Identification of climate change and sustainability texts using pre-trained Spanish language models

Gerardo Huerta Robles
Universidad Nacional De
Ingenieria, Nicaragua
gerardohuerta1705@gmail.com

Gabriela Zuñiga-Rojas
Universidad Nacional De
San Antonio Abad del Cusco
gabriela.zuniga@unsaac.edu.pe

Abstract

This study presents a method to identify texts related to climate change and sustainability in Spanish using pre-trained language models. The objective is to develop a tool that identify among different texts those that address topics such as climate change and its effects. A dataset labeled with Spanish texts was created, the model was trained with this data and evaluated using a validation set, achieving an accuracy of 0.916. F1 Score of 0.95 and Recall of 0.99. The results show the effectiveness of the proposed approach to identify relevant texts on climate change in Spanish. This model can be a useful tool to create and enrich repositories on the subject, thus contributing to a greater awareness and understanding of climate change and its effects.

1 Introduction

The amount of information that is generated in the field of climate change comes from many sources, which is why some non-profit organizations such as: Drawdown [1] try to create sites where they concentrate various solutions to combat climate change; although they have valuable and accurate information, these pages exclude practical examples, new research, and new ideas. [2]. or implementations of the solutions in real-life situations as success stories. In addition, most of these resources are in the English language, thus increasing the barrier of entry for interested persons who communicate mainly (or only) in Spanish.
Traditional NLP methods have been used and have demonstrated great capacity in solving this problem, according to Schober et al. (2018); Deep learning techniques have also been employed to enhance their efficiency. However, in recent years, large pre-trained language models (LLM) have revolutionized the field of natural language processing (NLP) and one of the most prominent language models is called BERT; from that model, De la Rosa et al. (2022) developed BERTIN which is a series of models based on BERT for Spanish.

To date, despite the increasing use of NLP for climate change-related research, a model with adaptive pre-training to the climate domain in Spanish has not yet been created and made publicly available. We present a language model that is specifically trained on corpora of Spanish texts related to climate change.

2 Objective

The work aims to create a tool capable of identifying texts such as: abstracts of papers, news headlines or articles in general; that talk about climate change, its solutions, its effects and recent studies, in such a way that allows us to have a basis for the development of more complex resources in Spanish as datasets or to enrich a repository of information on the subject.

3 Previous Work

Previous work such as the ClimateBERT project by Webersinke et al. (2022), developed a base model from English language documents related to climate change and performed fine-tuning for specific tasks on climate-related topics. For their work they created a model for identifying climate-related paragraphs. In this fine-tuning they used a smaller dataset of corporate disclosures obtaining a 32.64% better performance in terms of cross-entropy and a 35.71% reduction in F1 error rate compared to models made from the DistilRoBERTA model.

This work sets a clear precedent for using NLP for identifying climate-related information in corporate documents. However, some of the limitations of the model are mainly its training with data

[1] https://drawdown.org/

[2] By searching Scopus with "climate change" and "Sustainability" https://www.scopus.com/

entirely in English (including its base model), the dataset is based on corporate information which can generate biases and finally the dataset used is not balanced in its labels which can influence the performance of the model.

4 Method

We have used a Spanish base model from the Bertin project De la Rosa et al. (2022), this model uses a Spanish corpus for training, it also uses the token mask technique which makes it suitable for tasks such as text classification. The hyperparameters for the fine-tuning can be found in the table 1:

Hyperparameter	Set
learning_rate	2e-05
train_batch_size	16
eval_batch_size	16
seed	42
lr_scheduler_type	linear
epoch	2
optimizer	Adam

Table 1: Hyperparameters configuration

4.1 Creation of the Dataset [3]

The dataset formation process is based on three steps:

- Data selection: A dataset was developed from open data sources:
  - Spanish translation of the dataset climatebert/climate_detection Bingler et al. (2023).
  - Spanish News on topics not related to climate change extracted from the repository: Memoon (2024)
  - Translation of Opinions related to climate change extracted from the dataset: Flores (2016)

  Data sources with long texts (>140 characters) and short texts have been taken, because the original model did not contemplate short texts for training and testing.

- Preprocessing: In the creation of the dataset it is necessary to obtain texts related and not related to climate change with a corresponding binary label, which allows to identify with 1 if the text is related to climate change and 0 otherwise.
  In order to use the texts, a previous process of translation of the base dataset has been carried out using automatic translation tools, followed by a validation of the translations..
  For the news dataset in Spanish, the column containing news has been identified and the topics Macroeconomics, Innovation, Regulations, Alliances, Reputation have been labeled with (0).
  As for the opinions dataset, a data cleaning has been performed by removing hashtags, usernames, emojis and URLs to use only the textual content of the posts.

- Dataset formation and balancing: With the use of a script we took all the extracted data, balancing the amount of data for each label to avoid any bias, obtaining as a result the amount described in Table 2 for training and test data.

Train data			**Test Data**		
Data	Label	%	Data	Label	%
1600	1	55	480	1	62
1300	0	45	300	0	38

Table 2: Amount of data used for training and test

4.1.1 Dataset Structure:

- Question : Text
- Answer: binary label, if the text is related to climate change or sustainability (1) if the text is unrelated (0)

Training data (2900 data), of which; for **label 1**: 1000 are translated data from the original dataset, 600 are posts from X. And for **label 0**: 300 are from the original dataset and 500 from the news dataset and 500 from X posts.

Test data (780 data), of which; for **label 1**: 320 are translated data from the original dataset, 160 are posts from X. And for **label 0**: 80 are from the original dataset and 120 from the news dataset and 100 from X posts.

4.2 Validation Dataset

To test the performance of the model, a validation dataset was created from data not seen in the previ-

[3] https://huggingface.co/datasets/somosnlp/spa_climate_detection

ous datasets. These data were taken from; for label 1: the Drawdown project, abstracts of scientific articles taken from IEEE Xplore; for label 0: news in Spanish, and Wikipedia articles in Spanish. A total of 200 texts were taken from all the sources of which the final ratio was 50 - 50 of positive and negative examples.

5 Results

The results obtained from the training can be seen in the table 3. In this step we used train and test dataset.

Metrics	Results
Loss	0.1592
Accuracy	0.9705

Table 3: Results obtained in the training process

Model validation was then performed, yielding the results shown in table 4

Metrics	Results
Accuracy	0.95
Precision	0.916
Recall	0.99
F1 score	0.951

Table 4: Results obtained in the validation process

Where:
Recall: The proportion of relevant items that were retrieved correctly over the total relevant items that exist, focusing on the model's ability to capture all relevant items.
Precision: The proportion of retrieved items that are relevant over the total retrieved items, indicating the model's ability to retrieve only relevant items and avoid false positives.
F1 Score: A metric that combines precision and recall for evaluating a model's ability to classify correctly.
Accuracy: The proportion of correct predictions over the total predictions made by the model, indicating its overall ability to predict correctly.

6 Discussion

From the results obtained we can distinguish a decrease in performance using the validation dataset. We attribute this behavior to the fact that we did not used more generalized data such as those obtained from the Spanish Wikipedia in the training dataset, this has caused an increase in the inference of false positives. On the other hand, we have used abstracts of scientific articles related to climate change which have been detected without any false negative despite not being included as a source of the training dataset.

The following risks and limitations can be deduced.:

- Our model inherits the biases and limitations of the base model with which it was trained; however, they are not so obvious to find because of the type of task in which the model is being implemented, such as text classification.
- The use of high-level language in the dataset can complicate the identification of texts with low-level languages (e.g. colloquial).
- Although it performs well with short texts, the model can lower the performance, so it is more accurate with long texts.

7 Conclusions

We have obtained a model[4] in Spanish that can be useful as a tool for identifying texts related to climate change and thus create and enrich repositories that address the subject. Taking as a starting point the results obtained through this work, the following steps are considered:

- Generalization of the model by taking data from different sources and more data.
- Create a dataset with climate change and sustainability information based on sectors (electricity, agriculture, industry, transportation, etc.).
- Create an advanced model that allows to subclassify texts related to climate change based on sectors (token classification).
- Train a Q/A model that can provide relevant information on the topic of climate change.

Acknowledgments

This project was developed during the Hackathon Somos600M 2024 organized by SomosNLP[5]. We thank all event organizers and sponsors for their

[4] https://huggingface.co/somosnlp/bertin_base_climate_detection_spa
[5] https://somosnlp.org/

support during the event. We also want to acknowledge to Edison Jair Bejarano Sepulveda who was our translator and gave us valuable feedback.

References

Julia Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2023. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. Working paper, Available at SSRN 3998435.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Sofía Chávez Flores. 2016. Sentiment of climate change. https://data.world/crowdflower/sentiment-of-climate-change. Accessed: 2024-03-20.

Mohammad Memoon. 2024. Los angeles twitter news dataset. https://www.kaggle.com/datasets/muhammadmemoon/los-angeles-twitter-news-dataset. Accessed: 2024-03-29.

Andreas Schober, Christopher Kittel, Rupert J. Baumgartner, and Manfred Füllsack. 2018. Identifying dominant topics appearing in the journal of cleaner production. *Journal of Cleaner Production*, 190:160–168.

Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. ClimateBERT: A Pretrained Language Model for Climate-Related Text. *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.

Identificación de textos relacionados al cambio climático y sustentabilidad utilizando modelos de lenguaje preentrenados en español

Gerardo Huerta Robles
Universidad Nacional De
Ingenieria, Nicaragua
gerardohuerta1705@gmail.com

Gabriela Zuñiga-Rojas
Universidad Nacional De
San Antonio Abad del Cusco
gabriela.zuniga@unsaac.edu.pe

Abstract

Este estudio presenta un método para identificar textos relacionados con el cambio climático y la sustentabilidad en español mediante modelos de lenguaje preentrenados. El objetivo es desarrollar una herramienta que identifique, entre distintos textos, aquellos que aborden temas como el cambio climático y sus efectos. Se creó un dataset etiquetado con textos en español, se entrenó el modelo con estos datos y se evaluó utilizando un conjunto de validación, logrando una precisión de 0.916. F1 Score de 0.95 y Recall de 0.99. Los resultados muestran la efectividad del enfoque propuesto para identificar textos relevantes sobre el cambio climático en español. Este modelo puede ser una herramienta útil para crear y enriquecer repositorios sobre el tema, contribuyendo a una mayor conciencia y comprensión sobre el cambio climático y sus implicaciones.

1 Introducción

La cantidad de información que se genera en el campo del cambio climático proviene de muchas fuentes, por tal razón algunas organizaciones sin fines de lucro como: Drawdown [1] tratan de crear sitios en donde muestran soluciones para combatir el cambio climático; a pesar que tienen información valiosa y precisa, estas páginas excluyen ejemplos prácticos, nuevas investigaciones [2]. o implementaciones de las soluciones en situaciones de la vida real como casos de éxito. Además, la mayoría de estos recursos están en el idioma inglés, aumentando la barrera de entrada para las personas interesadas. Se han usado métodos tradicionales de NLP que han demostrado gran capacidad, Schober et al. (2018) ; asi también se usaron técnicas de Deep learning que mejoran su eficacia. Sin embargo los últimos años, los grandes modelos de lenguaje previamente entrenados (LLM) han revolucionado el campo del procesamiento del lenguaje natural (NLP) y uno de los modelos de lenguaje más destacados se llama BERT. A partir de ese modelo, De la Rosa et al. (2022) desarrollaron BERTIN, una serie de modelos basados en BERT para español.

Hasta la fecha, a pesar del creciente uso de NLP para investigaciones relacionadas con el cambio climático, todavía no se ha creado y puesto a disposición del público un modelo con preentrenamiento adaptativo al dominio climático en español. Presentamos un modelo de lenguaje que está específicamente entrenado en corpus de textos en español relacionados con el cambio climático.

2 Objetivo

El trabajo tiene como objetivo crear una herramienta capaz de identificar los textos como: resúmenes de articulos científicos, titulares de noticias o artículos en general; que hablen sobre el cambio climático, sus soluciones, sus efectos y estudios recientes, de tal forma que nos permita tener una base para el desarrollo de recursos más complejos en español como datasets o enriquecer un repositorio de información sobre el tema.

3 Antecedentes

Trabajos previos como el proyecto ClimateBERT de Webersinke et al. (2022), quienes desarrollaron un modelo base a partir de documentos en ingles relacionados al cambio climatico y realizaron fine-tuning para tareas específicas en temas relacionados al clima. Para su trabajo crearon un modelo de identificación de párrafos relacionados al clima. En este fine-tunning utilizaron un dataset más pequeño de divulgaciones corporativas obteniendo un rendimiento de 32.64% mejor en términos de cross-entropy y una reducción de 35.71% en la tasa de error F1 en comparación a modelos realizados a partir del modelo DistilRoBERTA.

[1] https://drawdown.org/

[2] Haciendo una consulta en Scopus con "cambio climatico " y "Sustentabilidad"https://www.scopus.com/

Este trabajo deja un precedente claro del uso de PLN para la identificación de información relacionada al clima en documentos corporativos. Sin embargo, algunas limitaciones del modelos son principalmente a su entrenamiento con datos totalmente en inglés (incluyendo su modelo base), el conjunto de datos está basado en información corporativa lo cual puede generar sesgos y finalmente los datos no estan balanceados en sus etiquetas lo que puede influir en el rendimiento del modelo.

4 Método

Hemos utilizado un modelo base en español del proyecto Bertin De la Rosa et al. (2022), este modelo utiliza un corpus en español para su entrenamiento, además utiliza la técnica "token mask" lo que lo hace adecuado para tareas como clasificación de texto. Los hiperparámetros para el entrenamiento los podemos encontrar en la tabla 1:

Hiperparámetro	**Set**
learning_rate	2e-05
train_batch_size	16
eval_batch_size	16
seed	42
lr_scheduler_type	linear
epoch	2
optimizer	Adam

Table 1: Hiperparámetros utilizados

4.1 Formación del dataset [3]

El proceso de formación del dataset, se basa en tres pasos:

- Eleccion de los datos:Se desarrolló un dataset a partir de fuentes (open-source):
  - Traducción al español del dataset climatebert/climate_detection Bingler et al. (2023).
  - Noticias en español de temas no relacionados al cambio climatico extraido del repositorio: Memoon (2024)
  - Traduccion de Opiniones relacionadas al cambio climatico extraido del dataset: Flores (2016)

  Se han tomado fuentes de datos con textos largos (>140 caracteres) y cortos, debido a que el modelo original no contemplaba textos cortos para su entrenamiento y pruebas.

- Preprocesamiento: Buscamos obtener textos relacionados y no relacionados al camblio climático con una etiqueta binaria correspondiente, que permita identificar con 1 si el texto es relacionado al cambio climático y 0 en caso contrario. Para poder usar los textos se ha realizado un proceso previo, de traducción del dataset base. Para el dataset de noticias en español se ha discriminado la columna con noticias y los temas Macroeconomia, Innovación, Regulaciones, Alianzas, Reputación han sido etiquetados con (0). En cuanto al dataset de opiniones se ha realizado una limpieza de datos quitando hashtags, nombres de usuarios, emojis y URLs.

- Formación y balanceo de dataset: Con el uso de un script se tomaron todos los datos extraidos, balanceando la cantidad de datos para cada etiqueta y evitar cualquier sesgo, obteniendo como resultado la cantidad descrita en la tabla 2.

Train data			**Test data**		
Data	Label	%	Data	Label	%
1600	1	55	480	1	62
1300	0	45	300	0	38

Table 2: Datos utilizados para entrenamiento y pruebas.

4.1.1 Estructura del dataset:

- Question : Texto
- Answer: etiqueta binaria, si el texto es relacionado a cambio climatico o sustentabilidad (1) si el texto no es relacionado (0)

Datos de Entrenamiento (2900 datos), de los cuales; para la **etiqueta 1**: 1000 son datos traducidos del dataset original, 600 son post de X. Y para la **etiqueta 0**: 300 son del dataset original y 500 del dataset de noticias y 500 publicaciones de X.

Datos de Pruebas (780 datos), de los cuales: para la **etiqueta 1**: 320 son datos traducidos del dataset original, 160 son post de X. Y para la **etiqueta 0**: 80 son del dataset original y 120 del dataset de noticias y 100 de posts de X.

[3] https://huggingface.co/datasets/somosnlp/spa_climate_detection

4.2 Dataset de validación

Para comprobar el funcionamiento del modelo se creó un dataset de validación a partir de datos no vistos en los dataset anteriores. Los datos fueron tomados de; etiquetas 1: proyecto Drawdown, resúmenes de articulos cientificos del IEEE Xplore; etiquetas 0: noticias en español y artículos de la Wikipedia en español. En total 200 ejemplos fueron extraídos de los cuales la proporción final era 50 - 50 de ejemplos positivos y negativos.

5 Resultados

Los resultados obtenidos del entrenamiento se pueden observar en la tabla 3.

Metrica	Resultado
Loss	0.1592
Accuracy	0.9705

Table 3: Resultados obtenidos en el entrenamiento

Posterior se realizó la validación del modelos obteniendo los resultados mostrados en la tabla 4

Métrica	Resultado
Accuracy	0.95
Precision	0.916
Recall	0.99
F1 score	0.951

Table 4: Resultados obtenidos en la evaluación

Donde
Recall: Proporción de elementos relevantes que se recuperaron correctamente sobre el total existente de estos, centrándose en la capacidad del modelo para capturar todos los elementos relevantes.
Precision: Proporción de elementos recuperados que son relevantes sobre el total de elementos recuperados, lo que indica la capacidad del modelo para recuperar sólo elementos relevantes y evitar falsos positivos.
F1 Score: Métrica que combina "Precision" y "Recall" para evaluar la capacidad de un modelo para clasificar correctamente.
Accuracy: Proporción de predicciones correctas sobre el total de predicciones realizadas por el modelo, lo que indica su capacidad general para predecir correctamente.

6 Discusión

De los resultados podemos diferenciar una disminución en el rendimiento utilizando el dataset de validación. Atribuimos este comportamiento al hecho que no se ha utilizado datos mas generalizados como lo son los obtenidos de la Wikipedia en español en el dataset de entrenamiento, esto ha causado un aumento en la inferencia de falsos positivos. Por otro lado, hemos utilizado artículos científicos relacionados al cambio climático los cuales los ha detectado si ningún falso negativo a pesar de no estar incluidos como fuente de los dataset de entrenamiento. Se deducen los siguientes riesgos y limitaciones:

- Hereda los sesgos y limitaciones del modelo base con el que fue entrenado; sin embargo, no son tan evidentes de encontrar por el tipo de tarea en el que se esta implementando el modelo como lo es la clasificacion de texto.
- El uso de lenguaje de alto nivel en el dataset pueden complicar la identificación de textos con lenguajes de bajo nivel (ejemplo: coloquial).
- A pesar de tener buenos resultados con textos cortos, el modelo puede bajar su rendimiento, por lo que es más preciso en textos largos.

7 Conclusiones

Se ha obtenido un modelo[4] en español que nos puede ser útil como herramienta para identificación de textos relacionados al cambio climatico y de esta forma crear y enriquecer repositorios que aborden del tema. Tomando como punto de partida los resultados obtenidos por medio de este trabajo se consideran los siguientes pasos a seguir:

- Generalizar aún más el modelo tomando datos de fuentes diversas y mayor cantidad de datos.
- Crear un dataset con información de cambio climatico y sustentabilidad basado en sectores.
- Crear un modelo avanzado que permita subclasificar los textos relacionados a cambio climatico en base a sectores (token classification), por ejemplo: Clasificar en base a electricidad, agricultura, industria, transporte, etc.
- Entrenar un modelo Q/A que pueda brindar información relevante en la tematica de cambio climatico.

[4] https://huggingface.co/somosnlp/bertin_base_climate_detection_spa

Agradecimientos

Este proyecto fue desarrollado durante el Hackathon Somos600M 2024 organizado por SomosNLP[5]. Agradecemos a todos los patrocinadores y organizadores del evento por todo el apoyo. De la misma forma, queremos agradecer a Edison Jair Bejarano Sepulveda quien fue voluntario y nos apoyó en la traducción de la versión en inglés de este documento.

References

Julia Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2023. How cheap talk in climate disclosures relates to climate initiatives, corporate emissions, and reputation risk. Working paper, Available at SSRN 3998435.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Sofía Chávez Flores. 2016. Sentiment of climate change. https://data.world/crowdflower/sentiment-of-climate-change. Accessed: 2024-03-20.

Mohammad Memoon. 2024. Los angeles twitter news dataset. https://www.kaggle.com/datasets/muhammadmemoon/los-angeles-twitter-news-dataset. Accessed: 2024-03-29.

Andreas Schober, Christopher Kittel, Rupert J. Baumgartner, and Manfred Füllsack. 2018. Identifying dominant topics appearing in the journal of cleaner production. *Journal of Cleaner Production*, 190:160–168.

Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. ClimateBERT: A Pretrained Language Model for Climate-Related Text. *Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges*.

[5] https://somosnlp.org/