Climate Change Awareness and Policies*

Mattia Cintioli, Emanuele Mezzi, Loriana Porumb, and Julio Prado Muñoz

University of Amsterdam, Amsterdam, Netherlands
https://www.uva.nl/en

Abstract. Every year the United Nations member states deliver statements during the General Debate (GD). These speeches provide invaluable insights about countries intentions and perspectives. In this paper, speeches from 1970 to 2020 from all countries are integrated with datasets about investments in renewable energy and green patents, to analyse how the general awareness about climate change and the problem perception from the USA and China, the two biggest CO2 emitters, varied over time. Moreover, this study responds to the question of whether it is possible to predict real awareness of climate change from these speeches, linking this awareness with the renewable energies consumption per capita.

Keywords: Pollution · Renewable energy · Green patents · China · USA · SVM · NLP

1 Introduction

Every September, the heads of states and other high-level country representatives gather in New York for the United Nations General Assembly (UNGA) and the General Debate marks the beginning of the UNGA. All member states deliver speeches discussing major issues in world politics. The statements gathered over time are an invaluable source of information on governments' intentions, worries and issues in international politics. Common topics that are usually discussed include terrorism, nuclear non-proliferation, development and aid, and climate change [1]. This study focuses on the general trends regarding the emergence, over time, of worries for climate change. Specifically, the study is divided in two parts: one of exploratory analysis and one of prediction. During the first phase it was possible, through the use of text-analysis, to argue that over time the general attention towards the climate change phenomena has increased, with a peak in 2019. Moreover, the focus was posed over China and the United States, to investigate how the two biggest CO2 emitters [2], differentiate in terms of attention towards this phenomena, and how over time they articulated their answers towards it, through the use of renewable energy and the development of green patents. For the predictive phase, speeches held at United Nations General Assembly were analysed utilizing a natural language processing pipeline, to predict in which of four categories a country will fall, in relation to the consumption of renewable energies per capita. The code can be consulted and downloaded at [3].

* Supported by University of Amsterdam.

2 Methodology

To conduct this study, speeches held at United Nations General Assembly were used. Specifically, speeches from all countries have been considered for sessions that span from 1970 to 2020, while for China and USA also the ones from 2021.

2.1 Exploratory Phase

To track the overall attention that was given to the climate change phenomena, a list of words associated to climate change (”climate”, ”global warming”, ”environmental disaster”, etc.), was created. Climate related words were counted for each country by each year. The more words found, the more attention regarding climate was given by that country in that specific year. To augment the possibility of finding words and to limit the list of words, the speeches were converted to lower-case. The first part of the exploratory analysis ended with the plotting of a line graph demonstrating the overall increase of attention towards climate change. In Fig. 1 it is possible to spot a peak in 2019 and a drastic drop the following year. The next step was to analyse the general debate speeches, comparing how countries differed in regards to the attention spent on climate change. For this comparison the year 2019 was chosen since it was the one with highest frequency of the climate related words. Fig. 2 illustrates a world map where the countries are coloured with different intensities of green according to the frequency of climate words found in their speeches in 2019. The three countries with the greatest attention towards climate change were France, Fiji, and Australia. The second part of the exploratory analysis required the collection and use of further data withdrawn from the Organisation for Economic Co-operation and Development (OECD). The data includes the percentage of renewable energy used by China and the United States from 1990 to 2019 and the percentage of patents related to green energies developed by the same countries. The aim is to show how the differences in attention to climate change are correlated to the differences in policies and investments.

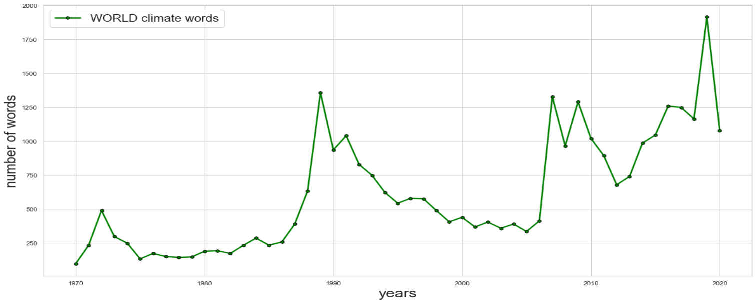


Fig. 1: Overall attention between 1970 and 2020 towards climate change

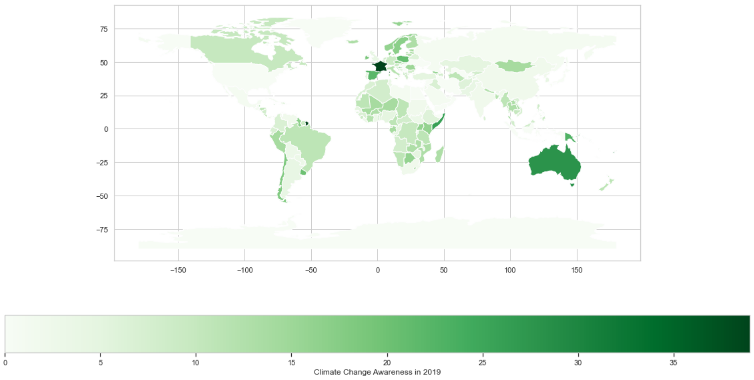


Fig. 2: Attention by each country in 2019

2.2 Inference Phase

For the inference phase of the study, the consumption of renewable energies per capita (in MWh) for every country and year was accessed and downloaded [4]. Nevertheless, this dataset did not contain all the statistics for every country and year for which speech data is available, resulting in a final working dataset comprising 51% of the initial data. In total, 4292 speeches spread over the time framework of this investigation (1970 - 2020) were used. In order to assign labels to our samples and transform the problem at hand into a multiclass classification task, we calculated the quantiles 0.25, 0.50 and 0.75 regarding the values of the renewable energy consumption per capita over the years. This technique segments the data into four equally distributed groups, giving the following thresholds (See Table 1) associated with the obtained quantiles. These groups were mapped to classes ”low”, ”medium”, ”high” and ”very_high” respectively.

0.25	**0.50**	**0.75**
0.17223	0.80609	2.85199

Table 1: Quantiles of renewable energies consumption per capita over the years

This part of the study required deeper pre-processing. The text pre-processing was conducted on every speech and aimed to present the data in a more legible and effective way to the subsequent machine learning pipeline. This was achieved by:

- Lowering the text.
- Removing stop words, new lines, punctuation and small words (less than three letters).

- Fixing possible encoding issues.
- Removing any resulting word that is not either a noun, a verb, a currency or a proper noun using spaCy tokens [5].

Scikit-learn module [6] was primarily used throughout the development and implementation of this experiment. After the text was cleaned, the goal was to find the best parameters to train the model. The method conducted to obtain these best parameters was grid search and more specifically its implementation in sklearn: GridSearchCV. This was executed on a sklearn Pipeline consisting of a CountVectorizer, TFidfTransformer and SVC model in combination with a k-fold cross-validation splitting strategy of five folds. SVC is based on libsvm [7] and is chosen for this experiment given the popularity and success of Support Vector Machines approaches on NLP text classification tasks in the past years as exposed in [8, 9]. The parametric space passed to the GridSearchCV module is described in Table 2. Due to time constraints, this parametric space is constructed from the default parameters of each element in the pipeline and extended with reasonable variations. Addressing potential performance issues, the maximum number of features to handle was strictly set to 10,000 n-grams using *max_features* option in CountVectorizer. From among the values in the parametric space, worth noting are the chosen kernels.

ngram range	**kernel**	**gamma**	**C**
(1,2), (1,3)	**rbf**, sigmoid	**1**, 1e-1, 1e-2, 1e-3, 1e-4, scale	1e-1, 1, **10**

Table 2: Grid search parametric space

rbf and *sigmoid* kernels were used to change the SVM modeling process into non-linear. These usually offer better performance than linear kernels for text classification tasks [10]. It is also important to note that the boundaries of the range of n-values for different word n-grams or char n-grams to be extracted were set to (1, 2) or (1,3). This decision was motivated by the idea of packaging together meaningful combination of words such as "United States" or "Global Warming". (1,1) was, therefore, discarded. Regarding *C* parameter, the default L2 regularization was applied along with one more restrictive and one more flexible value. The searching of the best parameters was applied on data from 2000 to 2018 also aiming to reduce the execution time as described above. The best parameters found for each element in the pipeline are also shown in Table 2 in bold type. The best parameters obtained were subsequently used to train the sklearn pipeline using the available data from 1970 to 2018. This model was k-fold cross-validated using five folds.

Following best practices [11] on splitting data into train and test set when handling time-dependant data and also considering that the main objective of the experiment is to predict renewable energies consumption per capita from new speeches, only speeches from 2019 and 2020 were left for testing. The researchers are aware of the possible discrepancy of this train/test proportion 0.96/0.04 compared to the canonical and recommended 0.80/0.20 train/test ratio. However, giving the fact that in the past ten years there has been an explosive increase in global warming concern and also that the objective is to effectively predict new speeches data, we opted for using only 2019 and 2020 data to test the model. The classification report on the test data (See Table 3) gives a really promising idea of how the classifier performs on new speeches as well as the corresponding confusion matrix (See Fig. 5).

3 Results and Discussion

3.1 Exploratory Phase

From the exploratory analysis it emerged that overall the attention towards the climate change phenomena increased, but not steadily. From (Fig. 1) two major peaks can be observed. The first, in 1989, when climate change began to become a major concern due to the first important alarms made by scientists. The second, in 2019, is far more recent and is followed by a drop one year later when the utmost global concern shifted from climate to the COVID-19 pandemic. China and the United States, which represent the biggest CO2 emitters, showed big difference in terms of attention towards the phenomena and also in terms of policies adopted towards the challenge. During its speeches, over time China dedicated substantially more words to the problem compared to the United States (Fig. 3), with a peak in 2007. Of great importance are also the differences in the employment of renewable energy and investments in green energies (Fig. 4). China demonstrated each year greater capacity in satisfying its energetic exigencies through renewable energy, and over time has recovered the industrial disadvantage towards the United States in relation to green patents.

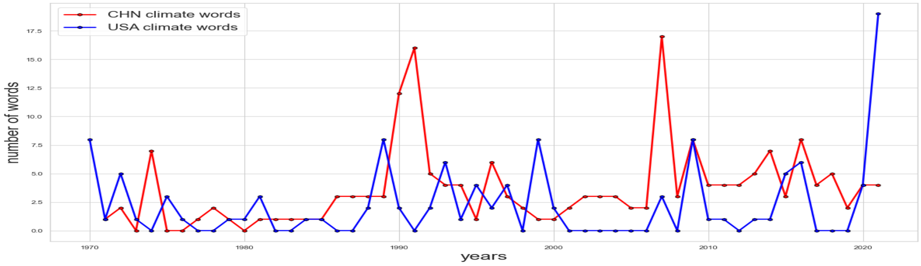
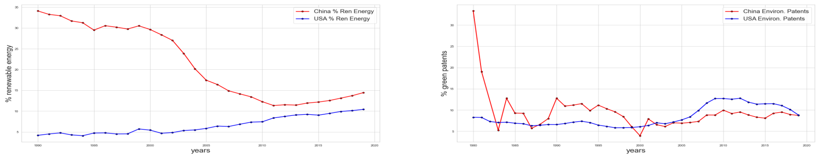


Fig. 3: Words spent by China and the USA about climate change



(a) Renewable energy from the USA and China

(b) Green patents from the USA and China

Fig. 4: Confrontation between the USA and China

3.2 Inference Phase

Following the same ”climate change” theme, we looked into how well we can predict a country per capita renewable energy usage, given the content of its UNGA speech. Our pipeline consisting of a CountVectorizer, TFidfTransformer, and a SVM classifier achieved a macro average precision score of 0.73 and a recall of 0.80. The highest precision achieved by the classifier, 0.95, was in predicting the *very high* class. However, data points belonging to this class are often misclassified as belonging to the *high* class, which leads to a lower recall score of 0.79. Generally, we can observe in the confusion matrix (Fig. 5) the tendency of the model to misclassify the ”higher” class as the ”lower” one. This could be due to the underlying relation between the classes that we are predicting. The classes capture degrees of renewable energy usage and can be ordered according to these, which makes the boundaries between consecutive classes less clear-cut. Another issue that possibly impacts our model performance is data drift, specially due to the time component, as the same country usually have several entries, belonging to different years, in the dataset. This could be problematic since the relation between countries' speeches and their environmental practices likely evolved over time. It is worth noting that the support of the classes in the evaluation dataset is unbalanced whilst in the training dataset every class represents 25% of the data. This fact clearly demonstrates an upward trend in the usage of renewable energies in the past years. Although we have trained the model on the ”earlier” part of the dataset, and tested it on the ”latest” entries, the cross-validation was year-agnostic. To provide a more robust evaluation of our model, a time-based cross-validation might be necessary.

4 Conclusion

From the exploratory phase we observed that the global attention towards climate change increased throughout the years, with highest peak in 2019 and a consequent drop in 2020. Moreover, China and the United States showed differences in climate change awareness with China being the more conscious. China also showed greater attention towards the use of renewable energy, and over time has recovered the industrial disadvantage towards the United States for green patents. We've seen that by relying only on nouns, proper nouns (like ”China”), verbs, and symbols present in a country's UNGA speech, our SVM classifier was able to predict that country's per capita renewable energy consumption rather precisely. This possibly points to a consistency between countries' speech content and their environmental practices. An interesting next step could be investigating which speech words were the most important features that the classifier relied on. This would challenge our assumptions that the most important features were climate-related words, as well as help us better understand the pattern behind the misclassifications.

5 Appendix

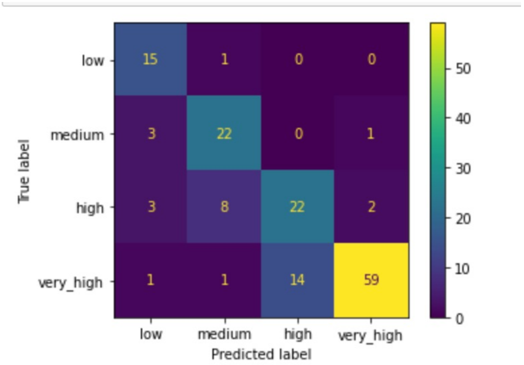


Fig. 5: Test data confusion matrix

	precision	**recall**	**f1-score**	**support**
low	0.61	0.63	0.62	35
medium	0.68	0.94	0.79	16
high	0.69	0.85	0.76	26
very_high	0.95	0.79	0.86	75
accuracy			0.78	152
macro avg	0.73	0.80	0.76	152
weighted avg	0.80	0.78	0.78	152

Table 3: Test data classification report

References

[1] Alexander Baturo, Niheer Dasandi, and Slava J. Mikhaylov. "Understanding state preferences with text as data: Introducing the UN General Debate corpus". In: (2017). URL: `https://doi.org/10.1177/2053168017712821`.

[2] PBL Netherlands Environmental Assessment Agency. *China now no. 1 in CO2 emissions; USA in second position.* 2006.

[3] M. Cintioli et al. *Climate Change Trends 2022.* URL: `https://github.com/juliopradom/climate-change-trends`.

[4] Our World In Data. *Per capita energy consumption from renewables, 2021.* URL: `https://ourworldindata.org/grapher/per-capita-renewables`. (accessed: 22.09.2022).

[5] ExplosionAI GmbH. *Spacy Documentation.* URL: `https://spacy.io/api/token`. (accessed: 22.09.2022).

[6] scikit-learn developers. *Scikit-learn Documentation.* URL: `https://scikit-learn.org/`. (accessed: 22.09.2022).

[7] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM – A Library for Support Vector Machines.* URL: `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`. (accessed: 22.09.2022).

[8] Bi-Min Hsu. "Comparison of Supervised Classification Models on Textual Data". In: (2020). URL: `https://www.mdpi.com/2227-7390/8/5/851/htm`.

[9] M. Trivedi et al. "Comparison of Text Classification Algorithms". In: (2015). URL: `https://www.ijert.org/research/comparison-of-text-classification-algorithms-IJERTV4IS020351.pdf`.

[10] A. Alves. "Support Vector Machines and Kernel Functions for Text Processing". In: (2013). URL: `https://www.researchgate.net/publication/332584408_Support_Vector_Machines_and_Kernel_Functions_for_Text_Processing`.

[11] Google. *Splitting Your Data.* URL: `https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/split`. (accessed: 22.09.2022).