An efficient Bayes classifier for word classification: an application on the EU Recovery and Resilience Plans

Limosani, Michele and Millemaci, Emanuele and Mustica, Paolo

University of Messina

2023

An efficient Bayes classifier for word classification: an application on the EU Recovery and Resilience Plans

Michele Limosani[a] **, Emanuele Millemaci**[a] **, Paolo Mustica**[a]*

[a] Department of Economics, University of Messina, Piazza Pugliatti 1, 98122, Messina, Italy
* Corresponding author
E-mail addresses: limosani@unime.it (Michele Limosani); emillemaci@unime.it (Emanuele Millemaci); fabri.must@gmail.com (Paolo Mustica).

Abstract

This paper proposes the Prior Adaptive Bayes (PAB) classifier, a new algorithm to assign words appearing in a text to their respective topics. It is an adaption of the Bayes classifier where, as the prior probabilities of classes, their posterior probabilities associated with the adjacent words are used. Simulations show an improvement of more than 20% over the standard Bayes classifier. The PAB classifier is applied to the Recovery and Resilience Plans (RRPs) of the 27 European Union member states to evaluate their alignment with the environmental dimension of the Sustainable Development Goals (SDGs) as compared to the socioeconomic one. Results show that the attention paid by the countries to the pro-environment SDGs increases with the funds per capita assigned, the gap in the environmental endowment and the touristic attractiveness. Finally, the environmental dimension appears associated positively with available GDP growth projections for the next few years.

JEL: C82, H22, O44

Keywords: textual analysis; Prior Adaptive Bayes classifier; Recovery and Resilience Plans; Sustainable Development Goals; pro-environment policy

Declarations of interest: none

1. Introduction

The work of a researcher is undoubtedly eased when he has access to information that is organized in a dataset rather than in a text composed of words. Numerical measures undoubtedly provide many advantages in comparison to a text, including the possibility to perform statistical tests. Textual analysis can provide the useful service of transforming the information of a text into a numerical one, thus making possible statistical analysis.

Picault and Renault (2017) developed a field-specific dictionary to measure the stance of the European Central Bank (ECB) monetary policy and the state of the Eurozone economy through the content of ECB press conferences. Starting from this lexicon, they computed a monetary policy indicator and an economic outlook indicator by analyzing the words appeared in each introductory statement of ECB press conferences. Their results show that the proposed dictionary explains future ECB monetary decisions and market volatility. Renault (2017) implemented a novel approach to derive investor sentiment from messages posted on social media. To do this, he constructed a lexicon of words used by stock market investors on social media and tested it on a test set of tagged messages. The accuracy achieved by this lexicon outperformed two well-known dictionaries usually adopted to measure sentiment in newspaper articles. Then, he used his lexicon to examine the relationship between the sentiment of stock market investors and intraday stock returns using a dataset of messages published by online investors on the microblogging platform StockTwits, finding that change in investor sentiment predicts positively the S&P 500 index ETF returns. Thorsrud (2020) constructed a daily business cycle index based on quarterly GDP growth and textual information contained in a daily business newspaper. For the analysis of newspaper data, he combined supervised and unsupervised methods, respectively a dictionary-based technique and a topic model belonging to the Latent Dirichlet Allocation class. He demonstrated that his index classifies the phases of the business cycle with almost perfect accuracy, outperforming coincident indexes based on more traditional economic variables. Alfano and Guarino (2022) analyzed the impact of text structure and given keywords in the announcements of house sales over the internet, finding that using many nouns and adjectives in writing a house sale announcement helps to sell the property at a higher price. More recently, Aprigliano *et al.* (2023) proposed a text-based sentiment index and an economic policy uncertainty index for forecasting Italian economic activity using a dictionary-based approach. They built a dictionary by downloading textual data from four popular national newspapers and assigned to each dictionary item a positive or negative polarity. Then, they used their dictionary to construct the two indices, which were therefore used in econometric analysis. The text-based models were usually able to produce more accurate forecasts of several macroeconomic indicators, such as the variation of the GDP, in comparison to the baseline model not accounting for the text indices.

Although simple textual analysis methodologies, such as the dictionary-based approach, have been used by most of the literature, they present some limitations. The first is that this approach usually needs a field-specific dictionary. Although in the literature there are some validated dictionaries to measure sentiment in the traditional media types, such as the Harvard-IV dictionary (Tetlock, 2007) and the LM dictionary (Loughran and McDonald, 2011), they might not be suitable in those cases where a specific jargon is predominant, such as the comments on financial issues reported in social networks. In these situations, the selection of the words to include in the dictionary may be a very subjective choice. Second, while the dictionary-based approach is usually used for a specific task, that of revealing a sentiment (usually positive or negative) in a text, its use is difficult in non-sentiment analyses. Third, the dictionary-based approach implies that the researcher uses only a few words or groups of words to assign a sentiment, discarding the majority of the words in a text and giving up their potentially interesting content (Hastie *et al.*, 2015).

A step forward in the analysis of unstructured data as textual data is the use of machine learning methods, which usually perform better than dictionary-based approaches (Kalamara *et al.* 2022). Several machine learning algorithms are available depending on the nature of the tasks to be implemented. Athey and Imbens (2019) surveyed machine learning methods that are very popular in the context of regression analysis in economics such as the LASSO and ridge regression (Hoerl and Kennard, 1970; Tibshirani, 1996), regression trees and random forests (Breiman *et al.*, 1984; Breiman, 2001), and neural networks and related deep learning methods (Hornik *et al.*, 1989; White, 1992). On the other hand, the classification of textual data in general requires the Bayes classifier (Sahami *et al.*, 1998; Wang, 2010). The Bayes classifier not only has the great advantage that it works well with textual data, but it is also easy to understand and transparent as compared to other complex methods, such as those based on neural networks (Ash and Hansen, 2022).

This paper proposes a new textual algorithm, an adaption of the Bayes classifier for the word classification task. Its main characteristic is that the prior probabilities of topics are not constant but adapt to the corresponding posterior probabilities associated with the adjacent words. Simulations show that this classifier achieves an improvement of more than 20% over the original classifier. As an example of application of the classifier, this paper focuses on the Recovery and Resilience Plans (RRPs) of the 27 European Union member states. Considering that the guiding principle for the realization of the RRPs is the new growth paradigm of competitive sustainability with which the SDGs are strongly associated (European Commission, 2019), the proposed algorithm is used to evaluate the alignment of the RRPs with the environmental Sustainable Development Goals (SDGs) as compared with the socioeconomic ones. The idea is to first train the classifier with documents related to each SDG and subsequently use it to assign each word contained in the name and description of each RRP project to the more appropriate SDG (or group of SDGs). Results show that the attention paid by the countries to the SDGs related to the environmental dimension increases with the funds per capita assigned, the gap in the environmental endowment, and the touristic attractiveness. Finally, the environmental dimension appears associated positively with available GDP growth projections for the next few years.

To the best of our knowledge, the contributions aiming at exploring the coherence of the RRPs with the SDGs are still few and focus on specific case studies.[1] The lack of numerical measures associating univocally funds of the RRPs and SDGs is important and difficult to get around. The projects contained in the RRPs with their respective costs are indeed associated with at least one of the six pillars described by the European Commission ("green transition", "digital transformation", "smart, sustainable and inclusive growth", "social and territorial cohesion", "health, social and institutional resilience" and "policies for the next generation"). However, this information is only available once a project is completed, and, in any case, the problem of identifying the corresponding SDG for each pillar remains unsolved.

The paper is structured as follows: section 2 presents a new adaptation of the Bayes algorithm for the word classification task, and tests its performance by simulation. As an application, section 3 investigates an economic and environmental issue and shows how the classifier contributes to a better understanding in cases where the only available information is of the textual type. Section 4 concludes.

[1] For instance, Rotondo *et al.* (2022) analyze the relationships between the domains of the SDGs and the Mission 2 of the Italian RRP. Recent studies not strictly related to the RRPs that tried to map the coverage of the SDGs in European documents are those realized by Borchardt *et al.* (2020) and Koundouri *et al.* (2021). In both cases, the authors manually defined some keywords associated with the SDGs, and mapped their presence in some European documents.

2. The algorithm

In a text, words are not randomly distributed but clustered in topics: a group of words generates a topic, another group generates another topic, etc. (Hildum, 1963). In the task of associating words to topics, looking at the adjacent words can help minimize mistakes. To better illustrate this point, consider the following sentence:

"*For me, the best animals are cats and dogs, while the best dishes are cheeseburgers and hot dogs*"

The sentence contains two topics: the first part is about animals, and the second one is about food. After excluding stop words (Alshanik *et al.*, 2020), the sentence becomes:

"*animals cats dogs dishes cheeseburgers hot dogs*"

The word "dogs" occurs twice but it does not belong to the same topic in both cases. To correctly associate each word "dogs" with its respective topic, the adjacent words can be used to improve prediction. One may consider it reasonable to assign the first "dogs" to the topic of animals, being it close to the words "animals" and "cats", and the second "dogs" to the topic of food, being it close to the words "dishes", "cheeseburgers" and "hot".

2.1. The Bayes classifier

A well-known algorithm used for document classification is the Bayes classifier (Mitchell, 2019). For the classification of individual words, the Bayes classifier can be written as follows:

$$\arg\max_j pr(c_j|x_i = w_k) = \frac{pr(x_i = w_k|c_j)pr(c_j)}{\sum_j pr(x_i = w_k|c_j)pr(c_j)}$$

where x_i is the i-th word in the text, w_k is the k-th word in the vocabulary and c_j is the j-th class. Using the framework of Bayes' theorem, $pr(c_j|x_i = w_k)$ is the posterior probability of the class j given the word i in the text, $pr(x_i = w_k|c_j)$ is the likelihood of c_j given a fixed x_i and $pr(c_j)$ is the prior probability of class j. According to the Bayes classifier, x_i will be classified in the class with the maximum posterior probability. For each i-th word in the text, the denominator – known as the normalizing constant – is the same for all classes. Since it is a constant, the denominator can be omitted and the maximization problem can be rewritten in the following way:

$$\arg\max_j pr(x_i = w_k|c_j)pr(c_j)$$

The likelihood and the prior probability are usually estimated with the frequentist approach, starting from a training set, or with the subjectivist approach, making assumptions about them. In particular:

- $pr(x_i = w_k | c_j)$ is given by the relative frequency of word k in the vocabulary labeled with class j. To avoid the zero-frequency problem for the probability of the intersection (see below), an observation is usually added for each $w_k | c_j$ before the corresponding relative frequency is calculated (Hae-Cheon *et al.*, 2020).
- $pr(c_j)$ is given by the relative frequency of all words in the vocabulary labeled with class j or assuming a uniform distribution for the distribution of classes (Peng *et al.*, 2004). The uniform distribution will be considered during the discussion, but the results are similar for the empirical distribution.

Taking the previous example, the Bayes classifier provides the following results (Table 1):

Table 1. Results of word classification with the Bayes classifier

pr(animals|A)pr(A)	0.11	pr(animals|F)pr(F)	0.05
pr(cats|A)pr(A)	0.11	pr(cats|F)pr(F)	0.05
pr(dogs|A)pr(A)	0.11	pr(dogs|F)pr(F)	0.10
pr(dishes|A)pr(A)	0.06	pr(dishes|F)pr(F)	0.10
pr(cheeseburgers|A)pr(A)	0.06	pr(cheeseburgers|F)pr(F)	0.10
pr(hot|A)pr(A)	0.06	pr(hot|F)pr(F)	0.10
pr(dogs|A)pr(A)	0.11	pr(dogs|F)pr(F)	0.10

The values reported in the Table are the products between the likelihood of a class given a fixed word and the prior probability. On the left are reported the values associated with the animal topic (A), on the right those associated with the food topic (F). The classifier classifies each word in the class with the maximum product. Words correctly classified and misclassified are highlighted in light green and light red, respectively.

Words that are correctly classified are highlighted in light green, while those that are misclassified are highlighted in light red. The Bayes classifier classifies words regardless of their position in the text. Only when a word is univocally associated with a topic, the classifier classifies correctly. In our example, the word "dogs" related to the topic of food is wrongly associated with the topic of animals.

2.2. The Bayes classifier with adaptive a priori probabilities

To address this problem, the algorithm exploits the rule that words in a text are clustered in topics (topic clustering). As a consequence, the prior probabilities of a topics are not constant for each word but vary according to the topic of the previous word or group of words. Going back to the example above, if the word "dogs" is preceded by words belonging to the topic of animals, the algorithm associates a higher prior probability to this topic.

To capture the topic clustering, the algorithm adapts the prior probabilities of topics to the corresponding posterior probabilities associated with the previous words. From a mathematical point of view, the priors are replaced by the corresponding posteriors associated with the previous p words:

$$\arg\max_j pr\left(x_i = w_k \middle| c_j\right) pr\left(c_j \middle| x_{i-1} \cap \ldots \cap x_{i-p}\right)$$

The posterior probabilities can be computed by assuming that words are independent (bag-of-words assumption; Ercan and Cicekli, 2012). Consequently, the probability of the intersection of p preceding words is given by the product of their probabilities:

$$pr\left(c_j \middle| x_{i-1} \cap \ldots \cap x_{i-p}\right) = \frac{pr\left(x_{i-1} \cap \ldots \cap x_{i-p} \middle| c_j\right) pr\left(c_j\right)}{\sum_j pr\left(x_{i-1} \cap \ldots \cap x_{i-p} \middle| c_j\right) pr\left(c_j\right)} = \frac{pr\left(x_{i-1} \middle| c_j\right) \ldots pr\left(x_{i-p} \middle| c_j\right) pr\left(c_j\right)}{\sum_j pr\left(x_{i-1} \middle| c_j\right) \ldots pr\left(x_{i-p} \middle| c_j\right) pr\left(c_j\right)}$$

Going back to the previous example, the classifier provides only correct results with $p = 1$ (Table 2), while it does not happen with $p = 2$ (Table 3). In the latter case, while the words "dogs" are still correctly classified, the word "dishes" is not anymore. This outcome is due to the existence of a trade-off: since words in a text are clustered in topics, the greater the number of previous words chosen associated with a topic, the higher the probability that the next word is classified in the same topic. This implies a lower sensitivity of the classifier to the new topic during the change of topic, as occurred for the word "dishes", which is the first word appeared in the topic of food. The optimal number of previous words to account for depends on the texts but, in general, it is likely to be a small one. The simulation reported in the subsection below provides evidence for the case of our example.

Table 2. Results of word classification with the proposed classifier ($p = 1$)

pr(cats|A)pr(A|animals)	0.15	pr(cats|F)pr(F|animals)	0.03
pr(dogs|A)pr(A|cats)	0.15	pr(dogs|F)pr(F|cats)	0.06
pr(dishes|A)pr(A|dogs)	0.06	pr(dishes|F)pr(F|dogs)	0.10
pr(cheeseburgers|A)pr(A|dishes)	0.04	pr(cheeseburgers|F)pr(F|dishes)	0.13
pr(hot|A)pr(A|cheeseburgers)	0.04	pr(hot|F)pr(F|cheeseburgers)	0.13
pr(dogs|A)pr(A|hot)	0.08	pr(dogs|F)pr(F|hot)	0.13

The values reported in the Table are the products between the likelihood of a class given a fixed word and the posterior probability associated with the previous word. The left-hand side reports the values associated with the animal topic (A), while the right-hand side reports the values associated with the food topic (F). The classifier classifies each word in the class with the maximum product. Words correctly classified and misclassified are highlighted in light green and light red, respectively.

Table 3. Results of word classification with the proposed classifier ($p = 2$)

pr(dogs|A)pr(A|cats∩animals)	0.19	pr(dogs|A)pr(A|cats∩animals)	0.03
pr(dishes|A)pr(A|dogs∩cats)	0.08	pr(dishes|A)pr(A|dogs∩cats)	0.06
pr(cheeseburgers|A)pr(A|dishes∩dogs)	0.04	pr(cheeseburgers|A)pr(A|dishes∩dogs)	0.12
pr(hot|A)pr(A|cheeseburgers∩dishes)	0.03	pr(hot|A)pr(A|cheeseburgers∩dishes)	0.15
pr(dogs|A)pr(A|hot∩cheeseburgers)	0.05	pr(dogs|A)pr(A|hot∩cheeseburgers)	0.15

The values reported in the Table are the products between the likelihood of a class given a fixed word and the posterior probability associated with the two previous words. The left-hand side reports the values associated with the animal topic (A), while the right-hand side reports the values associated with the food topic (F). The classifier classifies each word in the class with the maximum product. Words correctly classified and misclassified are highlighted in light green and light red, respectively.

2.3. Testing the classifier: data and main results

The Bayes classifier with adaptive a priori probabilities as described in the section above was tested on a well-known dataset usually used for testing the performance of text classification algorithms. The dataset is Reuters-21578, a collection of 21578 documents that appeared on the Reuters newswire in 1987 (Debole and Sebastiani, 2005; Pinheiro *et al.*, 2012; Zhang *et al.*, 2019). The documents were assembled and indexed with categories by personnel from Reuters Ltd. and Carnegie Group Inc. in 1987. Starting from the 90s, David D. Lewis and other researchers have formatted the documents, produced the associated data, and cleaned the collection.

Excluding units without content and units not labeled as training or test units, the number of documents is 18,323. 8,298 documents are labeled with a (unique) topic: this analysis considers textual data, or clusters of words, labeled with one topic. 25% and 45% of textual data are labeled with topics *acq* and *earn*, respectively. Thirty percent belongs to one of the 63 remaining categories. Due to the high fragmentation, these clusters of words are grouped into one broad class, named *other* (Table 4).

Table 4. Number of training and test data labeled as *acq*, *earn* and *other*

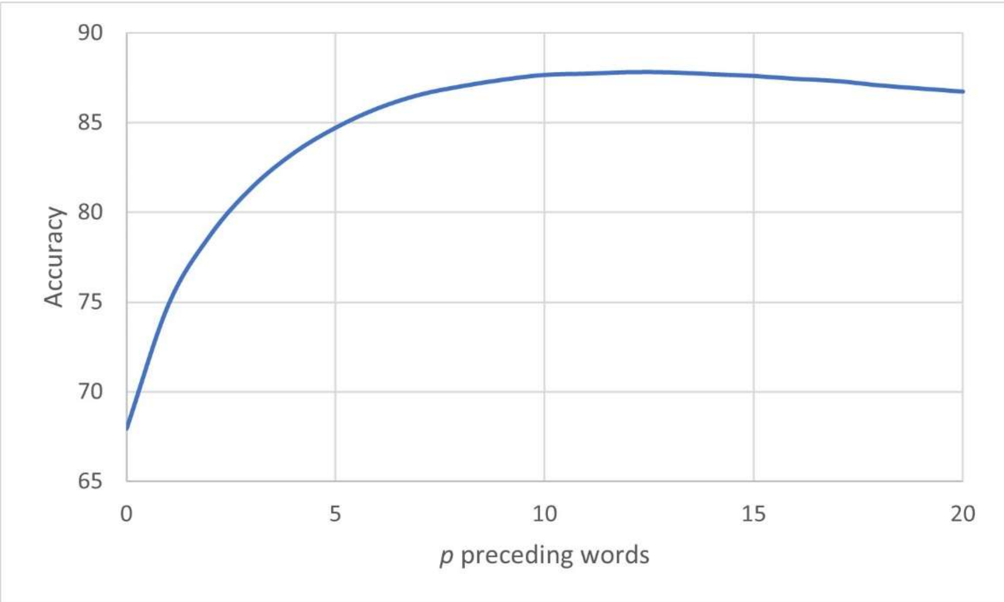
topic/set	training	test	total
acq	1435	620	2055
earn	2673	1041	3714
other	1841	688	2529
total	5949	2349	8298

Each cell contains the number of textual data labeled with one topic (*acq*, *earn* or *other*) in each set (training or test).

5,949 textual data trains the classifier, meaning that the words contained in these data are used to estimate the likelihood. Then, the classifier uses 2,349 textual data to predict the classes for the words contained in the test set. Before training the classifier, stop words, numbers, and stems are removed to reduce the dimensionality of data (Singh and Gupta, 2016) and, therefore, the problem of sparsity (Hastie *et al.*, 2015).

Textual data related to the test set are supposed to belong to a unique document. There are 2,349 clusters of words, corresponding to 128,948 words (a mean of 55 words per cluster). One way to evaluate the performance of a classifier is to analyze the accuracy, measured as the percentage of words correctly classified (Bramer, 2020). Figure 1 shows the accuracy for both the original classifier not considering preceding words ($p = 0$) and the proposed classifier considering a positive number of preceding words ($1 \leq p \leq 20$).

Figure 1. Accuracy by the number of preceding words



The original classifier provides an accuracy of 67.96%. As the number of preceding words increases, the accuracy of the proposed classifier increases but at a decreasing rate. Moreover, starting from a certain point, the accuracy starts decreasing slowly. The maximum accuracy is achieved at the number of 12 preceding words (87.81%), an improvement of about 20% over the original classifier. It is worth noting that the adaptive classifier already provides a noticeable improvement considering a few previous words. For example, in setting $p = 2$, the accuracy is 78.74% (+10.77%, as compared to $p = 0$).

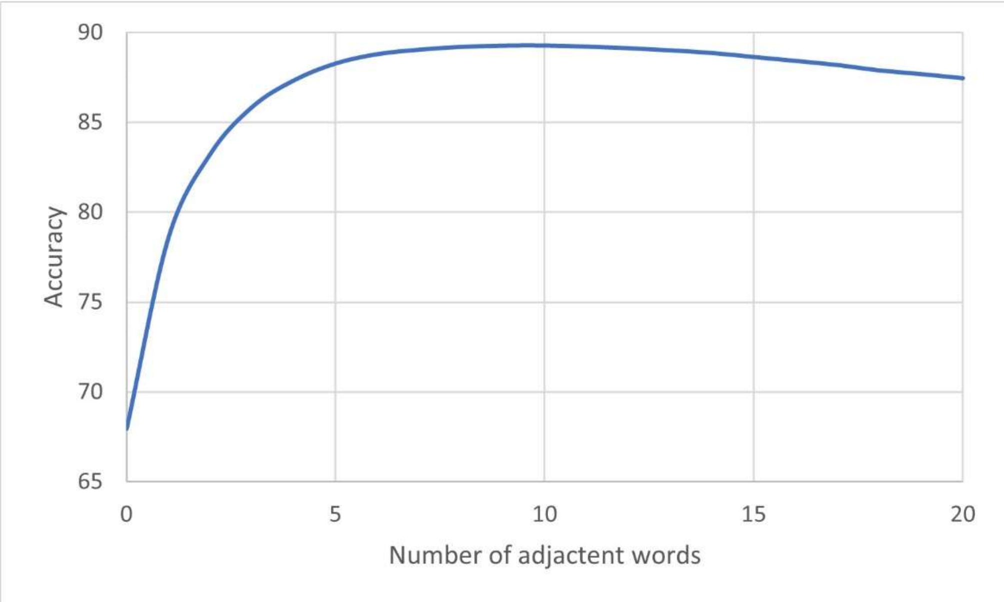
2.4. Mixed strategy

The strategy described so far is backward, meaning that each word is classified taking into account the previous word or group of words. On the contrary, by selecting both preceding and following words, the maximization problem becomes:

$$\arg\max_j pr(x_i = w_k|c_j)pr(c_j|x_{i-p} \cap \dots \cap x_{i-1} \cap x_{i+1} \cap \dots \cap x_{i+q})$$

where p and q are the numbers of preceding and following words, respectively. This methodology can be symmetrical as the classification of each word can rely on the same number of preceding and following words.

Figure 2. Accuracy by number of adjacent words (symmetrical approach)



This strategy reaches a higher accuracy than the backward method (Figure 2). While the maximum accuracy reached with the backward strategy was 87.81%, the mixed strategy reaches a maximum accuracy of 89.27%. In this case, the maximum improvement over the original classifier is 21.25%. As in the case of the adaptive method, the proposed classifier has an accuracy that increases with p and q at decreasing rates. Moreover, starting from a certain level of p and q, the accuracy begins to decrease slowly. For this reason, choosing a parsimonious number of adjacent words is preferable.

3. Using the classifier to determine the alignment of the RRPs with the pro-environment SDGs

To mitigate the economic and social impact of the coronavirus pandemic and make European economies and societies more resilient and sustainable, the European Parliament and the

Council of the European Union approved the regulation 2021/241, which establishes the Recovery and Resilience Facility (RRF). The RRF is a temporary recovery instrument that allows the European Commission to raise funds to finance member states' reforms and investments in line with the EU's priorities (Karaboytcheva, 2021). To benefit from the support of the RRF, member states submit to the European Commission their Recovery and Resilience Plans (RRPs), in which the reforms and investments to be implemented by end-2026 are set out (Dias *et al.*, 2021).

In what follows, the classifier is employed to measure the alignment of the RRPs with the pro-environment SDGs as compared with the socioeconomic ones, using a validated model to univocally assign each SDG to two broad categories. This analysis is structured as follows: the first subsection introduces the SDGs model, the second subsection presents the data, the third subsection introduces an index of pro-environment relative intensity and the fourth one shows the results.

3.1. The SDGs model

The SDGs are 17 goals defined by the UN in 2015 to address global challenges by 2030, such as inequality and poverty, climate change and environmental degradation, justice and peace (UN, 2022). These 17 SDGs can be grouped into a more compact number of categories or dimensions. In this regard, the Stockholm Resilience Centre introduced the Wedding Cake Model (Folke *et al.*, 2016), consisting of the absorption of all SDGs into three broad categories, namely biosphere protection, social cohesion, and economic growth (Figure 3). As SDG 17 (Partnership for the goals) is not associated with a specific dimension but shared by all three dimensions, it is excluded from the analysis. Figure 4 shows a furtherly simplified scheme: on the one hand, the biosphere protection (environment), and, on the other hand, social cohesion and economic growth put together (socioeconomic). This scheme allows us to make a direct comparison between the SDGs related to environmental issues and those SDGs not associated with this dimension. Note that, in contradiction with the assignment of the model, this study moves SDG 2 (Zero hunger) into the environmental dimension because, in the context of the EU, SDG 2 mainly asks for policies of sustainable food creation and resilient agricultural practices (European Commission, 2021a).

Figure 3. The Wedding Cake Model

Source: Folke *et al.*, 2016.

Figure 4. The environmental and socioeconomic dimensions

As mentioned above, the guiding principle for the realization of the RRPs is competitive sustainability, a new growth paradigm with which the SDGs are strongly associated. Since the RRPs are written as a function of the SDGs, the alignment of each RRP with the specified dimensions of the Wedding Cake Model is an interesting research question aiming at uncovering the priorities of national governments of the EU member states. Specifically, this classifier can help determine the number of words associated with the environmental and socioeconomic dimensions. Moreover, a measure of the relative intensity of the two dimensions can be used to identify the factors more intensely associated with them, including the environmental status quo and tourist attractiveness.

3.2. Data

The RRPs were assessed by the European Commission approximately two months after submission by the countries. After this assessment, the plan has been definitively approved by the European Council in an additional month. The attached document to the Council implementing decision (the annex) describes in detail the reforms and investments to be implemented by end-2026 (Dias *et al.*, 2022). Each annex of the RRPs contains the name, the description and the completion time of the projects. Overall, 6233 projects can be univocally attributed to a specific year in the range between 2020 and 2026.

A training set of documents related to each SDG has been used to compute the likelihood necessary to execute the classifier. To assure that these documents were coherent with the EU view and objectives, we considered the reading list on each SDG suggested by Eurostat in the 2020 report on progress towards the SDGs (Eurostat, 2020), selecting about 120 documents. Before training the classifier, we removed stop words and numbers. Moreover, we reduced the vocabulary to the one used in the projects of the RRPs to reduce the sparsity due to words not being used in the documents to be analyzed.

3.3. An index of the incidence of the environmental topics

As the experiment presented so far showed that two adjacent words are enough to obtain a noticeable improvement in accuracy, this application exercise makes use of the symmetrical approach with 1 backward lag and 1 forward lag.[2] Since each word contained in the projects of the RRPs is not labeled with its class, we cannot evaluate the accuracy of the classifier. As a validation tool, we rely on Zipf's law, for which very few words dominate the word count distribution (Manning and Schütze, 2003). This means that analyzing the most frequent words classified in a specific class is useful to understand if the classifier works correctly. In the supplemental data online, we reported the top 75 words of the environmental and socioeconomic dimensions. The list appears in line with expectations as we cannot identify words that are not reasonably used in the dimension in which they are classified.

[2] Notice that the use of 2 backward lags and 2 forward lags generates essentially the same results (available upon request from the authors).

Once the classifier estimated the number of words associated with each dimension, we built a measure of alignment of these projects with the two dimensions. Thus, for each project we calculated the relative intensity index as follows:

$$Alignment = \frac{n_{env} - n_{soceco}}{n_{env} + n_{soceco}}$$

where n_{env} and n_{soceco} are the number of words classified in the environmental dimension and the socioeconomic dimension, respectively. The index is constrained in the range –1 and 1. If alignment = –1, the project is fully aligned with the socioeconomic dimension, while if alignment = 1, the project is fully aligned with the environmental dimension.

The mean value over all projects is equal to –0.1835. The negative sign of the index may someway suggest that a relatively larger number of words are related to the socioeconomic dimension, as well as it may be simply a consequence of the fact that this dimension includes more SDGs than the environmental dimension. The standard deviation is 0.5121 suggesting that the observed series has a large variability.

As a further validation exercise, we plotted the index against the six pillars in which the RRPs are organized. Although the dimensions of the SDG model and the pillars do not overlap precisely, it is still possible to make predictions on the sign of the correlations. The figures in the supplemental data online display evidence in line with our expectations.

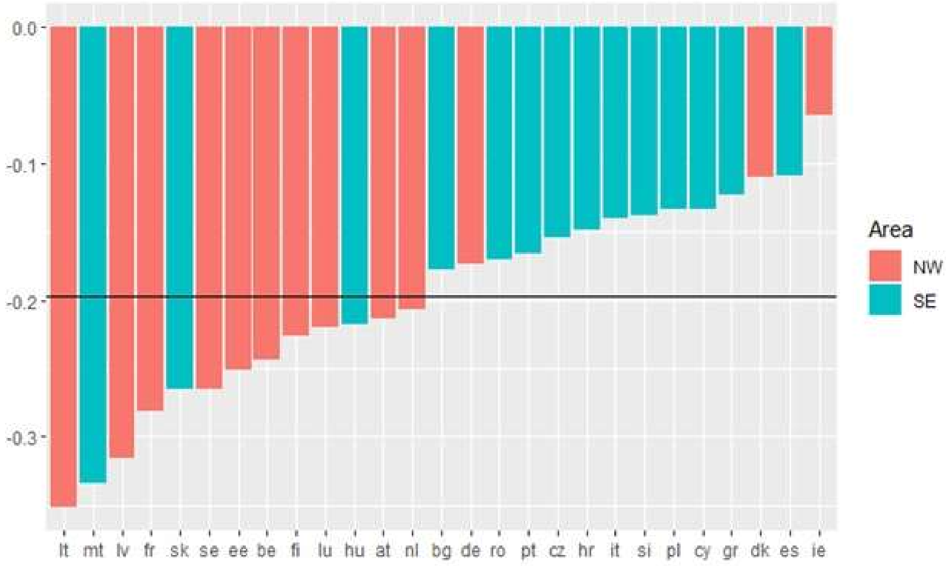
3.4. Results

3.4.1. The index and the geographical area

Is the index clustered by geographical area? If so, which cluster of countries is most aligned with the environmental dimension? Figure 5 shows the sorted distribution of the index for each country in the European Union.

The observations are grouped into two geographical clusters, according to the UN *geoscheme*: Northwest countries (red bars) and Southeast countries (blue bars); the horizontal line indicates the average value. On average, the cluster of Southeast countries tends to be more aligned with the environmental dimension with two noticeable exceptions (Denmark and Ireland).

Figure 5. Sorted distribution of the index

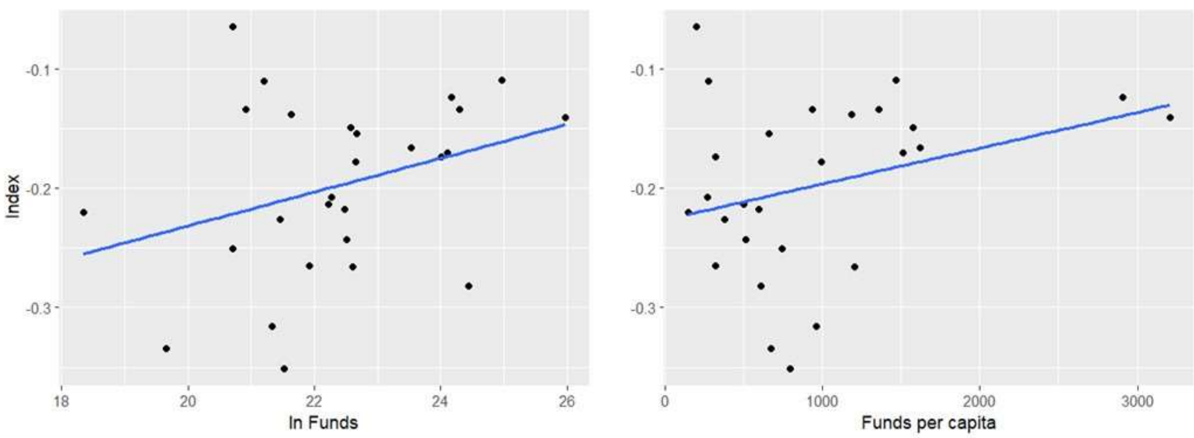


3.4.2. The index and the RRP funds

Are the countries that receive more funds from the RRF more interested in environmental or socioeconomic issues? We can expect that those countries receiving the larger financial support are those programming more balanced interventions across the two dimensions because they are not forced to focus only on the socioeconomic dimension that is generally considered higher on the political agenda.

Hypothesis 1: The countries receiving more funds show a higher value of the index (H1)

Figure 6. Scatter plots of the index against the log funds and the funds per capita



For each scatter plot the trend line is reported.

Both Figure 6, which reports the scatter plots of the index against the log funds and the funds per capita, and their correlation coefficients (0.33 and 0.31, respectively) give support to H1.

3.4.3. The determinants of the index

The index can be a useful tool to identify the main reasons behind different cross-country mixes of environmental vs. socioeconomic projects in the RRPs. For instance, countries lagging behind from an environmental perspective might want to fill this gap, by paying more attention to the environmental dimension. Moreover, countries where tourism is a key or growing economic sector should be more interested in investing more resources in the environmental dimension.

H2: The countries where the gap in the environmental endowment is larger show a higher value of the index

H3: The countries where the tourism sector plays a more important role show a higher index value

For this analysis, we selected measures for a country's current environmental status and tourism specialization. A brief description is provided in Table 5.

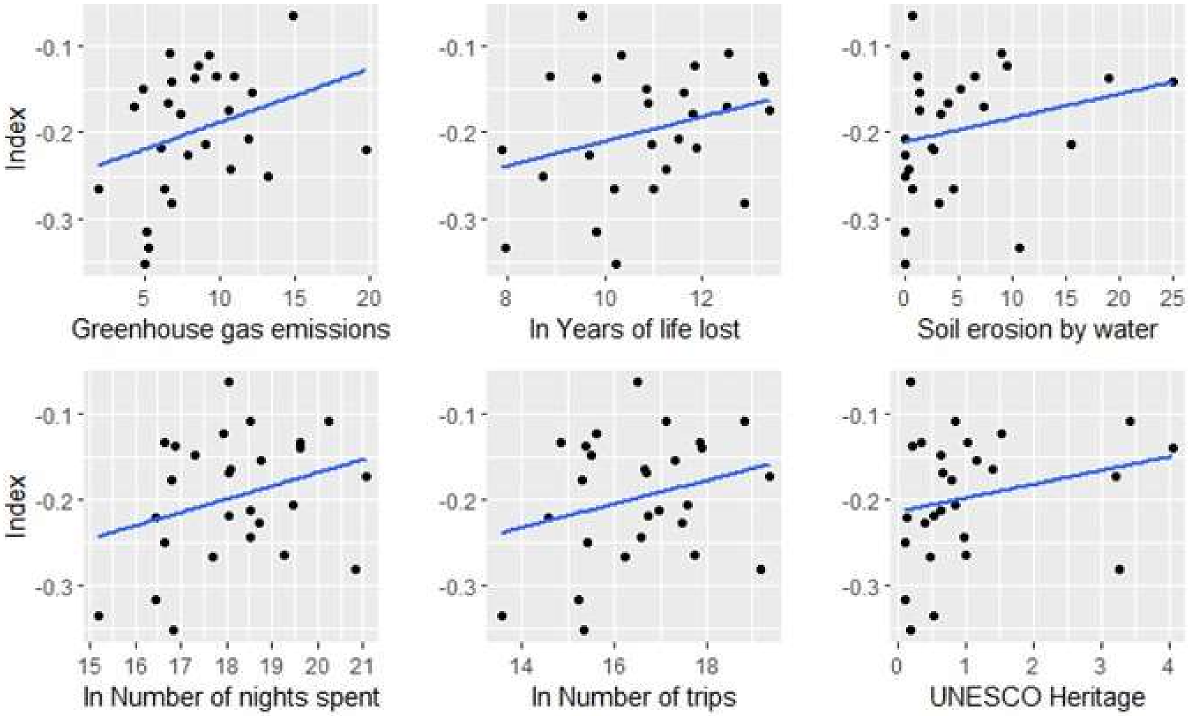
Table 5. Environmental and tourist indicators

Indicator	Description
Net greenhouse gas emissions (tonnes per capita)	The indicator measures total national emissions including international aviation of the so-called 'Kyoto basket' of greenhouse gases
ln Years of life lost due to PM 2.5 exposure	The indicator measures the log of years of life lost (YLL) due to exposure to particulate matter (PM 2.5). YLL is defined as the years of potential life lost as a result of premature death
Estimated soil erosion by water (%)	The indicator estimates the area potentially affected by severe erosion by water such as rain splash, sheet wash and rills
ln Number of nights spent	It is the log number of nights spent by country of destination
ln Number of trips	It is the log Number of trips by country of destination
UNESCO Heritage	The indicator represents the UNESCO heritage. It is given by the ratio between the number of World Heritage Sites in a country and the surface of that country

Source: EUROSTAT and UNESCO.

While Net greenhouse gas emissions, ln Years of life lost due to PM 2.5 exposure and Estimated soil erosion by water are environmental indicators, ln Number of nights spent, ln Number of trips and UNESCO Heritage are tourist indicators. Specifically, the indicators are the averages of the period between 2015 and 2019, five years before the start of the implementation of the RRPs. Figure 7 reports the scatter plots of each indicator against the index.

Figure 7. Scatter plots of the index against the environmental and tourism indicators



For each scatter plot the trend line is reported.

Evidence is in line with H2 and H3 if we both look at the scatterplots in Figure 7 and consider that the correlations of the index with the indicators are between 0.24 and 0.32. The positive relationship between the index and each indicator suggests that those countries experiencing a delay towards the environmental objectives and those countries with a more tourism-oriented economy have put higher (and more space to) the environmental dimension on their RRPs.

3.4.4. The association of the index with economic growth

Ideally, we would like to investigate whether the observed annual differences across countries in the reported index can be useful to predict their economic performance. However, it is too early to measure it as we need to wait at least until the end of 2026 to assess if the different performances could be associated in someway with the different prevalence of environmental and socioeconomic projects. What we can do at this moment is to compare the index with the economic performance as predicted by the main professional forecasters. For instance, we may consider the GDP growth forecasts elaborated by the European Commission staff and included in the working document of each RRP for the period between 2020 and 2026

(European Commission, 2021b). It is worth noting that these forecasts were made available after the approval and publication of the definitive version of the RRPs, while, on the other hand, the index is based on information (RRPs) before the time the forecasts were published.

H4: A higher value of the index is associated with a better or worse expected economic performance

We consider four different specifications to estimate the association between the index and the percentage change of the real GDP (Table 6). Firstly, we consider the simple univariate equation (model 1); secondly, we added the time trend as a control (model 2); thirdly, we added a Covid-19 time dummy as a control (the dummy takes 1 if the year is 2020 or 2021 and 0 otherwise; model 3). Finally, we included both the trend and the Covid-19 dummy (model 4). In all cases, we used the method of the pooled OLS estimator[3] with standard errors corrected for heteroskedasticity (Wooldridge, 2010).

Table 6. Dependent variable: RRP forecasts of Δ Real GDP (Pooled OLS estimates)

	1	2	3	4
Index	0.0716***	0.0634***	0.0610***	0.0616***
	(0.0161)	(0.0157)	(0.0145)	(0.0149)
Time trend		0.0041*		-0.0010
		(0.0015)		(0.0013)
Covid-19			-0.0216***	-0.0245***
			(0.0052)	(0.0045)
Observations	145	145	145	145
Adjusted R2	0.1701	0.2044	0.2527	0.2485

Robust standard errors are reported in brackets. The constant is not reported but is included in the estimated equation. *Significant at a 5% level, **significant at a 1% level, and ***significant at a 0.1% level

For all four specifications considered, the equation shows to fit adequately the data as the adjusted R2 ranges between 17 and 25 percent. The index is statistically significant at 0.1 percent of confidence level and shows a parameter stable between 0.06 and 0.07.[4] The positive association between the index and the expected variation in real GDP may reveal a positive effect of the environmental dimension on economic growth, even as compared with the socioeconomic one. The estimated economic effect suggests a small increase in growth when a country privileges the environmental dimension over the socioeconomic one, corresponding to 7 percentage points per year. This preliminary result can be considered in any case very interesting apart from surprising as the common view has always been that the socioeconomic dimension is more growth-enhancing than the environmental dimension.

[3] The Fixed effects and Random effects estimators provide similar results (available upon request from the authors).

[4] Robustness checks are available in the supplemental data online.

While several contributions in the literature show the positive impact of green policies on economic growth (Jouvet and de Perthuis, 2013; Mundaca and Markandya, 2016; Ringel *et al.*, 2016), the prevalence of the environmental policy measures as compared to the socioeconomic ones is new in the literature. If this evidence were confirmed in other empirical works focusing on observed data, it may provide support for the green policies and help mitigate the negative views of some parties and voters.

4. Conclusions

We proposed a new textual algorithm, an adaption of the Bayes classifier for the word classification task. Its main characteristic is that the prior probabilities of topics are not constant but adapt to the posterior probabilities associated with the adjacent words. The simulation has shown that the proposed classifier achieves an improvement of more than 20% over the original classifier.

The classifier has been used to identify how the countries in their respective RRPs sort in terms of importance attributed to the different SDGs. In particular, for each project contained in the RRPs, we used our classifier to estimate the number of words associated with the environmental and socioeconomic dimensions. Then, we built an index of alignment of these projects with one dimension as compared with the other. Three main results emerged by plotting the index against key variables. The environmental dimension plays a more important role in those countries that: 1) receive more funds from the RRF; 2) show a major delay toward the environmental objectives; and 3) are more involved in tourism. Finally, we examined the association between the index and the forecasts of the European Commission on the real GDP growth rates, finding that the percentage change of the real GDP tends to be positively influenced by the environmental dimension as compared to the socioeconomic one.

In contrast to other popular tools usually adopted in economics to extract quantitative information from textual data – the dictionary-based approach – the proposed classifier is more flexible, allowing us to perform a more objective analysis for tasks not necessarily related to sentiment analysis. Once the documents related to the classes to be studied have been collected, the researcher only needs to set the number of adjacent words to classify each word. Moreover, unlike other complex but less used methods, such as those based on neural networks, this classifier is not opaque in its functioning, allowing the researcher to easily understand the results obtained. Since it is neither too simple nor too complex, the proposed classifier has the potential for becoming a very useful tool for researchers of *all* fields in those cases where they need to extrapolate numerical information from documents only containing words.

Future refinements of the classifier may provide a built-in procedure allowing the researcher to automatically detect the optimal number of adjacent words to be considered to maximize the accuracy. Another future development of this classifier is to test a weighting scheme for the surrounding words, for example by giving more weight to the words that are closer to the word to be classified.

References

ALFANO V. and GUARINO M. (2022), A Word to the Wise Analyzing the Impact of Textual Strategies in Determining House Pricing, Journal of Housing Research, vol. 31, issue 1, pp. 88-112. doi: https://doi.org/10.1080/10527001.2021.2013058

ALSHANIK F., APON A., HERZOG A., SAFRO I. and SYBRANDT J. (2020), Accelerating Text Mining Using Domain-Specific Stop Word Lists, 2020 IEEE International Conference on Big Data (Big Data). doi: 10.1109/BigData50022.2020.9378226

APRIGLIANO V., EMILIOZZI S., GUAITOLI G., LUCIANI A., MARCUCCI J., MONTEFORTE L. (2023), The power of text-based indicators in forecasting Italian economic activity, International Journal of Forecasting, vol. 39, issue 2, pp. 791-808. doi: https://doi.org/10.1016/j.ijforecast.2022.02.006

ASH E. and HANSEN S. (2022), Text algorithms in economics, Annual Review of Economics, In press.

ATHEY S. and IMBENS G. W. (2019), Machine Learning Methods That Economists Should Know About, Annual Review of Economics, vol. 11, pp. 685-725. doi: https://doi.org/10.1146/annurev-economics-080217-053433

BORCHARDT S., BUSCAGLIA D., BARBERO VIGNOLA G., MARONI M. and MARELLI L. (2020), A sustainable recovery for the EU: A text mining approach to map the EU Recovery Plan to the Sustainable Development Goals, Publications Office of the European Union, Luxembourg. doi: 10.2760/030575

BRAMER M. (2020), Principles of Data Mining, Springer, Berlin.

BREIMAN L. (2001), Random Forests, Machine Learning, vol. 45, pp. 5-32. doi: https://doi.org/10.1023/A:1010933404324

BREIMAN L., FRIEDMAN J., STONE C. J. and OLSHEN R. A. (1984), Classification and Regression Trees, Chapman and Hall/CRC, New York.

COMMISSION DELEGATED REGULATION (EU) 2021/2106 of 28 September 2021 on supplementing Regulation (EU) 2021/241 of the European Parliament and of the Council establishing the Recovery and Resilience Facility by setting out the common indicators and the detailed elements of the recovery and resilience scoreboard.

DEBOLE F. and SEBASTIANI F. (2005), An analysis of the relative hardness of Reuters-21578 subsets, Journal of the American Society for Information Science and Technology, vol. 56, issue 6, pp. 584-596. doi: https://doi.org/10.1002/asi.20147

DIAS C., HECSER A. and TURCU O. (2022), Recovery and Resilience Plans - public documents, Economic Governance Support Unit.

DIAS C., ZOPPÈ A., GRIGAITĖ K., SEGALL R., ANGERER J., LEHOFER W., GOTTI G., KOMAZEC K. and TURCU O. (2021), Recovery and Resilience Plans – An overview, Economic Governance Support Unit.

ERCAN G. and CICEKLI I. (2012), Keyphrase extraction through query performance prediction, Journal of Information Science, vol. 38, issue 5. doi: https://doi.org/10.1177/0165551512448984

EUROPEAN COMMISSION (2019), Annual Sustainable Growth Strategy 2020, COM(2019) 650 final.
EUROPEAN COMMISSION (2021a), Annexes – Resilience dashboards for the social and economic, green, digital and geopolitical dimensions.
EUROPEAN COMMISSION (2021b), Commission staff working document – Guidance to member states Recovery and Resilience plans, SWD (2021) 12 final.
EUROSTAT (2020), Sustainable development in the European Union - Monitoring report on progress towards the SDGs in an EU context (2020 edition), Publications Office of the European Union, Luxembourg. doi: 10.2785/555257
FOLKE C., BIGGS R., NORSTRÖM A. V., REYERS B. and ROCKSTRÖM J. (2016), Social-ecological resilience and biosphere-based sustainability science, Ecology & Society, vol. 21, no. 3, art. 41. doi: http://dx.doi.org/10.5751/ES-08748-210341
HAE-CHEON K., JIN-HYEONG P., DAE-WON K. and JAESUNG L. (2020), Multilabel naïve Bayes classification considering label dependence, Pattern Recognition Letters, vol. 136, pp. 279-285. doi: https://doi.org/10.1016/j.patrec.2020.06.021
HASTIE T., TIBSHIRANI R. and WAINWRIGHT M. (2015), Statistical Learning with Sparsity: The Lasso and Generalizations, Chapman and Hall/CRC, New York. doi: https://doi.org/10.1201/b18401
HILDUM D. C. (1963), Semantic Analysis of Texts by Computer, Language, vol. 39, no. 4, pp. 649-653. doi: https://doi.org/10.2307/411960
HOERL A. E. and KENNARD R. W. (1970), Ridge regression: Biased estimation for nonorthogonal problems, Technometrics, vol. 12, pp. 55-67.
HORNIK K., STINCHCOMBE M. and WHITE H. (1989), Multilayer feedforward networks are universal approximators, Neural Networks, vol. 2, issue 5, pp. 359-366. doi: https://doi.org/10.1016/0893-6080(89)90020-8
JOUVET P. A. and DE PERTHUIS C. (2013), Green growth: from intention to implementation, International Economics, vol. 134, pp. 29-55. doi: https://doi.org/10.1016/j.inteco.2013.05.003
KALAMARA E., TURRELL A., REDL C., KAPETANIOS G. and KAPADIA S. (2022), Making text count: economic forecasting using newspaper text, Journal of Applied Econometrics, vol. 37, issue 5, pp. 896-919. doi: https://doi.org/10.1002/jae.2907
KARABOYTCHEVA M. (2021), Recovery and Resilience Facility, European Parliamentary Research Service.
KOUNDOURI P., DEVVES S. and PLATANIOTIS A. (2021), Alignment of the European green deal, the sustainable development goals and the European semester process: Method and application, Theoretical Economics Letters, vol. 11, issue 4, pp. 743-770. doi: 10.4236/tel.2021.114049
LOUGHRAN T. and MCDONALD B. (2011), When is a liability not a liability? Textual analysis, dictionaries, and 10-ks, The Journal of Finance, vol. 66, issue 1, pp. 35–65. doi: https://doi.org/10.1111/j.1540-6261.2010.01625.x
LOUGHRAN T. and MCDONALD B. (2016), Textual Analysis in Accounting and Finance: A Survey, Journal of Accounting Research, vol. 54, issue 4, pp. 1187-1230. doi: https://doi.org/10.1111/1475-679X.12123
MANNING C. D. and SCHÜTZE H. (2003), Foundations of Statistical Natural Language Processing, MIT Press, Cambridge.
MITCHELL T. M. (2019), Machine Learning, McGraw-Hill Education, New York.
MUNDACA L. and MARKANDYA A. (2016), Assessing regional progress towards a 'Green Energy Economy', Applied Energy, vol. 179, pp. 1372-1394.
OECD (2018), Economic Outlook No 103 - July 2018 - Long-term baseline projections.
OECD (2021), Economic Outlook No 109 - October 2021 - Long-term baseline projections. doi: https://doi.org/10.1787/cbdb49e6-en
PENG F., SCHUURMANS D. and WANG S. (2004), Augmenting Naive Bayes Classifiers with Statistical Language Models, Information Retrieval, vol. 7, pp. 317-345. doi: https://doi.org/10.1023/B:INRT.0000011209.19643.e2
PICAULT M. and RENAULT T. (2017), Words are not all created equal: A new measure of ECB communication, Journal of International Money and Finance, vol. 79, pp. 136-156. doi: https://doi.org/10.1016/j.jimonfin.2017.09.005

PINHEIRO R. H.W., CAVALCANTI G. D.C., CORREA R. F. and REN T. I. (2012), A global-ranking local feature selection method for text categorization, Expert Systems with Applications, vol. 39, issue 17, pp. 12851-12857. doi: https://doi.org/10.1016/j.eswa.2012.05.008

RENAULT T. (2017), Intraday online investor sentiment and return patterns in the U.S. stock market, Journal of Banking & Finance, vol. 84, pp. 25-40. doi: https://doi.org/10.1016/j.jbankfin.2017.07.002

RINGEL M., SCHLOMANN B., KRAIL M. and ROHDE C. (2016), Towards a green economy in Germany? The role of energy efficiency policies, Applied Energy, vol. 179, pp. 1293-1303. doi: https://doi.org/10.1016/j.apenergy.2016.03.063

ROTONDO F., PERCHINUNNO P., L'ABBATE S. and MONGELLI L. (2022), Ecological transition and sustainable development: integrated statistical indicators to support public policies, Scientific Reports, 12, article number 18513. doi: https://doi.org/10.1038/s41598-022-23085-0

SAHAMI M., DUMAIS S., HECKERMAN D. and HORVITZ E. (1998), A Bayesian approach to filtering junk e-mail, Learning for Text Categorization: Papers from the 1998 workshop 62, pp. 98-105.

SINGH J. and GUPTA V. (2016), Text Stemming: Approaches, Applications, and Challenges, ACM Computing Surveys, vol. 49, issue 3, pp. 1-46. doi: https://doi.org/10.1145/2975608

TETLOCK P. C. (2007), Giving Content to Investor Sentiment: The Role of Media in the Stock Market, The Journal of Finance, vol. 62, issue 3. doi: https://doi.org/10.1111/j.1540-6261.2007.01232.x

THORSRUD L. A. (2020), Words are the New Numbers: A Newsy Coincident Index of the Business Cycle, Journal of Business & Economic Statistics, vol. 38, issue 2, pp. 393-409. doi: https://doi.org/10.1080/07350015.2018.1506344

TIBSHIRANI R. (1996), Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, issue 1, pp. 267-288. doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

UN (2022), The Sustainable Development Goals Report 2022, United Nations Publications, New York.

WANG A. H. (2010), Don't follow me: Spam detection in Twitter, 2010 International Conference on Security and Cryptography (SECRYPT), pp. 1-10.

WHITE H. (1992), Artificial Neural Networks: Approximation and Learning Theory, Blackwell Publishers, Hoboken.

WOOLDRIDGE J. M. (2010), Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge (Massachusetts).

ZHANG W., HU HUA, HU HAIYANG and FANG J. (2019), Semantic distance between vague concepts in a framework of modeling with words, Soft Computing, vol. 23, pp. 3347-3364. doi: https://doi.org/10.1007/s00500-017-2992-x

Supplemental Data Online

Appendix A. Validation exercises

Appendix A.1. Zipf's law

If each word contained in the projects of the RRPs were labeled with its class, the performance of the classifier could be evaluated with the accuracy. However, since these words are not already labeled, an alternative way to evaluate the performance of the classifier is Zipf's law. According to Zipf's law, very few words dominate the word count distribution (Manning and Schütze, 2003). Consequently, these words can potentially have a large impact on the results (Loughran and McDonald, 2016). This means that analysing the most frequent words classified in a specific class is useful to understand if the classifier works correctly.

Figures A.1 and A.2 report the top 75 words of the two dimensions. The list appears in line with expectations as we cannot identify words that are not reasonably used in the dimension in which they are classified. Vice versa, there are some words that are clearly associated with environmental issues, such as environment, water, sustainable, renewable and climate, or socioeconomic terms, such as digital, education, health, development and school.

Figure A.1. Word count distribution for the environmental dimension

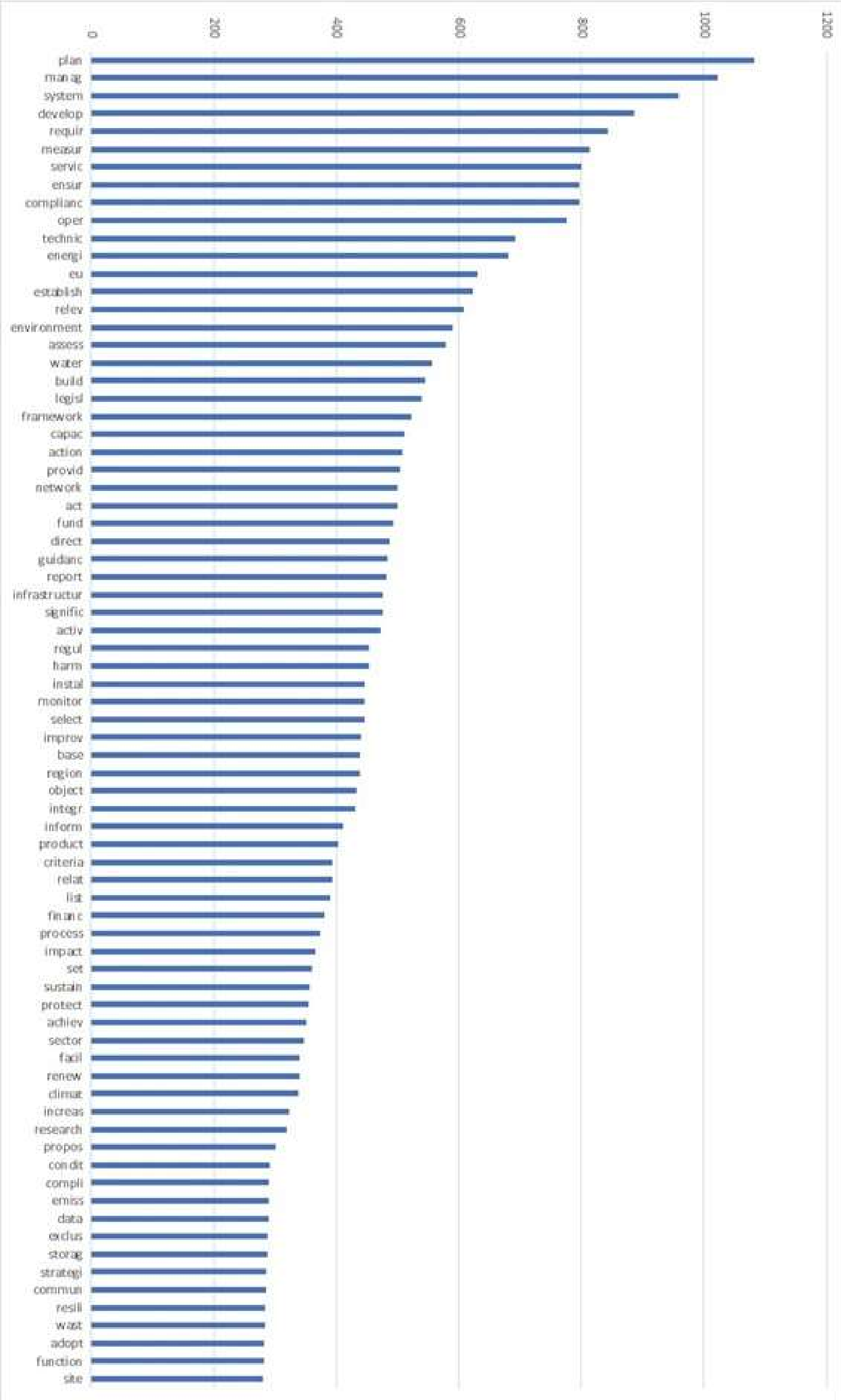
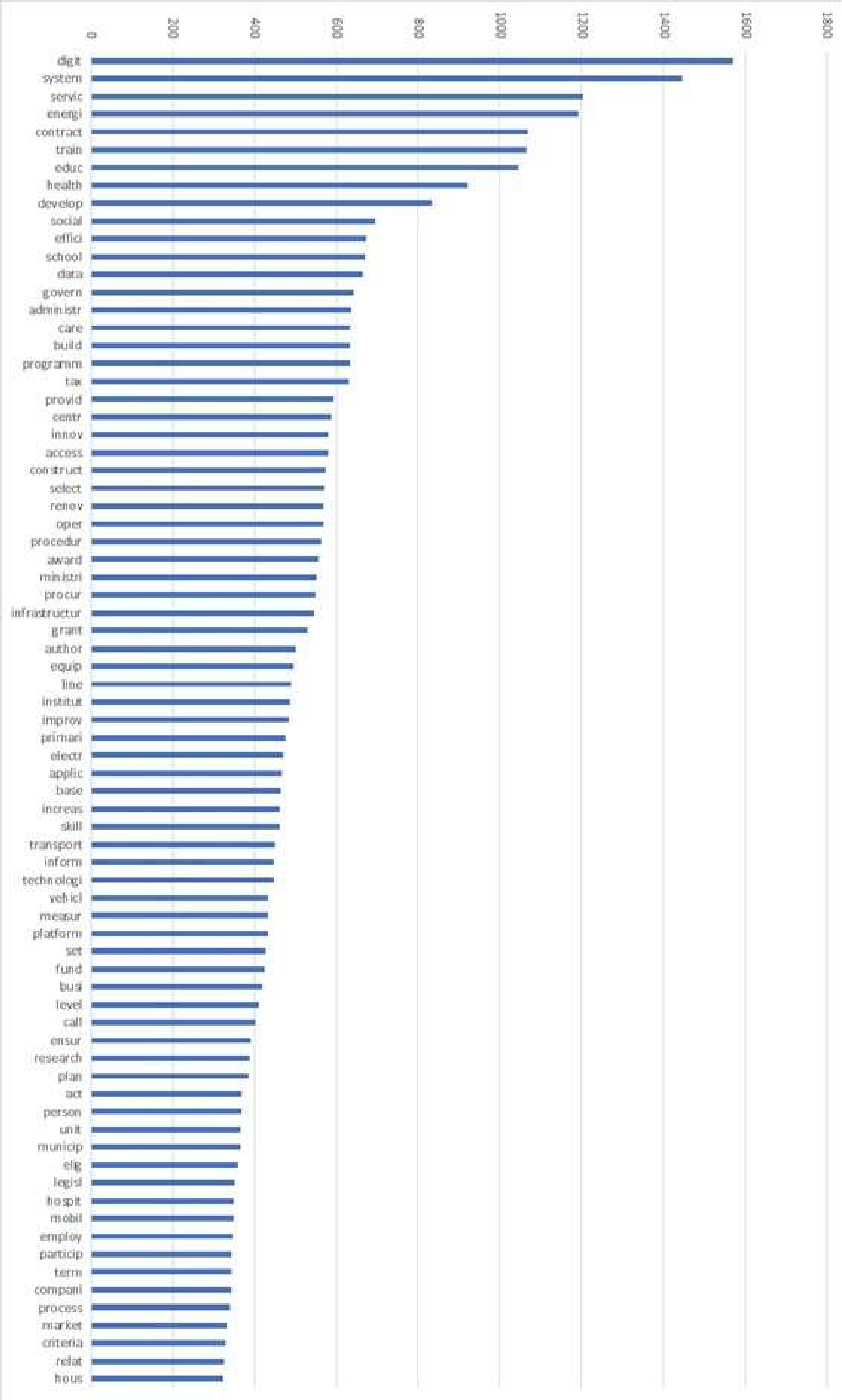


Figure A.2. Word count distribution for the socioeconomic dimension



Appendix A.2. The index and the RRP pillars

As a further validation exercise, we also plotted the index against the six pillars in which the RRPs are organized. We have already remarked that the SDGs and the pillars are different domains. The same is true if we extend the comparison to the dimensions of the Wedding Cake Model and the pillars. For example, the environmental dimension of the Wedding Cake Model does not correspond precisely to the green transition pillar.

By plotting the index against each pillar, our objective is to verify if the sign of these relationships is coherent with our expectations or not. Thus, a country deciding to allocate more funds to the green transition pillar should present a higher index, that is, more attention paid to the environmental dimension. On the other hand, funds for the pillars "digital transformation", "health, social and institutional resilience" and "policies for the next generation" are expected to be negatively correlated with the index, because they are more intensely related to the socioeconomic dimension, such as business support for the development of digital products and services, and capacity of educational and health facilities (Commission Delegated Regulation, 2021).

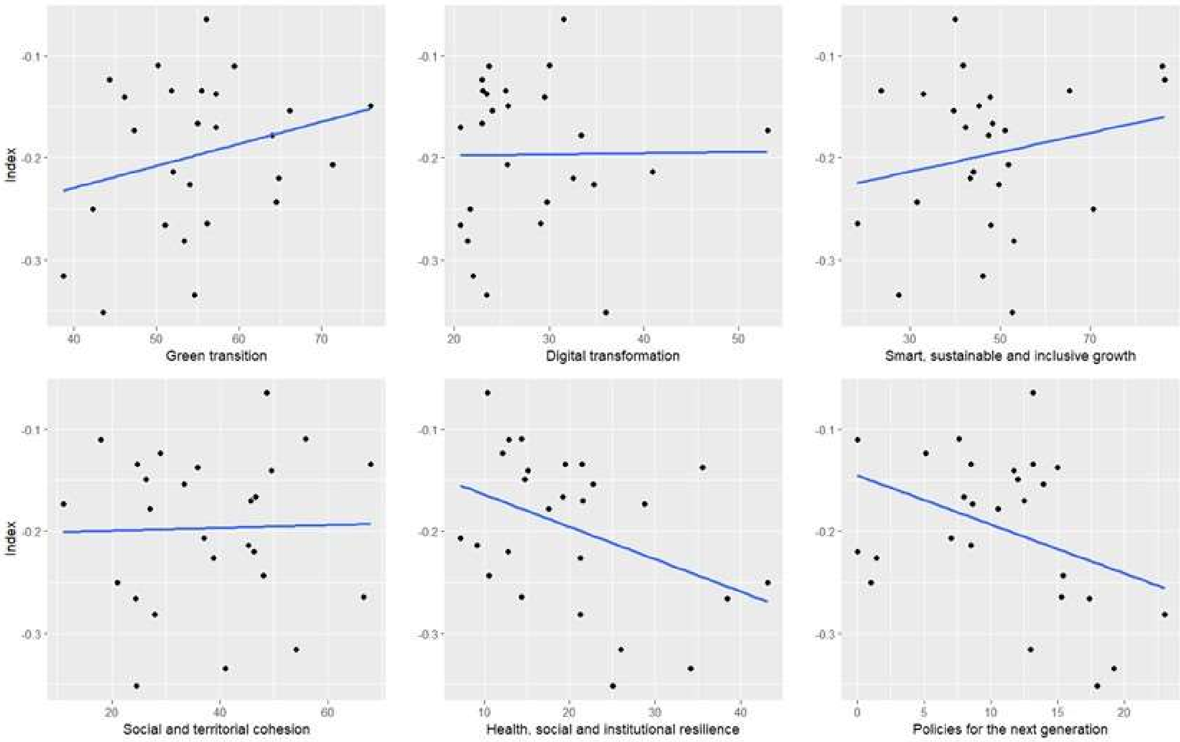
Hypothesis A.1: The countries allocating more funds to the green transition pillar show a higher value of the index (HA.1)

Hypothesis A.2: The countries allocating more funds to the pillars "digital transformation", "health, social and institutional resilience" and "policies for the next generation" show a lower value of the index (HA.2)

The relationships between the index and the pillars "smart, sustainable and inclusive growth" and "social and territorial cohesion" are not as easy to predict. These pillars deal with both environmental objectives (e.g., savings in energy consumption, renewable energy, infrastructure for alternative fuels, and benefits from protective measures against floods, wildfires and other climate-related natural disasters) and socioeconomic issues (e.g., support for firms in their activities and for people in finding a job, and inclusion of people in education or training activities) (Commission Delegated Regulation, 2021). Consequently, which of the two dimensions predominates can only be determined empirically.

Figure A.3 reports the scatter plots of the index against the percentages of funds allocated for the six pillars.

Figure A.3. Scatter plots of the index against the RRP pillars



For each scatter plot the trend line is reported.

The scatter plots show that HA.1 and HA.2 are reasonable if we except pillar "digital transformation", for which no particular correlation can be identified. Specifically, there is a positive relationship between the index and the green transition pillar (correlation 0.26) and a negative relationship between the index and the pillars "health, social and institutional resilience" and "policies for the next generation" (correlation –0.40 and –0.39, respectively). Finally, the correlations between the index and the pillars "smart, sustainable and inclusive growth" and "social and territorial cohesion", for which we were not able to predict a sign, are positive or almost zero (0.20 and 0.03, respectively).

Appendix B. Robustness check

To check the robustness of the estimated models, we also considered the forecasts of the variation of the real GDP made by another institution, the Organization for Economic Co-operation and Development (OECD). In particular, the data are from the long-term baseline projections made in 2021 (OECD, 2021), which take into account the effects of the Covid-19 pandemic and therefore the potential impact of the RRPs (table B.1):

Table B.1. Dependent variable: OECD forecasts (2021) of Δ Real GDP (Pooled OLS estimates)

	1	2	3	4
Index	0.0419**	0.0389*	0.0340*	0.0393**
	(0.0139)	(0.0151)	(0.0130)	(0.0133)
Time trend		0.0011		-0.0042***
		(0.0012)		(0.0009)
Covid-19			-0.0172**	-0.0307***
			(0.0051)	(0.0041)
Observations	145	145	145	145
Adjusted R2	0.1010	0.0937	0.1578	0.1861

Robust standard errors are reported in brackets. The constant is included in the estimated equation. *Significant at a 5% level, **significant at a 1% level and ***significant at a 0.1% level

The results are similar to the ones based on the RRPs forecasts. Indeed, the index is again significant. On average, the environmental dimension leads to an increase between 3 and 4 percentage points in the variation of the real GDP. Moreover, while the time trend is not significant or has a negligible impact on the dependent variable, the Covid-19 dummy is significant in all specifications and denotes that economic growth increases, on average, between 2 and 3 percentage points in the medium-long term.

We have carried out a final robustness check exercise. So far, we have considered the forecasts of economic growth after the start of the Covid-19 pandemic (and therefore after the realization of the RRPs). But what happens if we consider the forecasts before this shock starts? In this scenario, which does not consider the Covid-19 shock (and therefore the effects of the RRPs), it is reasonable to assume that neither the index nor the Covid-19 dummy should explain the dependent variable. To verify this hypothesis, we considered the long-term baseline projections made by the OECD in 2018 (OECD, 2018) (table B.2):

Table B.2. Dependent variable: OECD forecasts (2018) of Δ Real GDP (Pooled OLS estimates)

	1	2	3	4
Index	0.0022	-0.0002	0.0011	-0.0003
	(0.0052)	(0.0041)	(0.0044)	(0.0042)
Time trend		0.0009		0.0011
		(0.0008)		(0.0008)
Covid-19			-0.0023	0.0012
			(0.0026)	(0.0006)
Observations	145	145	145	145
Adjusted R2	0.0014	0.0043	-0.0065	-0.0020

Robust standard errors are reported in brackets. The constant is included in the estimated equation. *Significant at a 5% level, **significant at a 1% level and ***significant at a 0.1% level

The estimated models confirm our hypothesis, that is the index and the Covid-19 dummy do not affect the economic growth forecasts made before the start of the pandemic. This evidence is very useful to draw conclusion on the nature of the relationship between the index and the GDP growth forecasts. Indeed, we would have found a significant coefficient on the index if the positive association between these two variables had been driven by third unobserved factors that were already known at the time of the projections made by the OECD in 2018.