LREC-COLING 2024

The Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability (LREC-COLING 2024)

Workshop Proceedings

Editors
Federico Gaspari, Joss Moorkens, Itziar Aldabe, Aritz Farwell, Begoña Altuna, Stelios Piperidis, Georg Rehm and German Rigau

25 May, 2024
Torino, Italia

Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability

ISBN 978-2-493814-48-7
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Message from the Program Chairs

This volume includes the papers that were presented at the *Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability, co-located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* in Turin, Italy, on 25 May 2024. The event was a follow-up to the first *Workshop Towards Digital Language Equality (TDLE)*, held at the previous LREC Conference in Marseille (France) in June 2022.

We were very pleased with the continued interest in this workshop series from a broad and diverse community of developers, creators, vendors, distributors, brokers, users, evaluators and researchers of Language Resources and Technologies (LRTs) in various (combinations of) languages. Encouraged by this support as well as by the increasing body of work, research, publications, projects and initiatives, we are continuing our efforts dating back several years to promote Digital Language Equality, focusing this time on the crucial issue of sustainability. We believe that this is of great importance and interest to scholars and researchers, industry and commercial players, decision- and policy-makers at various local, regional, national and supranational levels, as well as to the broader public, at a time of unprecedented and fast-paced technological progress, especially in terms of language-centric applications and tools powered by artificial intelligence (AI).

Following a thorough peer-review process, six papers were eventually accepted for presentation at the workshop. The sequence of the papers gathered in these proceedings follows the rationale of going from the broad to the particular, with the first three contributions discussing various aspects and challenges of the ambitious path towards Digital Language Equality, and the remaining three papers focusing on a range of issues having to do with sustainability.

We are very grateful to the authors who submitted their paper proposals for review to the workshop – unfortunately, due to time constraints, it was not possible to accept all of them for presentation and inclusion in the proceedings. Many thanks are due to the Organizing Committee for generously contributing to the successful preparation and running of the workshop, and to the Programme Committee members for sharing their valuable expertise by providing comprehensive and helpful reviews of the submissions that were received. Finally, our heartfelt thanks to the evaluation committee of the general International LREC-COLING 2024 Conference for accepting this workshop as a co-located event, thus recognizing the importance and value to the community of striving for Digital Language Equality and of promoting sustainability with regard to LRTs.

Organizing Committee

Itziar Aldabe (HiTZ Center, UPV/EHU, Spain)
Begoña Altuna (HiTZ Center, UPV/EHU, Spain)
Aritz Farwell (HiTZ Center, UPV/EHU, Spain)
Federico Gaspari (Uni. Naples "Federico II", Italy & ADAPT Centre, DCU, Ireland – co-chair)
Joss Moorkens (SALIS/ADAPT Centre, DCU, Ireland – co-chair)
Stelios Piperidis (ILSP, Athena RC, Greece)
Georg Rehm (DFKI, Germany)
German Rigau (HiTZ Center, UPV/EHU, Spain)

Program Committee

Antonios Anastasopoulos (GMU, USA)
Steven Bird (CDU, Australia)
Fred Blain (Uni. Tilburg, Netherlands)
Franco Cutugno (Uni. Naples "Federico II", Italy)
Bessie Dendrinos (NKUA, Greece & ECSPM, Denmark)
Félix do Carmo (Uni. Surrey, UK)
Annika Grützner-Zahn (DFKI, Germany)
Ana Guerberof-Arenas (Uni. Groningen, Netherlands)
Davyth Hicks (ELEN, Belgium)
Monja Jannet (SALIS/ADAPT, DCU, Ireland)
John Judge (ADAPT, DCU, Ireland)
Dorothy Kenny (SALIS/CTTS/ADAPT, DCU, Ireland)
Sabine Kirchmeier (EFNIL, Luxembourg)
Teresa Lynn (MBZUAI, United Arab Emirates)
Maite Melero (BSC, Spain)
Helena Moniz (Uni. Lisbon, Portugal & EAMT)
Johanna Monti (UniOR, Italy)
Rachele Raus (UniBO, Italy)
Wessel Reijers (Uni. Paderborn, Germany)
Celia Rico Pérez (UCM, Spain)
Dimitar Shterionov (TU, Netherlands)
Carlos S. C. Teixeira (IOTA Localisation Services & Uni. Rovira i Virgili, Spain)
Antonio Toral (Uni. Groningen, Netherlands)
Vincent Vandeghinste (Instituut voor de Nederlandse Taal, Netherlands & KU Leuven, Belgium)

Table of Contents

Workshop Program

Saturday, May 25, 2024

14:00–14:30 ***Introduction to the TDLE Workshop***

14:30–15:00 *Surveying the Technology Support of Languages*
Annika Grützner-Zahn, Federico Gaspari, Maria Giagkou, Stefanie Hegele, Andy Way and Georg Rehm

15:00–15:30 *Which Domains, Tasks and Languages are in the Focus of NLP Research on the Languages of Europe?*
Diego Alves, Marko Tadić and Georg Rehm

15:30–16:00 *Fine-Tuning Open Access LLMs for High-Precision NLU in Goal-Driven Dialog Systems*
Lluís Padró and Roser Saurí

16:30–17:00 *Could We Have Had Better Multilingual LLMs if English Was Not the Central Language?*
Ryandito Diandaru, Lucky Susanto, Zilu Tang, Ayu Purwarianti and Derry Tanti Wijaya

17:00–17:30 *A Language Model Trained on Uruguayan Spanish News Text*
Juan Pablo Filevich, Gonzalo Marco, Santiago Castro, Luis Chiruzzo and Aiala Rosá

17:30–18:00 *Environmental Impact Measurement in the MentalRiskES Evaluation Campaign*
Alba M. Mármol Romero, Adrián Moreno-Muñoz, Flor Miriam Plaza-del-Arco, M. Dolores Molina González and Arturo Montejo-Ráez

Surveying the Technology Support of Languages

Annika Grützner-Zahn[1], Federico Gaspari[2], Maria Giagkou[3], Stefanie Hegele[1], Andy Way[2] and Georg Rehm[1]
[1]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
[2]Dublin City University (DCU), Ireland
[3]R. C. “Athena”, Institute for Language and Speech Processing (ILSP), Greece
Corresponding author: annika.gruetzner-zahn@dfki.de

Abstract

Many of the world’s languages are left behind when it comes to Language Technology applications, since most of these are available only in a limited number of languages, creating a digital divide that affects millions of users worldwide. It is crucial, therefore, to monitor and quantify the progress of technology support for individual languages, which also enables comparisons across language communities. In this way, efforts can be directed towards reducing language barriers, promoting economic and social inclusion, and ensuring that all citizens can use their preferred language(s) in the digital age. This paper critically reviews and compares recent quantitative approaches to measuring technology support for languages. Despite using different approaches and methodologies, the findings of all analysed articles demonstrate the unequal distribution of technology support and emphasise the existence of a digital divide among languages.

Keywords: Less-Resourced/Endangered Languages, Language Resources and Technologies, Projects/Policy issues, Quantitative Evaluation Methodologies

1. Introduction

The field of Language Technology (LT) and Natural Language Processing (NLP) has seen huge progress recently. Cutting-edge technology is integrated into our daily lives more and more and used by hundreds of millions of people on a regular basis. Still, many applications are able to handle only a fraction of the approximately 7,000 languages, excluding a large number of potential users.

The ability to monitor the progress of technology support, and to make comparisons across languages, is essential to encourage further development for languages that are lagging behind the so-called ‘major’ languages. This is particularly pertinent in multilingual societies all over the world, where members of language communities poorly supported by technologies face economic, cultural and social disadvantages due to language barriers; without dedicated intervention, this situation is bound to become worse, eventually leading to digital language extinction of many low-resource languages, while speakers of the ‘major’ languages benefit from unprecedented, increasing international connectivity and all related advantages (Rehm and Uszkoreit, 2012; Kornai, 2013; Daly et al., 2023).

To obtain a realistic picture of the state of digital readiness of the world’s languages, reliable indicators and agreed upon methods are needed to measure the level of technology support. The earliest investigations examining various languages in this regard found indicative evidence through qualitative methods, most notably the META-NET White Papers in 2012. Since then, and especially with the wide-scale adoption of neural methods, the field has made various breakthroughs.

Several quantitative approaches have recently been proposed to map what is happening in this space, usually with a view to tackling the observed inequalities by encouraging the LT/NLP community to address the identified gaps and shortcomings. However, these endeavours suffer from a lack of agreed upon methods of analysing the current state of affairs, so that comparisons across studies are essentially impossible. There are differences in the data analysed, a diverse range of concepts used, and different measures employed. This reflects not only a lack of agreement within the community, but also the possible different perspectives on the topic.

This paper analyses and compares recent approaches proposed to *quantitatively* measure the level of technology support of languages, based on a systematic review, following the PRISMA 2020 approach. This work aims at deepening our understanding of the many possible factors influencing the development of Language Resources and Technologies (LRTs) for different language communities and to provide a sound base for further examinations of consequences and solutions. Our results show that despite the heterogeneity of the approaches, the measures concerning LRTs can be differentiated between measures of quantity (how many LRTs are available?) and quality (how good are existing LRTs?). Most approaches consider socio-economic factors and examine dependencies on research and the broader economy. All surveyed studies demonstrate an unequal distribution of technology support, proving the existence of a digital divide.

Section 2 of this paper summarises related work. Sections 3 and 4 describe the methodology and the articles that were analysed and compared as part of this systematic review. The results are presented in Section 5, while Section 6 discusses selected aspects in more detail and Section 7 concludes the paper.

2. Related Work

Language death is a threat to many small communities. Bromham et al. (2021) examine the effects of a range of demographic and socio-economic aspects on the use and status of the world's languages, and conclude that language diversity is endangered since half of the languages are at risk of extinction. This trend also applies in the digital sphere. The vitality analysis by Kornai (2013) shows that at least 2,500 languages are considered to be endangered.

The preservation of languages is a key goal of future LRT development (Rehm and Uszkoreit, 2012; Meighan, 2021; Daly et al., 2023). Many publications advocate for implementing ethical principles such as equity or equality, fairness, and diversity for languages in the digital realm (Carew et al., 2015; Soria, 2017; Bender and Friedman, 2018; Birhane, 2021; Choudhury and Deshpande, 2021; Ramesh et al., 2023; Rehm and Way, 2023). With regard to the development of LRTs, the focus should shift from optimising performance to a more holistic, human-centred perspective in order to serve all user communities (Ethayarajh and Jurafsky, 2020). Emerging technologies are used by all kinds of language communities around the globe, from small to large, as well as in traditionally oral contexts or deaf communities (Prasad et al., 2018; van Esch et al., 2019). Focusing on primarily oral languages, Bird (2022) argues that a shift is required which builds on the participation of local communities to identify new opportunities for LRTs in low-resource scenarios, abandoning the assumption that all languages can be served by the same technologies.

Krauwer (2003) provided one of the first calls for action towards more emphasis on under-resourced languages. Subsequent qualitative analyses of the technology support of languages continued to indicate a trend towards a digital divide between a few dominant and widely-used languages and many other languages, which are far less supported (or not at all), often spoken by smaller language communities (Yin et al., 2021; Khanuja et al., 2023). In addition to the linguistic bias, Helm et al. (2023) talk about a techno-linguistic or 'design' bias (Santy et al., 2023), which is expressed through the inability of systems to adapt to the knowledge systems of non-Western language communities.

More recent research focuses on "digitally-disadvantaged languages". Corresponding language communities encounter the following three main challenges because of missing LRT support: "1. gaps in equitable access; 2. digital tools that negatively impact the integrity of their languages, scripts and writing systems, and knowledge systems; and 3. vulnerability to harm through digital surveillance and under-moderation of language content" (Zaugg et al., 2022, pp. 2–3).

The articles analysed in this work provide data-driven evidence for the current situation. They indicate which languages need further financial support and research efforts to be able to mitigate current inequalities and biases within LRTs.

3. Methodology

In order to analyse and compare the approaches used to measure the level of technology support, we conducted a systematic review of articles, with a view to pointing out common and unique features as well as shortcomings that require further investigation. We followed the PRISMA 2020 statement, which includes a checklist of items for systematic reviews emphasising transparency, accuracy and completeness (Shamseer et al., 2015; Page et al., 2021a,b, 2022).

The earliest relevant article for this systematic comparative review, Joshi et al. (2020), seeks to gauge the technology support of individual languages from a quantitative perspective. Since then, further quantitative research has been published. We performed an extensive search in Google Scholar, going through a total of 419 citations (as of August 2023) for Joshi et al. (2020). Qualitative papers, literature reviews and benchmark evaluations were excluded, since our goal was exclusively the evaluation of novel quantitative approaches. To ensure that Joshi et al. (2020) was indeed the first publication of our interest, the publications referenced in the papers collected were checked for missing papers written in languages other than English, but none were found. These steps led to a set of nine papers (see Table 1).[1]

Five key criteria were defined: C1 *Research question* examines the different perspectives on the topic of technology support for languages; C2 *Scope* compares the number of languages,

[1] P7a and P7b are two complementary publications from the same project. While listed separately in Table 1, in the rest of the paper they are considered jointly.

ID	Reference	Year
P1	The State and Fate of Linguistic Diversity and Inclusion in the NLP World, P. Joshi et al.	2020
P2	Systematic Inequalities in [LT] Performance across the World's Languages, D. Blasi et al.	2022
P3	Dataset Geography: Mapping Language Data to Language Users, F. Faisal et al.	2022
P4	Some Languages are More Equal than Others: […], S. Ranathunga	2022
P5	Assessing Digital Language Support on a Global Scale, G. F. Simons et al.	2022
P6	Evaluating the Diversity, Equity and Inclusion of NLP Technology: […], S. Khanuja et al.	2022
P7a	Introducing the Digital Language Equality Metric: Technological Factors, F. Gaspari et al.	2022
P7b	Introducing the Digital Language Equality Metric: Contextual Factors, A. Grützner-Zahn et al.	2022
P8	Writing System and Speaker Metadata for 2,800+ Language Varieties, D. van Esch et al.	2022

Table 1: Scientific articles included in our literature review

geographical and LRT areas covered and number of measurements reported; C3 *Conceptualisation and quantification* analyses the ways in which the items to be measured are conceptualised and quantified; C4 *Combination of factors* considers how different factors were put in relation to one another; C5 *Results* compares the results of the paper with those of the other articles under review.

We manually extracted the relevant information from each article, looking separately at independent measures in each one, i. e., for measurements related to separate factors. While we succeeded in analysing each paper for each criterion, the diversity required different types of comparison, in particular, for "Conceptualisation and quanitification" and "Combination of factors".

In terms of possible reporting biases, the article selection process may have resulted in this survey missing those articles that do *not* cite Joshi et al. (2020), but that still conduct relevant research. This would also apply to articles published before 2020. Additionally, there are some borderline cases, such as the well-known article by Kornai (2013), which could not cite Joshi et al. (2020) and which does not fit our survey fully because its goal was to assess the vitality of languages (i. e., decline, endangerment, and eventually death). Another borderline case was the PhD thesis by Berment (2004) which provides a framework for the quantification of the computerisation of languages, but the data used for this quantification is based on expert knowledge and a possible application of the framework is only shown for one language. Furthermore, during the analysis of the results (Section 5), the European affiliation of the authors may have resulted in a stronger emphasis and focus on European languages in the assessment of how the results match our knowledge (and potentially expectations) about a language community and its situation.

4. Materials and Data Sources

Our survey includes the most significant contributions in this area. Eight papers were published in 2022 compared to only one in 2020, which points to a dynamic and fast-progressing area of work, whose importance is likely to increase.

Joshi et al. (2020) investigate the relation between the world's languages and resources as well as their coverage in NLP conferences: their analysis reveals a severe disparity across languages in terms of available data sets and coverage in research fora. Blasi et al. (2022) assess the global utility of LTs in relation to demographic or linguistic demand and analyse over 60,000 NLP conference papers, showing evidence for the unequal development of LTs across languages.

Faisal et al. (2022) argue that the availability of data is the decisive factor for the quality of NLP systems, and investigate the geographical representativeness of datasets, gauging to what extent they meet the needs of the language communities, exploring especially geographical and socioeconomic factors that may explain the dataset distributions. Ranathunga and de Silva (2022) look at linguistic disparity in NLP, using a categorisation of languages based on speaker population and vitality. They examine the distribution of LRs, the amount of research, inclusion in multilingual platforms and models among the categories and analyse the role of some contextual factors.

Simons et al. (2022) evaluate the level of digital language support through the extraction of language names from the websites of over 140 tools, and propose a categorisation of the languages based on the number of tools per LT area. Khanuja et al. (2023) focus on Indian languages and discuss an approach to evaluating NLP technologies based on diversity, equity and inclusion to quantify the diversity of the users they can serve. The method aims at addressing gaps in LRT provisioning related to societal wealth inequality.

Gaspari et al. (2022) and Grützner-Zahn and Rehm (2022) present the Digital Language Equality metric, that quantifies the digital support of Europe's languages. Its technological factors measure the number of LRTs for each language within the European Language Grid (ELG, Labropoulou et al., 2020; Rehm et al., 2021; Rehm, 2023),[2] which is assumed to be representative. The contextual factors reflect the broad socio-economic ecosystem of the languages by taking into account a set of indicators considered to be relevant.

Finally, van Esch et al. (2022) describe an open-source dataset which covers 2,800 languages and their writing system(s), along with estimates of the speaker populations. They analyse the distribution of languages and writing systems in language models (LMs), comparing it to the coverage of the respective language families in NLP research, hoping that this dataset will help develop NLP research for under-researched languages.

5. Results

5.1. C1 Research Question

The authors of all papers measured the technology support of languages independently, with no common objective or framework, and so decided to consider a number of different factors, which are difficult to directly compare or relate to each other. Already the specific research questions target different aspects and focus on subsets of areas, such as data availability or scientific output in NLP research. The approaches distinguish between the measurement of LRs, LTs or socio-economic factors assumed to have an impact on the development of LRTs. We identified 20 different research questions or aims (see Table 4 in the Appendix) evenly distributed across these three areas. One difference between the approaches is whether the goals are defined to measure either availability, coverage and quantity or performance and quality of the LRTs. Next to the more specific aims like "distribution of available data" (Joshi et al., 2020) or "platform interface availability" (Ranathunga and de Silva, 2022), some authors tried to approximate notions such as "inclusion" (Ranathunga and de Silva, 2022; Joshi et al., 2020) or "equity" (Khanuja et al., 2023) in the realm of LRTs. Just like Simons et al. (2022), Gaspari et al. (2022) and Grützner-Zahn and Rehm (2022) define their own concepts. Interestingly, both approaches cover the largest number of LRT areas (see Section 5.2).

The analysis of scientific coverage of single languages as one of the most prevalent socio-economic factors is quite striking. The only other group of socio-economic factors directly mentioned as an object of measurement are geographic factors because of a specific focus on geographical representation (Faisal et al., 2022). In two papers, the socio-economic factors concerning the number of speakers (van Esch et al., 2022) and global demand (Blasi et al., 2022) are used as part of the ratios to LRT coverage or performance.

5.2. C2 Scope

The scope of an approach to measure the level of technology support can be thought of in different ways. The geographical coverage or number of languages investigated differs based on the focus of the research or data availability considered by each paper. Similarly, the numbers of languages addressed range from 15 to 7,829 (see Table 2). Faisal et al. (2022) propose a language-independent approach and do not state the number of languages in the datasets they analyse. Although six of the eight papers aim at surveying all languages of the world, only two actually cover more than 6,000 languages. The other approaches miss out on a huge number of languages that exist today in the world, despite the stated ambition to cover them all.

The influence of data availability is especially visible in the article by Blasi et al. (2022), where the number of languages under consideration varies dramatically depending on the task being evaluated, with values ranging between 15 and 630. A similar issue was observed for the socio-economic data (Grützner-Zahn and Rehm, 2022).

ID	Region	Number of Languages
P1	World	2,485
P2	World	Task-dependent (between 15 and 630)
P3	World	Not mentioned; language-independent
P4	World	6,420
P5	World	7,829
P6	India	22
P7	Europe	90
P8	World	2,800

Table 2: Targeted region and languages covered

The measurement approach can also reflect different aspects of technology support, such as quality of performance, quantity of LRTs or aspects such as efficiency. Table 3 shows that some papers focus on the creation of a single measurement concentrating on one aspect of technology support (e. g., Simons et al., 2022), while others establish a number of different, independent means (e. g., Ranathunga and de Silva, 2022, consider five). Those independent means can be mapped to different LRT areas (e. g., Ranathunga and de Silva, 2022), or a single area (Khanuja et al., 2023; van Esch et al., 2022); this raises the question of the extent to which the measurement of support for a language in a single LRT area can actually provide an accurate and reliable indication of the overall level of technology support that can also be compared across languages.

[2] https://www.european-language-grid.eu

In contrast, a small number of approaches are based on broader coverage. Simons et al. (2022) and Gaspari et al. (2022) cover a relatively high number of LRT areas, nine and ten. Half of the papers cover only one or two LRT areas, which are taken to be good enough indicators to reflect the overall state of technology support. This applies to Joshi et al. (2020) and Khanuja et al. (2023), which also define in their aim of measurement general LT performance or broad social concepts such as "inclusion", "equity" or "diversity" (see Section 5.1). Seven out of eight papers (i. e., all papers except Simons et al., 2022) also consider socio-economic or contextual factors, such as scientific coverage or Gross Domestic Product (GDP).

ID	Number of Approaches	Number of LRT Areas	Socio-economic Indicators
P1	4	2	yes
P2	2	6	yes
P3	3	2	yes
P4	5	3	yes
P5	1	9	no
P6	3	1	yes
P7	2	10	yes
P8	2	1	yes

Table 3: Scope of measurement

Another way of approaching this aspect is the analysis based on the size of the datasets used in the papers. We did not include this dimension in our review because it would have required a lot of additional work with potentially little actual gain, particularly due to the difficulty of directly combining, and comparing across languages, different measures of the size of the datasets. In addition, the papers do not always provide all details concerning the size of the datasets they discuss, often only referring readers to other sources. Preliminary attempts to gather the details from other cited papers, archives and platforms required substantial effort without necessarily leading to conclusive results that could be confidently analysed or compared for the purposes of this survey.

5.3. C3 Conceptualisation and Quantification

For C3, the measurement approaches were divided into LRs, LTs and socio-economic indicators.

5.3.1. Language Resources

Five papers measure the availability of data for a language or the representation of language (community) features in the available data (see Table 6). Joshi et al. (2020) is the only one that takes both dimensions into account. Three papers use the raw counts of datasets with labelled and unlabelled data in different repositories to approximate the availability of language data. Joshi et al. (2020) add the question of the distribution of these data resources. The distribution is exemplified through the classification of languages and further analysis of size. Ranathunga and de Silva (2022) focus solely on the coverage of languages. They reuse the approach from Joshi et al. (2020), but add another repository, Hugging Face, to the repositories used by Joshi et al. (2020), namely LDC and ELRA. A weighting of datasets based on their features is introduced by Gaspari et al. (2022) who use as data source the ELG which harvests several major repositories such as Zenodo and CLARIN. Gaspari et al. (2022) mention the problem of dataset size, which is difficult to measure because of missing data and incompatible descriptions and values, while recognising that it would be desirable to include this information.

Joshi et al. (2020), Faisal et al. (2022), and van Esch et al. (2022) analyse the representation of language features or local knowledge in the datasets. All three focus on different aspects of language representation which reflect the layers of diversity between languages and their communities. Joshi et al. (2020) examine which language features are not represented in the datasets. This typological conceptualisation is motivated by transfer learning and the idea that less-resourced languages can reach a better level of support if their features are covered in LMs. van Esch et al. (2022) also concentrate on a language's writing system. The share of a writing system in the vocabulary of multilingual models is calculated, from which the authors induce the representation in NLP. Both conceptualisations are motivated by language modelling (either through its learning mechanisms or the training data) and directly compare the languages with each other based on the chosen feature. Faisal et al. (2022) deviate from this through a focus on local knowledge contained in language data. The number of local entities in the dataset reflects the distance of the dataset from the users and their needs through language and LT-task-independent means. Nonetheless, only datasets designed for Named-Entity-Recognition (NER) and Question Answering (QA) are analysed in the paper. While Joshi et al. (2020) use a dataset that is independent of the LT area, Faisal et al. (2022) and van Esch et al. (2022) use LT-specific datasets and deduce the general concept of representation within NLP.

5.3.2. Language Technologies

Four papers include measures of LT performance, while three papers contain measures of LT availability, but none combine both perspectives (see

Table 7). The papers mainly use known performance measures for certain NLP tasks, such as reused Natural Language Inference (NLI) error rates (Artetxe and Schwenk, 2019; Joshi et al., 2020). Faisal et al. (2022) calculate F1 scores in the context of QA and the influence of geographical representation in the training datasets. Blasi et al. (2022) introduce a measure of utility which is quantified as performance divided by a theoretical maximum performance. The utility per LT area is added up to reach an overall score per language. Overall, the possibility to summarise LT performance of different LT tasks, despite the use of different performance measures, gives a broader picture than the approaches covering single LT areas. Khanuja et al. (2023) use different means to measure the performance of LMs. They reuse the utility metric from Blasi et al. (2022), but extend it through projected performance estimates for languages without available test data (otherwise set to 0) based on the performance of languages from the same family and the availability of unlabelled data. This extension is motivated by transfer learning and the possible performance increase, if language features are learnt from another language. Since the utility measure assigns the same scores to the languages covered by one model, despite different performance on different languages, another measure is proposed to account for the equity in model performance, namely the Gini-coefficient, which measures the inequality within a distribution (model performance on different languages). A third measure reflects inclusion through model efficiency concerning the use of computational resources. This last measure of efficiency shows that other methods of evaluating the "quality" may be of importance, even though they are considered by only one paper.

Similar to the LR availability measures, the LT availability measures are quantified as counts of services available for the languages. Ranathunga and de Silva (2022) collected the languages in which Google Translate and Facebook are available. Similarly, the languages covered by mBERT and XLM-R are counted, to provide an approximation for general model coverage. In the article by Simons et al. (2022), Digital Language Support is conceptualised as a scale with a strict support-level hierarchy, in which each level is quantified through the number of popular tools available for each LT area. For each LT area, the ten most popular tools globally and the five most popular tools of each of the ten largest countries in terms of population were selected. An approach to combining several LT areas into one metric was also chosen by Gaspari et al. (2022), based on the number of available LTs in ELG per language. Again, the authors included a weighting mechanism into the calculation of a score representing LT support, assuming that some LTs are more demanding to develop than others. While Ranathunga and de Silva (2022) analyse only two platforms and two models, Simons et al. (2022) and Gaspari et al. (2022) quantify the availability of LTs in a broader and more comprehensive way.

5.3.3. Socio-economic Indicators

The socio-economic factors are often approximated with indicators of scientific output or inclusion. A typical quantification of scientific output is the number of publications concerning a particular language (Joshi et al., 2020; Ranathunga and de Silva, 2022; Grützner-Zahn and Rehm, 2022; van Esch et al., 2022). Alternatively to the use of plain figures (Ranathunga and de Silva, 2022; van Esch et al., 2022), Joshi et al. (2020) use language occurrence entropy as a proxy for language diversity in NLP conferences. Another perspective is the use of reputation quantified as the number of citations (Blasi et al., 2022) or the prediction of proximity through embeddings in which authors, languages and conferences serve as entities and the title and abstract as context (Joshi et al., 2020).

In all cases, the economic situation is quantified with GDP, while the size of a language community is defined as the number of speakers, although often the information about how people qualify as speakers (acknowledged as a difficult issue) is missing. Blasi et al. (2022) quantify demographic demand using also the number of speakers, while contrasting it to linguistic demand. Faisal et al. (2022) introduce a geographical distance, reflecting the distance between user and producer based on entities in a dataset, and country size. The situation of a language and its speakers can be measured by a range of factors. Most authors focus on just a few, perceived as most relevant for the development of LRTs. Grützner-Zahn and Rehm (2022) present the only approach trying to combine socio-economic factors from different areas (such as science, education, economy, etc.) to paint an overall picture on a single scale of the specific contexts of Europe's languages as part of the Digital Language Equality concept.

5.4. C4 Combination of Factors

The approaches presented in the eight papers differ substantially in terms of how their indicators contribute to the bigger picture. While some only represent single indicators and their results, others create a metric in which the indicators are combined to a ratio (all except Ranathunga and de Silva, 2022). Some approaches measure the relation between two factors through co-occurrence or correlation measures (see Table 10 in the Appendix for the relevant details). While Blasi et al. (2022), Simons et al. (2022) and Gaspari et al. (2022) along with Grützner-Zahn and Rehm (2022) combine indicators of different types representing different LRT areas or even socio-economic factors into one overall score as a result, Joshi et al. (2020), Faisal et al. (2022) and van Esch et al. (2022) create ratios within one area of application. Khanuja et al. (2023) aim to measure the three concepts of "diversity", "equity" and "inclusion" creating ratios combining different aspects of model performance.

Joshi et al. (2020) and Simons et al. (2022) assign languages to classes. While in Joshi et al. (2020) the class of the language is derived from the data availability measure, Simons et al. (2022) distinguish classes of digital language support based on the coverage of available LTs. The step of including a classification on top of the scores is left out by Blasi et al. (2022) and Gaspari et al. (2022) along with Grützner-Zahn and Rehm (2022), although both approaches also result in overall scores per language.

The papers examining the relation between two factors use either basic occurrence measures searching for patterns or outliers (Joshi et al., 2020; Ranathunga and de Silva, 2022; van Esch et al., 2022) (see Table 9 in the Appendix) or correlation measures (Blasi et al., 2022; Faisal et al., 2022; Ranathunga and de Silva, 2022). Ranathunga and de Silva (2022) use the occurrence measures to identify the outlier, and analyse it through an additional correlation measure. In contrast, Blasi et al. (2022) and Faisal et al. (2022) use different kinds of correlation measures to examine which socio-economic factor (e. g., GDP or number of speakers) best predicts the result, such as the number of papers published on a language or the representation of language communities in a dataset.

5.5. C5 Results

All papers identified a digital divide between a few dominant languages and a majority of low-resourced languages. Not surprisingly, English is always, by far, the best supported language in all LRT areas when languages are compared directly, usually followed by Spanish, German and French (Joshi et al., 2020; Ranathunga and de Silva, 2022; Gaspari et al., 2022; Grützner-Zahn and Rehm, 2022), three official European Union languages with large numbers of speakers. Ranathunga and de Silva (2022) detect a bias towards Indo-European languages spoken in Europe and institutional languages[3] with large speaker populations. Still, even within Europe a huge imbalance was identified by Gaspari et al. (2022) and Grützner-Zahn and Rehm (2022). Regional and minority languages (RMLs) have mostly been ignored (with a few exceptions, such as Basque, van Esch et al., 2022). The authors conclude that much additional effort is needed to bridge the gap, although even most official languages lag way behind the 'major' languages mentioned.

Concerning data availability, more than half (Simons et al., 2022) or even 80% (Joshi et al., 2020) of the languages lack enough data to develop LT applications. Additionally, the size of the dataset decreases with the language class, meaning that even those datasets available for low-resourced languages are substantially smaller (Ranathunga and de Silva, 2022). Task-oriented datasets were found to have the highest counts for popular NLP tasks on large institutional languages, such as Machine Translation (MT) (Ranathunga and de Silva, 2022).

Most datasets exhibit biases towards the global west (Faisal et al., 2022) or linguistic feature representation of Indo-European or large official languages (Joshi et al., 2020; van Esch et al., 2022). Faisal et al. (2022) claim an unrepresentative number of entities in the data, but also find differences between datasets, such as MasakhaNER and Natural Questions from Google, which include a high proportion of entities from all over the world. Imbalances concerning linguistic features were detected to the extent that usually 2.86 categories per language feature are not represented in language data (Joshi et al., 2020). Combined with a measure showing higher error rates for languages containing these features, the results highlight the importance of language representation in data. Similarly, Faisal et al. (2022) show a decrease in performance on QA, if local knowledge is not included in training datasets.

For the development of LTs, Simons et al. (2022) find a correlation to higher categories of digital language support, implying that a strong basis of LRs and basic LT tools seems to be needed for the development of the higher categories, such as virtual assistants. Additionally, the results of the LT areas by Blasi et al. (2022) show that the majority of morphological or syntactic tools perform quite well if enough data is available. For MT and Text-to-Speech, the performance differs substantially among the languages with medium technology support. For complex tasks, such as NLI and QA, most systems perform poorly except for a few large official languages for which performance allows for actual use in operational settings. The performance of multilingual LMs shows a huge im-

[3] Ranathunga and de Silva (2022) introduce the term "institutional languages" as a class of languages. The term is used in this paper to avoid confusing the discussion of their results.

balance even for the best models, although region-specific tuning can counteract the limited transfer between languages in a multilingual model to a certain extent (Khanuja et al., 2023). Similar results were shown by Simons et al. (2022) and Gaspari et al. (2022). While basic LTs are available for a considerable number of languages, the number quickly decreases for less-resourced languages as the complexity of the tools grows.

The analyses of the socio-economic factors show a similar pattern. The scores for contextual factors (Grützner-Zahn and Rehm, 2022) describe an uneven distribution towards large official languages in Europe, while RMLs receive little attention from the economy, politics, etc. The results make national and regional efforts towards the support of RMLs visible, e. g., the co-official languages in Spain achieve relatively high scores compared to RMLs with similar numbers of speakers elsewhere. Correlation measures give indicative evidence that the GDP and/or geographical distance are the two socio-economic factors that best predict the amount of NLP research and development (Blasi et al., 2022; Faisal et al., 2022; Ranathunga and de Silva, 2022). The best predictor for geographical representation constitutes a ratio of the two factors, reflecting that potentially many socio-economic factors have an impact on LRT development (Faisal et al., 2022), as considered by Grützner-Zahn and Rehm (2022) in which a ratio of socio-economic factors is calculated. Although Blasi et al. (2022) and Faisal et al. (2022) show that the inclusion of speaker population causes the prediction to deteriorate, the number of speakers is considered by most papers analysing socio-economic factors (Blasi et al., 2022; Faisal et al., 2022; Khanuja et al., 2023; Grützner-Zahn and Rehm, 2022; van Esch et al., 2022).

The focus on NLP research in most papers shows a more fine-grained picture. Large European languages are considerably more often the subject of research, and more popular languages are in turn propagated more, making the existing imbalance even worse (Joshi et al., 2020). Additionally, the number of languages addressed in a publication does not predict the number of citations a paper is going to receive, i. e., there is no incentive for researchers to address a larger number of languages. Still, focused research communities have been detected for some non-European or non-official languages, such as Japanese, Turkish, Inuktitut, Hawaiian, etc. (Joshi et al., 2020; Blasi et al., 2022), and even among those languages with large speaker populations, some are under-represented (van Esch et al., 2022), showing that concentrated efforts are picked up by quantitative measures, and that it is not all about size.

Some authors classify languages based on the resulting scores. However, classifications create hard boundaries, introducing a distinction between languages which might otherwise be thought of as having similar levels of support, e. g., Simons et al. (2022) assign Hungarian the class "Thriving", while Latvian is "Vital". In Gaspari et al. (2022), though, Hungarian and Latvian achieve similar scores. In contrast, some languages which appear to have different levels of support are grouped together. Simons et al. (2022) classify Latvian, Occitan and Yiddish as "Vital", but they obtain very different scores in Gaspari et al. (2022). Comparing size proportions between the approaches using a taxonomy, Joshi et al. (2020) group 88% of the languages in the lowest class, while 50% of the languages constitute the lowest class in Simons et al. (2022). Overall, this paints different pictures. Moreover, Ramesh et al. (2023) show that adding another data source changes the classification for 87 languages based on data availability. They conclude that single classifications should be avoided.

6. Discussion

All papers use very diverse approaches to measure the level of technology support of languages. Some authors chose to use notions from other fields such as "demand" and "utility" (Blasi et al., 2022) or "inclusivity", "equity" and "accessibility" (Khanuja et al., 2023). These are used in different ways and can be ambiguous if not properly defined and operationalised with respect to the languages under consideration. For instance, the definition of demand depends on the background, e. g., in economics it is viewed as the need of goods by consumers and may or may not include the will to pay depending on the use case (Rinkinen et al., 2020). In Blasi et al. (2022), demand is conceptualised from two angles, resulting in different outcomes for the metric. Demographically, demand was quantified through the number of speakers, but who exactly counts as a speaker and which (type of) speaker needs which (kind of) LT can be debated. Further explorations of how different quantification of demand may influence the results of the metric would be desirable to better assess its impact and to argue for a specific way of quantifying demand. The same applies to the other concepts mentioned above.

When analysing large datasets, biases can have an impact on different levels of the study: • Dataset assessment: analysis of biases in a dataset or study reused; • Study assessment: analysis of what kind of biases may be introduced through the choice of quantification, methodology, etc.; • Outcome-level assessment: analysis of biases in the results; • Reporting bias assessment: detecting whether all relevant results are made available. Not all levels need to be analysed in all studies, but dataset assessment is applicable to all studies, because they all reuse data. Only Ranathunga and de Silva (2022) describe possible biases through their chosen methodology and data in the appendix.

One possible source for bias has to do with the Bender rule (Bender, 2019). Many authors do not explicitly mention the language(s) covered, which is why figures about number of publications per language inevitably miss relevant publications. Another question has to do with how a speaker of a language is defined. Are L2 speakers considered? And if so, how reliable are the figures? A closer look into Ethnologue shows that many figures concerning the number of speakers are outdated, only contain L1 speakers or derive the number of speakers from the citizenship of the individuals, which distorts the numbers, especially in certain countries and regions. The question of which tools to include when approximating the technology support of a language can also introduce biases. Meighan (2021) and Bird (2022) show that some smaller language communities develop their own LRTs. Criteria such as tool popularity miss these developments and may fail to detect smaller advancements, that however may be significant for the language communities in question.

In Section 5.5, only parts of the results of the eight papers could be covered since not all findings were published; only Faisal et al. (2022), Gaspari et al. (2022), Grützner-Zahn and Rehm (2022) and van Esch et al. (2022) published all results. Simons et al. (2022) published 10% of their results which facilitates traceability, but does not allow extensive comparisons with other research. Joshi et al. (2020), Blasi et al. (2022), Ranathunga and de Silva (2022) and Khanuja et al. (2023) do not provide their full results or datasets. Thus, only the results described in these papers could be included in this survey.

7. Conclusions and Future Work

The systematic comparison of the eight papers under examination has shown that despite the heterogeneous approaches and differences on all levels of analysis, the results clearly indicate a very uneven distribution of LRTs between large, official, mostly Indo-European languages and essentially all other languages. The papers highlight different aspects, such as the output of focused research communities on specific languages or the influence of local knowledge on the performance of LMs. Combining all results to assemble a bigger picture reveals the many dependencies between all areas of LRTs and socio-economic factors. Efforts are needed on all levels, starting with data collection, for at least half of the world's languages.

Future work needs to examine how to standardise and measure the size of LRs and, similarly, the scope of LT applications. Another open issue is the actual quality of LRTs. Moreover, biases need to be further analysed, especially concerning their influence on the quantitative measures. All approaches we analysed cover only parts of the relevant measures, which is why the development of a measure accounting for all qualitative and quantitative perspectives, and covering all LRT areas would be an important step forward. Based on such an all-encompassing approach, further steps towards evaluation and the examination of possible solutions could be conducted.

Ethical Statement

The research described in this paper does not require typical ethical considerations. Having said that, when considering ethical aspects, the authors believe that analysing equality or equity, fairness and diversity within the area of Language Resources and Technologies is a timely and crucial topic that, on a general level, deserves much more attention in our research field. Additionally, the authors are all located in Europe which may have resulted in a focus on European languages in the analysis of the results because the authors are more familiar with them.

Acknowledgements

The European Language Equality project has received funding from the European Union under the grant agreements no. LC-01641480 – 101018166 (ELE) and LC-01884166 – 101075356 (ELE2), see https://european-language-equality.eu.

Bibliographical References

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Vincent Berment. 2004. *Méthodes pour informatiser les langues et les groupes de langues “ peu dotées ”*. Ph.D. thesis, Université Joseph-Fourier - Grenoble I.

Steven Bird. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World’s Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2021. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2):163–173.

Margaret Carew, Jennifer Green, Inge Kral, Rachel Nordlinger, and Ruth Singer. 2015. Getting in Touch: Language and Digital Inclusion in Australian Indigenous Communities. *Language Documentation & Conservation*, (9):307 – 323.

Monojit Choudhury and Amit Deshpande. 2021. How Linguistically Fair Are Multilingual Pre-Trained Language Models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12710–12718.

Emma Daly, Jane Dunne, Federico Gaspari, Teresa Lynn, Natalia Resende, Andy Way, Maria Giagkou, Stelios Piperidis, Tereza Vojtěchová¡, Jan Hajič, Annika Grützner-Zahn, Stefanie Hegele, Katrin Marheinecke, and Georg Rehm. 2023. Results of the Forward-looking Community-wide Consultation. In Georg Rehm and Andy Way, editors, *European Language Equality: A Strategic Agenda for Digital Language Equality*, Cognitive Technologies, pages 245–262. Springer, Cham, Switzerland.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. 2022. Dataset Geography: Mapping Language Data to Language Users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3381–3411, Dublin, Ireland. Association for Computational Linguistics.

Federico Gaspari, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way. 2022. Introducing the Digital Language Equality Metric: Technological Factors. In *Proceedings of The Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 1–12, Marseille, France. European Language Resources Association.

Annika Grützner-Zahn and Georg Rehm. 2022. Introducing the Digital Language Equality Metric: Contextual Factors. In *Proceedings of The Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 13–26, Marseille, France. European Language Resources Association.

Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2023. Diversity and Language Technology: How Techno-Linguistic Bias Can Cause Epistemic Injustice. Publisher: arXiv Version Number: 1.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

András Kornai. 2013. Digital Language Death. *PLoS ONE*, 8(10):e77056.

Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*, Moscow, Russia. Moscow State Linguistic University.

Penny Labropoulou, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva. 2020. Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3421–3430, Marseille, France. European Language Resources Association (ELRA).

Paul J Meighan. 2021. Decolonizing the digital landscape: the role of technology in Indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17(3):397–405.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021a. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, page n71.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, and David Moher. 2021b. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, 134:103–112.

Matthew J. Page, David Moher, and Joanne E. McKenzie. 2022. Introduction to PRISMA 2020 and implications for research synthesis methodologists. *Research Synthesis Methods*, 13(2):156–163.

Manasa Prasad, Theresa Breiner, and Daan Van Esch. 2018. Mining Training Data for Language Modeling Across the World's Languages. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 61–65. ISCA.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in Language Models Beyond English: Gaps and Challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Georg Rehm, editor. 2023. *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer, Cham, Switzerland.

Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals. 2021. European Language Grid: A Joint Platform for the European Language Technology Community. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*, pages 221–230, Kyiv, Ukraine. Association for Computational Linguistics (ACL).

Georg Rehm and Hans Uszkoreit, editors. 2012. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages. Springer, Heidelberg etc.

Georg Rehm and Andy Way, editors. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer International Publishing, Cham.

Jenny Rinkinen, Elizabeth Shove, and Greg Marsden. 2020. *Conceptualising Demand: A Distinctive Approach to Consumption and Practice*, 1 edition. Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2021.

Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. Publisher: arXiv Version Number: 1.

L. Shamseer, D. Moher, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart, and the PRISMA-P Group. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*, 349(jan02 1):g7647–g7647.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Claudia Soria. 2017. What is Digital Language Diversity and why should we care? In *Digital Media and Language Revitalisation*, number 4 in Linguapax Review 2016, pages 13–.28.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing System and Speaker Metadata for 2,800+ Language Varieties. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5035–5046, Marseille, France. European Language Resources Association.

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Crew, Chieu Nguyen, and Françoise Beaufays. 2019. Writing Across the World's Languages: Deep Internationalization for Gboard, the Google Keyboard. Publisher: arXiv Version Number: 1.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Isabelle A. Zaugg, Anushah Hossain, and Brendan Molloy. 2022. Digitally-disadvantaged languages. *Internet Policy Review*, 11(2).

A. Appendix

ID	Aim of Measurement
P1	Distribution of available data over languages Typological features of languages represented in data and influence of missing representation on LT performance Language diversity and inclusion of NLP conferences Closeness of authors, conferences and languages
P2	To what degree is the global demand for LT met? Correlation of scientific production in NLP
P3	Geographical representativeness of NLP datasets Socio-economic correlates Geographical breakdown of models performance
P4	Annotated data availability Platform interface availability Model coverage Amount of research conducted for the languages
P5	Digital language support
P6	Diversity Equity Inclusion
P7	Digital language equality
P8	Representation of writing systems in NLP compared to their speaker numbers Distribution of published works that reference the languages

Table 4: Research questions or aims

ID	LRT Areas Covered
P1	Data; Natural Language Inference
P2	Morphological Inflection; Syntactic Parsing; Text-to-Speech; Machine Translation; Question Answering; Natural Language Inference
P3	Data; Language Modelling
P4	Data; Human-Computer-Interaction; Language Modelling
P5	Data; Encoding; Morphological Inflection; Syntactical Parsing; Text-to-Speech; Machine Translation; Question Answering; Natural Language Inference; Human-Computer-Interaction
P6	Language Modelling
P7	Data; Encoding; Morphological Inflection; Syntactical Parsing; Text-to-Speech; Machine Translation; Question Answering; Natural Language Inference; Human-Computer-Interaction; Language Modelling
P8	Data

Table 5: LRT areas covered

ID	Concept	Conceptualisation	Values Used
P1	Distribution of available data	Creation of language taxonomy based on available data	labelled data unlabelled data
	Representation of typological features	Number of typological features not represented in languages often covered by LRs	Language features language taxonomy
P3	Geographical representativeness	Occurrence of entities associated with countries	entities and geographical connection languages and geographical connection
P4	Coverage by resources	Resource containing data in respective language	languages covered by selected resources
P7	Digital Language Equality	LRs contained in ELG (Including LTs and Contextual Factors)	Resource type Subclass Linguality type Media type Annotation type Domain Conditions of use
P8	Representation of writing systems	Scripts represented in model vocabularies	proportional share of words in script

Table 6: Conceptualisation of approaches covering LRs

ID	Concept	Conceptualisation	Values Used
P1	LT performance	error rates	Reuse of error rates from Artetxe and Schwenk (2019)
P2	Utility	sum of proportions of language performance to theoretical maximum performance per task	actual language performance theoretical maximum performance
P3	Model performance	comparison of model accuracy on question-answering test dataset	f1-scores
P4	Platform interface availability	languages covered by platform	languages covered by platform
	Model coverage	languages covered by model	languages covered by model
P5	Digital language support	support of languages by specific software products covering digital language support categories	number of tools digital language support categories
P6	Diversity	Reuse of conceptualisation from paper 2	
	Equity	Gini-coefficient for LT performance	LT Performance
	Inclusion	model efficiency	Throughput (= number of instances it can process per second on a CPU) Memory saved (= size of model as a measure how expensive a model is to use in practise) benefit (= model performance)
P7	Digital Language Equality	LTs contained in ELG (Including LRs and Contextual Factors)	Language dependence Input type Output type Function type Domain Conditions of use

Table 7: Conceptualisation of approaches covering LTs

ID	Concept	Conceptualisation	Values Used
P1	Language diversity and inclusion in conferences	Language occurrence entropy	Number of conference papers mentioning respective language year
	Closeness	Prediction of entity embeddings based on the context	entities: author, language, conference
P2	demographic demand	Size of language community	Number of speakers
	linguistic demand	Always highest value	1
	Reputation gain in research	number of citations	Number of citations
	scientific production	Publications in NLP	Number of NLP conference papers
	economic gain	GDP	approximate GDP of number of users
P3	size of community	population of country	population of country
	economic gain	GDP	GDP of country GDP per capita of country
	Size of country	landmass	landmass
	Distance between user and dataset	geographical distance between entities referenced in dataset and respective language community	location of entities location of language community
P4	economic gain	GDP	GDP of country
	size of language community	population size	number of speakers
P7	Digital Language Equality	Contextual Factors (Including LRTs/ Technological Factors):	
		Size of economy	Size of economy, Size of the ICT sector
		Education	Students in LT/language, Inclusion in education
		Industry	Companies developing LTs
		Law	Legal status and legal protection
		Online	Wikipedia pages
		R & D & I	Innovation Capacity, Number of papers
		Society	Size of language community, Usage of social media
		Technology	Digital connectivity, internet access

Table 8: Conceptualisation of socio-economic indicators

ID	Factor 1	Factor 2	Method
P1	Error rates	Representation of typological features	Mapping features not included in datasets and their error rates
P2	Number of normalised citations	Number of languages covered	correlation calculated based on Bayesian generalised additive mixed effects models
	GDP	Numbers of papers published	regression calculated based on Bayesian generalised additive mixed effects models
	Number of speakers	Numbers of papers published	regression calculated based on Bayesian generalised additive mixed effects models
P3	geographical distribution	country population	Spearman's rank correlation coefficient
	geographical distribution	GDP	Spearman's rank correlation coefficient
	geographical distribution	GDP per capita	Spearman's rank correlation coefficient
	geographical distribution	land mass	Spearman's rank correlation coefficient
	geographical distribution	geographical distance	Spearman's rank correlation coefficient
P4	Wikipedia coverage	GDP	Pearson correlation
	Data availability	Geographical location	Count & Mapping
	Data availability	Language Family	Count & Mapping
	Data availability	language class based on size and vitality	Count & Mapping
	Language model coverage	Geographical location	Count & Mapping
	Language model coverage	Language Family	Count & Mapping
	Language model coverage	Language class based on size and vitality	Count & Mapping
	Platform interface availability	Geographical location	Count & Mapping
	Platform interface availability	Language Family	Count & Mapping
	Platform interface availability	Language class based on size and vitality	Count & Mapping
	language class	Number of papers published	calculation of proportional share in sample
P8	Number of speakers	Number of papers published	Calculation of number of papers per million speakers
	Per capita	Number of papers published	Calculation of the highest paper count per capita

Table 9: Co-occurrence of two factors

ID	Resulting Value	Combination Method	Values Used
P1	Language occurrence entropy	Calculation of probability distribution of papers mentioning the same language Calculation of entropy Calculation of a class-wise mean reciprocal Rank which orders the languages based on their frequency of being mentioned in a conference	Number of papers mentioning the language per conference Publication year of papers taxonomy of languages based on available data
	Closeness of entities	Entity Embedding Analysis: Creation of word vectors based on input from papers, Prediction of the context, which is here author, language and conference	Authors, languages and conferences per paper Title and abstract of papers
P2	Degree to which the global demand is met by available LT	Calculation of sum(demand per language x utility) of LT areas	demand per language utility per language
P3	geographical representativeness of NLP datasets	entity recognition and linking creation of dataset-country maps Calculation of percentage of all entities associated with the single countries Calculation of number of countries not represented in the dataset	entities geographical association of entities countries in which the language is spoken
P5	digital language support	Mokken Scale Analysis: Scaling the coverage of the languages per DLS category	top tools per DLS category Languages covered by tools
P6	Diversity	Reuse of utility metrics from paper 2	
	Equity	Calculation of Gini-Coefficient	cumulative task performance per language
	Inclusion	Measures the benefit per unit increase in cost (cost = decrease in throughoutput and memory saved) Calculation of a average benefit-cost ratio for each language per task	Throughput Memory saved benefit
P7	Digital Language Equality	Technological factors: Each language resource, dataset or tool in the ELG Catalogue for a given language obtains a score which corresponds to the sum of the weights of its relevant features; per language all scores are summed up	Tools Services Datasets Language models Computational grammars Lexical and conceptual resources
		Contextual Factors: Weighted mean based on the size of the language communities in different countries, normalisation of values to 0-1, mean of all contextual factors defined as the overall contextual score for a respective language	Annual GDP, GDP per capita; Perc. of the ICT sector in the GDP, ICT service exports in Balance of Payment Total no. of students in relevant area, Percentage of foreigners attaining tertiary education No. of enterprises in the field of I & C Scores extracted to represent the legal status of a language in different countries Number of articles in Wikipedia Innovation Index, Number of papers about the language Total number of speakers; Total number of social media users, Percentage of social media users Perc. of households with broadband
P8	Share of scripts	Calculation of proportional share of words in the specific scripts in the vocabulary of the model	Vocabulary per model scripts

Table 10: Approaches combining several factors

Which Domains, Tasks and Languages are in the Focus of NLP Research on the Languages of Europe?

Diego Alves[1], Marko Tadić[1], Georg Rehm[2]

[1] Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia

[2]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany

diego.alves@uni-saarland.de, marko.tadic@ffzg.hr, georg.rehm@dfki.de

Abstract

This article provides a thorough mapping of NLP and Language Technology research on 39 European languages onto 46 domains. Our analysis is based on almost 50,000 papers published between 2010 and October 2022 in the ACL Anthology. We use a dictionary-based approach to identify 1) languages, 2) domains, and 3) NLP tasks in these papers; the dictionary-based method using exact terms has a precision value of 0.81. Moreover, we identify common mistakes which can be useful to fine-tune the methodology for future work. While we are only able to highlight selected results in this submitted version, the final paper will contain detailed analyses and charts on a per-language basis. We hope that this study can contribute to digital language equality in Europe by providing information to the academic and industrial research community about the opportunities for novel LT/NLP research.

Keywords: European languages, language equality, language technology

1. Introduction

The fields of Natural Language Processing (NLP) and Computational Linguistics (CL) cover a wide range of topics. While CL draws from linguistics and NLP focuses more on computational methods, the terms are often used interchangeably. Language Technology (LT) is a neutral term encompassing both (Agerri et al., 2021). Today, Language Technology is integrated into various aspects of life. Recent progress has been driven by deep-learning models (Otter et al., 2020). Despite these advancements, challenges persist in achieving language equality, as outlined by a recent European Parliament (2018) resolution.

As the performance of machine learning and deep learning methods usually relies on large amounts of data, languages with smaller numbers of speakers are usually disadvantaged and endangered by digital extinction. With regard to Europe, the discrepancy regarding the availability of LT is highlighted by the reports of the European Language Equality (ELE) project describing the current status and challenges regarding LT for 39 European languages (Rehm and Way, 2023).

To promote digital language equality, it is crucial to understand individual language needs and by detecting their spot on the map of the NLP landscape. While initiatives like the European Language Grid (ELG, Rehm, 2023) contribute to the deployment of existing LT, it is also important to identify existing gaps concerning availability of resources designed for low-resourced languages.

We carried out a systematic analysis of current NLP research on Europe's languages with a specific emphasis on domains and NLP tasks. We analysed approx. 50,000 papers published in the ACL Anthology[1] between January 2010 and October 2022. Within this body of research, we identified the language, domain and NLP task a paper reports upon. One motivation behind this landscaping type of research was to identify popular domains and tasks as well as those that are very much under-researched. These gaps could potentially provided opportunities for novel research in the future. Our results provide a general overview into how NLP tools are used in different domains concerning Europe's languages and can be used by researchers to identify opportunities for future developments to promote language equality.

The remainder of this paper is structured as follows. First, Section 2 presents related work. Section 3 describes the methodology for information extraction based on a dictionary-based approach. Section 4 presents an evaluation of the dictionary-based approach and Section 5 highlights the general results regarding NLP tasks, domains, and languages. Section 6 describes a high-level overview of the results regarding the use of NLP tasks in different domains on a per-language basis. Section 7 concludes the paper.

2. Related Work

Current LT literature discusses technologies rather than domain-specific applications. Research papers describe new tools, methods and approaches and handbooks such as Mitkov (2022) provide

[1] https://www.aclweb.org/portal/

Proceedings of the Second TDLE Workshop @LREC-COLING-2024, pages 18–32
25 May, 2024.

an overview of existing areas and resources. Although presenting important findings regarding the status of LT for different languages, surveys such as the ones presented in the language reports of the ELE project (Rehm and Way, 2023) or the META-NET White Papers (Rehm and Uszkoreit, 2012) do not present detailed analyses how tools are deployed in different domains.

A few articles describe the use of LT in specific fields. For example, Osterrieder (2023) present a complete overview of LT in finance. Additionally, several research papers present tools and resources for a particular domain, for example, a chemical tagger (Hawizy et al., 2011).

In Web of Science[2] and Scopus,[3] users can filter for specific domains, making it possible to find NLP articles in these domains. However, it is impossible to generate a complete overview.

This article presents a detailed analysis, using a dictionary-based approach, of the development of LT by the NLP community concerning different domains and languages with a focus on 39 European languages. The analysis is based on the ACL Anthology, which is why research published elsewhere (or not at all) is excluded. We are aware of the fact that supervised machine learning outperforms dictionary-based classification (Kroon et al., 2022), which is our approach, however, due to the large number of domains, tasks and languages, and because of the lack of annotated data to train models regarding this specific task, we decided to use the dictionary-based approach to establish this groundwork that can be the base for more advanced studies in the future. Our work is based on the EuLTDom project report[4] with evaluation results regarding the dictionary based approach and further analysis.

3. Data and Methodology

The ACL Anthology is an important Open Access archive with Open Source components for the NLP community. It is the main source of CL and NLP scientific literature and offers both text and faceted search features of the indexed papers and also author-specific pages. It allows open access to the proceedings of all ACL-sponsored conferences and journal articles, also hosting literature from sister organisations and their national venues (Gildea et al., 2018).

We used the ACL Anthology Corpus repository (Rohatgi, 2022) which provides PDF files, full-text, references, and other details extracted from the PDF files using GROBID.[5] This repository contains 80,013 articles and posters from 1957 to October 2022. We analyse a subset of this data, a total of 49,466 articles published between January 2010 and October 2022.

To understand the use of LT in different domains for different languages, we implemented a dictionary-based approach. We count the number of research papers in the subset of the ACL Anthology Corpus (see above) that mention the defined terms concerning languages, domains, and NLP tasks at least twice. In the first step of the analysis we look at each of these three dimensions separately, while in the second step, we count the number of articles that mention the domain/language/NLP task triple to identify how different domains use specific LT for each language. The lists of languages, domains, and NLP tasks to be used in the dictionary-based approach were defined in a way to avoid certain possible biases and are described in the following subsections.

3.1. Languages

We analyse the texts of papers written in English from the ACL Anthology for those 39 European languages for which an ELE Language Report exists.[6] For the languages that have more than one name (i. e., Catalan/Valencian and Romanian/Moldavian/Moldovan), while searching for the number of mentions in each paper, all possible names were considered. The complete list of languages is presented in Appendix A.

3.2. Domains

The list of relevant domains was defined following the Fields of Research and Development classification (FORD), which is the basis of the Frascati Manual (2015). This approach is closely related to and consistent with UNESCO's Recommendation concerning the International Standardisation of Statistics on Science and Technology (Unesco, 1978). The FORD classification provides a more complete set of domains when compared with the list considered in the ELE language reports (e. g., Melero et al., 2022). Although similar, the ELE list is shorter and includes general terms such as "Technology", "Science", and "Innovation".

We customised the FORD classification as follows: 1. the list was completed with ELE fields not present in the FORD one, excluding generic terms previously mentioned; 2. the FORD elements that correspond to the label "Other" (e. g., "Other natural sciences") were excluded; 3. the Health and Media domains were excluded because they were the focus of a concurrent study; and 4. terms such

[2] https://www.webofknowledge.com

[3] https://www.scopus.com

[4] https://tinyurl.com/356xt6b5

[5] https://github.com/kermitt2/grobid

[6] https://european-language-equality.eu

as “Economic geography” and “Social Geography” were replaced with “Geography”.

Our final classification contains 46 domains, which are clustered into five broader classes as presented in Appendix B.

3.3. NLP Tasks

The list of NLP tasks includes the information provided by Mitkov (2022) complemented with tasks found in the Wikipedia article on NLP[7] and two other tasks mentioned on the IBM website[8]: “Spam detection” and “Virtual agents and chatbots”. While Mitkov (2022) divides NLP/LT into two classes (i. e., tasks and applications), Wikipedia has a more detailed classification. The complete list contains 51 tasks divided into seven classes and is fully displayed in Appendix C.

3.4. Text Processing

We first attempted to analyse each of the three dimensions separately and analysed if every term listed above appeared at least twice in each individual article within the collection, using the Python Regular Expression operations library[9]. Preliminary tests, with a qualitative evaluation, showed that the main differences in the overall comparison of the languages, NLP tasks, and domain did not change using the threshold of two, five, or 10 occurrences. However, the total number of articles classified according to them is reduced when the threshold was increased. Thus, to improve the recall, we decided to keep the rule of minimum of two occurrences per article.

Texts and query terms were converted to lowercase for uniformity and, for each text available in the ACL Anthology Corpus, its full text (i. e., from abstract to conclusion) was analysed. The idea of considering only those articles where each term is mentioned at least twice is due to the fact that a certain term may be mentioned in the article even if the text is not exactly focusing on this term specifically but only mentioning it in passing.

Our goal was to examine how the NLP community has developed LT for different domains and languages. Most articles describe tools and other resources, thus the main topic here are neither the languages nor the domains. An article or poster is relevant for a certain language and domain if it clearly describes a concrete resource or application of an NLP task.

We also examined languages separately. First, the articles were analysed to check if a language was mentioned at least twice. Then, we checked if the article mentioned each domain/NLP task pair. With these results, heat maps were generated using the statistical data visualization Python library Seaborn.[10] The query concerning domains and NLP tasks was performed with the identified terms including synonyms and alternative orthographic forms. Special attention was required for some terms in the list of domains that may be used in different contexts, not necessarily referring to the domain, for example, “literature” and “history”. In these cases, besides the noun, the respective adjective also had to be mentioned at least once for the article to be counted (e. g., literature and literary; history and historical). Special treatment also had to be implemented for the domain “Arts” (or “Art”). As many papers contain the term “state-of-the-art” or its variations, a way to verify the context of the regular expression match was implemented to guarantee that these phrases are not counted.

The code and the results are available in the project’s GitHub repository.[11]

4. Evaluation

The dictionary approach relies on counting the occurrences of specific terms. This approach has inherent weaknesses when compared to methods for topic classification based on supervised machine learning and embeddings (Kroon et al., 2022). Considering the lack of explicitly annotated training data as well as overall resource restrictions, we opted for the keyword-based approach. To validate the efficiency of the dictionary-based approach, we decided to conduct an evaluation focusing on its precision.

In total, 49,466 articles from the ACL Anthology Corpus were analysed. In order to have a result with a confidence level of 95% and a 5% margin of error, the set to be analysed for the evaluation must contain a minimum of 382 articles.[12] As three dimensions are examined, we decided to select a sample of 130 texts for each one, a total of 390.

We randomly selected texts from the categorised ones, guaranteeing that the evaluation data has at least two representative texts for each of the terms considered as matches.[13] Furthermore, we verified that the articles cover all the years of the ACL Anthology we looked at (January 2010 to October 2022).

For each article of the evaluation data, we checked if the term found in the article really corresponded to a language, domain or NLP task name.

[7] https://en.wikipedia.org/wiki/Natural_language_processing

[8] https://www.ibm.com/topics/natural-language-processing

[9] https://docs.python.org/3/library/re.html

[10] https://seaborn.pydata.org

[11] https://github.com/dfvalio/EuLTDom2023

[12] Value determined using Calculator.net.

[13] Those terms contained in the lists that could not be found in the data set were omitted in the evaluation.

We considered it as a true positive if the term was used in the context of a resource (i. e., tool, model, data set, etc.) or a real application (e. g., evaluation of existing tools, surveys, etc.). False positives corresponded to the cases where the term was used in the context of a future research direction or for incorrect matches due to problems with regular expressions. Table 1 presents the results for each dimension and the overall precision.

	Precision
Languages	0.86
Domains	0.74
NLP tasks	0.84
Overall	0.81

Table 1: Precision for each dimension and overall

With regard to our overall objective, we consider the results of the dictionary-based method satisfactory. In comparison with the analysis conducted by Kroon et al. (2022), our results are comparable to the best machine-learning techniques. The domain dimension is the most problematic one, with a precision of less than 0.75. Below we present a qualitative analysis of the encountered errors.

We would like to stress that only precision was considered in this evaluation. It does not provide information on articles that present contributions regarding the three defined dimensions using different terms than the one contained in the lists. It seems plausible to imagine that the domain dimension should be the one with the lowest recall as the text may describe an application in a certain domain using a different name.

4.1. Languages

The errors observed when languages were analysed correspond mainly to the sections of the papers that deal with related or future work (44.1% of the errors). We also encountered other types of false positives: 1. The language is present in the name of an Organisation (e. g., "Norwegian University of Science and Technology"); 2. the language is mentioned in the context of a translation; 3. the term is mentioned as being excluded from a study; and 4. the term refers to a nationality, not the language itself.

From all the terms used in the regular expressions, only "Romani" was problematic as it was considered a match with words such as "Romanian" and "Romanized". Thus, for this specific language, our results should be handled with care.

We did not consider abbreviations or language codes such as ISO 639-3. Thus, if a language is only mentioned using its name once and then using an abbreviation, it was not counted as a match.

4.2. Domains

Regarding domains, the most frequent error corresponds to using the term in example sentences (32.2% of the false positives), e. g., "Civil engineering" (presented as an example of a compound).

The other most common error (31.3%) is related to the mention of the term in organisation names. It is present mostly in the Acknowledgement section or in the main text when departments are referred to. The term "Government" was specifically problematic as it was mentioned in copyright-related parts of certain articles (e. g., "The U.S. Government"). Besides "Government", two other terms created errors repeatedly: "History" and "Arts". The first one was sometimes used in contexts such as "history-dependent", the second one was considered a match with words such as "parts" and "parts-of-speech".

The analysis of false positives concerning the three dimensions shows that some errors are recurrent and, thus, can be easily corrected. In the case of terms appearing in related or future work, a condition can be established to guarantee that the term should not be considered if it appears only in these specific sections. Concerning problematic terms, more precise rules could also be defined to exclude erroneous matches.

4.3. NLP Tasks

For NLP tasks, most false positives were linked to using the terms in related or future work sections (57.3%). In a few cases, the term was mentioned as a task that was, however, not used in the paper, for example, when it is proposed as an alternative way to process the data.

Regarding problematic terms, we encountered errors relative to the acronyms "OCR" and "QA". The first was identified in words such as "democratic", and the second in Arabic words written with the Latin script (e. g., "qarAr" and "qAmato"). Moreover, "parsing" is the term used for constituency and dependency parsing. Nevertheless, when used in this analysis, it also matches with "Semantic Parsing" which is a specific term in the list of the NLP tasks.

5. Results

Next up, we present overviews of the three separate analyses concerning languages, domains, and NLP tasks.

5.1. Languages

Nearly all languages were found in the data set, the only language which does not appear in our data is Tornedalian. Of the 49,466 texts in the ACL

Anthology Corpus (from 2010 to 2022), 45,737 (92.5%) mention at least twice one of the languages from our set. However, they are not distributed homogeneously (see Figure 1), mirroring the findings of others (Gaspari et al., 2023) and also indicating a strong digital language inequality.

As expected, the most mentioned language is English (i. e., more than 20,000 articles), followed by German, French, and Spanish (more than 3,000 articles each). These results are compatible with similar studies such as Joshi et al. (2020) who present an analysis in terms of entropy of the LT disparity between languages using an older version of the ACL Anthology. Italian, Czech, and Portuguese have 1,000 to 1,500 articles each, and the vast majority of languages are mentioned in a number of articles between 100 to 1,000. Languages with this level of development benefit from existing resources to improve the status of their technologies by adapting tools already available for more resourced languages.

The languages with the smallest representation in our data set (less than 100 articles each) are Galician, Welsh, Maltese, Bosnian, Faroese, Saami, Karelian, Yiddish, Luxembourgish, and Tornedalian. These languages seem to be the most endangered ones regarding digital language extinction, thus requiring more attention from the NLP community.

These are general numbers concerning the ACL Anthology. (Joshi et al., 2020) show that conferences such as LREC tend to have more linguistic diversity than others. The dominance of English is also favoured by the fact that, usually, NLP resources are developed for this language and then deployed to others, thus, English results are also presented as a baseline.

We do not consider conferences that are not part of the ACL Anthology. Thus, the bigger picture that emerges out of this survey does not correspond precisely to the LT reality of each language. For example, the ACL Anthology does not include the proceedings of the Baltic HLT conferences, which focus on the Baltic languages.

5.2. Domains

Only 6,179 ACL papers (12.5% of the total) explicitly mention at least one of the terms from the list of domains. This may be explained by the fact that the focus of many articles is on the development of the tools and resources themselves, and not on their applications in specific areas. Furthermore, it is possible that some papers may have certain domains in mind but not refer to them explicitly. The complete list of domains and the respective number of mentions is presented in Appendix D.

"Linguistics" is the most cited domain which is an expected result as our data concerns work published in Computational Linguistics conferences. However, "Computer Science" is not so prominent, even though ACL papers also deal with this domain. The top ten most mentioned terms are from the Social Sciences and Humanities and the arts (varying from 2,783 to 351 articles). The first domain from a different class is Biological sciences, followed by other Natural Sciences domains such as "Physics", "Chemistry", and "Mathematics". Engineering and technology is the class of domains with the least number of articles.

Figure 2 shows the distribution of certain classes. Social sciences and Humanities and the arts correspond to 89.9% of the mentions. The dominance of the class Humanities and the arts is partially explained by the elevated number of mentions of the term "Linguistics" and the bias identified in the search for the term "Arts" in the texts.

The following domains were never mentioned: "Agricultural biotechnology", "Veterinary", "Animal and dairy science", "Industrial biotechnology", "Environmental Biotechnology", "Environmental engineering", "Materials engineering", and "Electronic engineering". This does not mean that LT is not used in these areas but it indicates that LT is not primarily developed specifically for them. Some of these terms are more specific than others, such as "Industrial biotechnology", "Environmental Biotechnology", and "Environmental engineering", therefore, it is possible that papers dealing with them may use other terms in the text.

A more thorough understanding of the current use of LT in Natural Sciences, Engineering and Technology, and Agricultural and Veterinary sciences is necessary for the identification of new opportunities in terms of more directed NLP development for these fields.

5.3. NLP Tasks

In total, 32,154 (65.0%) articles mention one of the NLP tasks at least twice. This percentage is higher than the one for domains but smaller than the one for languages. One reason for this may be that the coverage of our list is not sufficient.

We can observe that Machine Translation has been one of the main areas of the NLP community between 2010 and 2022. The number of articles mentioning MT is approximately twice as high as the number concerning the second most frequently mentioned task (Parsing). Question answering is ranked third.

The term "Parsing" encompasses many NLP tasks, thus, it may explain this higher rank. Furthermore, we can observe that tasks that are not higher-level NLP applications such as parsing, word segmentation, part-of-speech tagging, and named-entity recognition are positioned in the top ten of the most frequent ones. This can be due to

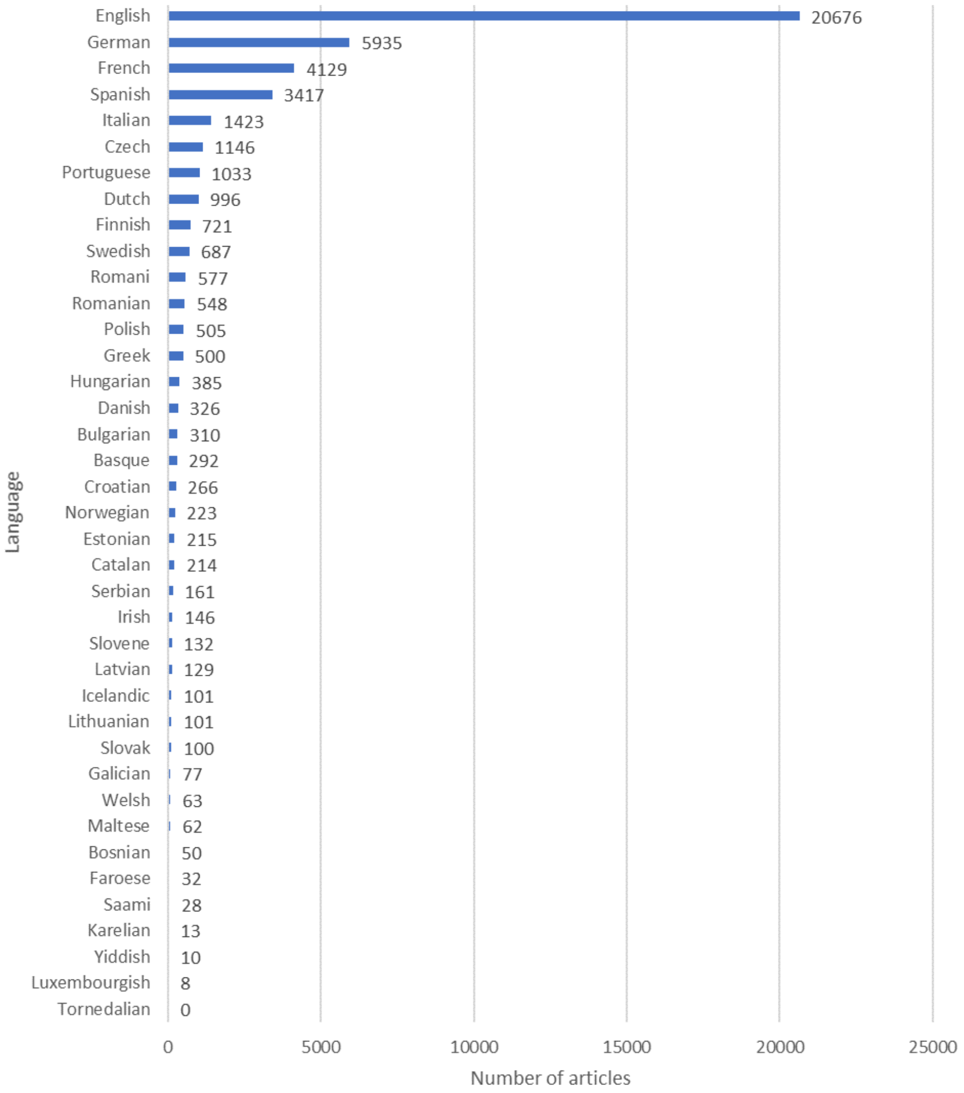


Figure 1: Mentions of European languages in the ACL Anthology (2010 until October 2022).

the fact that these tasks are part of more complex LT, being integrated into pipelines.

Of the 51 NLP tasks on our list, 39 (76.5%) are mentioned in less than 1,000 articles, thus, presenting a lot of potential for further development, e. g., deployment of existing architectures for languages other than English. Figure 3 presents the distribution of the NLP task classes. Almost half of the mentions correspond to higher-level NLP applications, due mostly to Machine Translation and Question answering.

Tasks with the lowest number of articles correspond to rather vague or very specific terms such as “Document AI” or “Implicit semantic role labeling”. The list of NLP tasks and the respective number of mentions is displayed in Appendix E. It would be useful to check if other names for these tasks are currently used by the NLP community to arrive at a more realistic view.

6. Results per Language

We present a detailed analysis concerning the use of LT in different domains per language (i. e., the number of articles where both domain and NLP task are mentioned at least twice each). The heat maps (x-axis: NLP tasks, y-axis: domains) provide a clear snapshot for each European language, and which can also be used as the basis of comparisons. All heat maps are available in the project’s GitHub repository.[14]

As expected, the languages with more mentions in the ACL Anthology result in more complete heat maps when compared to the languages with less mentions. However, we can clearly observe that not all domains and NLP tasks are not covered in recent research. Figure 4 shows the discrepancy in terms of technologies (i. e., data and tools) for different languages. Maltese is only mentioned in

[14] https://github.com/dfvalio/EuLTDom2023

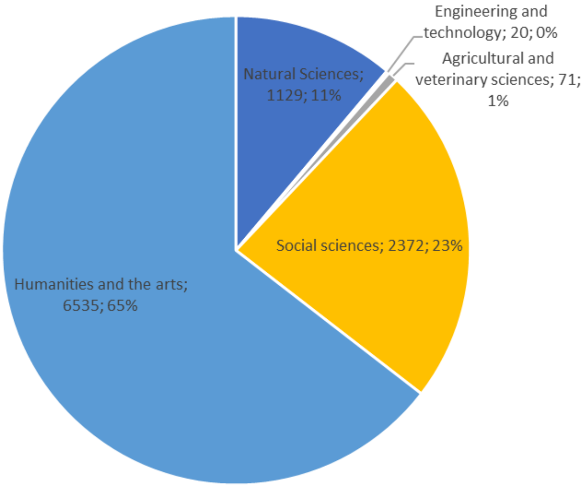


Figure 2: Number of articles presenting research about a certain class of domain.

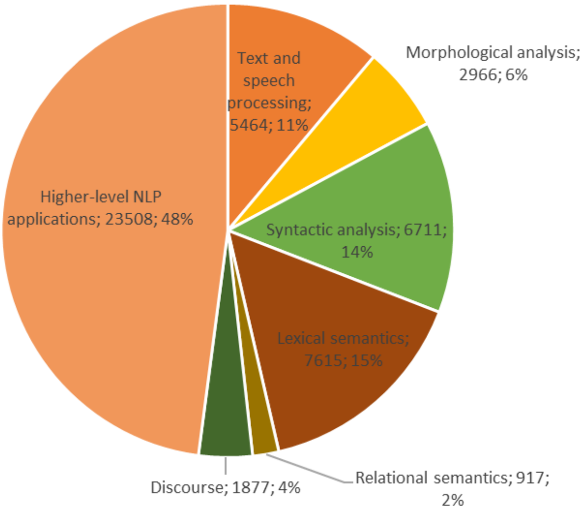


Figure 3: Number of articles presenting research about a certain class of NLP task.

62 articles, thus, its heat map is quite empty. On the other hand, for German (with 5,935 articles), the situation is much better, although still not comparable to the status of the NLP/LT development for English (with 20,676 articles).

Especially the gaps in the heat map of the English and other well-resourced languages can be used to identify new opportunities for the deployment of existing tools and algorithms. Furthermore, it is also possible to check what has been developed for closely related languages, which may facilitate cross-lingual transfer. We also generated a heat map with the overall use of NLP tasks by domains considering all European languages. As expected, "Linguistics" is the domain that has the highest number of associated NLP tasks.

Domains with relatively high usage of different types of LT (i. e., 20 articles or more) are "Arts", "Biological sciences", "Business", "Computer science", "Education", "Ethics", "Finance", "Government", "History", "Law", "Literature", "Physics", "Psychology", "Religion", "Sociology", and "Tourism". On the other hand, some domains use only specific NLP tasks. This is the case for "Ethics" with a predominance of articles on "sentiment analysis", "machine translation", and "question answering".

When examining the analysis regarding domains (except for Linguistics and Computer Science) that are most commonly associated with the top 10 tasks (i. e., tasks with at least 20 articles) we notice many similarities: "Business" and "Education" seem to be the domains that use most of the top 10 tasks. In Appendix F, we present these results in detail. The existence of more than 20 articles describing the use of LT in a specific domain seems to indicate that the specific application is well-developed and, thus, could represent an opportunity for low-resourced languages.

When we focus on the languages with less than 100 articles (excluding Tornedalian which was never mentioned), although the heat maps are very poorly populated, we can identify a few domains and tasks with at least some development. The "Business" and "Education" domains are usually associated with "Machine Translation" and "Natural Language understanding". "Education" is also sometimes mentioned in studies regarding "OCR", "part-of-speech tagging", and "speech-recognition". On the other hand, "History" is often associated with "speech recognition", "named-entity recognition", "machine translation", and "OCR". "Government" appears in association with "Natural Language understanding", "speech recognition", "question answering", and "machine translation", and "Biological Sciences" is usually associated with "information extraction", "named-entity recognition", "parsing", "machine translation", and "question answering".

Thus, it would be useful to check how the NLP data that was used in these papers can be applied to other tasks and deployed in other domains.

7. Conclusions and Future Work

We presented a mapping of NLP and Language Technology research onto 39 European languages and onto 46 domains. The analysis is based on almost 50,000 papers published between January 2010 and October 2022 in the ACL Anthology. The dictionary-based approach we use presents a satisfactory value of precision (i. e., higher than 0.80) when applied to identify how languages, domains, and NLP tasks are mentioned in articles contained in the ACL Anthology. We hope that this study can contribute to digital language equality in Europe by providing valuable information to the academic

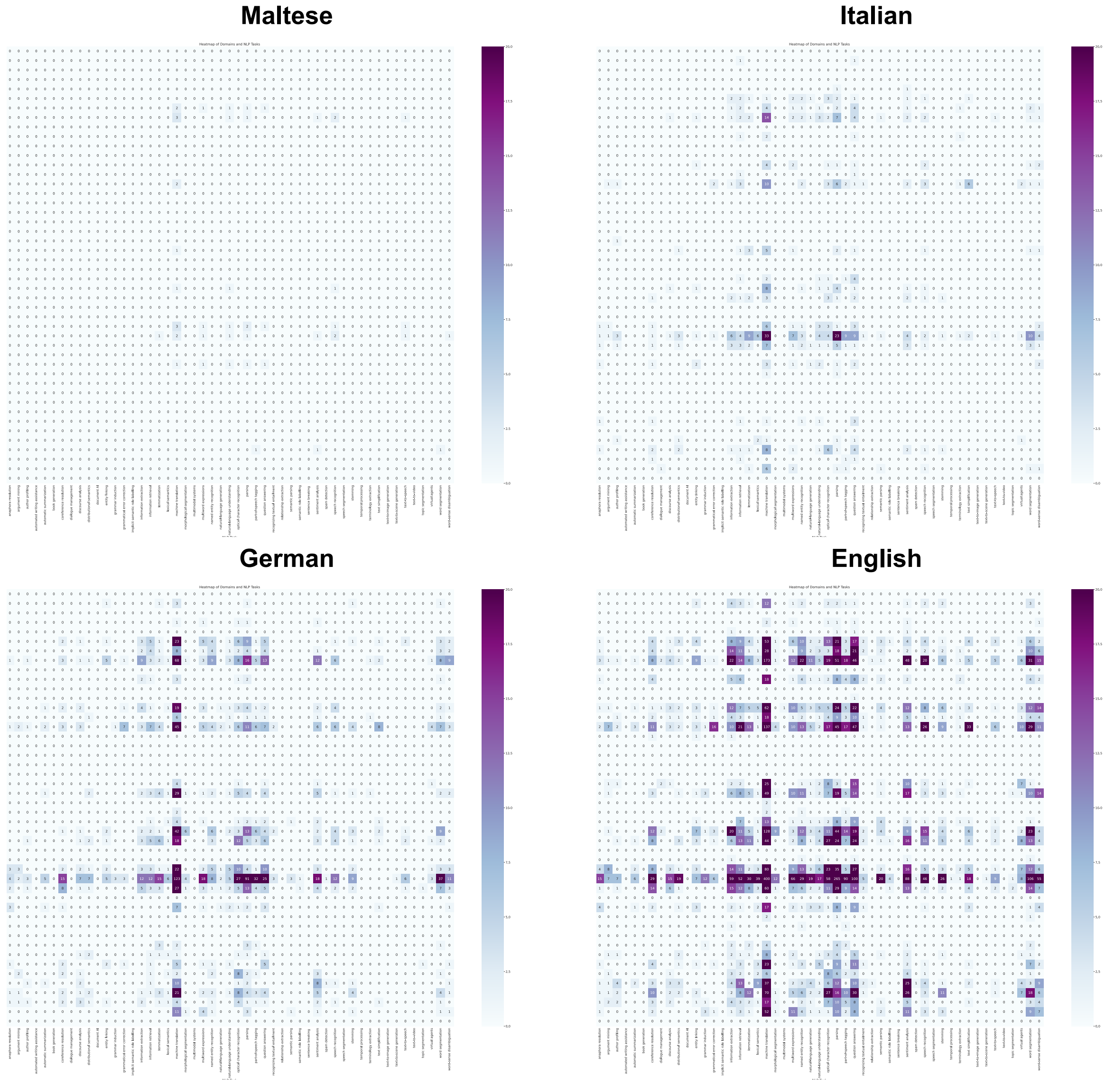


Figure 4: Comparison of four heat maps (Maltese, Italian, German, English).

and industrial research community about the opportunities for novel LT/NLP research.

This study only considers research published in the ACL Anthology. As a potential avenue for future work, a complementary study could be conducted considering other repositories such as Web of Science or Scopus, perhaps also fully structured repositories such as research knowledge graphs but these are too sparsely populated yet. Moreover, as ACL documents are only written in English, it would be useful to complete the analysis with the examination of papers written in the other listed languages. Furthermore, regular updates can be envisioned, for example, with new terms.

8. Ethics Statement

We affirm our commitment to conducting ethical research. We have followed established ethical guidelines and considered the broader societal implications of our work throughout the research process. We also respect copyright laws and intellectual property rights, giving proper attribution to the works of others in our research.

9. Acknowledgements

The European Language Equality project has received funding from the European Union under the grant agreements no. LC-01641480–101018166 (ELE) and no. LC-01884166–101075356 (ELE 2). The research presented in this article received funding from the project ELE 2 through the FSTP funding scheme under the title “European LT Domains 2023 (EuLTDom2023)”.

10. Bibliographical References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, et al. 2021. European Language Equality – Deliverable D1.2 – Report on the state of the art in Language Technology and Language-centric AI, September 2021.

European Parliament. 2018. Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI).

Frascati Manual. 2015. Guidelines for collecting and reporting data on research and experimental development.

Federico Gaspari, Annika Grützner-Zahn, Georg Rehm, Owen Gallagher, Maria Giagkou, Stelios Piperidis, and Andy Way. 2023. Digital Language Equality: Definition, Metric, Dashboard. In Georg Rehm and Andy Way, editors, *European Language Equality: A Strategic Agenda for Digital Language Equality*, Cognitive Technologies, pages 39–73. Springer, Cham, Switzerland.

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL Anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia. Association for Computational Linguistics.

Lezan Hawizy, David M Jessop, Nico Adams, and Peter Murray-Rust. 2011. Chemicaltagger: A tool for semantic text-mining in chemistry. *Journal of cheminformatics*, 3:1–13.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Anne C Kroon, Toni van der Meer, and Rens Vliegenthart. 2022. Beyond counting words: Assessing performance of dictionaries, supervised machine learning, and embeddings in topic and frame classification. *Computational Communication Research*, 4(2):528–570.

Maite Melero, Pablo Peñarrubia, David Cabestany, Blanca C. Figueras, Mar Rodríguez, and Marta Villegas. 2022. European Language Equality – Deliverable D1.32 – Report on the Portuguese Language.

Ruslan Mitkov. 2022. *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Joerg Osterrieder. 2023. A primer on natural language processing for finance. *Available at SSRN 4317320*.

Daniel W Otter, Julian R Medina, and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624.

Georg Rehm, editor. 2023. *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer, Cham, Switzerland.

Georg Rehm and Hans Uszkoreit, editors. 2012. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages. Springer, Heidelberg etc.

Georg Rehm and Andy Way, editors. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer, Cham, Switzerland.

Shaurya Rohatgi. 2022. ACL Anthology Corpus with Full Text. Github. Accessed: 2023-01-10.

Unesco. 1978. Recommendation concerning the international standardization of statistics on science and technology.

A. List of Languages

1. Bulgarian
2. Catalan/Valencian
3. Croatian
4. Czech
5. Danish
6. Dutch
7. English
8. Estonian
9. Finnish
10. French
11. German
12. Greek
13. Hungarian
14. Irish
15. Italian
16. Latvian
17. Lithuanian
18. Maltese
19. Polish
20. Portuguese
21. Romanian/Moldavian/Moldovan
22. Slovak
23. Slovene
24. Spanish
25. Swedish
26. Basque
27. Bosnian
28. Faroese
29. Galician
30. Icelandic
31. Luxembourgish
32. Norwegian
33. Serbian
34. Tornedalian
35. Welsh
36. Karelian
37. Romani
38. Saami
39. Yiddish

B. List of Domains based on FORD and ELE Classifications

Class	Domain
Natural sciences	Mathematics
	Computer and information sciences
	Physics
	Chemistry
	Environmental sciences
	Biological sciences
Engineering and technology	Civil engineering
	Electrical engineering
	Electronic engineering
	Information engineering
	Mechanical engineering
	Chemical engineering
	Materials engineering
	Medical engineering
	Environmental engineering
	Environmental biotechnology
	Industrial biotechnology
	Nano-technology
Agricultural and veterinary sciences	Agriculture
	Forestry
	Fisheries
	Animal and dairy science
	Veterenary science
	Agricultural biotechnology
Social sciences	Psychology
	Cognitive sciences
	Economics
	Business
	Finance
	Tourism
	Education
	Sociology
	Law
	Political Science
	Government
	Geography
Humanities and the arts	History
	Archeology
	Anthropology
	Literature
	Philology
	Linguistics
	Philosophy
	Ethics
	Religion
	Arts

Table 2: List of domains based on FORD and ELE classifications

C. List of NLP Tasks

Class	NLP Task
Text and speech processing	Optical character recognition
	Speech recognition
	Speech segmentation
	Text-to-speech
	Word segmentation (Tokenization)
Morphological analysis	Lemmatization
	Morphological segmentation
	Part-of-speech tagging
	Stemming
Syntactic analysis	Grammar induction
	Sentence breaking
	Parsing
Lexical semantics	Lexical semantics
	Distributional semantics
	Named entity recognition
	Sentiment analysis
	Terminology extraction
	Word-sense disambiguation
	Entity linking
	Multiword Expressions
Relational semantics	Relationship extraction
	Semantic parsing
	Semantic role labelling
Discourse	Coreference resolution
	Discourse analysis
	Implicit semantic role labelling
	Recognizing textual entailment
	Topic segmentation
	Argument mining
	Anaphora resolution
	Temporal processing
Higher-level NLP applications	Automatic summarization
	Grammatical error correction
	Machine translation
	Natural-language understanding
	Natural-language generation
	Book generation
	Document AI
	Dialogue management
	Question answering
	Text-to-image generation
	Text-to-scene generation
	Text-to-video
	Information retrieval
	Information extraction
	Multimodal systems
	Automated writing assistance
	Text simplification
	Author profiling
	Spam detection
	Virtual agents and chatbots

Table 3: List of NLP tasks

D. Number of Articles presenting Research about a certain Domain

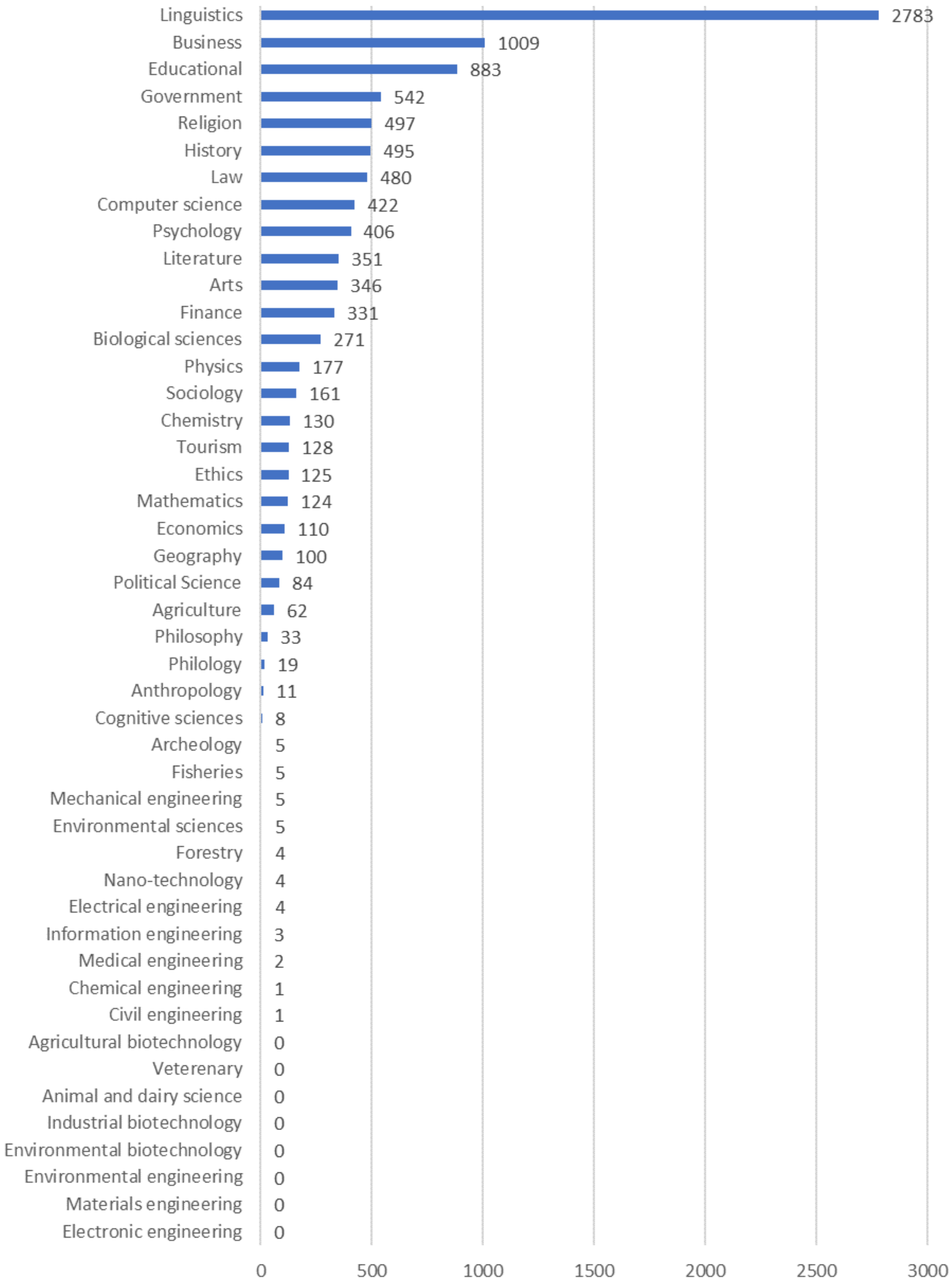


Figure 5: Number of articles presenting research about a certain domain

E. Number of Articles presenting Research about a certain NLP Task

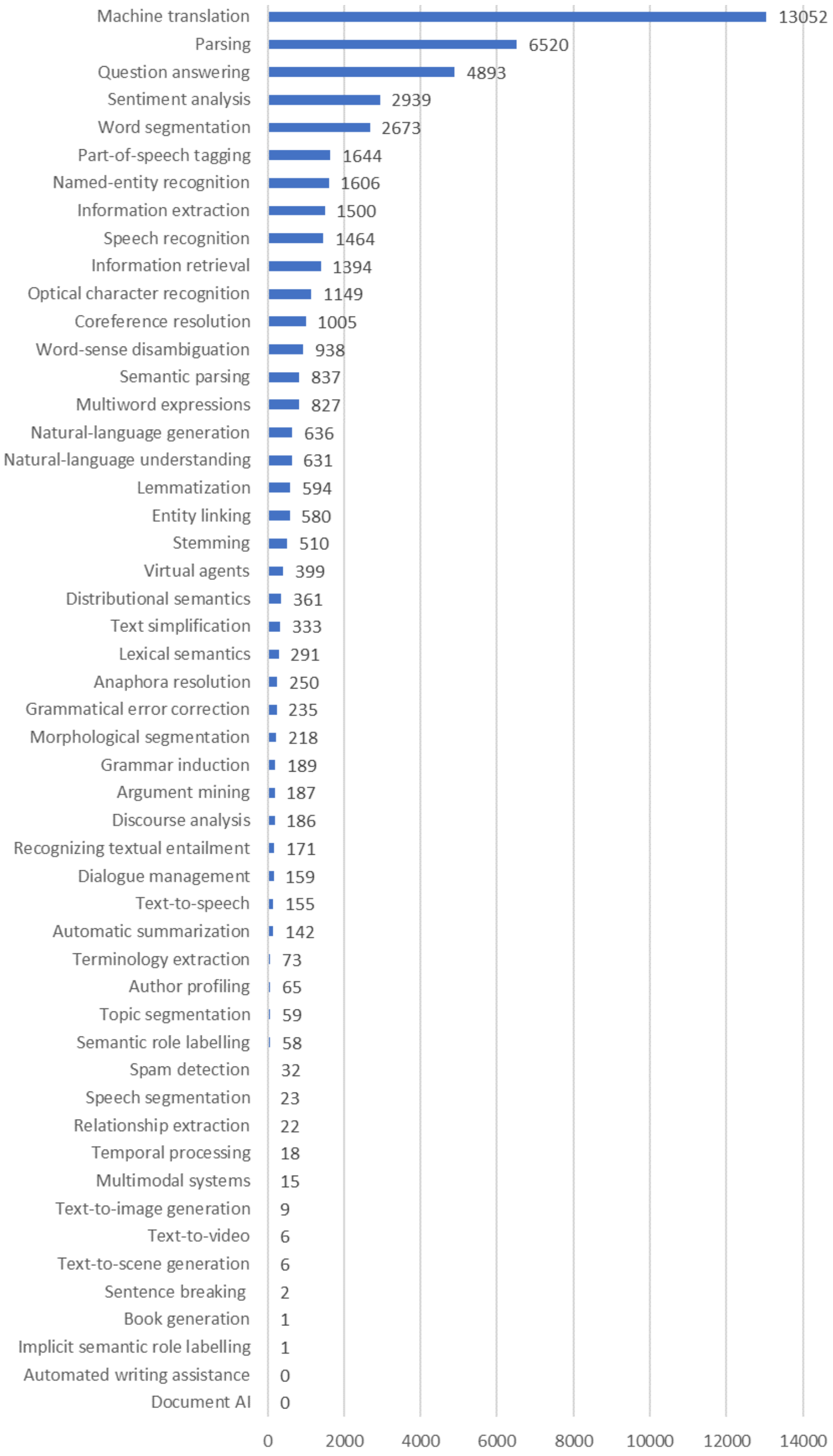


Figure 6: Number of articles presenting research about a certain NLP task

F. Domains which are mostly associated with the Top 10 NLP Tasks

NLP Task	Domains
Machine Translation	Arts, Biological Sciences, Business, Cognitive Sciences, Education, Ethics, Finance, Government, History, Law, Literature, Psychology, Religion, Sociology, and Tourism
Parsing	Arts, Biological Sciences, Business, Education, Finance, Government, History, Law, and Literature
Question Answering	Arts, Biological Sciences, Business, Education, Government, History, Law, and Religion
Sentiment Analysis	Business, Finance, Psychology, and Religion
Word Segmentation	Business, Education, Government, and Religion
Part-of-Speech tagging	Education
Named-entity recognition	Business
Information Extraction	Business and Government
Speech recognition	Business and Education
Information retrieval	Education

Table 4: Domains which are mostly associated with the top 10 NLP tasks

Fine-Tuning Open Access LLMs for High-Precision NLU in Goal-Driven Dialog Systems

Lluís Padró, Roser Saurí

Computer Science Department
Universitat Politècnica de Catalunya
C. Jordi Girona 1-3 – 08034 Barcelona,
Catalonia, Spain
{lluis.padro, roser.sauri}@upc.edu

Abstract

This paper presents a set of experiments on fine-tuning LLMs to produce high-precision semantic representations for the NLU component of a dialog system front-end. The aim of this research is threefold. First, we want to explore the **capabilities of LLMs on real, industry-based use cases** that involve complex data and strict requirements on results. Since the LLM output should usable by the application backend, the produced semantic representation must satisfy strict format and consistency requirements. Second, we also want to assess the **language scalability** of the LLMs in this kind of applications; specifically, whether a multilingual model is able to cast patterns learnt from one language to other ones –with special attention to underresourced languages–, thus reducing required training data and computation costs. Finally, we want to evaluate **the cost-benefit of open-source LLMs**, that is, the feasibility of running this kind of models in machines affordable to small-medium enterprises (SMEs), in order to assess how far this organizations can go without depending on the large players controlling the market, and with a moderate use of computation resources. This work was carried out within an R&D context of assisting a real company in defining its NLU model strategy, and thus the results have a practical, industry-level focus.

Keywords: Large Language Models, Natural Language Understanding, Fine Tuning, NL assistants, Goal-Driven Dialog Systems, LLMs carbon footprint, Underresourced languages

1. Introduction

Many NLP applications demand a Natural Language Understanding (NLU) component able to transform language utterances into structured representations satisfying the requirements of the application backend. Some examples are database natural language interfaces, domotic assistants, voice-activated computer desktop assistants, and, in general, any goal-oriented dialog system beyond mere Q&A or recreational chatbots, aiming at helping the user to accomplish complex goals such as booking a flight or paying taxes. All these applications do not require a plausible textual response, but a highy precise set of complex arguments for the user intent (which taxes should be paid, from which bank account, which light at home should be turned off and when, which file should be moved to what folder and under what name, etc.)

In this study, we delve into a series of experiments on tuning Large Language Models (LLMs) for generating precise semantic representations within a dialog system. The research is structured around three primary objectives:

First, we investigate the potential of LLMs to handle complex, real-world scenarios in the industry. The aim is to ensure that the semantic outputs from the LLMs meet strict standards of format and consistency for seamless integration into application backends.

Secondly, we explore the scalability of LLMs across diverse linguistic landscapes, particularly their ability to support low-resourced languages. We aim to ascertain whether a multilingual model can transfer knowledge from well-resourced languages to those with fewer resources, thereby reducing the need for extensive training data and computational resources, favoring environmental and economic sustainability.

Lastly, a significant portion of our research is dedicated to evaluate the viability of leveraging open-source LLMs in a way that is economically and environmentally sustainable for small to medium-sized enterprises (SMEs). This involves exploring how these companies can use advanced language models without exacerbating environmental impacts or depending on large market-dominating corporations.

Conducted within a R&D framework aimed at assisting a start-up company in formulating its Natural Language Understanding (NLU) model strategy, this investigation offers insights with a practical, industry-oriented focus, highlighting the environmental impact and the challenge of inclusivity for low-resourced languages.

2. Background

Dialog systems, personal assistants, and other applications requiring precise understanding of user commands or queries have become ubiquitous in

Proceedings of the Second TDLE Workshop @LREC-COLING-2024, pages 33–42
25 May, 2024.

various sectors, including healthcare, customer service, and business, among others. NLU is a crucial component in these systems, responsible for transforming unstructured human language into a format that can be understood and processed by the application backend.

The recent launch of Large Language Models (LLMs) such as OpenAI GPT (OpenAI, 2023), Llama (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), GPT-j (Wang and Komatsuzaki, 2021), GPT-neo (Black et al., 2021), Bloom (BigScience Workshop, 2022), or Mistral (Jiang et al., 2023), among others, has opened a large range of possibilities for all NLP applications. LLMs have shown to have powerful language "understanding" capabilities (Goldstein et al., 2023; Olney, 2023; Tsoutsanis and Tsoutsanis, 2024), being able to perform tasks such as entity recognition and classification (NERC), sentiment analysis, paraphrasing, summarization, or translation, with a quality close to human performance. Moreover, these models are able to generate syntactically (and often semantically) correct code in a variety of programming languages.

LLMs have been used as components in traditional NLP pipelines, proving able to perform NERC and relation extraction (Paolini et al., 2021; Ren et al., 2021), Semantic Role Labeling (Paolini et al., 2021), Coreference Resolution (Paolini et al., 2021), Event Extraction (Paolini et al., 2021; Du et al., 2021; Lu et al., 2021; Li et al., 2021), aspect-based sentiment analysis, or slot filling (Athiwaratkun et al., 2020; Rongali et al., 2020; Zhang et al., 2021). See (Min et al., 2023) for a detailed survey on the use of LLMs for NLP tasks.

However, using LLMs to perform partial NLP analysis has the same problems than traditional pipelines. On the one hand, the output is usually produced as annotated text, which requires a postprocessing step to integrate the relevant information. On the other, integrating the results from different stages into a unique semantic representation may lead to inconsistencies when outputs of different models are merged together.

LLMs' natural environment is end-to-end tasks involving natural language in both the input and the output: machine translation, summarization, sentiment analysis, question answering, and, obviously, chatbots. However, the completion-like chatbots that LLMs can successfully produce are far from being able to satisfy the strict formatting and semantic constraints needed by the backend of goal-oriented dialog systems.

Existing research on LLMs has focused either on performing low-level NLP tasks (NERC, coreference resolution, parsing, slot-filling), or on high-level user-oriented language tasks (translation, summarization, information extraction, question answering, etc.), but fewer efforts have been devoted to making LLMs produce actionable output. Noteworthy approaches in this direction include the elaboration of plans to achieve a goal (Huang et al., 2022) or the translation of commands into API calls (Patil et al., 2023). In a line similar to the latter, we aim to use LLMs to produce structured complex semantic representations from text that are suitable to the requirements of an application backend in a real-world industrial scenario.

Although LLMs are able to generate code, and thus they can provide a well-formatted semantic structure for a sentence when requested to do so, the resulting structure will not necessarily match the constraints of the backend, neither the produced representation will be consistent between different requests. Yet, LLMs can be fine-tuned with a reasonable effort to produce, with high precision, a semantic structure matching the specifications of a dialog system or assistant backend. The tuned models (even "light" versions –with about 6B parameters) are able to create correct structures even for very complex utterances where any classical NLU pipeline would fail at some point.

3. Target Application

In this paper, we approach the usability of LLMs at the core of a user interface NLU component for an **office assistant** in charge of automating administrative and management tasks of different complexity degrees, like sending messages via various means (email, SMS, telegram, etc.), scheduling meetings, or managing files.

We focused on the 7 basic *intents* presented in Table 1 (intents i01 to i07). Most of them are instructions for *actions* for the system to perform (e.g., *send an email*), except for intents i01 and i04, which can only be *events* that the system must be sensitive to (e.g. in intent i01 the user cannot command the system to receive a message, but only to be aware of whether a message is received, i.e. as a trigger for some other action). On the other hand, intents i02 and i03 can be both *actions* for the system to perform and *events* to be sensitive to (e.g. the system can be instructed either to send an email message or to monitor whether the user does it herself).

We also experimented with composite intents (intents i08 and i09) and included also a set of random sentences to train the system to disregard unrelated content (intent i00).

4. Data

4.1. Semantic Representation

The JSON schema for the target semantic representation specifies: (a) the appropriate class for

ID	Intent	Process type	Example
Basic Intents			
i00	Intent non-related content	N/A	*They went looking for you several times.* *His brother came looking for us.*
i01	Receiving a message	Event	*An email message from pepe@gmail.com arrives to my outlook account with subject ”invoice” and a PDF attachment.*
i02	Sending messages	Action	*Forward to my personal account any email from Lola arriving to my corporate account.*
		Event	*When anybody from my company replies with an attachment a message from a BSC employee*
i03	Creating calendar events	Action	*Invite Lola from BSC to a meeting called “weekly catchup” for every Wednesday at 9am, in office M3.*
		Event	*If I’m invited to a meeting on weekdays at 6pm on my Google Calendar*
i04	Scheduling a system action	Event	*Wait for 2 hours.*
		Event	*Every day at 3:30pm.*
		Event	*One week later.*
i05	Generating web forms	Action	*Create a form called “Personal information” asking for basic demographic and contact data, and email its URL to Lola.*
i06	Storing files	Action	*Store the new file in my cloud unit to folder MyDocs/customers/*
i07	Adding data to a spreadsheet	Action	*Add the values from the form fields “Name”, “DOB” and “Email” as the last row in spreadsheet users-data.xslx*
Composite Intents			
i08	Combination of Intents i04 & i02	Event + Action	*Every Friday at 3pm, send a message to Lola with the file ”summary.xls” attached and subject ”weekly report”*
i09	Combination of Intents i03 & i02	Event + Action	*If Lola invites me to a meeting on Monday, morning, send her a message via Teams with text ”Sorry, I can not make it”.*

Table 1: Intents targeted by the NLU model

each intent, together with its relevant parameters; (b) the appropriate type for each of these parameters (string, integer, array, object); (c) any constraint on the possible values for these parameters (e.g., an integer value must be within a given range, an object value can only be of a certain class); and (d) the optionality for each parameter. The job of the NLU model is to identify the intent in the user utterance and convert it into a JSON structure compliant with the schema used by the assistant backend. For instance, Figure 1 shows the representation for the following sentence, which belongs to intent i09:

> *If Lola invites me to a meeting on Monday morning in room S1.207, send her a message via Teams with text ”Sorry, I’m booked”*

The semantic representation for this sentence must be an object of class CalendarEventAdded, followed by an object of class Send. The former requires slots process-type and event-object. The latter is in turn instantiated by an object of class CalendarEvent with parameters organizer, attendees, subject, location, start-time (and optionally end-time and duration). Event parameters organizer and attendees are instantiated by objects of class User, and parameter start-time is instantiated by an object of class TxSet. The second Send object also has its own requirements on the expected parameters. Note how the coreference between *her* and *Lola* is resolved setting Lola as the recipient of the TeamsMessage.

Our schema manages 8 classes for modeling actions/events and 19 for entities of different sorts: messages, calendar events, users, forms, files, spreadsheet data, time expressions, etc. Some of those classes have also subclasses (e.g., class Message can be further specified into EmailMessage, SMSMessage, TelegramMessage, TeamsMessage, etc.). Finally, there are also a few classes that represent grammatical aspects of the

```
[{”class”: ”CalendarEventAdded”,
  ”process-type”: ”Event”,
  ”event-object”: {
     ”class”: ”CalendarEvent”,
     ”location”: ”room␣S1.207”,
     ”subject” : ”_unknown”,
     ”attendees”: [
         {”class”: ”User”,
          ”lemma”: ”me”,
          ”org-name”: ”_mine”,
          ”user-id”: ”_unknown”,
          ”user-name”: ”_me”
         }
     ],
     ”organizer”: {
          ”class”: ”User”,
          ”lemma”: ”lola”,
          ”org-name”: ”_unknown”,
          ”user-id”: ”_unknown”,
          ”user-name”: ”Lola”
     },
     ”start-time”: {
          ”class”: ”TxDateTime”,
          ”when”: {
              ”partofday”: ”MORNING”,
              ”weekday”: ”MONDAY”
          }
     }
   }
 },
 {
  ”class”: ”Send”,
  ”process-type”: ”Action”,
  ”sent-object”: [ {
     ”class”: ”TeamsMessage”,
     ”subject”: ”Sorry,␣I 'm␣booked”
     ”recipient”: [
         { ”class”: ”TeamsUser”,
           ”lemma”: ”lola”,
           ”org-name”: ”_unknown”,
           ”user-id”: ”_unknown”,
           ”user-name”: ”Lola”
         }
       ],
     ”sender”: {
         ”class”: ”TeamsUser”,
         ”lemma”: ”.implicit.”,
         ”org-name”: ”_mine”,
         ”user-id”: ”_unknown”,
         ”user-name”: ”_me”
     }
   }
  ]
 }
]
```

Figure 1: JSON representation for instruction: *If Lola invites me to a meeting on Monday morning in room S1.207, send her a message via Teams with text ”Sorry, I'm booked”*

Process Classes			
None	663	ScheduledEvent	3758
ProcessCalEvent	3862	SendMessage	3971
ProcessSpreadsheet	579	SendForm	781
ReceiveMessage	1279	StoreFile	942
Entity Classes			
Attachment (ms)	1969	Form (fo)	781
CalendarEvent (ev)	3862	Message (ms)	6031
DatumField (sp)	300	DateTime (tx)	4195
DatumFrame (sp)	579	Duration (tx)	838
DatumLocation (sp)	579	DurationLen (tx)	838
DatumPosition (sp)	710	Set (tx)	3280
Field (fo)	1143	SetRepeat (tx)	3280
FieldValidation (fo)	1143	TxWhen (tx)	6807
File (fi)	4432	User (us)	19459
FileLocation (fi)	1316		
Classes for grammatical information			
CorefLocat (co)	926	CorefStep (co)	926
CorefObj (co)	1041	Cardinality (ca)	707

Table 2: Parent classes, their frequencies and the kind of information they encode. Legend: (ca) entity cardinality, (co) coreference, (ev) events, (fi) files, (fo) forms, (ms) messages, (sp) spreadsheet data, (tx) time expressions, (us) users.

utterance, namely coreference and entity cardinality. Table 2 presents the frequencies in the dataset of the top classes in the hierarchy.

Note that many intent parameters require a value that is an object, which, in turn, may also have parameters requiring other objects. Thus, the resulting semantic structures can be quite complex, with several nesting levels. As a result, the needed NLU model is multi-level: it not only must discern among the 8 types of basic intents –which would be a simple task for a classical ML classifier– but also to identify the relevant language fragments expressing their parameters, and properly combine the detected objects in compliance with the representation schema constraints.

4.2. Datasets

To train and test our models, we used sets of utterances expressing instructions for the intents presented above, together with their representation in JSON. That data is developed and owned by a startup company dedicated to build NLU-based office assistants. The dataset was created following a semi-automatic process that combines steps of manual curation and AI-based synthetic data augmentation, completed with a final phase for a fully manual check to ensure optimum data quality.

We used data in 3 different languages: Catalan, English, and Spanish. The number of total sentences used for each intent (train and test) is provided in Table 3. In addition, for evaluating the benefits of multilingual vs. monolingual models we also used a smaller subset with only English and Spanish data for intents i01 and i02. See Section

		Basic intents								Composite intents		
Language	Task	i00	i01	i02	i03	i04	i05	i06	i07	i08	i09	Total
Catalan	Train	0	56	71	58	704	37	33	112	46	56	1173
	Test	0	23	36	30	80	12	16	38	31	30	296
English	Train	270	529	597	625	829	308	397	252	507	481	4795
	Test	61	90	87	108	101	70	54	52	102	73	798
Spanish	Train	256	502	542	584	682	305	393	99	484	480	4327
	Test	76	79	107	107	90	49	49	26	99	107	789

Table 3: Number of sentences used for each intent and language (alphabetically sorted).

5.2 for a more detailed explanation.

5. Experiments and Results

We used the dataset described above to carry out different experiments aiming to shed light on three main questions: (1) whether LLMs are able to transform complex user instructions into a JSON semantic representation satisfying strict syntactic and semantic constraints required by the application backend, (2) whether a single multilingual model is better than tuning language-specialized models, (3) whether this multilingual model is able to process new languages with none or small data, and thus easing the support to under-resourced languages, and (4) whether existing open access LLMs are an effective alternative to existing proprietary LLM services, reducing the dependence on large models with large carbon footprint.

In all cases, JSON structures produced by the model where compared to a gold standard and evaluated both at the slot and sentence levels:

- For slots, we compute precision, recall, and F1. A slot is considered to be rightly extracted if it has the right value and it is in the right location inside the JSON structure.
- At sentence level, we compute the percentage of sentences with 100% accuracy (extracted JSON identical to the gold standard) and the percentage of sentences with an unusable output (non-parseable JSON).

The pre-trained LLMs that we analyzed include, on the one hand, five proprietary models owned by OpenAI: **Ada** (350M parameters), **Babbage** (1.3M parameters), **Curie** (6.7B), **Davinci** (175B) and **gpt-3.5-turbo** (20B)[1], and on the other hand, four open access LLMs: **GPT-j** (Wang and Komatsuzaki, 2021), **Falcon** (Almazrouei et al., 2023), **Mistral** (Jiang et al., 2023), and **Flor** (BSC, 2023)[2]. For all open access models, the version around 6-7 billion parameters was used.

5.1. Preliminary explorations

First trials involved using zero-shot and few-shot via prompting, where the model was asked to produce a JSON structure for a sentence after being given a few examples of the expected output.

As can be expected, the complexity of the required output structures and the variety of targeted intents is too wide for the models to grasp with only a few examples, and they behaved creatively with respect to which slots the JSON structure must contain and where to locate them, producing results unusable by the backend component.

Thus, fine-tuning was selected as the strategy to follow, since it allows to provide a larger number of examples and to adjust the model to the specific needs of the application.

Also, initial fine-tuning experiments with OpenAI proprietary models showed that Ada and Babbage had a performance under the minimum usability (under 70% F1 at slot level, under 50% sentences with perfect structure, over 10% sentences with invalid JSON output). Davinci had the best results, followed by Curie. Since the performance difference between them was under two percent points and Curie's economic cost was 10 times smaller, we chose Curie as our reference proprietary model. This allowed us to perform more thorough experimentation and to use larger tuning datasets. Later replacement of Curie and Davinci with gpt-3.5-turbo allowed as to include this newer model in the study.

5.2. Language Scalability

Firstly, we explored whether a multilingual model would be able to cast the patterns learnt from one language to another, or if instead a monolingual model for each target language was better. Table 4 shows the results of tuning different models for each target language versus tuning a single multilingual model. We ran that on the subset of English and Spanish data for intents i01 and i02.[3] We used Curie with 4 epochs, LR multiplier of 0.1, and default batch size (8).

[1] The size of gpt-3.5-turbo is not officially disclosed by OpenAI but it is assumed to be around 20B parameters.

[2] Flor is a Bloom version reinforced with additional Spanish and Catalan data.

[3] Since this piece of work was part of defining a language model strategy for the company, those were the only datasets available at that point.

Secondly, we explored if the inclusion of new languages into the system would benefit from the datasets for the already available languages, or would require extending the dataset. We speculated on such a feasibility due to the proximity of 2 of the languages involved: Spanish (already present in the multilingual model) and Catalan (the new language to be incorporated). We ran an experiment in which the new Catalan dataset was used only for testing, and a second one in which we split that dataset into 75% for training and 25% for testing. Results are presented in Table 5.

	P	R	F1	Perf	Fail
English					
Mono	88.0	88.1	88.1	38.4	0.0
Multi	92.4	92.6	92.5	42.0	0.0
Spanish					
Mono	91.6	91.2	91.4	37.9	0.7
Multi	92.3	93.0	92.6	38.6	0.0

Table 4: Results for fine-tuning with monolingual vs. multilingual models. Using English and Spanish data for intents 01 and 02.

	P	R	F1	Perf	Fail
No Catalan training data					
Catalan	88.0	87.2	87.6	28.0	0.4
English	94.8	94.7	94.8	61.6	0.0
Spanish	97.7	97.7	97.7	78.0	0.0
Some Catalan training data					
Catalan	96.5	96.8	96.6	61.2	0.0
English	95.2	95.4	95.3	63.4	0.1
Spanish	98.3	98.4	98.3	82.8	0.0

Table 5: Results of the multilingual model when fine-tuned only with English and Spanish data (top) or also including Catalan data – 11.4% of the total training dataset (bottom). Data for intents 01 and 02 is used.

5.3. Fine-tuning experiments

To compare proprietary and open access models, we tuned all of them with the same dataset and compared the results. Different combinations of learning rate, epoch number, and batch size were tried to select the best for each model.

Best overall results for each model are shown in Table 6. For each language, slot-wise precision/recall/F1 is reported, as well as percentage of perfect sentences and unusable JSON cases. Best parameterization for Curie is 4 epochs, 0.2 learning rate multiplier, and batch size 8. For the open access models, 2 epochs, 10^{-5} learning rate, and batch size 4 (for GPU memory limitations).

As shown in Table 6, Curie and Mistral obtain the best results. Curie is slightly better in English, and Mistral wins by a narrow margin in Catalan and Spanish. However, the difference is not statistically significant. The other open access models do not achieve the same performance and are all in a similar range of results.

It is noteworthy that despite being much larger than Curie and Davinci –and thus supposed to be a better model– gpt-3.5-turbo obtains results similar to those of the worst open access models. The reason seems to be that gpt-3.5-turbo is too oriented to chat and it tends to get too creative in the produced JSON structures and often fails to respect the output requirements. An second possibility could be that it has a stronger resilience to fine-tuning.

These results prove the ability of fine-tuned LLMs to produce strict constraint-compliant semantic representations of complex user utterances, therefore allowing to be used in an application backend such that for an advanced office assistant.

With regard to the usability of open access models, Table 7 shows performance results of the two best models (Curie and Mistral) detailed at intent level. Intent-wise, the differences are small in most intents, but in some cases (such as intents i03 and i04), there is a significant difference in either one or the other direction. Results for Spanish are better than for the two other languages because the Spanish dataset sentences are less complex.

6. Discussion

Given the results above, we can evaluate our initial questions: What are the capabilities of LLMs on real industry-based use cases requiring high precision NLU (Section 6.1); what is the model scalability to new languages (Section 6.2); and finally, what is the cost-benefit relation of comercial vs. open-source LLMs for SMEs (Section 6.3).

6.1. Usability of LLMs for high-precision NLU

As seen in Table 6, the average percentage of unusable output sentences (%Fail) per language is at most 0.3% for both Curie and Mistral; in fact, it is 0% for most intents in both cases. These are remarkably positive results considering the strict format required by the office assistant backend.

Moreover, the percentage of perfect sentences (%Perf) is reasonably acceptable, as it is around 65% for Catalan and English, and even in a much higher rate for Spanish: 79%. The fact that a sentence is not classified as perfect does not preclude the dialog system to process it. It just means that the JSON structure contains extra slots or misses some expected ones, which can often be managed by the backend or by the users interacting through the system's GUI. The difference between Spanish (79%) and Catalan and English (65-68%) has to do with the nature of the user sentences in our Spanish dataset, which are in general syntactically simpler and more homogeneous than the

	Catalan					English					Spanish				
model	**P**	**R**	**F1**	**Perf**	**Fail**	**P**	**R**	**F1**	**Perf**	**Fail**	**P**	**R**	**F1**	**Perf**	**Fail**
curie	96.3	95.7	96.0	65.2	**0.0**	**95.8**	**95.7**	**95.7**	62.5	**0.3**	97.5	97.2	97.3	78.6	**0.0**
gpt-3.5	90.4	88.0	89.2	58.4	7.8	93.6	90.4	92.0	51.3	4.6	95.5	94.0	94.7	71.2	3.7
Mistral	**96.7**	**96.7**	**96.7**	**67.9**	**0.0**	95.7	95.3	95.5	**65.2**	0.4	**97.6**	**97.7**	**97.7**	**79.3**	0.3
Falcon	89.6	89.5	89.5	39.5	0.3	92.5	92.3	92.4	48.7	0.8	95.4	95.2	95.3	67.3	0.4
GPT-j	90.6	89.9	90.2	42.6	0.3	92.5	92.6	92.6	51.5	0.8	95.2	94.7	94.9	66.5	0.6
Flor	95.6	94.5	95.1	61.8	1.4	94.1	89.3	91.6	53.6	3.5	96.4	90.1	93.1	70.5	4.3

Table 6: Results for different LLMs fine-tuned and evaluated on our target dataset. Curie and gpt-3.5-turbo are openAI proprietary models, and the rest are open access models. Columns P, R, F1 show slot-wise precision, recall and F1. Column *Perf* shows the percentage of senteces with perfect JSON. Column *Fail* shows the percentage of sentences with unusable ill-formed json.

Catalan										
	Curie					**Mistral**				
intent	P	R	F1	Perf	Fail	P	R	F1	Perf	Fail
intent 01	**94.9**	**94.9**	**94.9**	**73.9**	**0.0**	**94.9**	**94.9**	**94.9**	**73.9**	**0.0**
intent 02	95.4	95.4	95.4	66.7	**0.0**	**96.6**	**97.3**	**96.9**	**75.0**	**0.0**
intent 03	**98.0**	**97.0**	**97.5**	**80.0**	**0.0**	96.2	96.0	96.1	66.7	**0.0**
intent 04	93.5	94.2	93.9	70.0	**0.0**	**96.5**	**97.4**	**96.9**	**83.8**	**0.0**
intent 05	96.7	95.3	96.0	**58.1**	**0.0**	**97.7**	**97.3**	**97.5**	54.8	**0.0**
intent 06	97.1	96.3	96.7	43.3	**0.0**	**97.9**	**97.6**	**97.7**	**53.3**	**0.0**
intent 07	**99.6**	**99.6**	**99.6**	**91.7**	**0.0**	97.9	97.9	97.9	75.0	**0.0**
intent 08	98.0	**98.0**	**98.0**	68.8	**0.0**	**98.7**	96.3	97.5	81.2	**0.0**
intent 09	**94.4**	**94.6**	**94.5**	**50.0**	**0.0**	93.6	94.1	93.9	39.5	**0.0**
TOTAL	96.3	95.7	96.0	65.2	**0.0**	**96.7**	**96.7**	**96.7**	**67.9**	**0.0**

English										
	Curie					**Mistral**				
intent	P	R	F1	Perf	Fail	P	R	F1	Perf	Fail
intent 00	88.2	98.4	93.0	98.4	1.6	**100.0**	**100.0**	**100.0**	**100.0**	**0.0**
intent 01	**95.3**	**95.3**	**95.3**	**68.9**	**0.0**	91.1	91.1	91.1	58.9	**0.0**
intent 02	96.2	95.2	95.7	63.2	1.1	**97.7**	**97.7**	**97.7**	**70.1**	**0.0**
intent 03	91.9	92.5	92.2	47.2	**0.0**	**92.7**	**93.5**	**93.1**	**49.1**	**0.0**
intent 04	96.2	96.1	96.2	78.2	**0.0**	**98.2**	**98.7**	**98.4**	**89.1**	**0.0**
intent 05	**96.2**	**95.8**	**96.0**	**46.1**	**0.0**	95.5	94.2	94.9	42.2	2.0
intent 06	**97.9**	**97.7**	**97.8**	52.1	**0.0**	**97.9**	97.6	97.7	**58.9**	**0.0**
intent 07	**98.3**	**97.9**	**98.1**	**72.9**	**0.0**	98.1	96.8	97.5	71.4	1.4
intent 08	93.8	93.6	93.7	50.0	**0.0**	**94.3**	**94.0**	**94.1**	**57.4**	**0.0**
intent 09	96.3	95.8	96.1	55.8	**0.0**	**97.0**	**97.0**	**97.0**	**67.3**	**0.0**
TOTAL	**95.8**	**95.7**	**95.7**	62.5	**0.3**	95.7	95.3	95.5	**65.2**	0.4

Spanish										
	Curie					**Mistral**				
intent	P	R	F1	Perf	Fail	P	R	F1	Perf	Fail
intent i00	**97.4**	**99.1**	**98.3**	**98.7**	**0.0**	90.7	97.8	94.1	97.4	1.3
intent i01	97.9	97.8	97.8	82.3	**0.0**	**98.7**	**98.7**	**98.7**	**84.8**	**0.0**
intent i02	96.3	96.5	96.4	72.0	**0.0**	**96.4**	**97.2**	**96.8**	**75.7**	0.9
intent i03	98.0	**98.0**	**98.0**	**84.1**	**0.0**	**98.1**	97.9	**98.0**	**84.1**	**0.0**
intent i04	**98.7**	**98.5**	**98.6**	90.0	**0.0**	**98.7**	**98.5**	**98.6**	**91.1**	**0.0**
intent i05	97.4	96.9	97.1	**72.7**	**0.0**	**97.5**	**97.4**	**97.5**	70.7	**0.0**
intent i06	97.6	96.9	97.2	**65.4**	**0.0**	**98.0**	**97.5**	**97.8**	62.6	**0.0**
intent i07	98.9	**99.0**	98.9	**87.8**	**0.0**	**99.0**	**99.0**	**99.0**	**87.8**	**0.0**
intent i08	97.7	97.7	97.7	**77.6**	**0.0**	**98.3**	**98.3**	**98.3**	**77.6**	**0.0**
intent i09	92.0	94.5	93.2	34.6	**0.0**	**92.9**	**95.0**	**93.9**	**53.8**	**0.0**
TOTAL	97.5	97.2	97.3	78.6	**0.0**	**97.6**	**97.7**	**97.7**	**79.3**	0.3

Table 7: Results for Curie and Mistral on different languages. Columns P, R, F1 show slot-wise precision, recall and F1. Column *Perf* shows the percentage of senteces with perfect JSON. Column *Fail* shows the percentage of sentences with unusable ill-formed json.

sentences for the other two languages.

A qualitative analysis of the results shows that many of the errors concentrate in slots related to grammatical properties of the input sentences, such as properly identifying entity cardinality (e.g., *sending all the emails* vs. *3 emails* vs. *an email*) or representing coreference information (e.g. *the meeting that Pepa set up in the previous step* or *the email that I just sent*) so that the backend can retrieve the refered entity.

Another source of error involves time expressions (e.g., *within 2 hours*, *every Monday*, or *Wednesday at 15:30h*). Here, the most challenging language for both LLMs is Catalan, which suggests a scarce presence of Catalan time expressions in the pretraining data for both models.

Finally, a further area of error has to do with identifying named entities, both prototypical (people and organization names) and expressions such as names for Teams/Slack channel (e.g. *#dev-team*), folders (e.g., *MyDocuments/Invoices*) and drives (e.g., *the C unit*, *our cloud drive*, etc.).

6.2. Cross-language generalization

A second conclusion from our exploration is that the multilingual model takes advantage of cross-linguistic information, obtaining better results than the models tuned on single languages. Results in Table 4 show that the multilingual model yields an F1 between 1 and 4 points higher than separated monolingual models.

With regard to the extension to new languages, Table 5 (top) shows that the multilingual model delivers quite acceptable results for Catalan data when it is unseen in the fine-tuning data. However, these results for Catalan are still far from the great performance for English (around 94.7%) and especially from Spanish (around 97.7%), from which it should supposedly benefit the most, not to mention the poor score of only 28% Perfect parsed sentences for Catalan, as opposed to scores over 60% for the other two languages.

While it is obvious that the multilingual model is capable to generalize over a third unseen language, the advantage of the multilingual model over monolingual ones seems to be mainly due to the fact that it is fine-tuned with twice as much data. The benefit of larger datasets for fine-tuning can be also attested in the bottom half of Table 5. Note that the results for English and Spanish also slightly improve when an additional small dataset of Catalan training data is incorporated (containing 1173 datapoints, which amounts to only 11.4% of the multilingual training dataset).

6.3. Open-source LLMs as an alternative

Although the performance of Mistral in terms of output quality is comparable, or even slightly better than that of Curie, processing speed is another key issue to be considered when planning to develop an open access LLM-based app or service. Inference on Curie via OpenAI API runs at about 400 tokens/second, and processes one average utterance in 2.5 seconds, including network latency. By contrast, Mistral inference runs locally on a Nvidia RTX-3090 GPU (24Gb) at 17 tokens per second, with an average of 15 seconds per utterance, which is not suitable for real-time applications. However, the same Mistral model quantized to 4-bits, runs on the same RTX-3090 at a speed similar to that offered by OpenAI models, with a very small loss in performance, which definitely opens the door to in-house usage of open access LLMs in applications developed by start-ups and SMEs, enabling not only an economic cost reduction, but also a lighter carbon footprint.

7. Conclusions & Further Work

Our experiments point out that fine-tuned LLMs are a good choice for the NLU component of goal-driven dialog systems. Also, evaluated open access models are able to compete with proprietary models in output quality and speed. However, if a multi-user app or a SaaS application attending many customers simultaneously are envisioned, the cost of dedicated hardware may rise very fast and pay-per-use may be a cheaper option. Technological independence must also be taken into account. Big companies such as openAI not only have a larger carbon footprint, but also take strategic decisions that may negatively impact the performance of applications based on their models[4].

Finally, multilingual models can deal with unseen languages to an acceptable degree, although adding even a small amount of data for the new language contributes to an overall improvement.

Future lines of research include a wider exploration on quantization to increase speed and reduce carbon footprint, while maintaining as much quality as possible, as well as exploring new lighter open access models that may run locally or even in a phone or tablet (Google, 2023; Microsoft, 2023).

On the dataset front, we want to improve the degree of sentence heterogeneity, in particular concerning Spanish. A second line of data improvement has to do with incorporating more sentences displaying those features for which the models tended to performed the worst; in particular, entity cardinality, coreferences, and time expressions of different kinds. Last but not least, we plan to widen the range of supported intents by incorporating more office-related tasks, as well as to integrate a larger variety of languages.

[4] OpenAI recently deprecated Curie leaving gpt-3.5-turbo as the only available alternative, which in our case yields significantly worse results at a higher cost.

Acknowledgements

This research has been funded by the Spanish Science and Innovation Ministry, via grant TED2021-131257B-I00, supported by Next Generation European Funds. We thank the anonymous reviewers for their thorough and insightful comments.

8. Bibliographical References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance. https://huggingface.co/tiiuae/falcon-40b.

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020. Augmented natural language for generative sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 375–385, Online. Association for Computational Linguistics.

BigScience Workshop. 2022. BLOOM (revision 4ab0472).

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. https://doi.org/10.5281/zenodo.5297715.

BSC. 2023. FLOR-6.3B. https://huggingface.co/projecte-aina/FLOR-6.3B. Projecte AINA, Language Technology Unit, Barcelona Supercomputing Center. Barcelona, Spain.

Xinya Du, Alexander Rush, and Claire Cardie. 2021. Template filling with generative transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 909–914, Online. Association for Computational Linguistics.

Alon Goldstein, Miriam Havin, Roi Reichart, and Ariel Goldstein. 2023. Decoding stumpers: Large language models vs. human problem-solvers. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Google. 2023. Gemini. https://deepmind.google/technologies/gemini/.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. https://arxiv.org/abs/2310.06825.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Microsoft. 2023. Phi-2: The surprising power of small language models. https://www.microsoft.com/en-us/research/blog/.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

Andrew M. Olney. 2023. Generating multiple choice questions from a textbook: Llms match human performance on most metrics. *Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation at the AIED'23 Conference*.

OpenAI. 2023. Gpt-4 technical report. https://arxiv.org/abs/2303.08774.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations (ICLR'21)*.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large language model connected with massive apis.

Liliang Ren, Chenkai Sun, Heng Ji, and Julia Hockenmaier. 2021. HySPA: Hybrid span generation for scalable text-to-graph extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4066–4078, Online. Association for Computational Linguistics.

Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. Don't parse, generate! a sequence to sequence architecture for task-oriented semantic parsing. WWW '20, page 2962–2968, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. `https://arxiv.org/abs/2302.13971`.

Panagiotis Tsoutsanis and Aristotelis Tsoutsanis. 2024. Evaluation of large language model performance on the multi-specialty recruitment assessment (msra) exam. *Computers in Biology and Medicine*, 168:107794.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Could We Have Had Better Multilingual LLMs If English Was Not the Central Language?

Ryandito Diandaru[♠], Lucky Susanto[♢], Zilu Tang[♣],
Ayu Purwarianti[♠], Derry Wijaya[♣,♡]
Bandung Institute of Technology[♠], University of Indonesia[♢], Boston University[♣],
Monash University Indonesia[♡]
13519157@std.stei.itb.ac.id, lucky.susanto@ui.ac.id, zilutang@bu.edu, ayu@itb.ac.id,
derry.wijaya@monash.edu

Abstract

Large Language Models (LLMs) demonstrate strong machine translation capabilities on languages they are trained on. However, the impact of factors beyond training data size on translation performance remains a topic of debate, especially concerning languages not directly encountered during training. Our study delves into Llama2's translation capabilities. By modeling a linear relationship between linguistic feature distances and machine translation scores, we ask ourselves if there are potentially better central languages for LLMs other than English. Our experiments show that the 7B Llama2 model yields above 10 BLEU when translating into all languages it has seen, which rarely happens for languages it has not seen. Most translation improvements into unseen languages come from scaling up the model size rather than instruction tuning or increasing shot count. Furthermore, our correlation analysis reveals that syntactic similarity is not the only linguistic factor that strongly correlates with machine translation scores. Interestingly, we discovered that under specific circumstances, some languages (e.g. Swedish, Catalan), despite having significantly less training data, exhibit comparable correlation levels to English. These insights challenge the prevailing landscape of LLMs, suggesting that models centered around languages other than English could provide a more efficient foundation for multilingual applications.

Keywords: Llama2, machine translation, linguistic distances

1. Introduction

Large Language Models (LLMs) have been a popular research topic in Natural Language Processing (NLP) due to their remarkable performance on various tasks including machine translation (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b). Extensive evaluations on machine translation of the popular GPT model family (OpenAI, 2023) have suggested that they can translate high-resource languages (Robinson et al., 2023; Hendy et al., 2023). However, it is rarely the case for low-resource or underrepresented languages (Robinson et al., 2023; Hendy et al., 2023; Stap and Araabi, 2023; Kadaoui et al., 2023).

A straightforward approach for the lack of training data in low-resource translation is to collect more labeled data. However, investing in data creation is nontrivial as it comes with challenges, including the cost of such endeavors. For example, Aji et al. (2022) described the absence of Wikipedia articles on Indonesian regional languages and the challenges of labeled data collection for them, which includes the lack of speakers, the diversity of dialects, and the lack of a writing standard. In addition, training large language models on more data brings environmental consequences (Strubell et al., 2019). In the long run, more training data may require longer GPU compute hours, which will release more greenhouse gas emissions.

Aside from data creation, other techniques are often employed as an alternative. A popular approach for multilingual or low-resource NLP is to leverage other languages to benefit from cross-lingual transfer. These approaches include using them as pivot (Wijaya et al., 2017; Xia et al., 2019), transfer learning (Gu et al., 2018; Nguyen and Chiang, 2017), and joint training (Neubig and Hu, 2018; Johnson et al., 2017). Improvements from such methods indicate a strong influence of the presence of other languages in the training data. Given that including related languages alongside the low-resource language can improve performance (Xia et al., 2019; Poncelas and Effendi, 2022; Gu et al., 2018; Nguyen and Chiang, 2017; Neubig and Hu, 2018; Johnson et al., 2017), it is beneficial to include proximity measurements between these languages on evaluations, which can be done using the vectors from the URIEL database (Littell et al., 2017). The utilization of the URIEL database has made evaluating multiple languages more explainable by leveraging linguistically aware feature vectors from which linguistic distances can be computed. These vectors have been utilized by previous works in various ways including determining which language to use as transfer or pivot language (Lin et al., 2019; Nambi et al., 2023) and measuring language diversity (Ruder et al., 2021).

Proceedings of the Second TDLE Workshop @LREC-COLING-2024, pages 43–52
25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

It has been established that there are benefits to using other languages in the training process. However, multilingual labeled data creation is challenging. In this paper, we aim to provide hints to narrow down future data collection strategies by evaluating an existing LLM family. A constraint in previous studies that assess the GPT model series (Hendy et al., 2023; Robinson et al., 2023) has been the fact that these models are proprietary, closed-source systems that do not disclose information regarding their training data. This presents a challenge as it remains unclear which languages are included in the training of the models. On the other hand, open-source LLMs such as Meta's Llama2 (Touvron et al., 2023b), is more transparent about its training process, including the languages that are included in its training data. This makes the model more suitable as a subject for our evaluation.

In this work, we are evaluating Llama2 (Touvron et al., 2023b) for machine translation to highlight its multilingual capability in languages it has or has not seen during training. We also model a linear relationship (through correlation scores) between the linguistic feature distances and the translation metrics and use these scores as a basis for language importance analysis. The goal of the analysis is to narrow down the data investment effort by shedding light on which language(s) may improve the translation of other languages when included in the training data. An efficient data collection strategy will result in future multilingual LLMs that can be trained and deployed more efficiently, thus promoting sustainability. In summary, our contributions are as follows:

1. We evaluate Llama2 and provide machine translation scores of this model for 41 languages, 15 of which were not seen during its training.
2. We reveal that increasing model parameters is more effective in improving translation over instruction tuning and few-shot learning.
3. Our research reveals that syntactic similarity between languages is not the only linguistic aspect that is strongly linked to machine translation performance. Surprisingly, these strong correlations between linguistic feature distances and machine translation performances extend beyond English and hold true across various languages, therefore opening up the possibility of other better central languages for multilingual LMs

2. Methodology

2.1. Machine Translation Evaluation

We experiment with languages reported in the training data of Llama2 (Touvron et al., 2023b), the list of which and their respective ISO 639-3 codes can be found in Table 1. We refer to this set of languages as **inllama**. We also pick 15 languages not reported in the training data which we will refer to as **outllama**, presented in Table 2. It is important to highlight that languages not explicitly mentioned in Llama2 might still be present in the training data, albeit at a minuscule proportion of less than 0.005% of its training data (Touvron et al., 2023b). Languages in **outllama** cover various language genera and writing systems. The machine translation evaluation is conducted using the FLORES-200 (Guzmán et al., 2019) benchmark as it is available for numerous low-resource languages. We exclude X→English translation directions to mitigate the risk of potential data leakage, given that FLORES-200 uses Wikipedia for its English sentences. We also exclude zero-shot translation as LLMs often get the language wrong in this prompting setup as reported by Robinson et al. (2023). We measure translation quality using machine translation scores. Translation quality is measured with the BLEU score (Papineni et al., 2002) and a model-based machine translation metric (COMET-22 (Rei et al., 2022)) where applicable. COMET-22 is used to compensate for the drawbacks of BLEU and vice-versa.

Language	Genus	BLEU	COMET-22
German (deu)	Germanic	33.68	0.83
Swedish (swe)	Germanic	37.71	0.87
Dutch (nld)	Germanic	27.45	0.84
Norwegian (nor)	Germanic	29.54	0.86
Danish (dan)	Germanic	36.21	0.86
French (fra)	Romance	42.4	0.84
Spanish (spa)	Romance	28.54	0.84
Italian (ita)	Romance	28.78	0.85
Portuguese (por)	Romance	43.21	0.87
Catalan (cat)	Romance	35.92	0.84
Romanian (ron)	Romance	31.58	0.84
Russian (rus)	Slavic	28.21	0.85
Polish (pol)	Slavic	22.34	0.83
Ukrainian (ukr)	Slavic	26.03	0.83
Serbian (srp)	Slavic	23.96	0.81
Czech (ces)	Slavic	24.94	0.82
Bulgarian (bul)	Slavic	29.57	0.83
Croatian (hrv)	Slavic	21.3	0.81
Slovenian (slv)	Slavic	19.51	0.77
Chinese (zho)	Chinese	19.79	0.82
Japanese (jpn)	Japanese	17.02	0.84
Vietnamese (vie)	Vietic	28.77	0.82
Korean (kor)	Korean	11.08	0.78
Indonesian (ind)	Malayo-Sumbawan	31.15	0.86
Finnish (fin)	Finnic	18.08	0.82
Hungarian (hun)	Ugric	18.4	0.78

Table 1: List of **inllama** languages along with their ISO 639-3 codes, genus, and machine translation scores obtained using one-shot Llama2-7B.

We aim to experiment with open-source LLMs

Language	Genus	Writing System
Afrikaans (afr)	Germanic	Latin
Galician (glg)	Romance	Latin
Macedonian (mkd)	Slavic	Cyrillic
Slovak (slk)	Slavic	Latin
Armenian (hye)	Armenian	Armenian
Basque (eus)	Basque	Latin
Georgian (kat)	Kartvelian	Georgian
Icelandic (isl)	Germanic	Latin
Igbo (ibo)	Igboid	Latin
Javanese (jav)	Javanese	Latin
Sinhala (sin)	Indic	Sinhala
Tagalog (tgl)	Greater Central Philippine	Latin
Tamil (tam)	Dravidian	Tamil
Telugu (tel)	Dravidian	Telugu
Welsh (cym)	Celtic	Latin

Table 2: List of **outllama** languages and their ISO 639-3 codes. We also include in this table additional language information retrieved from WALS (Dryer and Haspelmath, 2013)

that replicate proprietary models such as ChatGPT (OpenAI, 2023) in terms of usability and safety. At the time the Llama2 model was released and the experiment design for this paper was constructed, none of the open-source models are suitable substitutes for production models as they may not have been aligned to match human preferences and there may be a performance gap (Touvron et al., 2023b). On account of this, we decided to move forward only with the Llama2 model family. The machine translation evaluation begins with one-shot translations for both languages in **inllama** and **outllama** using the vanilla 7B model. From this experiment, we categorize languages that yield under 10 BLEU as **unlearned** languages[1]. For the **unlearned** languages, we experiment further with model scale, chat version, and adding the shot count to maximize the potential of in-context learning. Our choice of randomly picking 5 shots from the validation set of FLORES-200 is motivated by the experimental setup used by Hendy et al. (2023) which states that increasing beyond 5 shots does not result in meaningful improvement and shows that selected quality shots do not always improve more than 1 BLEU compared to random selections for GPT (text-davinci-003) model. For translation with chat models, we use the prompt by Robinson et al. (2023) which follows the recommendation of Gao et al. (2023) for designing prompts for translation using instruction-tuned models. The prompts used in our experiments are given in Table 3

2.2. Correlation Score Analysis

We consider several language subsets. For every language subset, we calculate the Pearson correlation score between the linguistic similarity scores of each language in the subset to a language in **inllama** and their respective translation scores. We assume that a certain language is important if we observe a positive correlation. For example, consider the language **A** and the language subset {**B**, **C**, **D**, **E**}. When the similarity of **A** with each language in the subset {**B**, **C**, **D**, **E**} and the respective machine translation scores for {**B**, **C**, **D**, **E**} exhibit a positive correlation, i.e. the closer they are to **A** the better their machine translation scores, **A** is deemed as a valuable language and is therefore hypothesized to be more optimal for a central language when developing multilingual language models. **A** is checked for each language in **inllama**. Similarity scores are calculated on five dimensions: GENETIC, GEOGRAPHICAL, INVENTORY, PHONOLOGY, and SYNTACTIC as per the URIEL typological database (Littell et al., 2017). We exclude the FEATURAL distances to focus on each dimension as FEATURAL distances are combinations of all the other feature distances[2]. Language subsets considered are **inllama** languages only, **outllama** languages only, both **inllama** and **outllama** languages, only **Germanic** languages, only **Romance** languages, only **Slavic** languages, and languages belonging to **Other genera**.

3. Results and Analysis

3.1. Machine Translation Evaluation Results

One-shot 7B Llama2 translation results are presented in Table 1 and Table 4. From Table 1, we observe that none of the languages included in **inllama** produce a BLEU score below 10. This suggests that we can reasonably assume that Llama2 is capable of translating into all the languages it has encountered during training. However, many languages in **outllama** yield a BLEU score under 10, this is expected as Llama2 is presumably unfamiliar with these languages. On the other hand, we hypothesize that there are two possibilities for the high-performing **outllama** languages; (1) those languages are indeed included in the training data i.e. included in the 0.005% of the training data, or (2) similar languages in **inllama** indeed boosted their performance.

We move forward with languages in **outllama** that yield a BLEU score below 10 and experiment

[1] Based on "Almost useless" interpretation from https://cloud.google.com/translate/automl/docs/evaluate

[2] For more detailed explanation of these distances, consult https://www.cs.cmu.edu/~dmortens/projects/7_project/

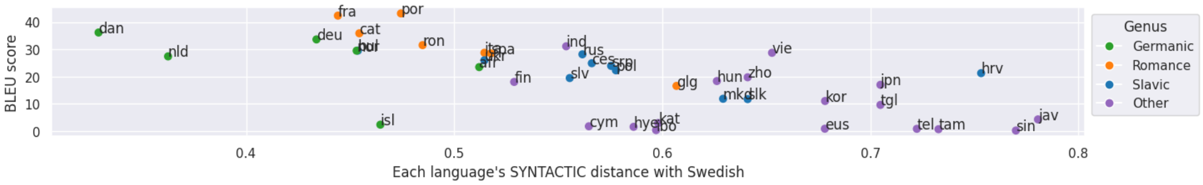


Figure 1: Scatter plot for **inllama** and **outllama** languages against the SYNTACTIC distance to **Swedish**. The correlation score is -0.67 and the p-value is 3.16×10^{-6}. The negative correlation here implies that the smaller the SYNTACTIC distance of a language to Swedish, the better is its MT performance

Model	Prompt
Non-chat	[SRC]: [src-sentence] [TGT]: [tgt-sentence] ... [SRC]: [src-sentence] [TGT]:
Chat	This is an English to [TGT] translation, please provide the [TGT] translation for these sentences: [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] ... Please provide the translation for the following sentence. Do not provide any explanations or text apart from the translation. [SRC]: [src-sentence] [TGT]: [tgt-sentence] [TGT]

Table 3: Prompts used in our experiments to translate languages using the non-chat and chat versions of Llama2

Languages in outllama	BLEU	COMET-22
Afrikaans	23.52	0.74
Galician	16.62	0.76
Macedonian	11.90	0.67
Slovak	11.77	0.68
Armenian	**1.6**	0.31
Basque	**0.91**	0.33
Georgian	**2.99**	0.31
Icelandic	**2.39**	0.35
Igbo	**0.39**	0.41
Javanese	**4.33**	0.59
Sinhala	**0.25**	0.29
Tagalog	**9.65**	0.60
Tamil	**0.73**	0.30
Telugu	**0.87**	0.33
Welsh	**1.8**	0.35

Table 4: Llama2-7B one-shot translation results for languages in **outllama**. Languages with results in boldface are considered **unlearned** languages

with other variations of Llama2. We explore the effect of scale, chat version, and adding shot count and present the results in Table 5. Due to our limited compute resources we excluded the 70B and 70B-chat versions of Llama2.

Scaling up the model enhances translation ability. However, improvements from instruction–tuning and adding shot count remain inconclusive. Results presented in Table 5 demonstrate that the 13B versions of Llama2 outperform the smaller 7B versions for all **unlearned** languages. However, larger models do not seem to yield the same number of gains for every language. In best cases, 13B models increase on average as high as 2.53 BLEU with a standard deviation of 1.64. For instruction-tuning (chat) models, we observed both performance increase and decrease. The best improvements are observed in Igbo and Javanese, which improves as much as 3.16 and 2.87 respectively, and a decrease is observed in Tagalog, which performs worse on chat models with a decrease as severe as 2.64. Adding the shot count generally improves performance although it is less drastic than model scale and instruction-tuning with a mean increase of 0.47 and 0.08 for non-chat and chat Llama-13B respectively. While these model variations appear to enhance Llama2's capacity to translate into some languages greatly, there are languages where the prospects are limited. For instance, for Sinhala and Tamil, scaling up the model/adding shot count/using chat models results in less than 1 BLEU score increase.

3.2. Language Importance Analysis

We use the results from Table 1 and Table 4 for the linguistic proximity analysis. We first retrieve pre-computed distances[3] from the URIEL database and retrieve only the distances between the languages we are translating into and the languages reported in Llama2. Self or identity distances e.g. Igbo-to-Igbo distance are excluded in the Pearson correlation calculation. This correlation analysis aims to model the linear relationship between language proximity and machine translation scores to

[3] http://www.cs.cmu.edu/~aanastas/files/distances.zip

Language	7B 1S	7B 5S	7B-chat 1S	7B-chat 5S	13B 1S	13B 5S	13B-chat 1S	13B-chat 5S
Armenian	1.6	1.95	2.26	2.43	2.52	**3.03**	2.89	**3.03**
Basque	0.91	1.08	2.98	3.11	1.52	1.9	3.72	**3.88**
Georgian	2.99	3.44	4.41	4.7	5.57	**6.19**	5.97	5.88
Icelandic	2.39	3.06	3.9	3.86	4.72	5.21	**5.24**	5.04
Igbo	0.39	0.59	1.77	2.04	0.56	0.67	**3.72**	3.49
Javanese	4.33	3.71	4.94	5.06	3.15	3.76	5.92	**6.63**
Sinhala	0.25	0.38	0.57	0.52	0.48	**0.63**	**0.63**	0.62
Tagalog	9.65	10.98	10.8	10.97	16.1	**16.91**	13.82	14.27
Tamil	0.73	1.01	0.82	1.09	1.79	**2**	1.7	1.56
Telugu	0.87	1.04	1.02	0.86	2.29	**2.45**	1.77	1.68
Welsh	1.8	å 2.37	4.38	3.93	5.68	**6.8**	6.45	6.6

Table 5: BLEU scores with various Llama2 versions and shot count for languages considered **unlearned** by Llama2 (Table 4). **1S**/**5S**=one-shot/five-shot. Best result for each language is bolded.

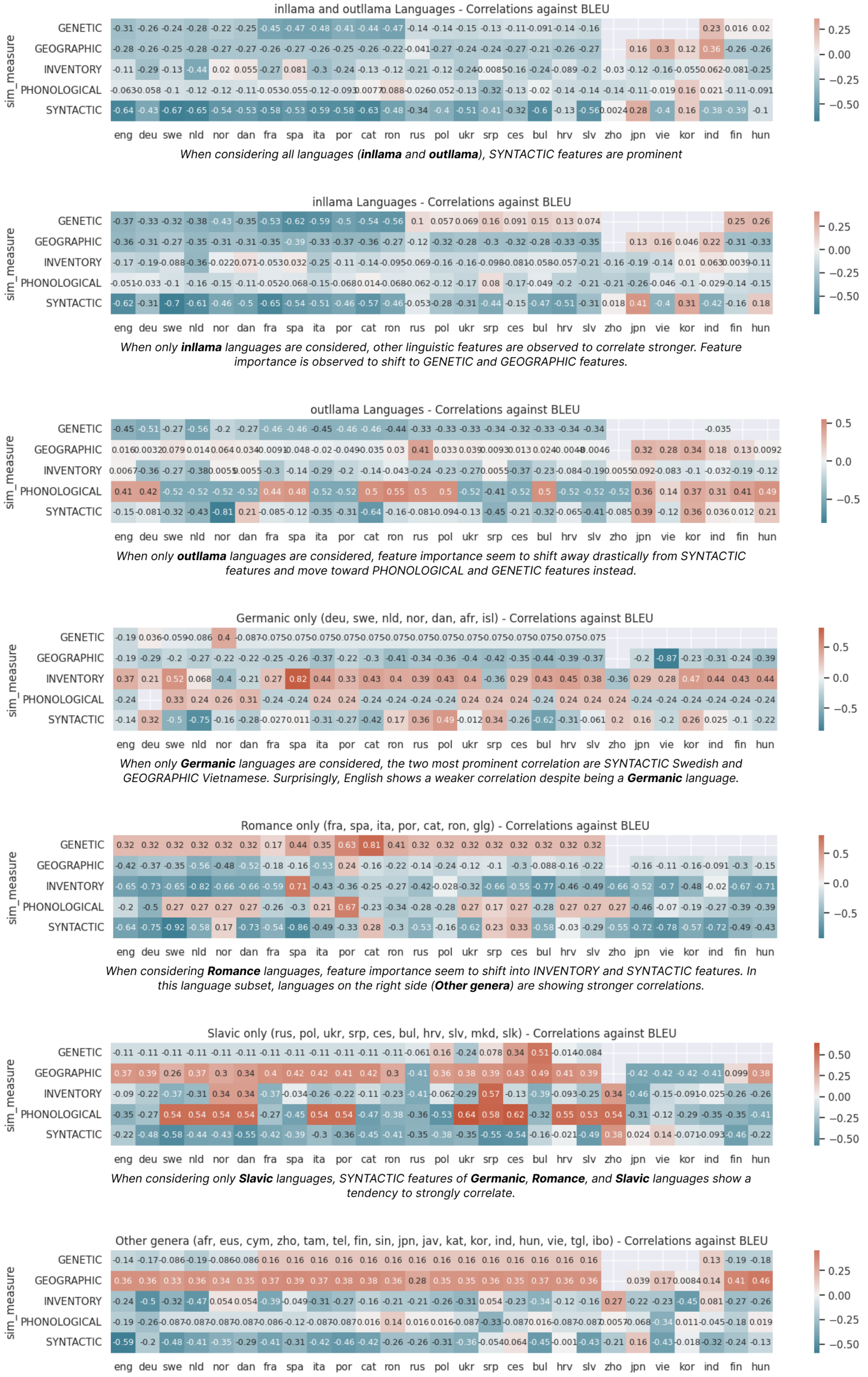
identify languages whose data may be beneficial for multilingual training.

We present our analysis as heatmaps in Figure 2 and 3 for correlations with BLEU and COMET-22 respectively. To help understand where each number came from in the heatmap, a scatter plot visualization for SYNTACTIC distance to Swedish for the combined **inllama** and **outllama** language subset against BLEU scores is presented in Figure 1 as an example. We create several different heatmaps according to the subset considered. It is important to highlight that *distance* is used as a similarity score. Therefore, a negative correlation between linguistic distance and MT scores would imply that the closer (i.e., the *smaller* the linguistic distance) a language is to this language, the higher the MT score is likely to be. In addition, since Wikipedia is a permanent fixture of LLMs' training data, we observe that there is a positive correlation between MT scores and Wikipedia article counts[4], as high as **0.64** using BLEU and **0.55** using COMET-22.

Syntactic similarity may be an important feature, but other linguistic dimensions can be too. When including every language, i.e. the **inllama** and **outllama** subset, BLEU complemented with COMET-22 scores show consistently strong correlations with syntactic features, especially with Germanic and Romance languages. This finding may not be particularly groundbreaking, as we already understand that the languages in **inllama** predominantly belong to these language genus. However, when considering only **outllama** language subset, translation performance seems to have higher correlations (either positive or negative) with GENETIC and PHONOLOGICAL distances. When considering languages in **outllama**, only SYNTACTIC similarities to certain languages e.g. Norwegian and Catalan display a strong correlation with MT performances. Correlation with features other than SYNTACTIC is also observed when considering languages in **other genera**, in which the proximity with the INVENTORY feature of Vietnamese, Dutch, German, and French are shown to correlate with COMET-22 scores.

English is not always the most syntactically important. When considering languages from **other genera** English demonstrates the most substantial syntactic correlation with MT performance, although there are other languages, such as Swedish and Vietnamese, that also display some degree of correlation. However, despite having the highest amount of training data, English is often not in the first place when considering languages by genus (e.g., **Germanic**, **Slavic**, and **Romance**). Similar to when we observe that syntactic proximity to Norwegian and Catalan have higher correlations with MT scores than syntactic proximity to English when considering only **outllama** languages, this phenomenon is accentuated when calculating correlations by genus. Among **Germanic** languages, syntactic proximity to English surprisingly shows little to no correlation with either BLEU or COMET-22 scores. Instead, **Germanic** languages' MT scores appear to correlate more with syntactic proximity to Dutch, Swedish, Catalan, and Bulgarian. This is also observed in **Slavic** languages where the MT scores generally correlate with syntactic proximity to most Germanic and Romance languages *except English*. With **Slavic** languages, syntactic proximity to English has the lowest correlation on BLEU and almost no correlation on COMET-22 scores. Finally, when focusing exclusively on **Romance** languages, it is interesting to observe that proximities to languages situated on the right side of the heatmaps i.e. **other genera**, exhibit higher correlations while they show no correlation when only considering other language subsets (Figure 2 and 3).

[4] Retrieved from https://meta.wikimedia.org/wiki/List_of_Wikipedias_by_language_group on October 2023



*When considering all languages (**inllama** and **outllama**), SYNTACTIC features are prominent*

*When only **inllama** languages are considered, other linguistic features are observed to correlate stronger. Feature importance is observed to shift to GENETIC and GEOGRAPHIC features.*

*When only **outllama** languages are considered, feature importance seem to shift away drastically from SYNTACTIC features and move toward PHONOLOGICAL and GENETIC features instead.*

Germanic only (deu, swe, nld, nor, dan, afr, isl) - Correlations against BLEU
sim_measure
GENETIC -0.19 0.036 -0.059 -0.086 0.4 -0.087 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075 -0.075
GEOGRAPHIC -0.19 -0.29 -0.2 -0.27 -0.22 -0.22 -0.25 -0.26 -0.37 -0.22 -0.3 -0.41 -0.34 -0.36 -0.4 -0.42 -0.35 -0.44 -0.39 -0.37 -0.2 -0.87 -0.23 -0.31 -0.24 -0.39
INVENTORY 0.37 0.21 0.52 0.068 -0.4 -0.21 0.27 0.82 0.44 0.33 0.43 0.4 0.39 0.43 0.4 -0.36 0.29 0.43 0.45 0.38 -0.36 0.29 0.28 0.47 0.44 0.43 0.44
PHONOLOGICAL -0.24 0.33 0.24 0.26 0.31 -0.24 -0.24 0.24 0.24 -0.24 -0.24 -0.24 -0.24 0.24 0.24 0.24 -0.24 0.24 0.24 0.24 -0.24 -0.24 -0.24 -0.24 -0.24 -0.24
SYNTACTIC -0.14 0.32 -0.5 -0.75 -0.16 -0.28 -0.027 0.011 -0.31 -0.27 -0.42 0.17 0.36 0.49 -0.012 0.34 -0.26 -0.62 -0.31 -0.061 0.2 0.16 -0.2 0.26 0.025 -0.1 -0.22
eng deu swe nld nor dan fra spa ita por cat ron rus pol ukr srp ces bul hrv slv zho jpn vie kor ind fin hun
0.5 0.0 −0.5

*When only **Germanic** languages are considered, the two most prominent correlation are SYNTACTIC Swedish and GEOGRAPHIC Vietnamese. Surprisingly, English shows a weaker correlation despite being a **Germanic** language.*

Romance only (fra, spa, ita, por, cat, ron, glg) - Correlations against BLEU
sim_measure
GENETIC 0.32 0.32 0.32 0.32 0.32 0.32 0.17 0.44 0.35 0.63 0.81 0.41 0.32 0.32 0.32 0.32 0.32 0.32 0.32 0.32
GEOGRAPHIC -0.42 -0.37 -0.35 -0.56 -0.48 -0.52 -0.18 -0.16 -0.53 0.24 -0.16 -0.22 -0.14 -0.24 -0.12 -0.1 -0.3 -0.088 -0.16 -0.22 -0.16 -0.11 -0.16 -0.091 -0.3 -0.15
INVENTORY -0.65 -0.73 -0.65 -0.82 -0.66 -0.66 -0.59 0.71 -0.43 -0.36 -0.25 -0.27 -0.42 -0.028 -0.32 -0.66 -0.55 -0.77 -0.46 -0.49 -0.66 -0.52 -0.7 -0.48 -0.02 -0.67 -0.71
PHONOLOGICAL -0.2 -0.5 0.27 0.27 0.27 0.27 -0.26 -0.3 0.21 0.67 -0.23 -0.34 -0.28 -0.28 0.27 0.17 0.27 -0.28 0.27 0.27 0.27 -0.46 -0.07 -0.19 -0.27 -0.39 -0.39
SYNTACTIC -0.64 -0.75 -0.92 -0.58 0.17 -0.73 -0.54 -0.86 -0.49 -0.33 0.28 -0.3 -0.53 -0.16 -0.62 0.23 0.33 -0.58 -0.03 -0.29 -0.55 -0.72 -0.78 -0.57 -0.72 -0.49 -0.43
eng deu swe nld nor dan fra spa ita por cat ron rus pol ukr srp ces bul hrv slv zho jpn vie kor ind fin hun
0.5 0.0 −0.5

*When considering **Romance** languages, feature importance seem to shift into INVENTORY and SYNTACTIC features. In this language subset, languages on the right side (**Other genera**) are showing stronger correlations.*

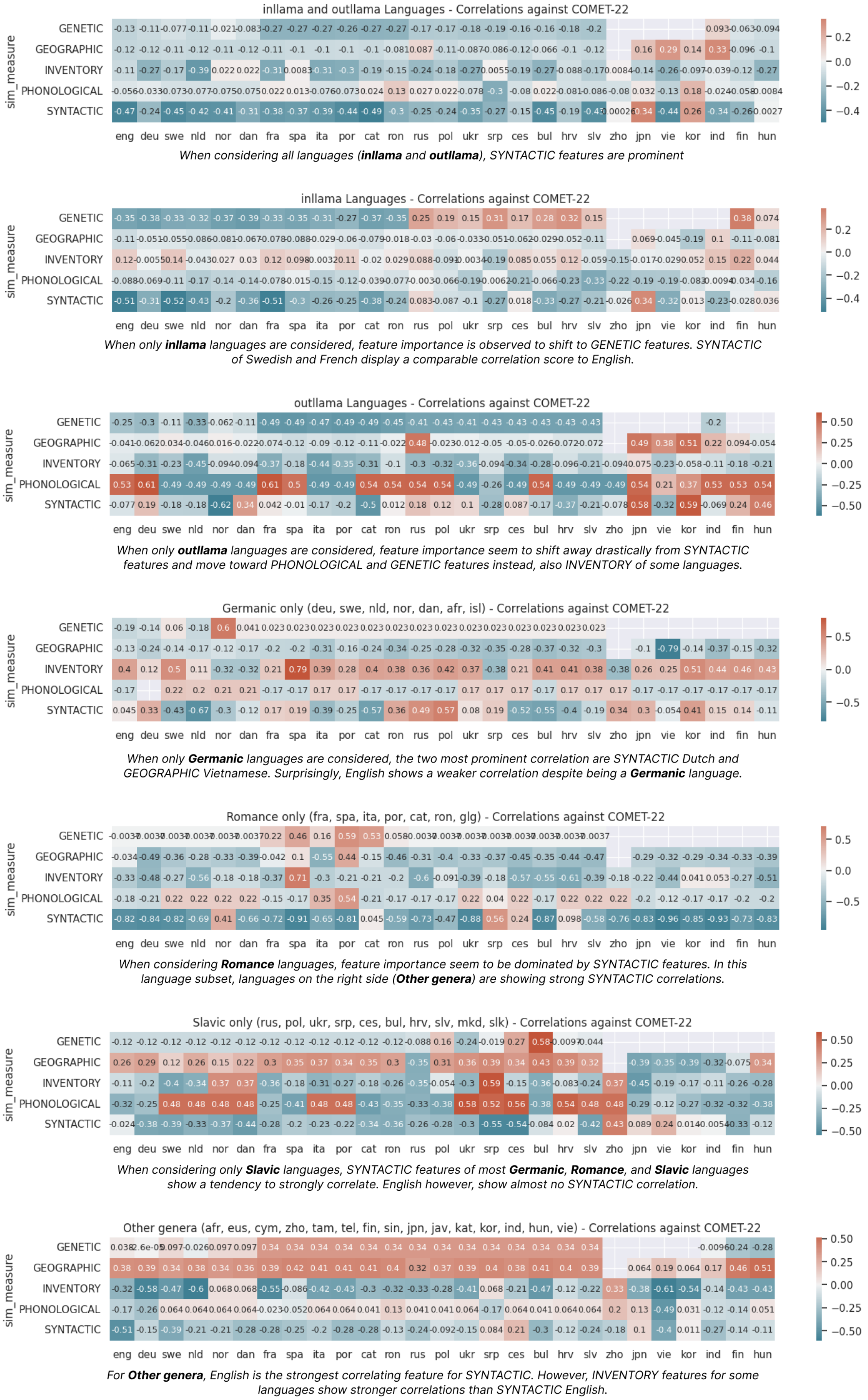
Slavic only (rus, pol, ukr, srp, ces, bul, hrv, slv, mkd, slk) - Correlations against BLEU
sim_measure
GENETIC -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.11 -0.061 0.16 -0.24 0.078 0.34 0.51 -0.014 -0.084
GEOGRAPHIC 0.37 0.39 0.26 0.37 0.3 0.34 0.4 0.42 0.42 0.41 0.42 0.3 -0.41 0.36 0.38 0.39 0.43 0.49 0.41 0.39 -0.42 -0.42 -0.42 -0.41 0.099 0.38
INVENTORY -0.09 -0.22 -0.37 -0.31 0.34 0.34 -0.37 -0.034 -0.26 -0.22 -0.11 -0.23 -0.41 -0.062 -0.29 0.57 -0.13 -0.39 -0.093 -0.25 0.34 -0.46 -0.15 -0.091 -0.025 -0.26 -0.26
PHONOLOGICAL -0.35 -0.27 0.54 0.54 0.54 0.54 -0.27 -0.45 0.54 0.54 -0.47 -0.38 -0.36 -0.53 0.64 0.58 0.62 -0.32 0.55 0.53 0.54 -0.31 -0.12 -0.29 -0.35 -0.35 -0.41
SYNTACTIC -0.22 -0.48 -0.58 -0.44 -0.43 -0.55 -0.42 -0.39 -0.3 -0.36 -0.45 -0.41 -0.35 -0.38 -0.35 -0.55 -0.54 -0.16 -0.021 -0.49 0.38 0.024 0.14 -0.071 -0.093 -0.46 -0.22
eng deu swe nld nor dan fra spa ita por cat ron rus pol ukr srp ces bul hrv slv zho jpn vie kor ind fin hun
0.50 0.25 0.00 −0.25 −0.50

*When considering only **Slavic** languages, SYNTACTIC features of **Germanic**, **Romance**, and **Slavic** languages show a tendency to strongly correlate.*

Other genera (afr, eus, cym, zho, tam, tel, fin, sin, jpn, jav, kat, kor, ind, hun, vie, tgl, ibo) - Correlations against BLEU
sim_measure
GENETIC -0.14 -0.17 -0.086 -0.19 -0.086 -0.086 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.16 0.13 -0.19 -0.18
GEOGRAPHIC 0.36 0.36 0.33 0.36 0.34 0.35 0.37 0.39 0.37 0.38 0.38 0.36 0.28 0.35 0.35 0.36 0.35 0.37 0.36 0.36 0.039 0.17 0.0084 0.14 0.41 0.46
INVENTORY -0.24 -0.5 -0.32 -0.47 0.054 0.054 -0.39 -0.049 -0.31 -0.27 -0.16 -0.21 -0.21 -0.26 -0.31 0.054 -0.23 -0.34 -0.12 -0.16 0.27 -0.22 -0.23 -0.45 0.081 -0.27 -0.26
PHONOLOGICAL -0.19 -0.26 -0.087 -0.087 -0.087 -0.087 -0.086 -0.12 -0.087 -0.087 0.016 0.14 0.016 0.016 -0.087 -0.33 -0.087 0.016 -0.087 -0.087 0.0057 -0.068 -0.34 0.011 -0.045 -0.18 0.019
SYNTACTIC -0.59 -0.2 -0.48 -0.41 -0.35 -0.29 -0.41 -0.31 -0.42 -0.46 -0.42 -0.26 -0.26 -0.31 -0.36 -0.054 0.064 -0.45 -0.001 -0.43 -0.21 0.16 -0.43 -0.018 -0.32 -0.24 -0.13
eng deu swe nld nor dan fra spa ita por cat ron rus pol ukr srp ces bul hrv slv zho jpn vie kor ind fin hun
0.25 0.00 −0.25 −0.50

*For **Other genera**, SYNTACTIC English is the strongest correlating feature.*

Figure 2: Heatmaps of correlations between linguistic distances with BLEU scores of the Llama2-7B one-shot prompting setup (language subset considered is written above each heatmap)



*When considering all languages (**inllama** and **outllama**), SYNTACTIC features are prominent*

*When only **inllama** languages are considered, feature importance is observed to shift to GENETIC features. SYNTACTIC of Swedish and French display a comparable correlation score to English.*

*When only **outllama** languages are considered, feature importance seem to shift away drastically from SYNTACTIC features and move toward PHONOLOGICAL and GENETIC features instead, also INVENTORY of some languages.*

*When only **Germanic** languages are considered, the two most prominent correlation are SYNTACTIC Dutch and GEOGRAPHIC Vietnamese. Surprisingly, English shows a weaker correlation despite being a **Germanic** language.*

*When considering **Romance** languages, feature importance seem to be dominated by SYNTACTIC features. In this language subset, languages on the right side (**Other genera**) are showing strong SYNTACTIC correlations.*

*When considering only **Slavic** languages, SYNTACTIC features of most **Germanic**, **Romance**, and **Slavic** languages show a tendency to strongly correlate. English however, show almost no SYNTACTIC correlation.*

*For **Other genera**, English is the strongest correlating feature for SYNTACTIC. However, INVENTORY features for some languages show stronger correlations than SYNTACTIC English.*

Figure 3: Heatmaps of correlations between linguistic distances with COMET-22 scores of the Llama2-7B one-shot prompting setup (language subset considered is written above each heatmap)

4. Related Work

Our work aligns with previous studies that assess LLMs for translation, resembling the work by Hendy et al. (2023) and Robinson et al. (2023). We aim to extend such evaluations further by investigating the influence of the languages included in the training data of the model, which was previously underexplored due to the lack of transparency of LLMs used. Our method of analysis, similar to the work of Robinson et al. (2023), investigates feature importance. Our objective is to extend that exploration by encompassing other linguistic features obtained from the URIEL typological database (Littell et al., 2017). We are interested in the phenomenon observed in the work of Lin et al. (2019) which shows that however important dataset statistics are compared to linguistic features, there are cases where using them alone to choose transfer languages results in poor performance. This phenomenon drove us to conduct a more comprehensive exploration of linguistic features.

5. Conclusion

We provide a comprehensive evaluation of machine translation in Llama2 for languages seen or unseen in its training data. In this work, we provide English→X machine translation scores of Llama2 7B for 26 languages reported to be in the training data of Llama2 models. We also evaluated 15 additional languages that are not reported to be in Llama2 training data using the 7B, 7B-chat, 13B, and 13B-chat Llama2 models. Our results show that Llama2 is capable of translating into languages it is unfamiliar with, although this phenomenon is observed only in some languages. We demonstrate that model scaling has the most substantial impact when compared to instruction tuning and adding shot count, whose improvements vary by language. We also modeled the linear relationship of linguistic distances and translation quality through correlation scores and revealed that syntactic similarity is not the only feature that displays strong correlations with machine translation scores. Furthermore, despite English having the most training data, there are other languages (e.g. Swedish, Catalan) whose linguistic distances exhibit comparable correlation scores to English albeit having much fewer training data. Our findings pose a unique perspective on the current landscape of language models, suggesting that the prevailing focus on English-centered models may not be the most optimal setup for multilingual models. We hope to open doors toward more effective and training-data-efficient multilingual systems that are shaped by languages other than English, thus promoting digital language equality and sustainability.

Limitations

Our research heavily depends on the language distances obtained from the URIEL typological database, as introduced by Littell et al. (2017). The original authors noted that many languages in the database may have missing features, which means the accuracy of our findings is constrained by the methods used to compensate for these missing features. Our evaluation with the COMET-22 metric is only done for languages supported in their models. However, the model may not be equally reliable for all languages, thus the COMET-22 correlations are only as accurate as the COMET-22 model. Furthermore, there are other ways to model the relationship between language feature distances and machine translation scores. We leave such investigations for future work. We also left out positively correlated features in our analysis as they are not readily interpretable in the context of our analysis.

In an ideal scenario, it would be advantageous to include all languages from the FLORES-200 benchmark and all available versions of Llama2 and other multilingual models to provide more evidence of the effectiveness of scaling parameter count and the overall generalizability of our findings. Unfortunately, our research is constrained by limited computational resources, preventing us from achieving this comprehensive coverage. We exclude X→English translation directions as Llama2 is likely trained on English Wikipedia. We also exclude prompting languages in **outllama** using various dictionary-based prompting techniques due to the challenging work required to collect accurate dictionary entries for low-resource languages. However, we leave this for future work.

We are also aware that the chat versions of Llama2 have been intentionally trained to prevent the generation of harmful or toxic content, and this protective design may affect the quality of translations. Moreover, the chat versions of the model generate numerous artifacts in addition to the translated sentences. We have made diligent efforts to automate the output parsing process to ensure that metrics are calculated fairly. The task of human evaluation and manual parsing of the outputs is left for future work.

Acknowledgements

Authors from Indonesian institutions are supported by the Indonesian Ministry of Education, Culture, Research, and Technology (MoECRT) ACE Open Research program. This work is also supported in part by the Indonesia-US Research Collaboration in Open Digital Technology grant funded by the Indonesian Ministry of Education, Culture, Research, and Technology. The authors would like to thank the Indonesian government for their funding and Boston University for providing essential computing resources through the Shared Computing Cluster (SCC).

6. Bibliographical References

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Akshay Nambi, Vaibhav Balloli, Mercy Ranjit, Tanuja Ganu, Kabir Ahuja, Sunayana Sitaram, and Kalika Bali. 2023. Breaking language barriers with a leap: Learning strategies for polyglot llms.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages

311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alberto Poncelas and Johanes Effendi. 2022. Benefiting from language similarity in the multilingual MT training: Case study of Indonesian and Malaysian. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 84–92, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463, Copenhagen, Denmark. Association for Computational Linguistics.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

7. Language Resource References

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

A Language Model Trained on Uruguayan Spanish News Text

Juan Pablo Filevich[1], Gonzalo Marco[1], Santiago Castro[2], Luis Chiruzzo[1], Aiala Rosá[1]

[1]Universidad de la República – Uruguay [2]University of Michigan – Ann Arbor, USA
{juan.filevich,gonzalo.marco.mohotse}@fing.edu.uy

Abstract

This paper presents a language model trained from scratch exclusively on a brand new corpus consisting of about 6 GiB of Uruguayan newspaper text. We trained the model for 30 days on a single Nvidia P100 using the RoBERTa-base architecture but with considerably fewer parameters than other standard RoBERTa models. We evaluated the model on two NLP tasks and found that it outperforms BETO, the widely used Spanish BERT pre-trained model. We also compared our model on the masked-word prediction task with two popular multilingual BERT-based models, Multilingual BERT and XLM-RoBERTa, obtaining outstanding results on sentences from the Uruguayan press domain. Our experiments show that training a language model on a domain-specific corpus can significantly improve performance even when the model is smaller and was trained with significantly less data than more standard pre-trained models.

Keywords: Uruguay, News Corpus, Pre-trained Language Model

1. Introduction

In recent years, the Natural Language Processing community has witnessed considerable improvements in several areas – including Question Answering (Izacard et al., 2022; Zhang et al., 2021), Machine Translation (Takase and Kiyono, 2021; Liu et al., 2020a), and Sentiment Analysis (Raffel et al., 2020; Yang et al., 2019) – largely due to the advances in the pre-training methodology and the availability of data and pre-trained models to build upon (Jia et al., 2022; Liu et al., 2020b; Tian et al., 2020). Even though most of these advances have focused on English (Brown et al., 2020; Devlin et al., 2019; Liu et al., 2019), several efforts have considered multiple languages, including Spanish (Cañete et al., 2020; Pérez et al., 2022; De la Rosa et al., 2022; Xue et al., 2021; Conneau et al., 2020).

Current approaches that employ the Spanish language focus on pre-training on data dominated by Spanish varieties from countries with the most speakers or the most resources (i.e., Mexico, USA, and Spain). For example, the corpus used for BETO (Cañete et al., 2020; Cañete, 2019) employs many European source texts, hinting at a strong presence of Peninsular Spanish. Low-resource Spanish varieties have been broadly left behind, even when the Spanish language, like other languages, varies significantly from country to country (and even by region) in aspects such as grammar and vocabulary (Lipski, 2012). In addition to linguistic diversity, there are culture-related aspects that are unique to each country and region, which are typically underrepresented in low-resource communities. Such aspects are present in the training data used by today's pre-trained language models, albeit typically dominated by high-resource languages.

This work compiles a corpus of Uruguayan texts and presents models trained using this data. As far as we are concerned, these are the first data and general-purpose models tailored to conducting Natural Language Processing research with Uruguayan-specific texts. The dataset features 900,000 documents obtained from four Uruguayan news outlets with 400 million tokens in total in 6 GiB of uncompressed data. The data has been meticulously filtered and cleaned for quality purposes.

In the current context of NLP and AI, access to computational resources has become increasingly more difficult, especially in Global South countries. In particular, this type of language model is significantly resource-intensive to train. Considering this, besides creating a model specifically tailored to Uruguayan text, our motivation is also to create a model that is smaller and, hence, less computationally intensive to train and use than the available ones. Instead of fine-tuning an already pre-trained larger model, in this work, we train our model from scratch to tailor its size to make it appropriate for limited-resource settings.

Another motivation for developing specific resources for processing local texts, particularly news texts, is the growing interest in their automatic analysis by Uruguayan researchers in areas such as Sociology, Economics, and Communications. We believe it is necessary to have a language model that represents this type of text as well as possible.

We show the quality of the data by training BERT-based models on it through ablations on Uruguayan-related tasks and also by

Proceedings of the Second TDLE Workshop @LREC-COLING-2024, pages 53–60

comparing them with other pre-trained models such as BETO (Cañete et al., 2020), and XLM-RoBERTa (Conneau et al., 2020). We also perform a qualitative analysis of the knowledge captured by such models. The dataset and the pre-trained models are publicly available at `https://huggingface.co/pln-udelar/rouberta-base-uy22-cased`.

2. Related Work

Several works have compiled corpora in Spanish for research. Cañete (2019) compiled a 3-billion-word training corpus by combining multiple sources, including subtitles and news stories, an updated version of the one compiled by Cardellino (2019). Pérez et al. (2022) collected 622 million tweets in Spanish. Gutiérrez-Fandiño et al. (2022) built a massive corpus of 135 billion words from the Spanish Web Archive. Other works, such as (Wenzek et al., 2020; Conneau et al., 2020), have built multilingual datasets by leveraging efforts such as Common Crawl. As far as we are aware, our work is the first one to build a Spanish corpus dedicated to studying the Uruguayan variety and cultural references in the text.

Regarding pre-trained models, there have been efforts to build both multilingual models and Spanish-specific ones. Several multilingual models have originally been implemented from model architectures that had been used to train models for English first, such as Multilingual BERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), mT5 (Xue et al., 2021), and mBART (Liu et al., 2020b). More recently, other efforts have focused on building multilingual large language models, such as BLOOM (Scao et al., 2022), GPT-4 (OpenAI, 2023) and PaLM (Chowdhery et al., 2022). A line of work has focused on Spanish-specific models (Cañete et al., 2020; De la Rosa et al., 2022) and specific domains, such as RoBERTuito (Pérez et al., 2022), specifically for Spanish tweets. Unlike previous works, this paper presents models trained on Uruguay-specific Spanish data, which can capture its particular linguistic and cultural features.

The idea of training a domain-specific LM has been attempted in the past. Still, these generally start from large models or are trained with significantly more data, making them computationally intensive. Some existing domain-specific LMs are created by fine-tuning a general language model (e.g. BioBERT, Lee et al. (2019), FinBERT, Araci (2019); MatSciBERT, Gupta et al. (2022)), or are trained from scratch as our model, but using a larger corpus (e.g. SciBERT, Beltagy et al. (2019); RoBERTuito, Pérez et al. (2022)). In this work, our main goal was to obtain good performance with a model significantly smaller than that of the usual language models, trained on a relatively small data set, and with shorter training times due to limitations in computational resources. So, we are testing not only the usefulness of having a domain-specific model but also the performance of a small model trained with few resources. Verifying the usefulness of such a model is notoriously relevant for us since we usually work in both model training and inference in low-resource contexts.

3. A Uruguayan News Corpus

We scraped four of some of the most important media outlets in Uruguay: *El Observador*, *El País*, *Montevideo Portal*, and *La Diaria*. The first three were scraped from the Internet, while the latter provided us with their articles. For every article, we retrieved the main text (i.e., the article's body) and some potentially useful metadata such as the URL, date, category, title, keywords, and a front picture or cover (if any). After one month of scraping (carried out between November and December 2022), we collected more than 6 GiB of uncompressed data, with articles spanning from the early 2000s up to December 2022. We call our new corpus *UY22*. Table 1 shows the distribution of articles for each website.

We conducted a data quality assurance process based on stripping the HTML tags, trimming and removing duplicate whitespaces, the normalization of strange characters using the *Unidecode* Python library, the removal of emojis, and converting the links into the string "<link>". We also deleted any article with fewer than sixteen words. We split the texts by document and split them into sentences. We refer interested readers to a more in-depth explanation of the scraping and preprocessing phases of this corpus to this project repository[1]. We made the raw and clean versions publicly available (the latter is about 4 GiB uncompressed).

4. ROUBERTa: a Uruguayan LM

We employ a RoBERTA-base (Liu et al., 2019) architecture and train it on the clean version of our data using HuggingFace's Transformers library (Wolf et al., 2019). We use a BPE (Sennrich et al., 2016) tokenizer with a vocabulary size of 30,000 tokens. The model is trained for 30 days on ClusterUY (Nesmachnow and Iturriaga, 2019) with one NVIDIA P100 (12 GiB) for about 6 million steps using RoBERTa's training objective (Masked Language Modeling – cross-entropy loss on the prediction of a masked token, where each token has a 15% masking probability). We show in Figure 1 the training loss curve we obtained. We

[1] `https://gitlab.fing.edu.uy/uy22/uy22`

Name	Website	# Articles	# Words	From	To	Share
El Observador	elobservador.com.uy	314,821	150,007,925	2011	2022	37%
El País	elpais.com.uy	147,004	92,605,424	2013	2022	23%
Montevideo Portal	montevideo.com.uy	433,244	145,422,666	2000	2022	36%
La Diaria	ladiaria.com.uy	20,000	14,079,916	2009	2021	4%

Table 1: UY22 corpus statistics. The share of each website is computed based on the number of words.

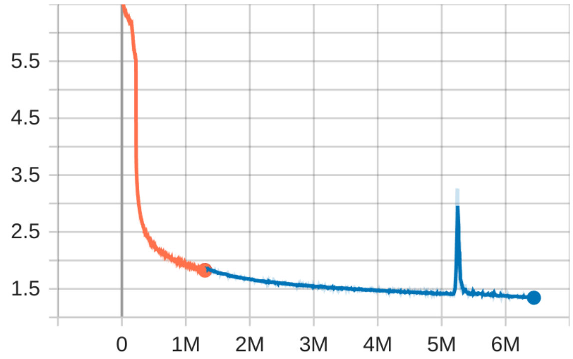


Figure 1: Loss curve for the cased variant of our model with an exponential moving average smoothing value of 0.6. The x-axis shows the number of training steps. The y-axis shows the loss. The color change shows when we restarted the training with a smaller max context length and a larger batch size.

trained for this number of steps due to our limited computational resources and the fact that the loss value was still converging. We chose to train the model from scratch instead of fine-tuning a more general pre-trained model for Spanish, seeking to obtain a model of an appropriate size for use with medium-end computers.

We note a loss spike between 5M and 6M training steps. This could be due to multiple reasons (Takase et al., 2023; Wortsman et al., 2024), including a large amount of consecutive bad-quality data (Soldaini et al., 2024), a large beta2 parameter when using Adam, or a very high learning rate for the batch size we employed. However, we still need to conduct further analyses to understand what is happening in this case.

Figure 2 shows the performance of the model on a Sentiment Analysis task (see Section 5 for details) at different moments during training. We employed a Dynamic Masked Language Modelling task, following Liu et al. (2019). Most hyperparameter values were similar to those used to train RoBERTa, including a max sequence length of 384. However, to cope with GPU memory limitations, we decided to stop it early during training and continue with a batch size of 32 and a max sequence length of 128 (plus two for the special tokens). The learning rate started from 1e-4 and was linearly decayed during training.

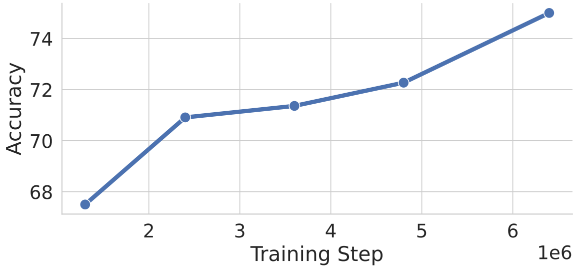


Figure 2: Performance of the cased variant of the model on a Sentiment Analysis task concerning the number of training steps.

The model is named *ROUBERTa* (after ROU-based RoBERTa, where ROU stands for “República Oriental del Uruguay” – the target country’s full name in its native language). We train both cased and uncased variants for our model, although, as we will see in Section 6, the cased variant generally has better results.

5. General Evaluation

This section and the following present the evaluation of our model. This first section will compare only the best-performing model against external baselines. At the same time, in Section 6, we present a deeper evaluation of some interesting cases with examples, and we show ablation tests to see how our design choices affected the model’s performance.

We evaluate our model on two in-domain tasks: Question Answering and Sentiment Analysis on Uruguayan news articles. We perform the experiments for our cased model and compare it with two strong baselines for Spanish: the BETO and XLM-Roberta models, using the cased versions of the models in all cases. We describe the two benchmarks hereafter.

Sentiment Analysis The first benchmark is a sentiment analysis dataset (Dufort y Álvarez et al., 2016), which is composed of a collection of short spans of text that contain an opinion (i.e., a statement by some actor about some topic) and its sentiment polarity in one of three classes (“POS”, “NEG” and “NEU”). The dataset contains 1261 examples and was split into 80% for training and 20% for

Model	Sentiment Analysis Accuracy	Question Answering EM	 F1
XLM-RoBERTa	73.4	26.8	**36.4**
BETO	74.6	24.6	29.4
ROUBERTa	**75.0**	**28.1**	32.3

Table 2: Results of the main experiments.

testing. The model was fine-tuned for 6 epochs with a learning rate of 1e-5.

Question Answering The second benchmark is the QuALES question answering task (Rosá et al., 2022), which contains questions from Uruguayan news articles about the COVID-19 pandemic. All articles, questions, and answers are in Spanish. The QuALES dataset is rather small compared to other QA datasets, containing around 3,600 question-answer pairs (for comparison, SQuAD (Rajpurkar et al., 2016) has more than 100,000), and only 1,000 of those comprise the training set. The dataset format is similar to SQuAD's, which enabled us to experiment with a widely used strategy for Question Answering tasks, starting from a pre-trained BERT-based model and fine-tuning with SQuAD data. In this case, we swapped the SQuAD data with the QuALES data and used the following models as starting points: BETO (`bert-base-spanish-wwm-cased`), XLM-Roberta (`xlm-roberta-base`), and our ROUBERTa-base-cased. We fine-tuned the models for 3 epochs with a batch size of 16 and a max context length of 384. The learning rate was 2e-5, and the weight decay was 0.01. We employ the evaluation metrics Exact Match and F1.

Table 2 shows the results for both experiments. The ROUBERTa_cased model outperforms the other baselines for the sentiment analysis task and the exact match metric of the QA task while getting a solid second place considering the F1 metric of the QA task. In this second task, ROUBERTa_cased performs slightly better than BETO and outperforms XLM-RoBERTa by almost two points. However, in the QA task, ROUBERTa_cased only outperforms XLM-RoBERTa in the Exact Match metric and outperforms BETO in both analyzed metrics. Overall, we can say that our model achieves a more even performance, always within the top ranks among the three compared models. It is worth mentioning that XLM-RoBERTa was chosen for these experiments instead of comparing us with a Spanish version of RoBERTa, based on the results of the original QuALES competition (Rosá et al., 2022), in which the top performing systems used XLM-RoBERTa.

ID	Source	# Masked Sentences	XLM-RoBERTa	BETO	ROUBERTa
text01	La Diaria 06/21/2023	533	151	160	**219**
text02	La Diaria 06/21/2023	170	54	42	**70**
text03	Montevideo Portal 06/22/2023	235	63	75	**123**
text04	Montevideo Portal 06/08/2023	245	78	95	**114**
text05	Búsqueda 07/06/2023	335	98	107	**151**
text06	La Diaria 02/28/2024	304	77	66	**112**
text07	Montevideo Portal 02/20/2024	546	116	96	**193**
	Total	2368	637 (27%)	641 (27%)	**982** (**42**%)

Table 3: Evaluation based on the word-masking task. For each model, we show the number of masked words that were correctly predicted. The dates are in mm/dd/yyyy format.

6. Ablation Study

In this section, we perform an empirical justification of the decisions we have made to build our corpus and train our models.

6.1. Predicting Words in Unseen Texts

We are interested in inquiring if the trained model captures aspects of the Uruguayan culture. We evaluate our model and others on a masking task using five recent texts from three different Uruguayan media outlets. The objective is to evaluate whether the model captures country-specific information such as names of public people, locations, organizations, and the style of the local press. We analyze examples where the masked words contain such information, and, as we will show, our model tends to perform better than general models. In the following analysis, we include some examples to illustrate this behavior.

Note that the model has not seen any text employed in this evaluation since they belong to more recent news articles than the training data. Furthermore, one of the media outlets employed here, Búsqueda[2], is not part of the four media outlets used by our training data. Consequently, this evaluation measures the generalization capacity of our

[2] busqueda.com.uy

trained model on unseen in-domain data. The texts were selected based on different criteria. Some texts are about usual topics in the local press: text03 is about judicial issues, text05 is about insecurity, and texts 06 and 07 are about Carnival, a popular cultural activity in Uruguay. Other texts are about current topics that are not frequent in the country: text01, text02, and text04 are about a severe drought that caused issues with the drinking water distribution in 2023.

To carry out this evaluation, we proceed as follows. For each text t_i and each sentence s_{ij}, we generate different versions of the sentence by masking each word with more than four letters. Except for the masked word, each new sentence is the same as s_{ij}. By these means, we obtain an extended set of sentences for each text, $ExtSent_i$, where each original sentence s_{ij} has multiple versions, one for each masked word. Then, for each sentence in $ExtSent_i$, we obtain candidates for each masked word, using our model ROUBerta_cased, and three other strong models: BETO (Cañete et al., 2020), trained specifically for Spanish[3]; multilingual BERT (Devlin et al., 2019)[4], and the multilingual model XLM-RoBERTa (Conneau et al., 2020)[5]. Table 3 shows the results of this evaluation, except for multilingual BERT, which performed significantly worse than the rest of the models and was therefore not included in the table. As shown in the table, our model, trained exclusively with Uruguayan press texts, gives the best results for the five evaluated texts. It correctly predicts the masked word with a 42% top-1 accuracy, which is a high gap compared to the 27% accuracy obtained by the other models.

Analyzing the results, we observe some interesting examples. In texts about current topics not usually found in the press, such as the drought suffered this year, proper nouns related to our country are correctly predicted by our model, such as the name of the capital of Uruguay in the following example: *Es decir, el agua que sale por las canillas, sale con gusto salado, al menos en <mask> y el área metropolitana. || That is, the water that comes out of the taps, comes out tasting salty, at least in <mask> and the metropolitan area.*

Predictions
ROUBERTa: **Montevideo**
BETO: México
XLM-RoBERTa: Bogotá

On the other hand, the style of the texts also seems to have been captured by our model, as shown in the following example, where it correctly predicts a verb form very usual in the local press: *Luego de que radio Universal <mask> sobre la adjudicación de una vivienda bajo la modalidad de alquiler con opción a compra a una militante de Cabildo Abierto (CA) sin pasar por sorteo || After radio Universal <mask> about the awarding of a house under the rent-to-buy modality to a Cabildo Abierto (CA) militant without going through a raffle*

Predictions
ROUBERTa: **informara**
BETO: ##a
XLM-RoBERTa: informó

It can also be seen that our model incorporated the lexical preferences of the local press, as seen in the following example: *Así lo anunció la titular de la <mask>, Karina Rando, este jueves en conferencia de prensa. || This was announced by the head of the <mask>, Karina Rando, this Thursday at a press conference.*

Predictions
ROUBERTa: **cartera**
BETO: cadena
XLM-RoBERTa: entidad

Finally, for the Carnival theme, our model correctly predicts the word *murga*, which is a typical artistic expression of the Uruguayan carnival: *Desde que el Carnaval volvió tras la pandemia, solo una <mask> obtuvo el primer premio y fue Asaltantes con Patente. || Since Carnival returned after the pandemic, only one <mask> won first prize and that was Asaltantes con Patente.*

Predictions
ROUBERTa: **murga**
BETO: persona
XLM-RoBERTa: empresa

6.2. Is the Uruguayan Data Necessary?

To study the effect of using Uruguayan-specific data compared to a general-Spanish dataset, we trained a new RoBERTa (Liu et al., 2019) model with the corpus used for training BETO (Cañete et al., 2020; Cañete, 2019). RoBERTa models, ours, and the one trained with the BETO corpus were fine-tuned for the Sentiment Analysis task on Uruguayan news, following the steps described in Section 5. Table 4 shows the performance of both models on this task. We can see that our model achieves better results than the one trained with the BETO corpus, even when the latter is five times larger.

6.3. Whole-Word Masking

We consider the model's performance when using the whole-word masking technique introduced in BERT (Devlin et al., 2019) code repository. For this evaluation, we consider the same sentiment analysis. Table 5 shows the results. Not employing

[3] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[4] https://huggingface.co/bert-base-multilingual-cased

[5] https://huggingface.co/xlm-roberta-large

Training data	Size (GiB)	Acc.
BETO's (Cañete, 2019)	20	65.0
UY22 (ours)	4	**68.6**

Table 4: The model's performance on a Sentiment Analysis task when varying the training data. The uncased variant is employed.

Whole-word masking	Accuracy
Yes	35.0
No	**68.6**

Table 5: The model's performance on a Sentiment Analysis task when using the whole-word masking technique. The uncased variant is employed.

whole-word masking proved to be superior in our case, which is, on the one hand, inconsistent with BERT experiments but, on the other hand, consistent with what was reported by Dai et al. (2022).

6.4. Case Sensitivity

We study the effect of case sensitivity in the tokenization. These refer to the cased and uncased variants of the model. We present the results in Table 6. Similarly to other works, such as BERT (Devlin et al., 2019), the cased variant performs better than the uncased one.

7. Conclusion

In this work, we present a dataset specific to Uruguayan Spanish based on news articles and RoBERTa-based models pre-trained on it. We demonstrate the value of our new corpus and the pre-trained models through quantitative and qualitative evaluations employing Uruguayan-news-based tasks. We make both publicly available and hope to enable further research on Uruguayan Natural Language Processing. At the same time, we encourage other community members to replicate our efforts on other Spanish language varieties.

Our model performs better for the analyzed tasks, but most importantly, it did so using a smaller context length, a smaller corpus, and less GPU VRAM than usual. This shows that it is possible to achieve competitive metrics using fewer resources and smaller models. When comparing our results with the ones reported by (Agerri and Agirre, 2023), we observe our model was trained with a corpus significantly smaller than those considered in that paper and with a much smaller parameter count. Despite this consideration, our model achieves better results than others, particularly when compared to XLM-RoBERTa (except in F1 for the QA task presented in Section 5), which was the best model in the mentioned work. This is particularly relevant to researchers in this region, where we usually work in low-resource contexts for model training and subsequent use.

Variant	Accuracy
Uncased	68.6
Cased	**75.0**

Table 6: The model's performance on a Sentiment Analysis task when varying the case sensitivity.

The most important takeaway from this work is that we built a much smaller language model, trained on much less data and requiring much less computational power, and that still keeps up or outperforms other baselines for relevant tasks. This is essential in research labs with limited access to computational resources.

Ethics Statement

Even if we employed a small model, which requires considerably less power than larger models like RoBERTa, language model training typically requires significant energy consumption. However, the carbon footprint associated with our model's training was at least partially reduced given that we employed ClusterUY's infrastructure, which during the time of our experiments used more than 90% renewable energy sources[6]. Still, further analysis is needed to measure how big the impact is.

8. Bibliographical References

Rodrigo Agerri and Eneko Agirre. 2023. Lessons learned from the evaluation of spanish language models.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[6] https://www.gub.uy/ministerio-industria-energia-mineria/comunicacion/noticias/uruguay-logra-90-energias-renovables-matriz-electrica-contexto-tres-anos

Cristian Cardellino. 2019. Spanish Billion Words Corpus and Embeddings.

José Cañete. 2019. Compilation of large spanish unannotated corpora.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yong Dai, Linyang Li, Cong Zhou, Zhangyin Feng, Enbo Zhao, Xipeng Qiu, Piji Li, and Duyu Tang. 2022. "Is whole word masking always better for chinese bert?": Probing on chinese grammatical error correction.

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*, 68:13–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Guillermo Dufort y Álvarez, Fabián Kremer, and Gabriel Mordecki. 2016. Determinación de la orientación semántica de las opiniones transmitidas en textos de prensa. Bachelor's thesis, Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay.

Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodríguez Penagos, Aitor Gonzalez-Agirre, and Marta Villegas Montserrat. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Robin Jia, Mike Lewis, and Luke Zettlemoyer. 2022. Question answering infused pre-training of general-purpose contextualized representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 711–728, Dublin, Ireland. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

John M Lipski. 2012. Geographical and social varieties of spanish: An overview. *The handbook of Hispanic linguistics*, pages 1–26.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020a. Very deep transformers for neural machine translation. *arXiv preprint arXiv:2008.07772*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sergio Nesmachnow and Santiago Iturriaga. 2019. Cluster-UY: Collaborative scientific high performance computing in Uruguay. In *Supercomputing*, pages 188–202, Cham. Springer International Publishing.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Aiala Rosá, Luis Chiruzzo, Lucía Bouza, Alina Dragonetti, Santiago Castro, Mathias Etcheverry, Santiago Góngora, Santiago Goycoechea, Juan Machado, Guillermo Moncecchi, et al. 2022. Overview of QuALES at IberLEF 2022: Question answering learning from examples in spanish. *Procesamiento del Lenguaje Natural*, 69:273–280.

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*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Sho Takase and Shun Kiyono. 2021. Rethinking perturbations in encoder-decoders for fast training. *arXiv preprint arXiv:2104.01853*.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. 2023. Spike no more: Stabilizing the pre-training of large language models. *arXiv preprint arXiv:2312.16903*.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Mitchell Wortsman, Tim Dettmers, Luke Zettlemoyer, Ari Morcos, Ali Farhadi, and Ludwig Schmidt. 2024. Stable and low-precision training for large-scale vision-language models. *Advances in Neural Information Processing Systems*, 36.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14506–14514.

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
25 May, 2024.

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.

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

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

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.

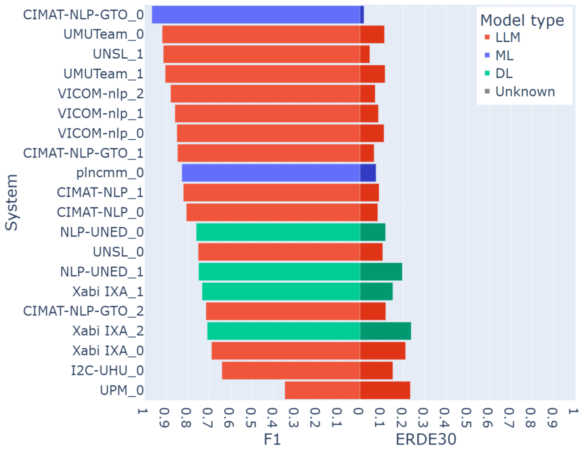
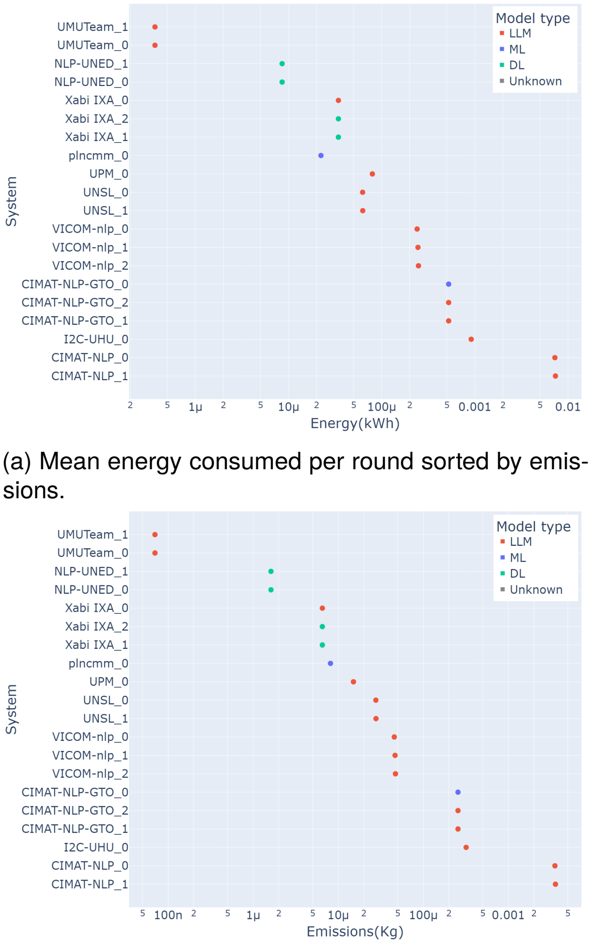


Figure 1: Macro-F1 and ERDE30 scores obtained by the systems sorted by the F1 metric. The y-axis represents the team's name followed by a number representing the system used.

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.



(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.

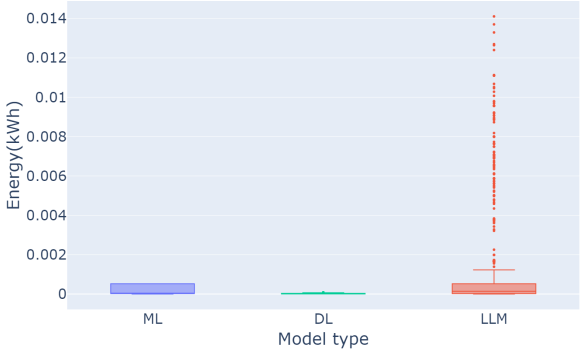


Figure 3: Boxplot of energy consumption (KWh) of each prediction in task 1a by model type.

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 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

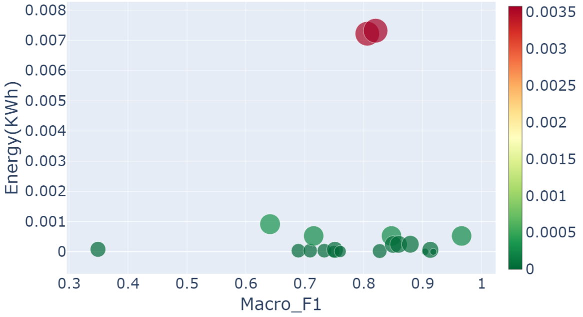


Figure 4: Distribution of the systems for task 1a. The size of the marks is by the emissions produced. The emissions were scaled and a logarithmic normalisation in base 2 was performed for better visualisation. The colour scale corresponds to the actual values of CO2 emissions in kilograms.

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

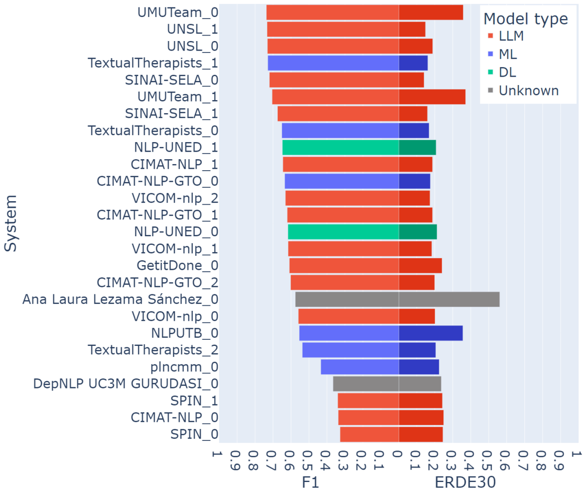
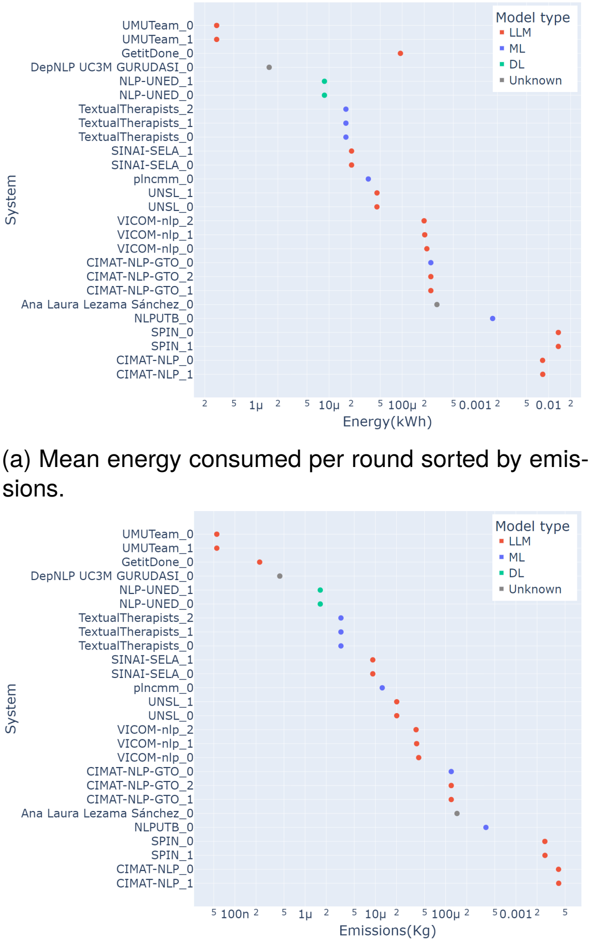


Figure 5: Macro-F1 and ERDE30 scores obtained by the systems sorted by the F1 metric. The y-axis represents the team's name followed by a number representing the system used.

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.



(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.

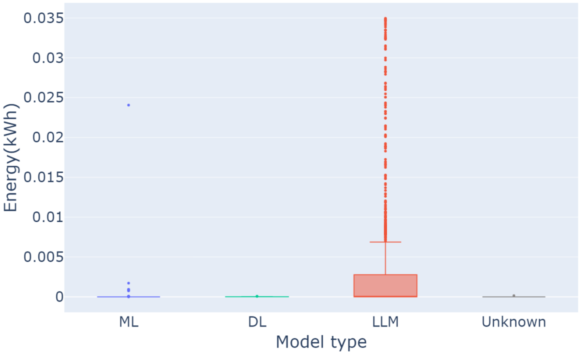


Figure 7: Boxplot of energy consumption (KWh) of each prediction in task 2a by model type.

In Figure 8 systems are visualised according to 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.

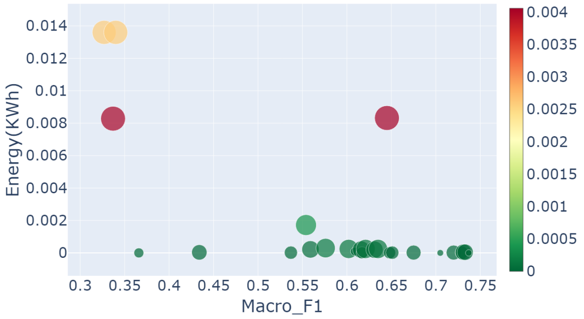


Figure 8: Distribution of the systems for task 2a. The size of the marks is by the emissions produced. The emissions were scaled and a logarithmic normalisation in base 2 was performed for better visualisation. The colour scale corresponds to the actual values of CO2 emissions in kilograms.

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.

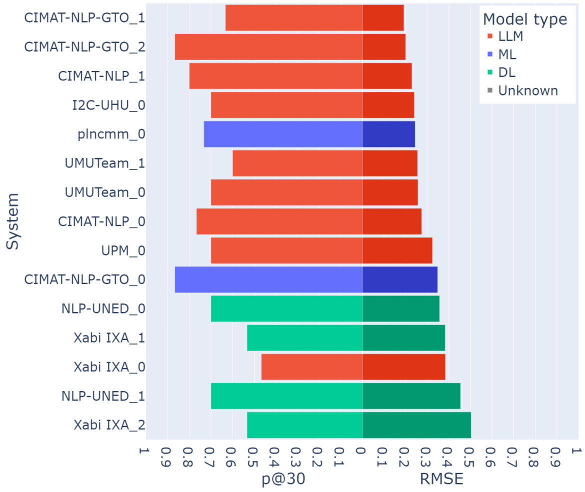


Figure 9: Precision at 30 and RMSE scores obtained by the systems sorted by RMSE metric. The y-axis represents the team's name followed by a number representing the system used.

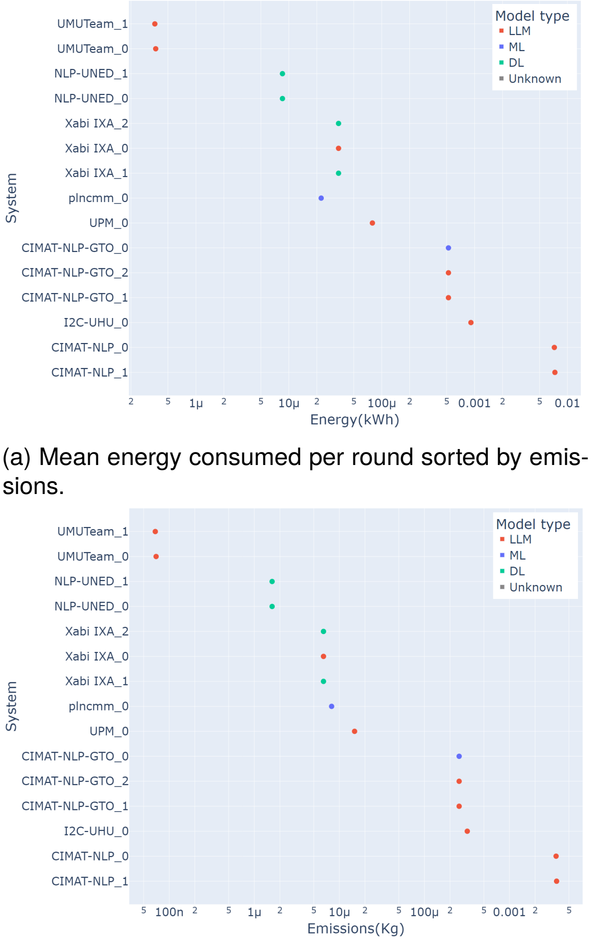
In Figure 12 systems are visualised according to their ranking, energy consumed and emissions produced.

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.

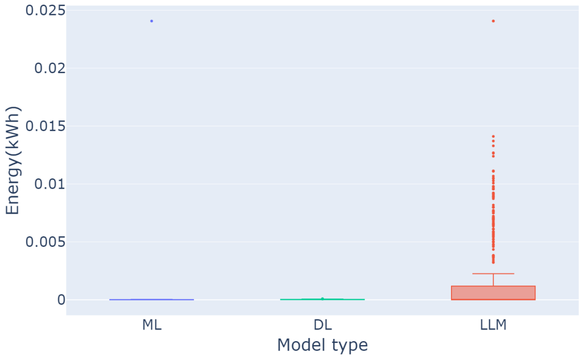


Figure 11: Boxplot of energy consumption (KWh) of each prediction in task 1b by model type.

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.

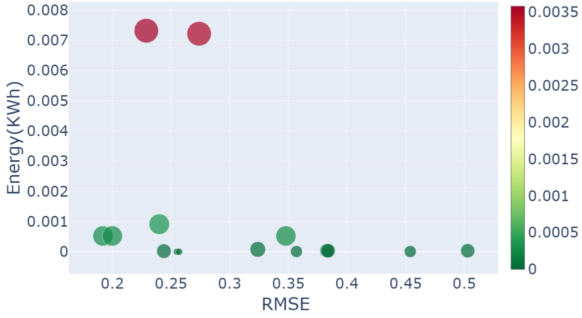


Figure 12: Distribution of the systems for task 1b. The size of the marks is by the emissions produced. The emissions were scaled and a logarithmic normalisation in base 2 was performed for better visualisation. The colour scale corresponds to the actual values of CO2 emissions in kilograms.

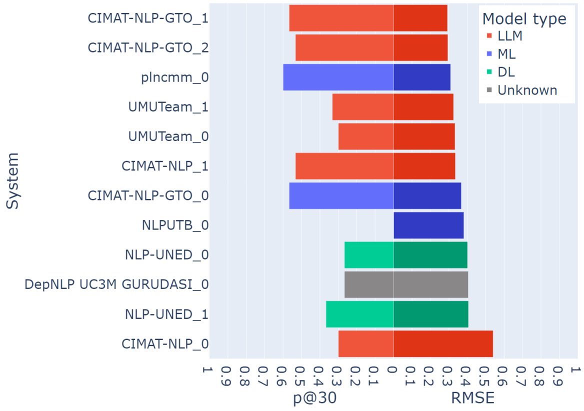
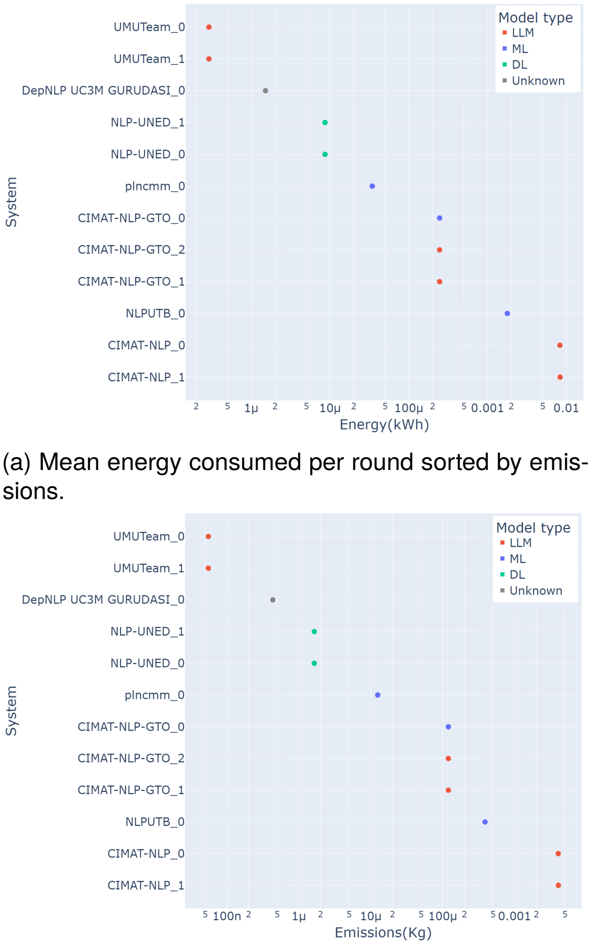


Figure 13: Precision at 30 and RMSE scores obtained by the systems sorted by RMSE metric. The y-axis represents the team's name followed by a number representing the system used.

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.

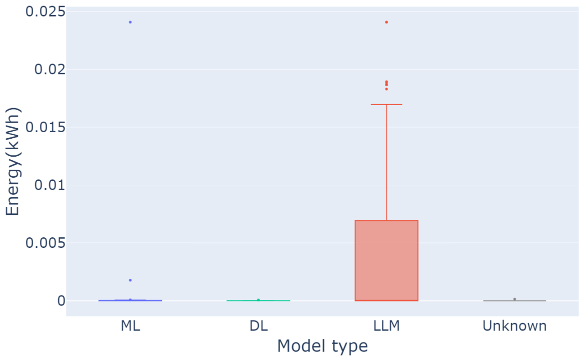


Figure 15: Boxplot of energy consumption (KWh) of each prediction in task 2b by model type.

- 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.

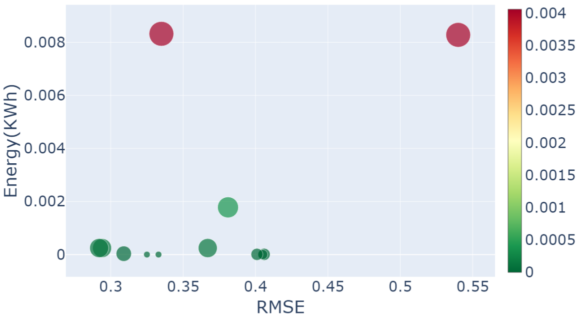


Figure 16: Distribution of the systems for task 2b. The size of the marks is by the emissions produced. The emissions were scaled and a logarithmic normalisation in base 2 was performed for better visualisation. The colour scale corresponds to the actual values of CO2 emissions in kilograms.

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 Marıa 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.

Author Index