests, boosting, XGBoost, and Support Vector Machines (SVM) (Singh, Thakur and Sharma, 2016).

For example, if we had multiple years of text data labelled as illiberal, we could classify new texts on whether they are illiberal or not. Being our problem very likely non-linear, decision trees, random forests, boosting, XGBoost, and Support Vector Machines (SVM) could be potential good classifiers. However, as we will explore more in detail later, the currently popular approach for this kind of problem would be to directly train a deep neural network to understand the context of a sentence beyond the 'simple' vectorisation of terms used in both supervised and unsupervised

machine learning. Both these families of machine learning transform the text into vectors to make algorithms able to process it.

Currently, in NLP, we have four main supervised learning classification techniques: Topic Modelling, Classification Modelling, Sentiment Analysis, Named Entity Recognition, and Part-of-Speech-Tagging. It is important to note that the first three can also be used in unsupervised machine learning. Classification Modelling and Sentiment Analysis are the only two methodologies currently used in supervised learning that could potentially be used for our purpose that prevalently supervised learning. Classification modelling classifies observations into categories. It can predict wins versus losses or whether a text is aggressive or friendly. It can also label parts of text into a single category such as relevant or not relevant.

Sentiment analysis, on the other hand, can identify whether a document expresses positive or negative emotions, whether it uses active or passive terms, or whether words aim at evoking emotions, among others. In political science, its initial usages focused on understanding the sentiment of voters (Laver, Benoit and Garry, 2003; Mullen, 2006; Hopkins et al., 2007), politicians positions (Thomas, Pang and Lee, 2012), and opinions regarding pending policies or proposals (Zavestoski, 2005). In its most frequent use, Sentiment Analysis uses a dictionary of words with pre-defined values or words (Young and Soroka, 2012). A dictionary usually assigns a score to each word based on whether it is positive or negative. It can also assign a specific sentiment to each word. We describe Topic Modelling in the following section as it is prevalently used in unsupervised learning.

Option 2: Unsupervised Learning

In unsupervised machine learning the training data is labelled by a statistical technique instead of an expert or a researcher. A common way to do so is by clustering data with techniques such as K-Nearest Neighbors (KNN) or Density-based spatial clustering of applications with noise (DBSCAN). In NLP classical unsupervised learning deals with extracting distributions of co-occurring terms from a training text corpus and using it to perform a task via an algorithm. The main

method to do this is Topic Modelling, which identifies the most frequent topics by clustering text into groups or types based on similar words, usually an algorithm called Latent Dirichlet Allocation (LDS).

In our case, we could use unsupervised learning to look for patterns in our data and see whether we can find combinations of elements that partly or completely overlap with the existing definitions of illiberalism. However, the same problems that would make us prefer deep learning over supervised learning apply as well to unsupervised machine learning.

Option 3: Deep Learning

Deep learning models are 'evolved' machine learning models that use neural networks to complete their tasks. Neural networks are architectures made of multiple layers of interconnected nodes. Nodes are activated by the input they receive and then pass transformed signals either to another layer or to the final output. In Figure 1, we can find the simple representation of a Neural Network, and more specifically of a BERT-base model. In the upper part of the picture, the circles represent the nodes and how information can be exchanged between them in such architecture. Neural networks are inspired by human brains in the way the different nodes mimic the way different neurons are connected in constructing complex thoughts. The multiple layers and nodes of a neural network are what allow a deep learning model to identify more complex patterns such as the semantic meaning of words.

One simple way to understand deep learning and neural networks is via an example of how these models perform image classification. In image classification, the layers of a neural network can learn sequentially more complex features about the input data. For example, the first layer recognises the edges of the pictures, while the second understands the subject of the picture, and the third classifies it.

Simple neural networks like the ones used in image classification are called feedforward neural networks, and they are networks where information is fed from one layer to the next until the

end. There are many more neural network models than feedforward neural networks. Multi-Layer Perceptrons (MLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) are just some of the most popular types. Just for reference, it is important to know that Recurrent Neural Networks (RNN) are the most popular neural networks used in NLP, even if CNNs have been used to predict policy domains using electoral manifestos (Koh and Boey, 2021).

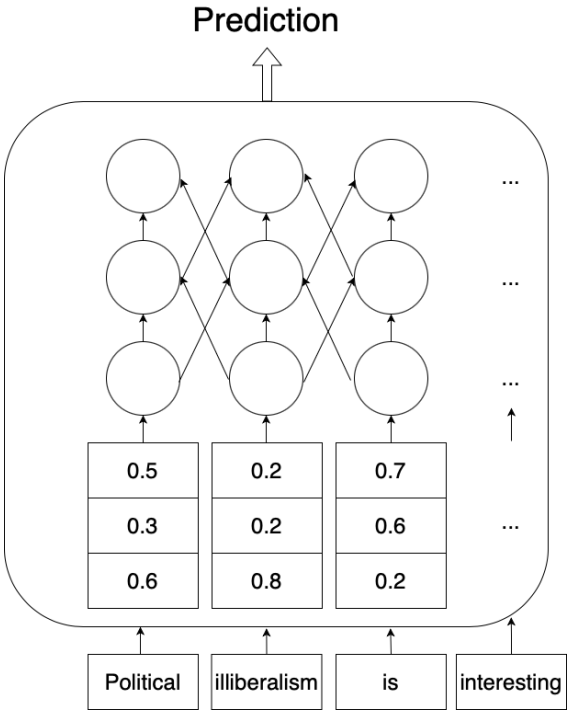

Figure 1. Simple representation of a BERT-base transformer architecture, an architecture that uses both vectors and neural networks to understand text

However, RRN became obsolete with the invention of Long Short Term Memory networks (LSTM) (Hochreiter and Schmidhuber, 1997) and subsequently attention. Attention is a mechanism that mimics cognitive attention. The idea behind attention is to let every step of an RNN pick information to look at some larger collection of information (Xu et al., 2016). It calculates "soft" weights for each word, more precisely for representing words in vector spaces. Attention is used in RNNs to calculate “soft” weights in a non-sequential way. Its most important evolution, the parallel multi-head attention mechanism, makes neural networks able to compute multiple parallel processes at the same time, significantly reducing computation time and increasing accuracy. The attention

mechanism was further improved by the transformer architecture, which is the most recent advance in the NLP field also available to use (Vaswani et al., 2023).

Transformer architecture is, as the name might suggest, based on transformers. Transformers contain different Encoders and Decoders organised in layers that can have multiple configurations. The Encoder generate embeddings, also known as low-dimensional space into which you can translate high-dimensional vectors. In the case of words, an Encoder generates embeddings for each word, and it creates a system where similar words are closer to each other. On the other hand, the Decoder takes the Encoder embeddings and uses them to generate a new text. Overall, encoders vectorize the inputs while decoders generate outputs based on the vectorized inputs. In a simple English-to-French translation, the Decoder learns the sentence grammar and its context by relating the words to each other, while the Decoder uses the structure created by the Decoder to create a text in French. Multi-head attention expands the transformer architecture by repeating the computation of the transformers multiple times in parallel, and each of these parallel computations is called a head. The final goal of multi-head attention is to give the transformer greater power to encode multiple relationships and nuances for each word and calculate the attention which is also what makes models using it computationally more demanding.

Transfer learning by definition stores ‘language knowledge’ and ‘task knowledge’ in the parameters of a model to be able to learn new tasks faster and better. Transformer architecture, as summarised in Figure 1 for a BERT-base model representation, relies not only on the concepts of neural networks but also on vectorisation. In a transformer architecture, each word is transformed into a vector. Vectors are groups of data elements of the same type, numbers in this case, stored in a sequence. Words have similar vectors between each other depending on whether they are surrounded by similar words. Traditional rule-based models or simple vectorisation techniques such as Bag-of-words and Term Frequency–Inverse Document Frequency use this kind of vectorisation of words to find information about a text.

However, in a transformer architecture like the one of BERT models, the vectors are passed to a neural network. This is represented in the upper part of the picture in Figure 1. In this process, a vector of a specific word is multiplied by all the vectors of all other words. This way the vector 'pays attention' to all the other words, using the attention mechanism. The same is repeated for all other vectors representing each word. The process for all words is also repeated across the different layers of the neural network. At the end of the process, each vector will be close to certain specific words according to the context of the inputted text. This way a BERT-base model can understand the context of text. This process summarised in Figure 1 is repeated for potentially millions of iterations.

BERT

BERT stands for Bidirectional Encoder Representations from Transformers. The model was first created in 2017 by a team of Google scientists (Devlin et al., 2018). In just a few years multiple versions of models were created based on the architecture in Figure 1, and they became the basis for text classification. Its architecture is fast, reliable, easy to use, and has an outstanding performance besides being innovative. We first explore the components of its name before moving into more technical applications.

1. Bidirectional. BERT is a bidirectionally trained model. It means that it applies its self-attention mechanism to learn information from a text in two directions: from the right to the left and vice versa. This gives a deeper understanding of a word by considering both the words that exist on its left side and its right side.
2. Encoder representations. Its transformer architecture only contains encoders and not decoders. This means that the model cannot generate new data based on its encoders but only performs tasks such as including text classification and sequence labelling.
3. Transformer. BERT uses a partial transformer architecture, only using encoders, to apply its self-attention mechanism.

The original BERT architecture is called BERT-base, and it can only perform two tasks: Masked Language Modelling (MLM) and Next Sentence Prediction. (NSP). MLM is a simple self-supervised task in which the model takes a sentence and randomly masks a word out of it. The output is a guess of the masked word, and the accuracy of the prediction helps BERT understand the context within a sentence. In the case of Next Sentence Prediction, the model takes two sentences, and it estimates the probability of the second sentence following the first one. This helps the model understand the context among sentences as well. All the vectors in this family of models are optimised for these two tasks until you find the best parameters (or vectors) able to understand the contest based on the other words.

BERT-base models are only good at MLM and NSP and they need to learn any other task from scratch. However, this task is so comprehensive that two families of BERT models specialised in two tasks were born from it: Natural Language Inference (NLI) and next-word prediction. Next-word prediction predicts the next word or token based on the previous ones. It calculates the probability for each of the words in the vocabulary it is trained on and then uses the most likely one for its prediction. This family of models is what is known today as GPT models.

The second main family of BERT models, the Natural Language Inference ones, are the ones at the basis of the most advanced classification models we have today. Natural Language Inference is a task that in its most basic form determines whether a hypothesis in the form of a sentence is true, false, or neutral given the context-sentence. The output of a model is the probability of whether the hypothesis is true, false, or neutral. This process can also be used for labelling. Any other classification task we have today is an adaptation of this universal task.

This task specialization is one of the three reasons that makes BERT models so interesting. The second one is the fact that these models come pre-trained, as described below, and the last one is fine-tuning, as explored later in this section and practised in the model construction part of the paper. Pre-training means that a classical BERT model already comes trained on a large corpus of texts to perform next-word prediction or natural language inference. Standard BERT models for example

come pre-trained for the understanding of the English language using the Toronto BookCorpus and Wikipedia as a text basis. Today different models are pre-trained on different languages according to different languages and their task specialization.

Fine-tuning is the last element that makes BERT models appealing for a text classification task. It is known as the process of 'adjusting' the layers of a deep learning neural network to optimally perform a specific task. In practical terms, it means changing some of the parameters of the model by adding task-specific layers to the later layers of the neural network. The parameters are usually created via a supervised learning task-specific model. In simpler terms, we take the output layer of the BERT model and replace it with a small set of labelled examples to perform supervised learning. The more the parameters created from this task, the better the performance.

The fine-tuning process requires minimal labelled data. The model is therefore able to have an extremely deep understanding of the language on which is trained and to classify the text with extreme accuracy using only a small number of examples. For general NLI BERT models, there are many examples of labelled datasets created for this purpose (Bowman et al., 2015; Williams, Nangia and Bowman, 2018; Nie et al., 2020). The probability created in this model is a data-rich task usually trained over a million annotated sentences. However, creating such large datasets requires a considerable number of resources, time, and effort. When it comes to text classification of a specific concept, like in our case, it is also better to pre-train and train on a specific set of examples using a smaller BERT model (Laurer et al., 2023).

4. Text Data Sources and Problems

Data Sources

Our goal of classifying illiberalism using text data is intrinsically linked to the possibility of getting enough data to build a reliable corpus for our model. More specifically, we want to use a corpus of text in which part of the text is created by illiberal actors, and another part is not. The actors

linked to such text could be individuals, politicians, or political institutions. For our purpose, it would be interesting to explore if text produced by different actors gives us a similar conceptualization of illiberalism. For this reason, we explore the data availability for what concerns social media text (such as Instagram, Facebook, and X posts, also called Tweets), parliamentary documents, and manifestos as a first option.

Ideally, we would like to include data from the target countries in our study: Austria, Czechia, France, Hungary, Italy, Poland, and the United Kingdom. In Table 1 we include the country coverage for the datasets investigated for the analysis. We also decided to focus our analysis on the year 2000 because of reliability concerns of data in Central and Eastern Europe before this date. In the following sections, we briefly elaborate on our considerations regarding the three categories of data we could use.

Table 1. Data sources and characteristics

Dataset Name	Type Data	Country Coverage	Problems
Chapel Hill Expert Survey (CHES)	Complementary Variables	Yes	Not a text source
The Comparative Agendas Project (CAP)	Parliamentary Documents	Partial (Italy, France, Hungary, United Kingdom)	Not enough data for a large language model
The PopuList	Complementary Variables	Yes	Not a text source
ParlGov Project	Complementary Variables	Yes	Not a text source
ParlSpeech	Parliamentary Documents	Partial (Austria, Czechia, United Kingdom)	Not good enough coverage for Europe
V-Dem V-Party Dataset	Complementary Variables	Yes	Not a text source
Manifesto Project	Manifestos	Yes	Manifestos might be limited examples
Observatory for Political Texts in European Democracies (OPTED)	Complementary Variables	Not Published Yet	Not a text source
The Populism and Political Parties Expert Survey (POPPA)	Complementary Variables	Yes	Not a text source
Minet Tweets corpus	Social Media Data	Yes	Difficult to access due to recent API restrictions
Global Populism Database	Speeches of chief executives (presidents and prime ministers)	Yes	Smaller corpus than the Manifesto Project and not a text source per se
MAPLE Parliamentary Datasets	Parliamentary Documents	No	Comprehensive dataset for Belgium, Germany, Greece, Ireland, Portugal and Spain.

Social media text

Social media is a reference for classification models. This kind of text includes posts from famous platforms such as Instagram, Facebook, and X, previously known as Twitter. Especially X has a long tradition of making its tweets accessible to researchers via its API. Our preferred way of doing that would be by using a mechanism such as Crowdtangle and Minet. The process involves manually inputting all the relevant parties or official pages in Crowdtangle and then using Minet to download each tweet written by each party or official page of the list. At present, we are currently struggling to build this corpus because of the API recently becoming closed. However, we could also use already existing corpora of tweets such as the ones available at the WhatEvery1Says (WE1S) Project.

Parliamentary documents

A corpus of parliamentary documents could be manually built from the parties of interest for the relevant countries. It is a time-consuming and expensive option that we are still evaluating how to implement.

Manifestos

Political science articles often focus on electoral manifestos, which serve as a valuable source of information for researchers regarding the priorities, goals, and intentions of political parties. The importance of manifestos is so widespread that the Manifesto Project systematically collects them and codes them on a quasi-sentence level (Volkens, 2002). It also makes them available for public use via their online database. It also codes most of the collected documents on a semi-sentence level. Merz, Regel and Lewandoski explain in detail the possibilities linked to this data (Merz, Regel and Lewandowski, 2016).

The way we use manifestos to understand the priorities, goals, and intentions of political parties has come a long way. The text analysis of these documents has a long qualitative tradition,

starting in the 1970s and formalised for the first time in 1979 with the Manifesto Research Group. According to Slapin and Proksch (Slapin and Proksch, 2008), only documents that deal with the issue of interest should be compared to defining a concept.

Table 2. Manifesto project coverage of countries of interest

Country	Number of Distinct Manifestos	Time Range
Austria	7	2002-2019
Czechia	11	2002-2021
France	6	2002-2022
Hungary	6	2002-2022
Italy	14	2001-2022
Poland	6	2001-2019
United Kingdom	8	2001-2019

Training Data

For a text classification BERT model, we need a small but reliable set of training data. Labelled data is data that has been tagged with a label that represents a specific metric, property or class identification. The way you label the data has a significant effect on the reliability of the model. In a recent investigation (Feng et al., 2023), it was discovered that pre-trained language models, particularly those trained on extensive and varied datasets, have the potential to unintentionally replicate the biases ingrained within their training data.

As a consequence, this phenomenon can result in biased predictions in crucial domains such as identifying illiberalism. In such cases, the model might erroneously categorize an innocuous statement as illiberalism due to the biases it has absorbed. Moreover, pre-trained language models exhibit diverse perspectives concerning matters of social and economic significance. The BERT family of models seem to be more sensitive in classifying text as authoritarian compared to the GPT family. Also, the corpus on which the model is pre-trained seems to affect the task.

Consequently, it becomes important to evaluate different ways we can label our training data. In our work we evaluated three main options to label our training data: (i) using existing datasets providing annotation, (ii) using experts' knowledge to annotate the data, and (iii) using Chat GPT-4

to annotate the data. The last option is particularly problematic because Chat GPT-4 is indeed a generative deep learning model that is trained on extensive and varied datasets. It is consequently at a very high risk of replicating existing biases, as highlighted by the model's creators as well (OpenAI, 2023). Furthermore, in terms of replicability and reliability, it would be useful to avoid 'black box' models. Neural networks are widely known for being difficult to understand in the way they train and pre-train data. Therefore, they are not appropriate for many applications where transparency is important.

Chat GPT-4 could be useful to temporary label data in the first attempts of the analysis. However, in this case, it would make more sense to code sentences and quasi-sentences as illiberal using expert coding. For the Manifesto Project, it would make sense to select quasi-sentences as illiberal according to the dataset's existent coding. It is important to notice that the Manifesto Project coding is ambiguous, so we would first rely on the selection of a combination of codes based on the definition of illiberalism in the first section. Based on the Manifesto Project codebook and the definition in the first section, we code quasi-sentences containing negative mentions of individual rights and multicultural pluralism as illiberal.

This method however would not work for social media data, and it is in general reductive and ambiguous, as we would never be able to encompass all the fundamental characteristics of illiberalism. For this reason, the third option, using experts' knowledge to annotate the training data, seems the most appropriate. In practice, we would ask our AUTHLIB experts to hand-label approximately one hundred sentences based on the literature's definition of illiberalism. Experts would rate between one to three parties each according to the items list created by WP2 and using a Likert Scale. The idea is to have approximately one hundred labelled sentences. We chose this amount of sentences as it has been proved across different models to be the right amount to maximise performance. (Laurer et al., 2023).

The choice to use experts to label data, even if the most appropriate, still relies on the fundamental problem of how we can identify the minimum elements constituting illiberalism in text.

Therefore, we will use existing coding for manifesto data while we validate a set of characteristics, and we create the experts' dataset for the training data for our different data sources (tweets, speeches, debates, manifestos etc.).

Language Problem

Our target countries are Austria, Czechia, France, Hungary, Italy, Poland, and the United Kingdom. This mixture comes with the inherent problem of having to deal with multiple languages. Furthermore, our deep learning model would come pre-trained in different languages. So far, we considered three solutions to be able to translate our data into one language and build one model based on that.

The first option relies on the use of APIs such as the DeepL API and the Google Cloud Translation API. This solution is however expensive, and it does not make us able to assess the accuracy of such translation. The second option would be to use a model such as BERTopic to translate our text before inputting it into the model. BERTopic works great on small text snippets and that can use a multi-lingual embedding before clustering the quasi-sentences. The third option would be using a pre-trained multilingual sentence embedding model as the basis for training classifier models. We are currently working on using the last two options while temporarily using the first one.

5. Model(s) Construction Roadmap

Our paper reviews measurements and methodological innovations regarding illiberalism and using text data. In our literature we explored the conceptual problem of measuring illiberalism, why BERT models are the most innovative approach for a text classification problem like ours, and the data options available. In the following sections, we review the choices we are planning to undertake to put all these elements together. Our final goal is to build a model that will enable us able to create a map of 'illiberalisms' throughout Europe. In this section we explore the key decisions we will make:

the data we will use, the BERT model we will choose, and the fine-tuning process we will adopt. We will also describe the outputs created by such a model and how these translate into a 'map of illiberalisms'.

No matter the data we choose, we need to input data in our model that is related to our task. We decided to first use the Manifestos Project database because of its completeness and availability. The database provides annotated quasi-sentences for most parties in Europe and with good coverage for our years of interest. Overall, the Manifesto Project provides 4882 manifestos for 1280 parties in 61 countries. This provides a good pre-training basis for a BERT model. Providing a training dataset of approximately one hundred sentences also seems feasible. The already existing annotations also provide elements related to illiberalism such as negative mentions of multiculturalism.

As previously mentioned, we are currently also considering social media data and parliamentary documents as data sources. We will try different options for scraping social media texts, especially from X and Facebook. On the other hand, it is unlikely that we will be able to build a corpus of parliamentary documents to pre-train our model. We will also assure data reliability by using CleanLab, Galileo, LabelStudio or Argilla, which are common packages for data quality in machine learning using messy, real-world data and labels. They are commonly used to clean datasets and to identify low-level label quality to have clean data for training purposes. We will also potentially deal with class imbalance when fine-tuning by downsampling our data or changing the loss function. One way to do that is by giving more weight to the minority classes categorised.

We settle on the BERT family of models because they are built to find reliable new patterns in a vast amount of data. This family of models is also at the forefront of its field for what concerns text classification and the use of multiple languages. More interestingly, it has also never been used to classify illiberalism or political concepts more in general. A transformer architecture like the one in BERT models relies on two main components before using the training data. The first is a tokenizer that prepares the text in a clean format, which the model understands. The second is a model that

processes the tokenizer's output and returns a prediction, for example, which class an input belongs to.

These two components are the same independently of the model tasks (classification, translation, summarisation, etc.). To choose the best one for our purpose, we need to load specific classes from the transformers HugginFace library to perform specific tasks. The original BERT model is currently considered out of date in the NLP community. Being our purpose of multilingual classification with limited computational ability and a relatively small corpus, we could choose models such as DistilBERT (small and efficient), DeBERTa (a newer version of the model), or MdBERTa (multilingual by nature). We start by using the MdBERTa base because besides being multilingual is also a new, easy-to-use model that requires a small computational capacity. Also, it is widespread opinion in the literature that a small properly fine-tuned model not only uses less GPU computation but always performs better than a large non-fine-tuned model (Laurer et al., 2023).

Overall, we will first use a MdBERTa multilingual model trained on the whole corpus of manifestos and trained on approximately one hundred sentences. We will use the already existing Manifesto Project coding while in parallel we will develop a coding scheme based on the work of WP2. We will fine-tune the model hyperparameters and we will eventually repeat the process for the other potential data sources.

Once we have our first reliable results, we will also try to see why the BERT classifier classified things in a specific way. This way we could be able to uncover new underlying dynamics regarding illiberalism. At this point, we should be able to 'open' the model and see what led to a specific decision. Eventually, we would also consider building our classifier using PyTorch and Tensorflow instead of the HugginFace tensor model.

Once we finished this process, the expected output would be a large number of quasi-sentences accurately classified as illiberal. Looking at the sentences classified as illiberal would make us able to "re-establish" kind of words, and topics that are more common among actors where illiberalism is predominant. Options like topic modelling, cluster analysis, and word embeddings

could be used to uncover specific patterns. This will allow us to place actors among the individual conceptual components and consider any relationships between their placements.

6. Conclusion

In this work, we reviewed the underlying concepts related to the most recent advances in text classification. We will use a MdBERTa multilingual model trained on the whole corpus of manifestos available for our countries and years of interest to classify texts characterised by illiberalism. Using this model will make us able to be methodologically innovative and sound while using a new approach to classify illiberalism. Despite the multiple problems and decisions that will have to be taken along the way, from the creation of the training data to the use of different data sources, this will still give us the options to explore new underlying dynamics and potential new associations of words and topics to illiberalism. This will in turn make us able to create a map of text concepts related to illiberalism and we will be able to measure how different political actors relate to it.

7. References

Bowman, S.R. *et al.* (2015) 'A large annotated corpus for learning natural language inference', in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2015*, Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642. Available at: https://doi.org/10.18653/v1/D15-1075.

Devlin, J. *et al.* (2018) 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. Available at: https://arxiv.org/abs/1810.04805 (Accessed: 23 April 2023).

Feng, S. *et al.* (2023) 'From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models', in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. *ACL 2023*, Toronto, Canada: Association for Computational Linguistics, pp. 11737–11762. Available at: https://doi.org/10.18653/v1/2023.acl-long.656.

Halmai, G. (2021) *Illiberalism in East Central Europe*. Routledge. Available at: https://cadmus.eui.eu/handle/1814/74480 (Accessed: 29 August 2023).

Hawkins, K.A. *et al.* (2021) 'Measuring Populist Discourse : The Global Populism Database'.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, 9(8), pp. 1735–1780. Available at: https://doi.org/10.1162/neco.1997.9.8.1735.

Hopkins, D. *et al.* (2007) ‘Extracting systematic social science meaning from text’.
illiberalism.org (2021) ‘Whats is illiberalism? Definition of illiberalism | illiberalism.org’, *https://www.illiberalism.org/*. Available at: https://www.illiberalism.org/definition-of-illiberalism/ (Accessed: 29 August 2023).
Koh, A. and Boey, D.K.S. (2021) ‘Student Projects Showcase: Predicting Policy Domains and Preferences with BERT and Convolutional Neural Networks’. Available at: https://hertie-data-science-lab.github.io/student-projects/posts/predicting-policy-domains-and-preferences-with-bert-and-convolutional-neural-networks/ (Accessed: 12 April 2023).
Laruelle, M. (2022) ‘Illiberalism: a conceptual introduction’, *East European Politics*, 38(2), pp. 303–327. Available at: https://doi.org/10.1080/21599165.2022.2037079.
Laurer, M. *et al.* (2023) ‘Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI’, *Political Analysis*, pp. 1–17. Available at: https://doi.org/10.1017/pan.2023.20.
Laver, M., Benoit, K. and Garry, J. (2003) ‘Extracting Policy Positions from Political Texts Using Words as Data’, *American Political Science Review*, 97(2), pp. 311–331. Available at: https://doi.org/10.1017/S0003055403000698.
Merz, N., Regel, S. and Lewandowski, J. (2016) ‘The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis’, *Research & Politics*, 3(2), p. 2053168016643346. Available at: https://doi.org/10.1177/2053168016643346.
Mudde, C. (2010) ‘The Populist Radical Right: A Pathological Normalcy’, *West European Politics*, 33(6), pp. 1167–1186. Available at: https://doi.org/10.1080/01402382.2010.508901.
Mullen, T. (2006) ‘A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse’. Available at: https://www.academia.edu/66280569/A_Preliminary_Investigation_into_Sentiment_Analysis_of_Informal_Political_Discourse (Accessed: 20 April 2023).
Nasteski, V. (2017) ‘An overview of the supervised machine learning methods’, *HORIZONS.B*, 4, pp. 51–62. Available at: https://doi.org/10.20544/HORIZONS.B.04.1.17.P05.
Nie, Y. *et al.* (2020) ‘Adversarial NLI: A New Benchmark for Natural Language Understanding’, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, Online: Association for Computational Linguistics, pp. 4885–4901. Available at: https://doi.org/10.18653/v1/2020.acl-main.441.
OpenAI (2023) ‘GPT-4 Technical Report’. arXiv. Available at: https://doi.org/10.48550/arXiv.2303.08774.
Sajó, A., Uitz, R. and Holmes, S. (2021) *Routledge Handbook of Illiberalism*. Routledge.
Singh, A., Thakur, N. and Sharma, A. (2016) ‘A review of supervised machine learning algorithms’, in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 1310–1315.
Slapin, J.B. and Proksch, S.-O. (2008) ‘A Scaling Model for Estimating Time-Series Party Positions from Texts’, *American Journal of Political Science*, 52(3), pp. 705–722. Available at: https://doi.org/10.1111/j.1540-5907.2008.00338.x.
Thomas, M., Pang, B. and Lee, L. (2012) ‘Get out the vote: Determining support or opposition from Congressional floor-debate transcripts’. arXiv. Available at: https://doi.org/10.48550/arXiv.cs/0607062.

Vaswani, A. *et al.* (2023) ‘Attention Is All You Need’. arXiv. Available at: https://doi.org/10.48550/arXiv.1706.03762.
Volkens, A. (2002) ‘Manifesto Coding Instructions’, *Berlin* [Preprint].
Williams, A., Nangia, N. and Bowman, S.R. (2018) ‘A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference’. arXiv. Available at: https://doi.org/10.48550/arXiv.1704.05426.
Xu, K. *et al.* (2016) ‘Show, Attend and Tell: Neural Image Caption Generation with Visual Attention’. arXiv. Available at: https://doi.org/10.48550/arXiv.1502.03044.
Young, L. and Soroka, S. (2012) ‘Affective News: The Automated Coding of Sentiment in Political Texts’, *Political Communication*, 29(2), pp. 205–231. Available at: https://doi.org/10.1080/10584609.2012.671234.
Zavestoski, S. (2005) ‘Language processing technologies for electronic rulemaking: a project highlight’. Available at: https://www.academia.edu/7459648/Language_processing_technologies_for_electronic_rulemaking_a_project_highlight (Accessed: 20 April 2023).