Environmental Impact Measurement in the MentalRiskES Evaluation Campaign

Alba María Mármol-Romero[1], **Adrián Moreno-Muñoz**[1], **Flor Miriam Plaza-Del-Arco**[2], **M. Dolores Molina-González**[1], **Arturo Montejo-Ráez**[1]
[1]Universidad de Jaén, [2]Bocconi University
[1]Campus Las Lagunillas, 23071, Jaén, Spain
[2]Via Sarfatti 25, 20100, Milan, Italy
[1]{amarmol, ammunoz, mdmolina, amontejo}@ujaen.es
[2]flor.plaza@unibocconi.it

Abstract

With the rise of Large Language Models (LLMs), the NLP community is increasingly aware of the environmental consequences of model development due to the energy consumed for training and running these models. This study investigates the energy consumption and environmental impact of systems participating in the MentalRiskES shared task, at the Iberian Language Evaluation Forum (IberLEF) in the year 2023, which focuses on early risk identification of mental disorders in Spanish comments. Participants were asked to submit, for each prediction, a set of efficiency metrics, being carbon dioxide emissions among them. We conduct an empirical analysis of the data submitted considering model architecture, task complexity, and dataset characteristics, covering a spectrum from traditional Machine Learning (ML) models to advanced LLMs. Our findings contribute to understanding the ecological footprint of NLP systems and advocate for prioritizing environmental impact assessment in shared tasks to foster sustainability across diverse model types and approaches, being evaluation campaigns an adequate framework for this kind of analysis.

Keywords: mental disorder detection, NLP systems, energy consumption, environmental impact

1. Introduction

With the advent of Large Language Models (LLMs), the Natural Language Processing (NLP) community is increasingly recognizing the importance of addressing and mitigating the environmental impact of these models. The lifecycle of an NLP model, including data ingestion, pre-training, fine-tuning, and inference, significantly contributes to energy consumption and emissions.

This concern amplifies when developing shared tasks, i.e., competitions where different teams are encouraged to develop different systems to address a specific NLP task. For instance, MentalRiskES (Mármol-Romero et al., 2023) is a recent task on early risk identification of mental disorders in Spanish comments from Telegram users. The organizers of this shared task encourage teams to submit their energy and environmental impact consumption alongside their prediction systems. This shared task consists of an online problem where participants detect a potential risk (eating disorders (EDs), depression, and anxiety) as early as possible in a continuous stream of data. A total of 16 teams participated in submitting more than 130 runs.

In this work, we perform an empirical study to quantify the energy consumption and environmental impact of the systems participating in the MentalRiskES shared task. While this may seem like it should be a straightforward calculation, several variables can influence compute time and energy consumption, ranging from (1) the type of model architecture used for addressing the tasks; (2) the type of task and the type of computation required to carry it out; and (3) intrinsic characteristics of the dataset, such as average sequence length, number of users, etc.

In this paper, we are among the first to study the environmental impact of the different systems developed for a shared task. We focus on the MentalRiskES shared task, as it stands out as one of the few reporting the energy consumption of participants. The systems submitted for this task range from traditional ML models to state-of-the-art LLMs. Our study aims to comprehensively evaluate the ecological footprint across all model types involved in the competition.

Furthermore, we advocate for shared task organizers to prioritize and promote the crucial practice of environmental impact measurement. This proactive approach fosters sustainability in the NLP community and encourages environmentally conscious methodologies across diverse model types.

2. The Environmental Cost of NLP Systems

Digitization has sometimes been seen as a green solution, mainly because of the reduction of physical resources, like paper. But any software sys-

Proceedings of the Second TDLE Workshop @LREC-COLING-2024, pages 61–72

tem is undoubtedly linked to hardware, the physical counterpart, and, even more, to the amount of energy needed to power these systems. Computing already demands 1% of the total energy generated in the world according to a recent report (IEA, 2023), which also found that current Artificial Intelligence (AI) advancements have come with the side effect of a high increase in power consumption and, therefore, an impact on greenhouse gases emissions. This is significant, especially considering that these systems are primarily operated in the cloud, meaning they often run in data centres specifically designed for energy efficiency (Dodge et al., 2022).

When dealing with LLMs, the related impact on CO2 emissions can be significant. It has been estimated that the training of a large model like the BLOOM model (Le Scao et al., 2022) emitted about 24.7 tonnes of CO2 considering only power consumption, and more than 50 tonnes if all processes involved are considered (from equipment manufacturing to energy-based operational consumption) (Luccioni et al., 2023). That is equivalent to 300,000 km drive of a diesel car. BLOOM has 176 billion parameters, so we can imagine the equivalent emissions to train GPT-4, which is estimated to be around 1.76 trillion (1,000 diesel cars over their whole lifetime).

The concept *Sustainable AI* has emerged to discuss, in the words of Van Wynsberghe (2021), "how to develop AI that is compatible with sustaining environmental resources for current and future generations". As such, it is more a matter of being sure that AI advances are sustainable, rather than finding sustainable means to maintain AI technologies.

Therefore, AI systems must be limited in their carbon footprint and every research activity where deep learning is involved should report on this issue. Fortunately, several libraries have emerged to help in the measurement of the environmental impact of the execution of deep learning models, like ML CO2 Impact Tools (Lacoste et al., 2019) or the more recent Eco2AI tool (Budennyy et al., 2022). But one that has been found to be very effective is the CodeCarbon[1] tool, as it considers where executions take place so energy sources can be better estimated. This tool has been designed according to the work by (Kirkpatrick, 2023).

3. Objectives

In this paper, we address three main different objectives related to environmental impact:

1. Estimate how different ML approaches (mainly shallow learning vs. deep learning ones) impact the overall demand for computing resources and power consumption when dealing with early risk prediction over the internet.

2. Evaluate the amount of greenhouse gases associated with an evaluation campaign for a better understanding of the environmental cost of this kind of scientific and research forums in the scope of artificial intelligence applied to mental health.

3. Promote a responsible design of algorithms and techniques to mitigate or reduce the energy and emissions associated, identifying the most promising solutions with a balanced trade-off between performance and efficiency.

4. Data acquisition process

MentalRiskES (Mármol-Romero et al., 2023) is a task on early risk identification of mental disorders in Spanish comments from Telegram users. Given a history of messages about a user, the goal is to identify whether the user suffers from the disorder or not, and his/her attitude to it. The task must be resolved as an online problem, that is, messages per subject are provided in a sequence of rounds and the systems must submit a prediction for each round. Therefore, the performance not only depends on the accuracy of the systems but also on how fast the problem is detected. For this shared task edition, the disorders considered are eating disorders (task 1), depression (task 2), and an unknown one (task 3) which later revealed itself as anxiety. In this paper, we focus on tasks 1 and 2 and subtasks 1a, 1b, 2a, 2b.

For task 1, eating disorder detection, teams had to detect if the user suffered from anorexia or bulimia (task 1a - binary classification) and provide a probability for the user to suffer anorexia or bulimia (task 1b - simple regression). For task 2, depression detection, teams had to detect if the user suffered from depression (task 2a - binary classification) and provide a probability for the user to suffer depression (task 2b - simple regression).

In addition, as early detection is the main goal of this evaluation campaign, teams were provided with access to a server to which they had to connect to read messages and send predictions simulating a system that aims to predict mental problems in social networks and in real-time. Therefore, predictions are sent per round, each round being the access to a new message from the subjects' history. Therefore, the later the round, the more user messages the teams will have available, with the first round being the first message of each subject's history.

[1] https://codecarbon.io/

4.1. How CodeCarbon Works

To conduct the CO2 tracking analysis, the CodeCarbon[2] package in its 2.1.4 version is used. CodeCarbon calculates the carbon intensity of the consumed electricity as a weighted average of the emissions from the different energy sources. Each way of generating electricity (fossil fuels coal, petroleum, natural gas, and renewable or low-carbon) is associated with specific carbon intensities. Based on the mix of energy sources in the local grid, CodeCarbon calculates the carbon intensity of the electricity consumed.

4.2. Sending Dynamics

In the MentalRiskES competition, for each task, participants were asked to submit some information to measure the impact of their systems in terms of resources needed and environmental issues, with the aim of recognizing those systems that can perform the task with minimal resource demand.

In particular, participants submitted the following metrics as part of the metadata in every prediction (for each round):

- Duration: Duration of the compute, in seconds.
- Emissions: System emissions as CO2 equivalents [CO2eq], in kg.
- Energy used per CPU: Power consumption per CPU in kWh.
- Energy used per GPU: Power consumption per GPU in kWh.
- Energy used per RAM: Power consumption per RAM in kWh.
- Total energy used: Total power consumption in kWh. The sum of CPU, GPU, and RAM energy used in kWh.

The participants submitted together with their system predictions the accumulated of each metric, that is, each submission of the different rounds has the previous one added, so the difference gives the measurement for the interval. In this way, the metrics are known in each submission and we can calculate the mean and standard deviation for each metric. The participants also submitted information about their hardware:

- CPU count: number of CPU.
- GPU count: number of GPU.
- CPU model: example Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz.
- GPU model: example 1 x NVIDIA GeForce GTX 1080 Ti.
- RAM total size: total RAM available.

[2] https://mlco2.github.io/codecarbon/

5. Results and Discussion

This section presents the results obtained in some subtasks of the MentalRiskES (Mármol-Romero et al., 2023) competition (binary classification and simple regression) for tasks 1 and 2 (eating disorders and depression detection), as well as the corresponding environmental values. This section also includes the comparison and analysis of these two aspects of the systems grouped according to the type of algorithm used: (1) Classical ML for systems using algorithms such as Support Vector Machine (SVM) or Random Forest (RF), Deep Learning (DL) for systems using algorithms such as LSTMs or CNNs, and LLM for systems using large language models such as BERT or GPT.

Throughout this section, performance metrics such as Macro-F1 and Early Risk Detection Error (Losada and Crestani, 2016) at round 30 (ERDE30) established in the competition will be discussed. In addition, the energy values refer to the sum of CPU, GPU, and RAM energy in kWh (Energy) and the average emissions as CO2 equivalents, in kg per round (Emissions).

These values are obtained from those provided by the competition organizers (Mármol-Romero et al., 2023) and from all papers published by the sixteen teams. Note that teams had the possibility to submit predictions from three different systems. In some cases, teams submitted predictions from the same system as three different systems, which led to their consolidation within the same line in some of the following tables. This conclusion was reached after seeing that all three alleged systems provided the same predictions and gave the same emission values.

5.1. Binary Classification

The results obtained with the environmental data for subtask 1a (ED) and subtask 2a (depression) are compared below. For this type of task, most systems used LLM to resolve the problem although not always obtain the best scores.

For subtask 1a, about eating disorders (ED), 10 teams participated and there are 20 systems. There were only two systems that used classical ML and four that used DL. The Macro-F1 and ERDE30 results obtained by the teams' systems are shown in Figure 1. Despite the popularity of the use of LLM systems, fourteen systems in total, the best score was obtained by the team CIMAT-NLP-GTO (Echeverría-Barú et al., 2023), with the

system that used a classical Naïve Bayes algorithm with a value of 0.966 for the Macro-F1 metric, 0.048 points ahead of the second-best system, UMUTeam (Pan et al., 2023), that used a LLM, MarIA model (Gutiérrez-Fandiño et al., 2021). Also, the best system in the F1 metric obtained the best score in the ERDE30 metric with the lowest value of 0.018.

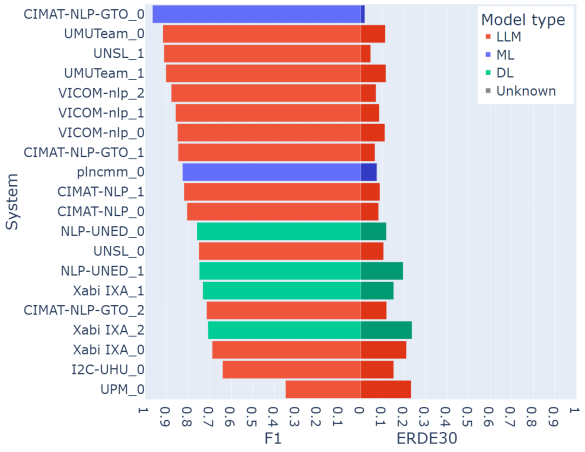
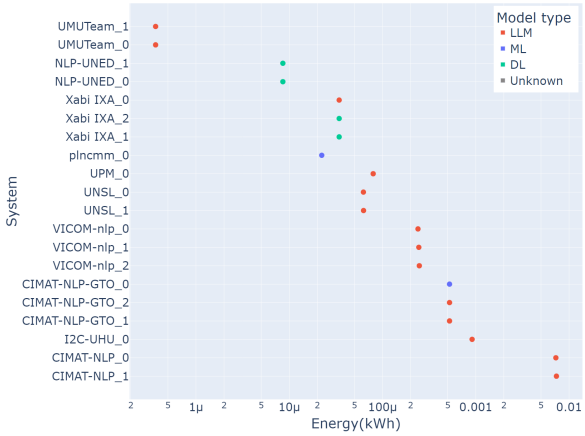


Figure 1: Macro-F1 and ERDE30 scores obtained by the systems sorted by the F1 metric. The y-axis represents the team's name followed by a number representing the system used.

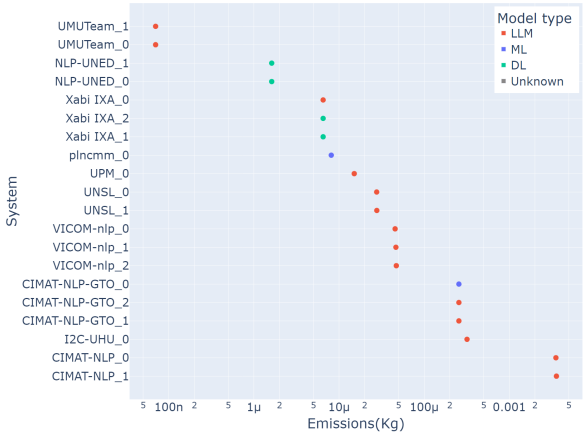
Subfigures 2a and 2b show the emissions and energy values of the systems across the different teams. Although, in general, the systems do not consume excessive energy in calculating predictions in a single round, LLMs occupy lower positions in the graph, which translates into higher values of energy consumption. This is very clear to see in Figure 3 (which shows the energy consumption per model type). The last four systems used the RoBERTuito model (Pérez et al., 2021) followed by a Naive Bayes algorithm which obtained the best Macro-F1 and ERDE30 scores. On the other hand, the second-best system in the F1 score is in the second place in the emissions ranking which shows that it is possible to have a good prediction and be friendly with the environment. The average value obtained by the teams' systems in task 1a for several metrics the organisers asked for related to environmental impact are shown in Table 1, Appendix A.1.

In Figure 4 systems are visualised according to their ranking, energy consumed and emissions produced. In this case, the systems that consume the most energy are also the ones that produce the most emissions. Moreover, some systems consume very little and with a very small sphere (low emissions) that obtains a very high F1 value.

For subtask 2a, about depression, 14 teams participated and there are 26 systems. In this subtask six systems used classical ML, two used DL and



(a) Mean energy consumed per round sorted by emissions.



(b) Mean emissions emitted per round sorted by emissions.

Figure 2: Average values obtained per round by the systems for subtask 1a on environmental friendliness. The y-axis represents the team's name followed by a number representing the system used.

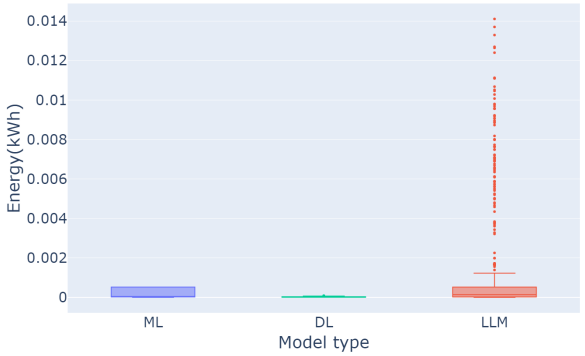


Figure 3: Boxplot of energy consumption (KWh) of each prediction in task 1a by model type.

sixteen LLM systems. In this case, the first five systems have a similar score in the Macro-F1 metric although the first, obtained by UMUTeam, has a very high ERDE30 score (0.358) compared to the best (0.140) in the fifth position, SINAI-SELA

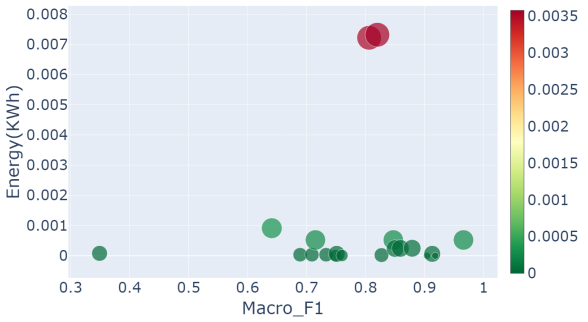


Figure 4: Distribution of the systems for task 1a. The size of the marks is by the emissions produced. The emissions were scaled and a logarithmic normalisation in base 2 was performed for better visualisation. The colour scale corresponds to the actual values of CO2 emissions in kilograms.

(González-Silot et al., 2023), that used a BERT-based model (Devlin et al., 2018). These values are shown in Figure 5.

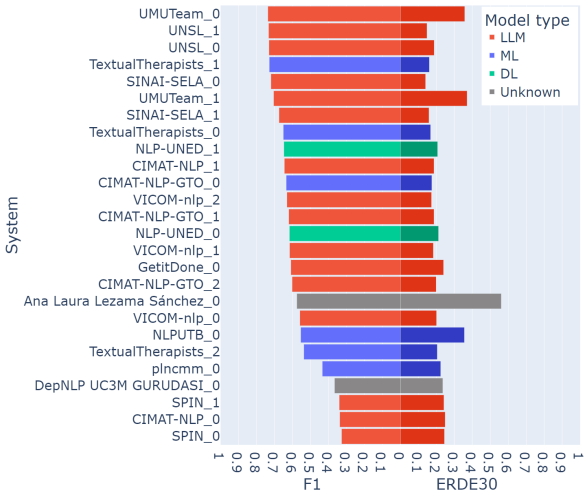
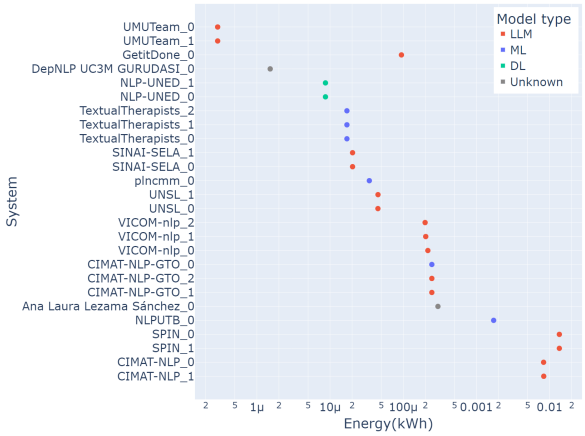


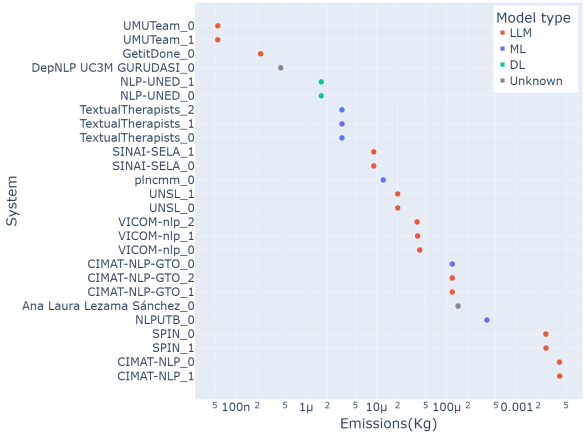
Figure 5: Macro-F1 and ERDE30 scores obtained by the systems sorted by the F1 metric. The y-axis represents the team's name followed by a number representing the system used.

The values obtained in this subtask show that there is no relationship between emissions and energy used in the prediction because models based on BERT like Bertin (De la Rosa et al., 2022), trained with Spanish language data, consumed a lot of energy but were not among the systems that emitted more CO2. This is represented in Figure 7. RoBERTuito-based systems again occupy the lowest position in the charts shown in Subfigures 6a and 6b. The average value obtained by the teams' systems in task 2a for several metrics the organisers asked for related to environmental impact are shown in Table 2, Appendix A.1.

In Figure 8 systems are visualised according to



(a) Mean energy consumed per round sorted by emissions.



(b) Mean emissions emitted per round sorted by emissions.

Figure 6: Average values obtained per round by the systems for subtask 2a on environmental friendliness. The y-axis represents the team's name followed by a number representing the system used.

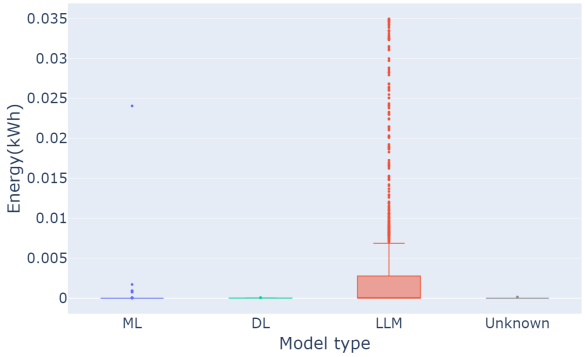


Figure 7: Boxplot of energy consumption (KWh) of each prediction in task 2a by model type.

their ranking, energy consumed and emissions produced. This image clearly shows how energy consumption does not necessarily have to be directly related to CO2 emissions produced, as the source of this energy can be renewable.

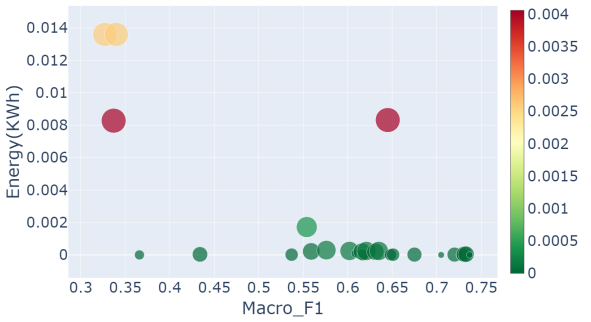


Figure 8: Distribution of the systems for task 2a. The size of the marks is by the emissions produced. The emissions were scaled and a logarithmic normalisation in base 2 was performed for better visualisation. The colour scale corresponds to the actual values of CO2 emissions in kilograms.

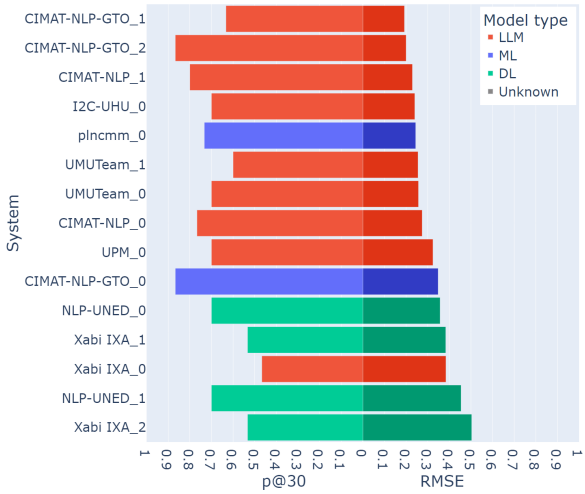


Figure 9: Precision at 30 and RMSE scores obtained by the systems sorted by RMSE metric. The y-axis represents the team's name followed by a number representing the system used.

In general terms, the application of LLM has been the most predominant in this type of task (binary classification). In general, the energy consumption needed to make the predictions can be considered low and the CO2 emissions emitted per round to make the prediction have not been very high either, with a few exceptions. It is shown that an environmentally friendly system can achieve good results in the experiments and that LLMs, in general, consume more energy as shown in Figures 3 and 7.

5.2. Simple Regression

The results obtained with the environmental data for subtask 1b (ED) and subtask 2b (depression) are compared below. For this task, most systems used, again, LLM to resolve the problem.

For subtask 1b, about ED, there was precision at 30 (P@30) and Root Mean Square Error (RMSE) results obtained by teams are shown in Figure 9. Eight teams participated in this subtask and there are fifteen different systems. Two systems used classical ML systems (the best uses Gradient Boost Regressor (GBR) and the second Naive Bayes), four apply DL techniques and the rest, nine, use LLM systems. For this regression task, LLM-based systems seem to perform better as they are at the top of the ranking except for the system based on Sentence-BERT (SBERT), used by team Xabi IXA (Larrayoz et al., 2023).

Energy consumption and emissions, shown in Subfigures 10a and 10b respectively, for this task, are similar to those of the binary classification task. As in the previous figures, DL-based systems are always at the top of the graph, showing their low environmental impact. This is very noticeable in the graph in Figure 11. The average value obtained by the teams' systems in task 1b for several metrics the organisers asked for related to environmental impact are shown in Table 3, Appendix A.1.

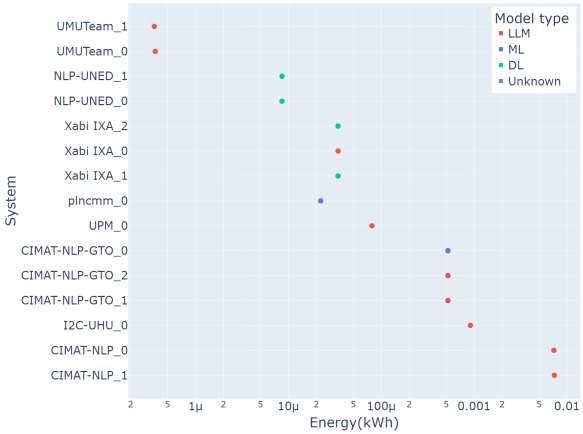
In Figure 12 systems are visualised according to their ranking, energy consumed and emissions produced.

For subtask 2b, about depression, there was P@30 and RMSE results obtained by teams are shown in Figure 13. For this subtask, 7 teams participated and there are 12 different systems. Three systems use classical ML, and the best of them, obtained by the PLN-CMM team (Guerra et al., 2023), applies a Linear Regression. Moreover, two systems of the NLP-UNED team (Fabregat et al., 2023) used DL systems (ANN) and six applied LLM. There is a system that we do not know the type of algorithm they apply located between the two DL-based systems. The ML system that obtains the best results is in the top 3 systems that emit the most emissions, as shown in Subfigures 14a and 14b. Again, Figure 15 shows that LLM uses much more energy than other types of systems.

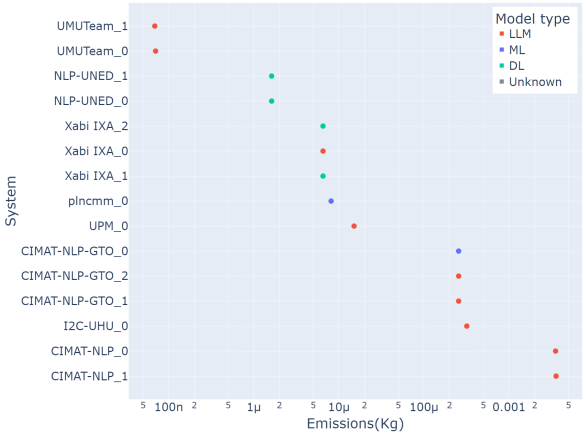
The average value obtained by the teams' systems in task 2b for several metrics the organisers asked for related to environmental impact are shown in Table 4, Appendix A.1.

In Figure 16 systems are visualised according to their ranking, energy consumed and emissions produced. This figure shows that there is a system that consumes more energy than the rest and emits more kilograms of CO2 per prediction and that also obtains the worst result according to the RMSE metric.

LLM seems to have triumphed for this type of task (simple regression), in addition to being the most energy-consuming and emission-intensive for forecasting, although there are also environmen-



(a) Mean energy consumed per round sorted by emissions.



(b) Mean emissions emitted per round sorted by emissions.

Figure 10: Average values obtained per round by the systems for subtask 1b on environmental friendliness. The y-axis represents the team's name followed by a number representing the system used.

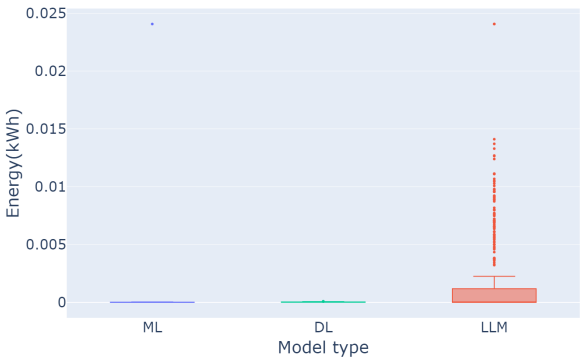


Figure 11: Boxplot of energy consumption (KWh) of each prediction in task 1b by model type.

tally friendly systems that apply LLM. It is possible to use these models without emitting large amounts of CO2 as shown in Figures 10b, 14b, 12 and 16.

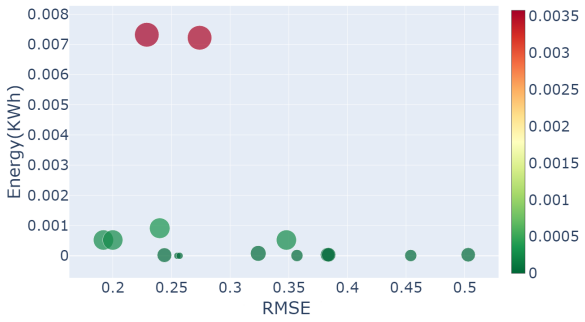


Figure 12: Distribution of the systems for task 1b. The size of the marks is by the emissions produced. The emissions were scaled and a logarithmic normalisation in base 2 was performed for better visualisation. The colour scale corresponds to the actual values of CO2 emissions in kilograms.

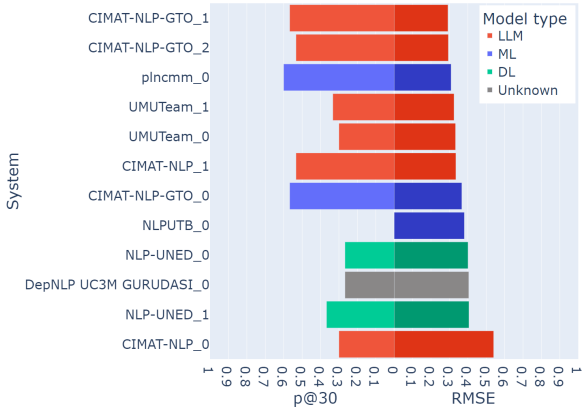
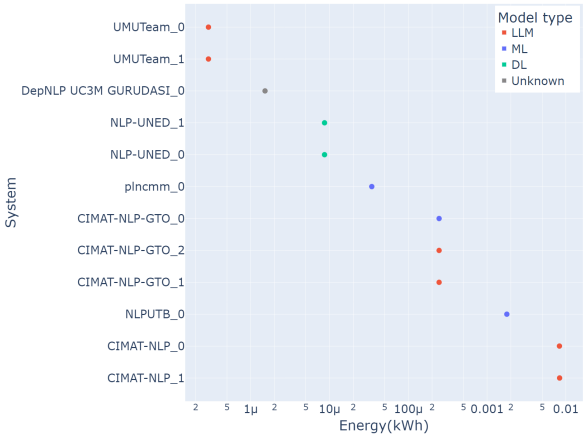


Figure 13: Precision at 30 and RMSE scores obtained by the systems sorted by RMSE metric. The y-axis represents the team's name followed by a number representing the system used.

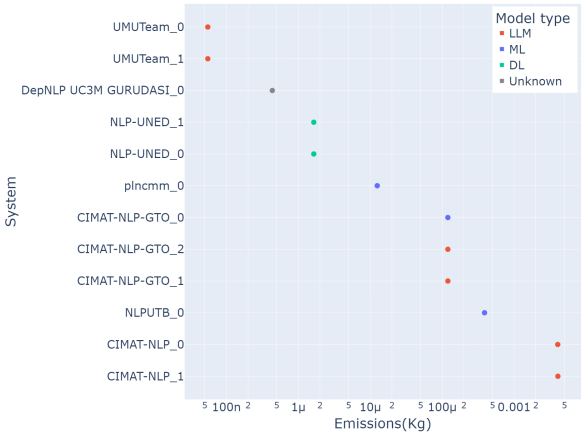
6. Discussion

From the previous analysis, it is clear that LLMs are among the solutions with a more demanding need of power consumption and, therefore, associated CO2 emissions. The RoBERTuito model was found to be the one with a major impact in terms of emissions, but this fact has to be considered carefully for two main reasons: (1) we cannot guarantee the confidence of the reported values by participants, and (2) the validity of the measurements computed by Code Carbon may be biased according to location. In any case, despite this potential weakness in our methodology, if we trust the data, some interesting facts arise:

- Performance is not always linked to complexity. Some classical machine learning systems with very low carbon footprint exhibited superior results.



(a) Mean energy consumed per round sorted by emissions.



(b) Mean emissions emitted per round sorted by emissions.

Figure 14: Average values obtained per round by the systems for subtask 2b on environmental friendliness. The y-axis represents the team's name followed by a number representing the system used.

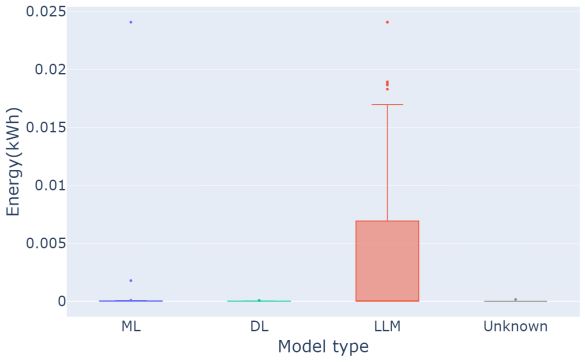


Figure 15: Boxplot of energy consumption (KWh) of each prediction in task 2b by model type.

- Similar systems may lead to very different energy consumption values or emissions. The source of the energy and the efficiency of the computing infrastructure may play a crucial role here.

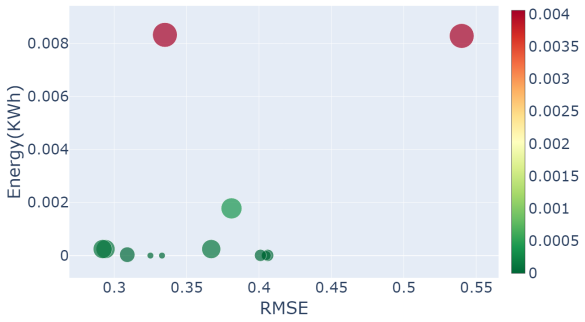


Figure 16: Distribution of the systems for task 2b. The size of the marks is by the emissions produced. The emissions were scaled and a logarithmic normalisation in base 2 was performed for better visualisation. The colour scale corresponds to the actual values of CO2 emissions in kilograms.

7. Conclusions

This work is, to the best of our knowledge, the first attempt to introduce environmental impact analysis in an evaluation campaign. In this paper, we focus on performing this analysis in the MentalRiskES shared task (Mármol-Romero et al., 2023), a competition about detecting early mental risk disorders in Spanish. Participants reported several efficiency metrics when submitting their results. We use these metrics to conduct our analysis. Based on our results, we found that systems based on DL models, as expected, count for the major impact in terms of carbon dioxide emissions. This is even more dramatic when LLMs are involved. Nonetheless, the source of the energy consumed or the efficiency of the computing infrastructure can mitigate this negative impact Besides, in many cases there exist alternatives based on less demanding approaches that can produce high performances in the task of early prediction of mental disorders.

Given the importance of assessing the environmental impact of NLP systems, we strongly advocate for shared task organizers to prioritize the essential practice of environmental impact measurement. This proactive stance not only promotes sustainability within the NLP community but also encourages the adoption of environmentally conscious methodologies across a wide range of model types.

8. Acknowledgements

This work has been partially supported by projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government

A. Appendix

A.1. Emissions values

This section contains the official competition environmental impact data tables for the tasks addressed in this document.

B. Bibliographical References

Semen Andreevich Budennyy, Vladimir Dmitrievich Lazarev, Nikita Nikolaevich Zakharenko, Aleksei N. Korovin, O.A. Plosskaya, et al. 2022. Eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI. In *Doklady Mathematics*, volume 106, pages S118–S128. Springer.

Javier De la Rosa, Eduardo G Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Maria Grandury. 2022. BERTIN: Efficient Pre-training of a Spanish Language Model using Perplexity Sampling. *arXiv preprint arXiv:2207.06814*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the Carbon Intensity of AI in Cloud Instances. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1877–1894, New York, NY, USA. Association for Computing Machinery.

Franklin Echeverría-Barú, Fernando Sanchez-Vega, and Adrián Pastor López-Monroy. 2023. Early Detection of Mental Disorders in Spanish Telegram Messages using Bag of Characters and BERT Models. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Hermenegildo Fabregat, Andres Duque, Lourdes Araujo, and Juan Martinez-Romo. 2023. NLP-UNED at MentalRiskES 2023: Approximate Nearest Neighbors for Identifying Health Disorders. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Santiago González-Silot, Eugenio Martínez-Cámara, and L. Alfonso Ureña-López. 2023. SINAI at MentalRisk: Using Emotions for Detecting Depression. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Rodrigo Guerra, Benjamín Pizarro, Claudio Aracena, Carlos Muñoz-Castro, Andrés Carvallo, Matías Rojas, and Jocelyn Dunstan. 2023. CMM PLN at MentalRiskES: A Traditional Machine Learning Approach for Detection of Eating Disorders and Depression in Chat Messages. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. MarIA: Spanish Language Models. *arXiv preprint arXiv:2107.07253*.

O IEA. 2023. Tracking clean energy progress 2023. IEA Paris, France.

Keith Kirkpatrick. 2023. The Carbon Footprint of Artificial Intelligence. *Commun. ACM*, 66(8):17–19.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint arXiv:1910.09700*.

Xavier Larrayoz, Nuria Lebeña, Arantza Casillas, and Alicia Pérez. 2023. Eating Disorders Detection by means of Deep Learning. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*.

David Losada and Fabio Crestani. 2016. A Test Collection for Research on Depression and Language Use. In *International conference of the cross-language evaluation forum for European languages*, volume 9822, pages 28–39.

Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. 2019. Energy usage reports: Environmental awareness as part of algorithmic accountability.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research*, 24(253):1–15.

Alba María Mármol-Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-del Arco, María Dolores Molina-González, Maria Teresa Martín-Valdivia, Luis Alfonso Ureña-López, and Arturo Montejo-Raéz. 2023. Overview of MentalRiskES

at Iberlef 2023: Early Detection of Mental Disorders Risk in Spanish. *Procesamiento del Lenguaje Natural*, 71:329–350.

Ronghap Pan, José Antonio García-Díaz, and Rafael Valencia-García. 2023. UMUTeam at MentalRiskES2023@IberLEF: Transformer and Ensemble Learning Models for Early Detection of Eating Disorders and Depression. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.

Juan Manuel Pérez, Damián A Furman, Laura Alonso Alemany, and Franco Luque. 2021. RoBERTuito: a pre-trained language model for social media text in Spanish. *arXiv preprint arXiv:2111.09453*.

Aimee Van Wynsberghe. 2021. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, 1(3):213–218.

	team_run	algorithm	duration (s)	emissions (kg)	cpu_E (kWh)	gpu_E (kWh)	ram_E (kWh)	E_consumed (kWh)
2	UMUTeam_0	LLM	11.51	7.01E-08	2.28E-07	1.40E-07	1.02E-09	3.69E-07
4	UMUTeam_1	LLM	11.51	7.01E-08	2.28E-07	1.40E-07	1.02E-09	3.69E-07
12	NLP-UNED_0	DL	0.61	1.62E-06	2.98E-06	5.45E-06	1.13E-07	8.54E-06
14	NLP-UNED_1	DL	0.61	1.62E-06	2.98E-06	5.45E-06	1.13E-07	8.54E-06
15	Xabi IXA_1	DL	2.04	6.51E-06	3.36E-05	0.00E+00	6.50E-07	3.43E-05
17	Xabi IXA_2	DL	2.04	6.51E-06	3.36E-05	0.00E+00	6.50E-07	3.43E-05
18	Xabi IXA_0	LLM	2.04	6.51E-06	3.36E-05	0.00E+00	6.50E-07	3.43E-05
9	plncmm_0	ML	2.80	8.11E-06	2.10E-05	1.24E-06	1.46E-07	2.23E-05
20, 21, 22	UPM_0	LLM	303.05	1.51E-05	7.93E-05	0.00E+00	1.49E-07	7.94E-05
13	UNSL_0	LLM	4.63	2.77E-05	6.11E-05	0.00E+00	1.34E-06	6.24E-05
3	UNSL_1	LLM	4.64	2.78E-05	6.13E-05	0.00E+00	1.34E-06	6.26E-05
7	VICOM-nlp_0	LLM	3.63	4.56E-05	8.86E-05	1.50E-04	1.41E-06	2.40E-04
6	VICOM-nlp_1	LLM	3.62	4.66E-05	8.85E-05	1.55E-04	1.41E-06	2.45E-04
5	VICOM-nlp_2	LLM	3.61	4.71E-05	8.83E-05	1.58E-04	1.42E-06	2.48E-04
1	CIMAT-NLP-GTO_0	ML	3.29	2.56E-04	1.80E-04	3.42E-04	5.67E-07	5.23E-04
8	CIMAT-NLP-GTO_1	LLM	3.29	2.56E-04	1.80E-04	3.42E-04	5.67E-07	5.23E-04
16	CIMAT-NLP-GTO_2	LLM	3.29	2.56E-04	1.80E-04	3.42E-04	5.67E-07	5.23E-04
19	I2C-UHU_0	LLM	75.73	3.19E-04	8.94E-04	0.00E+00	1.90E-05	9.13E-04
11	CIMAT-NLP_0	LLM	35.01	3.53E-03	2.59E-03	4.59E-03	3.58E-05	7.22E-03
10	CIMAT-NLP_1	LLM	35.45	3.58E-03	2.63E-03	4.66E-03	3.63E-05	7.32E-03

Table 1: Emission values obtained for task1a ranked according to the average emitted emissions. The first column indicates the ranking obtained according to the value of Macro-F1. The team_run column is the team's name followed by a number representing the system it used. Some teams such as UPM seem to have used the same system on different runs as they have the same values in all metrics and variables.

	team_run	algorithm	duration (s)	emissions (kg)	cpu_E (kWh)	gpu_E (kWh)	ram_E (kWh)	E_consumed (kWh)
1	UMUTeam_0	LLM	19.49	5.52E-08	1.02E-07	1.88E-07	7.73E-10	2.91E-07
6	UMUTeam_1	LLM	19.49	5.52E-08	1.02E-07	1.88E-07	7.73E-10	2.91E-07
16	GetitDone_0	LLM	11.73	2.25E-07	7.33E-05	2.01E-05	1.16E-06	9.45E-05
25, 26, 27	DepNLP UC3M GURUDASI_0	Unknown	15.07	4.35E-07	5.70E-07	9.28E-07	2.50E-08	1.52E-06
9	NLP-UNED_1	DL	0.73	1.64E-06	3.56E-06	4.96E-06	1.37E-07	8.65E-06
14	NLP-UNED_0	DL	0.73	1.64E-06	3.56E-06	4.96E-06	1.37E-07	8.65E-06
4	TextualTherapists_1	ML	25.78	3.23E-06	1.67E-05	0.00E+00	3.15E-07	1.70E-05
8	TextualTherapists_0	ML	25.77	3.23E-06	1.67E-05	0.00E+00	3.15E-07	1.70E-05
23	TextualTherapists_2	ML	25.78	3.23E-06	1.67E-05	0.00E+00	3.15E-07	1.70E-05
5	SINAI-SELA_0	LLM	30.58	9.17E-06	8.68E-06	1.13E-05	3.01E-07	2.03E-05
7	SINAI-SELA_1	LLM	31.06	9.17E-06	8.68E-06	1.13E-05	3.01E-07	2.03E-05
24	plncmm_0	ML	4.27	1.25E-05	3.20E-05	2.16E-06	2.34E-07	3.44E-05
3	UNSL_0	LLM	3.35	2.01E-05	4.42E-05	0.00E+00	9.98E-07	4.51E-05
2	UNSL_1	LLM	3.35	2.01E-05	4.42E-05	0.00E+00	9.98E-07	4.52E-05
12	VICOM-nlp_2	LLM	3.01	3.79E-05	7.35E-05	1.25E-04	1.18E-06	1.99E-04
15	VICOM-nlp_1	LLM	3.27	3.86E-05	7.90E-05	1.23E-04	1.27E-06	2.03E-04
19	VICOM-nlp_0	LLM	3.38	4.13E-05	8.13E-05	1.35E-04	1.35E-06	2.17E-04
11	CIMAT-NLP-GTO_0	ML	1.54	1.20E-04	8.46E-05	1.61E-04	2.66E-07	2.46E-04
13	CIMAT-NLP-GTO_1	LLM	1.54	1.20E-04	8.46E-05	1.61E-04	2.66E-07	2.46E-04
17	CIMAT-NLP-GTO_2	LLM	1.54	1.20E-04	8.46E-05	1.61E-04	2.66E-07	2.46E-04
18	Ana Laura Lezama Sánchez_0	Unknown	9.72	1.45E-04	1.42E-04	1.52E-04	3.77E-06	2.98E-04
20, 21, 22	NLPUTB_0	ML	105.80	3.74E-04	1.69E-03	0.00E+00	2.72E-05	1.72E-03
30	SPIN_0	LLM	184.64	2.58E-03	0.00E+00	1.35E-02	8.63E-05	1.36E-02
28	SPIN_1	LLM	185.12	2.59E-03	0.00E+00	1.36E-02	8.65E-05	1.36E-02
29	CIMAT-NLP_0	LLM	41.21	4.04E-03	2.83E-03	5.41E-03	4.20E-05	8.28E-03
10	CIMAT-NLP_1	LLM	41.42	4.06E-03	2.84E-03	5.44E-03	4.22E-05	8.32E-03

Table 2: Emission values obtained for task2a ranked according to the average emitted emissions. The first column indicates the ranking obtained according to the value of Macro-F1. The team_run column is the team's name followed by a number representing the system it used.

	team_run	algorithm	duration (s)	emissions (kg)	cpu_E (kWh)	gpu_E (kWh)	ram_E (kWh)	E_consumed (kWh)
6	UMUTeam_1	LLM	11.27	6.86E-08	2.23E-07	1.37E-07	9.93E-10	3.61E-07
7	UMUTeam_0	LLM	11.51	7.01E-08	2.28E-07	1.40E-07	1.02E-09	3.69E-07
13	NLP-UNED_0	DL	0.61	1.62E-06	2.98E-06	5.45E-06	1.13E-07	8.54E-06
16	NLP-UNED_1	DL	0.61	1.62E-06	2.98E-06	5.45E-06	1.13E-07	8.54E-06
14	Xabi IXA_1	DL	2.04	6.51E-06	3.36E-05	0.00E+00	6.50E-07	3.43E-05
15	Xabi IXA_0	LLM	2.04	6.51E-06	3.36E-05	0.00E+00	6.50E-07	3.43E-05
17	Xabi IXA_2	DL	2.04	6.51E-06	3.36E-05	0.00E+00	6.50E-07	3.43E-05
5	plncmm_0	ML	2.80	8.11E-06	2.10E-05	1.24E-06	1.46E-07	2.23E-05
9, 10, 11	UPM_0	LLM	303.33	1.51E-05	7.93E-05	0.00E+00	1.49E-07	7.94E-05
1	CIMAT-NLP-GTO_1	LLM	3.29	2.56E-04	1.80E-04	3.42E-04	5.67E-07	5.23E-04
2	CIMAT-NLP-GTO_2	LLM	3.29	2.56E-04	1.80E-04	3.42E-04	5.67E-07	5.23E-04
12	CIMAT-NLP-GTO_0	ML	3.29	2.56E-04	1.80E-04	3.42E-04	5.67E-07	5.23E-04
4	I2C-UHU_0	LLM	75.73	3.19E-04	8.94E-04	0.00E+00	1.90E-05	9.13E-04
8	CIMAT-NLP_0	LLM	35.01	3.53E-03	2.59E-03	4.59E-03	3.58E-05	7.22E-03
3	CIMAT-NLP_1	LLM	35.45	3.58E-03	2.63E-03	4.66E-03	3.63E-05	7.32E-03

Table 3: Emission values obtained for task1b ranked according to the average emitted emissions. The first column indicates the ranking obtained according to the value of Root Mean Square Error (RMSE). The team_run column is the team's name followed by a number representing the system it used.

	team_run	algorithm	duration (s)	emissions (kg)	cpu_E (kWh)	gpu_E (kWh)	ram_E (kWh)	E_consumed (kWh)
4	UMUTeam_1	LLM	19.49	5.52E-08	1.02E-07	1.88E-07	7.73E-10	2.91E-07
5	UMUTeam_0	LLM	19.49	5.52E-08	1.02E-07	1.88E-07	7.73E-10	2.91E-07
12, 13, 14	DepNLP UC3M GURUDASI_0	Unknown	15.07	4.35E-07	5.70E-07	9.28E-07	2.50E-08	1.52E-06
11	NLP-UNED_0	DL	0.73	1.64E-06	3.56E-06	4.96E-06	1.37E-07	8.65E-06
15	NLP-UNED_1	DL	0.73	1.64E-06	3.56E-06	4.96E-06	1.37E-07	8.65E-06
3	plncmm_0	ML	4.27	1.25E-05	3.20E-05	2.16E-06	2.34E-07	3.44E-05
1	CIMAT-NLP-GTO_1	LLM	1.54	1.20E-04	8.46E-05	1.61E-04	2.66E-07	2.46E-04
2	CIMAT-NLP-GTO_2	LLM	1.54	1.20E-04	8.46E-05	1.61E-04	2.66E-07	2.46E-04
7	CIMAT-NLP-GTO_0	ML	1.54	1.20E-04	8.46E-05	1.61E-04	2.66E-07	2.46E-04
8, 9, 10	NLPUTB_0	ML	109.60	3.88E-04	1.75E-03	0.00E+00	2.87E-05	1.78E-03
16	CIMAT-NLP_0	LLM	41.21	4.04E-03	2.83E-03	5.41E-03	4.20E-05	8.28E-03
6	CIMAT-NLP_1	LLM	41.42	4.06E-03	2.84E-03	5.44E-03	4.22E-05	8.32E-03

Table 4: Emission values obtained for task2b ranked according to the average emitted emissions. The first column indicates the ranking obtained according to the value of Root Mean Square Error (RMSE). The team_run column is the team's name followed by a number representing the system it used.