International Journal on Natural Language Computing (IJNLC) Vol.13, No.1, February 2024

Performance, energy consumption and costs: a comparative analysis of automatic text classification approaches in the Legal domain

Leonardo Rigutini [1], Achille Globo [1], Marco Stefanelli [2], Andrea Zugarini [1], Sinan Gultekin [1], Marco Ernandes [1]

[1] Department of Hybrid Linguistic Technologies - expert.ai spa - Italy
[2] Department of Information Engineering and Mathematics - University of Siena - Italy

Abstract

The common practice in Machine Learning research is to evaluate the top-performing models based on their performance. However, this often leads to overlooking other crucial aspects that should be given careful consideration. In some cases, the performance differences between various approaches may be insignificant, whereas factors like production costs, energy consumption, and carbon footprint should be taken into account. Large Language Models (LLMs) are widely used in academia and industry to address NLP problems. In this study, we present a comprehensive quantitative comparison between traditional approaches (SVM-based) and more recent approaches such as LLM (BERT family models) and generative models (GPT-2 and LLAMA2), using the LexGLUE benchmark. Our evaluation takes into account not only performance parameters (standard indices), but also alternative measures such as timing, energy consumption and costs, which collectively contribute to the carbon footprint. To ensure a complete analysis, we separately considered the prototyping phase (which involves model selection through training-validation-test iterations) and the in-production phases. These phases follow distinct implementation procedures and require different resources. The results indicate that simpler algorithms often achieve performance levels similar to those of complex models (LLM and generative models), consuming much less energy and requiring fewer resources. These findings suggest that companies should consider additional considerations when choosing machine learning (ML) solutions. The analysis also demonstrates that it is increasingly necessary for the scientific world to also begin to consider aspects of energy consumption in model evaluations, in order to be able to give real meaning to the results obtained using standard metrics (Precision, Recall, F1 and so on).

Keywords

NLP, text mining, green AI, green NLP, carbon footprint, energy consumption, evaluation.

1. Introduction

In the field of NLP, there has been a significant paradigm shift in the past decade. The rise of end-to-end approaches has led to the development of a wide range of Large Language Models (LLMs) with varying neural network architectures and billions of parameters. These massive models are typically accessible only to a few global companies like Google, Microsoft or Meta AI, due to their substantial training and deployment costs. They are usually offered as pre-trained models

DOI: 10.5121/ijnlc.2024.13102

and require fine-tuning to meet specific customer requirements. However, their operation demands extensive hardware and energy resources.

Despite their significance, energy consumption aspects are often overlooked by academics, data scientists, and industry insiders. Nevertheless, the escalating trend of energy-intensive computations raises important concerns. From an ethical and societal perspective, we are witnessing the severe consequences of pollution and CO_2 emissions through climate change. Moreover, from an economic and industrial standpoint, energy costs have skyrocketed in recent years, making lightweight and energy-efficient Machine Learning solutions crucial for companies.

This work presents a comparative analysis of some commonly used families of text classification models, focusing on their performance and power consumption. The main objective is to investigate the trade-off between performance, energy consumption and carbon footprint in the context of vertical domain classification, simulating a typical use case in industry. On the performance front, the widely adopted F1 classification metric is considered, while the environmental impact is evaluated based on energy consumption (KWh), estimated costs (€), and CO_2 production. In particular, we extends the investigation reported in [10] by introducing also the most recent generative approaches. The experiments are conducted using the LexGLUE benchmark, and the results demonstrate that lightweight models often achieve excellent performance at significantly lower costs.

These findings highlight the importance of conducting further in-depth studies on the application of Deep Learning approaches in industry. Moreover, they emphasize the need to consider various aspects beyond prediction quality when selecting the most suitable Machine Learning solution for NLP projects.

The paper is organized as follows. Section 2 reports the related works and provides some of the reasons that led us to carry out this analysis and experimentation. In Section 3, the details of the investigation are described, such as the models and the datasets employed, while in Section 4, we report the results of the experiments and outline the emerging considerations. Finally, Section 5 draws conclusions and possible ideas for future works.

2. Related Work and Motivation

The cost associated with training and deploying deep neural networks has witnessed a significant surge in the past decade, pushing modern ML models towards an energy-intensive trajectory. As a result, researchers have increasingly focused on optimizing models' efficiency and exploring potential adaptations. Numerous studies have tackled the challenge of compressing model size through various techniques, including knowledge distillation [20], pruning [33], quantization [11], and vocabulary transfer [9, 8]. Nevertheless, while a green-friendly communication strategy is gaining traction in many sectors, such as the initiatives taken by Googlee [1] and Amazon [2], the importance of environmental considerations has not yet gained significant attention in the field of Artificial Intelligence (AI) research. Over the past few years, there has been an emerging focus on the eco-sustainability of artificial intelligence. Although there have been attempts to raise awareness about the significance of environmental considerations, only a limited number of studies are found in the existing literature [19]. In [27], the authors conduct a comparative study on the energy consumption and CO_2 production of various neural network models employed in NLP. Notably, they highlight the substantial amount of CO_2 emitted during a single training cycle of a transformer-based NLP model, which surpasses the average annual CO_2 emissions of an individual. However, it's worth

[1] https://sustainability.google/carbon-free/

[2] https://sustainability.aboutamazon.com

mentioning that this analysis does not encompass lightweight methods like SVM and does not establish a correlation between costs and performance. In [22], the authors present a thoughtful examination of the eco-sustainability of AI, emphasizing the prevalent dominance of Red AI over Green AI in the scientific community. They conduct an analysis on a sample of papers published in top AI conferences, revealing the rarity of discussions on efficiency within the field. The findings of this study are summarized in figure 1. Simultaneously, the literature has introduced various

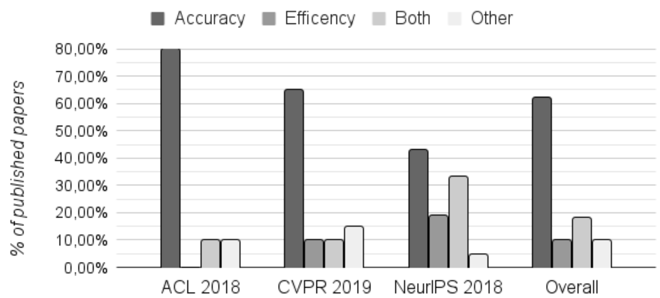

Figure 1: Trend of accuracy and efficiency in AI papers. The charts were recreated with data from [22].

tools for assessing the Carbon-Footprint, such as those presented by Luccioni et al. [14] and CodeCarbon [13]. Notably, the CodeCarbon library has gained significant popularity and is currently one of the most widely utilized tools for quantifying the energy consumption and carbon footprint associated with algorithms.
Recently, in Gultekin et al. [10], the authors present an in-depth study to verify how complex and energy-hungry models are actually necessary to obtain high performances in real and industrial use cases. With their work, the authors underline that for typical industrial use cases (such as the categorization of texts in the LEGAL field), the use of very complex models does not produce advantages significant enough to justify their use, given their high energy consumption values and high costs. In fact, the results show that, in many cases, the use of classical models such as SVMs returns comparable or even better performances compared to models based on LLM, with extremely low energy consumption and costs. This work extends the comparative analysis to the most recent generative LLM models. In particular, following the idea of a comparison based on quality and energy consumption metrics, we added the results obtained on the same dataset from models such as GPT-2 and LLAMA2.

Our motivation for this comprehensive investigation arose from the observation that very few existing studies in the AI literature comprehensively address the combined analysis of performance, energy consumption, costs and carbon footprint in a real-world business context. We firmly believe that such analysis is vital when evaluating AI solutions, given the worrying trend towards ever-larger models requiring energy-intensive computations. In particular because, from the results obtained, it is clear that in many practical cases, this trend is not necessary and raises considerable concerns.

Considering the ethical and social perspectives, we are all witnesses of the serious consequences of climate change caused by pollution, especially CO_2 emissions. Many countries are actively exploring alternative solutions to fossil fuels, but it is equally essential to promote a more conscientious and sustainable use of resources. The world of AI-driven businesses has a responsibility to prioritize environmentally friendly technologies and solutions, while ensuring that performance levels are not compromised. Furthermore, there is the question of democratic access to resources. The search for larger neural network models has created a situation where only a handful of global IT companies have access to them, excluding numerous universities and private research labs, as well as small companies. This phenomenon is often referred to as the "rich get richer" effect. On the contrary, considering the economic and industrial perspective, it is evident that energy costs have increased significantly in recent years. As a result, discovering lightweight AI solutions can result in significant cost savings, which are vital to the sustenance of businesses. For these compelling reasons, we believe that the presented analysis can play a vital role in recommending taking additional aspects beyond performance into consideration when selecting an AI solution, especially in practical scenarios.

3. The investigation

In this article, we present a comprehensive analysis comparing three widely utilized families of text classification models in terms of their performance and power consumption. Our investigation intends to examine the trade-off between performance and the carbon footprint exhibited by different models, specifically focusing on (1) classic Support Vector Machines (SVM), (2) the first generation of Large Language Models (LLMs) and (3) the most recent generative models (GenLLM), when applied within a vertical domain. Our objective is to replicate a typical real-world scenario where the analyzed documents predominantly pertain to a specific domain of interest, such as finance, law, or healthcare. In this study, we specifically concentrate on the legal sector and employ a standard benchmark for this industry, namely the LexGLUE dataset.

3.1 The benchmark

With the proliferation of multitask benchmarks in the NLP domain, such as GLUE and SuperGLUE, there has been a recent release of the LexGLUE Benchmark [5]. The LexGLUE (Legal General Language Understanding Evaluation) benchmark is specifically curated for evaluating the performance of models across a wide range of legal NLP tasks, comprising seven datasets that focus on the legal domain. Initially, the benchmark [3] predominantly covers the English language, offering a foundation for evaluating legal NLP models. However, future iterations of LexGLUE are anticipated to include additional datasets, tasks, and languages as more legal NLP resources become available.

Dataset	Data Type	Task	Train/Validation/Test	Classes
ECtHR (Task A)	ECHR	Multi-label classification	9,000/1,000/1,000	10+1
ECtHR (Task B)	ECHR	Multi-label classification	9,000/1,000/1,000	10+1
SCOTUS	US Law	Multi-class classification	5,000/1,400/1,400	14
EUR-LEX	EU Law	Multi-label classification	55,000/5,000/5,000	100
LEDGAR	Contracts	Multi-class classification	60,000/10,000/10,000	100
Unfair ToS	Contracts	Multi-label classification	5,532/2,275/1,607	8+1
CaseHOLD	US Law	Multiple choice QA	45,000/3,900/3,900	n/a

Table 1: Statistics about the seven datasets included in the LexGLUE benchmark.

The collection of seven datasets within the LexGLUE Benchmark is constructed using various legal sources. These sources include the European Court of Human Rights (ECtHR), the U.S. Supreme Court (SCOTUS), European Union legislation (EUR-LEX), the U.S. Security Exchange Commission (LEDGAR), Terms of Service extracted from popular online platforms (Unfair-ToS), and Case Holdings on Legal Decisions (CaseHOLD). Further information about each dataset can be found in Table 1 and a more comprehensive description can be found in the original paper by

[3] https://github.com/coastalcph/lex-glue

Chalkidis et al. [5]. The ECtHR [2] dataset was constructed by collecting 11K cases from the European Court of Human Rights (ECtHR) public database [4]. The dataset has two variations in itself, in the first one, task A, a model takes an input of list of facts from the case description and gives the set of violated articles as output. On the other hand, in task B, a list of facts is fed as input, however different point is the output. In this variant, the output is the set of allegedly violated articles [5]. Since both outputs are a set of articles, the task is considered multi-label classification. The SCOTUS [26] dataset was released by collecting information from US Supreme Court [5]. 7.8K cases are provided from the metadata, and each case is classified in 14 issue areas [5], which makes the task multi-class classification. The EUR-LEX [3] dataset was published in the European Union legislation portal [6]. Annotated EU-Law is gathered, around 65K documents, in 100 most frequent concepts [5] as a multi-label classification task. The LEDGAR [30] dataset was presented at the LREC 2020 conference [7]. It consists of 80K clauses extracted from contracts downloaded from the EDGAR [8] site of the U.S. Security Exchange Commission [9]. Each clause is classified into a taxonomy of about 100 categories in a multi-class categorization task. The Unfair ToS [12] dataset was a collection of 50 Terms and Services from different online services. Each document is split into its sentences, a total of 9.4K sentences, and each sentence is classified from 8 unfair contractual terms(if any) [5]. The CaseHOLD [32] dataset was collected by US Court cases from the Harward Law Library. It is a Question-Answering (Q&A) oriented dataset and we did not used it in the experimentation since it differ very deeply from the Text Categorization task.

3.2 Models

In our study, we focus on three families of models widely used for automatic text analysis: classic SVM, first generation of LLM and the more recent generative LLMs. During the experiments, we made efforts to replicate the same configurations as those reported in the original LexGLUE experimentation, ensuring consistency wherever possible. [5].

SVM-based approaches

Support Vector Machines (SVMs) [6] are established Machine Learning models that have been extensively employed in text categorization tasks for several decades [18, 24]. They function by identifying an optimal subset of training examples that effectively define a separation hyperplane. Moreover, SVMs utilize kernels to enable the identification of nonlinear separation hyperplanes. For our SVM-based approach, we initially selected a straightforward and basic configuration, employing a linear kernel SVM with a Bag-Of-Word (BoW) representation. This combination has long been the most commonly utilized approach for text categorization problems [18].

Moreover, we incorporated an approach that combines the standard Bag-Of-Word (BoW) text representation with supplementary linguistic and semantic features. This combined approach has been extensively utilized in previous years and has consistently exhibited promising results in various text classification problems [1, 23, 31]. In our analysis, we included this approach to examine whether integrating external linguistic knowledge into the feature space can effectively reduce

[4] https://hudoc.echr.coe.int/eng/
[5] http://supremecourt.gov/
[6] https://eur-lex.europa.eu/
[7] https://lrec2020.lrec-conf.org/en/
[8] https://www.sec.gov/edgar/search/
[9] https://www.sec.gov

model complexity (and subsequently lower energy consumption) without significantly compromising performance. This approach involves an initial NLP step that generates a set of linguistic and semantic features, such as lemmas, Part-Of-Speech tags, and concepts. These features are then combined with the standard Bag-Of-Word representation. The resulting augmented feature space is subsequently utilized to train Machine Learning models. For the NLP analysis, we used the expert.ai hybrid natural language platform, while a linear SVM was used as the on-top ML classifier. The expert.ai natural language platform consists in an integrated environment for deep language understanding and provides a complete natural language workflow with end-to-end support for annotation, labeling, model training, testing and workflow orchestration [10].

In the paper we will refer to these two approaches as SVM_{bow} and SVM_{nlp} , respectively.

BERT-based models

BERT [7] is a widely recognized Large Language Model (LLM) that operates on the transformer architecture. It is renowned for its pre-training on a vast collection of general-purpose documents, making it a strong contender as a generic language model. BERT has consistently demonstrated remarkable performance in the realms of text analysis and natural language processing (NLP). However, due to its large and deep neural network structure, substantial computational resources are required for its execution. Additionally, when dealing with a specific domain, the availability of a language model that captures the linguistic statistics and terminology peculiar to that domain can be highly advantageous. As a result, literature proposes various BERT variants that have been retrained on domain-specific documents. Considering our focus on the legal domain, we include LegalBERT [4] in our comparative analysis. LegalBERT is a derivative of the BERT model, pretrained on legal corpora encompassing legislations, contracts, and court cases.
Lastly, since our analysis delves into energy consumption, closely associated with the model's size, we also incorporate DistilBERT [21] into our evaluation. DistilBERT represents a compact version of the original BERT model, achieved through the utilization of distillation techniques.

Generative models

Generative Large Language Models (GenLLM) have revolutionized the field of natural language processing (NLP). These models have paved the way for advancements in various applications such as text completion, dialogue generation, and story writing. Generative LLMs are built on transformer architectures and they are trained on massive amounts of text data to learn the underlying patterns, statistical regularities, and contextual dependencies within language. This allows them to generate human-like text outputs that can be indistinguishable from those written by humans in many cases. These models excel at producing coherent and contextually relevant sequences of words, making them highly useful in diverse NLP tasks. Most of the numerous models published in the last year refer to two main families: GPT (Generative Pre-trained Transformer) and LLAMA (Language Models for the Advancement of Machine Learning and Artificial intelligence).

GPT [16] is a family of models developed by OpenAI [11] based on a proprietary transformer architecture which allows to capture long-range dependencies in sequences of words and to generate

[10] https://www.expert.ai/products/expert-ai-platform/
[11] https://openai.com/

coherent and contextually relevant text. Although the most recent model is GPT4 [15] (but it has not been made available), we selected GPT2 [17, 25] for our experiments since it has a feasible number of parameters. GPT-2 has been trained on an extensive corpus of diverse text data, showing remarkable performance in a variety of NLP tasks, including text completion, language translation, and question-answering systems.

LLAMA [28] is a collection of language models released by Meta AI [12] under open license. The models range in size and complexity, allowing researchers to select the most appropriate model for their specific needs. Recently, Meta AI developed LLaMA-2[29], the next generation of models which have been released in three model sizes: 7, 13, and 70 billion parameters [13]. For our experiments, we selected two LLAMA2 based models: LLAMA2 with 7 billions of parameters (LLAMA2-7b) and LLAMA2 with 13 billions of parameters (LLAMA2-13b). We made this choice to also investigate how the performances and energy consumption of the same model change based on its size.

3.3 Experimental setup

The comparative analysis encompassed both performance-oriented metrics and eco-friendly indicators. Performance was evaluated using the standard F1 score, including both micro mF1 and macro MF1 metrics. Eco-friendly considerations involved estimating the energy consumption (KWh), costs (€), and carbon footprint (CO_2) associated with each approach.

To assess energy consumption, we utilized the widely adopted ”codecarbon” library [14], which enables the measurement of energy usage during the execution of a sequence of instructions, including GPU utilization [13]. To ensure consistency, for the models reported in the LexGLUE article, we replicated the experiments detailed in the article by incorporating instructions from the ”codecarbon” library directly into the authors’ code.

When evaluating the SVM_{nlp} approach, we also considered the energy required by the NLP analysis phase. Particular attention should also be paid to BERT-based and generative models. While svm-based models are natively classifiers, BERT (and its derivatives) are encoders, i.e. language models pre-trained to find a semantically informative representation of the input test. To be used in specific tasks (such as text classification), a final neural layer has been added and a training phase (fine-tuning) is performed to refine the parameters of the entire model (BERT + final layer). Similarly, as their name suggests, generative models were designed primarily to generate text and not for old-style tasks such as Text Categorization. Thus, also in this case, we adapted them to the text classification tasks of the LexGLUE Benchmark by adding a dense neural layer to the model in order to project the LLM outputs into the class label space. This layer is trained together with the LLM in the fine-tuning phase.

To gain better insight into the cost-effectiveness of the examined approaches, we conducted a separate evaluation of the energy cost specifically pertaining to the prediction phase. In fact, a typical industrial use case consists of two distinct and very different phases: the Research and Development (R&D) phase, in which analysts and scientists execute a large series of experiments in search of the best solution and configuration, and the production phase, in which the optimal solution now identified is put into production and used massively by the customer. The two phases evidently have

[12] https://ai.meta.com/

[13] https://about.fb.com/news/2023/07/llama-2/

[14] https://github.com/mlco2/codecarbon/

different characteristics and therefore were addressed separately in our experiments (sections 4.1 and 4.2).

The experiments were carried out on an Intel Xeon processor-based server with 503GB of RAM equipped with 4 NVIDIA RTX A6000 GPU with 48GB of dedicated Graphic RAM each one (a total of 192GB of Graphic RAM). We excluded the CaseHOLD dataset from our evaluation since it was designed for a Question Answering (QA) task that significantly differs from text classification, unlike the other datasets included in the study.

4. Experimental results

In the development of NLP projects, there are typically two primary phases: (a) model training and evaluation, involving iterative training-validation-test steps to assess the solution during the research and development (R&D) phase, and (b) final delivery and production, where the chosen model is deployed and utilized in a production environment. Hence, we conducted two distinct investigations. Firstly, we compared models in terms of their performance and energy consumption throughout a typical train/validation/test procedure. Secondly, we compared the energy and time requirements of the models when making predictions on a fixed number of documents.

4.1 R&D Scenario

In our initial analysis, we emulated the research and development (R&D) phase of a project. This phase involves the initial setup of the system and often requires multiple iterations. The number of trials can vary depending on project characteristics and intricacies, with considerable variation that can make effort estimations unreliable. In the subsequent sections, we present a detailed comparative analysis for each dataset, considering (a) performance metrics using the F1 score with both micro (mF1) and macro (MF1) averaging, and (b) energy consumption (KWh), costs (€), and carbon footprint (CO_2) estimated for each experimental scenario.

ECtHR Datasets

The findings from the tests conducted on the two European Court of Human Rights (ECtHR) datasets [2] are presented in Table 2. Across both datasets, the SVM_{nlp} approach emerges as the most environmentally friendly option while maintaining comparable performance to SVM_{bow} . Notably, both SVM-based models exhibit lower performance compared to BERT and LegalBERT. However, the energy consumption of the latter is significantly higher, ranging from 40 to 75 times greater than that of the SVM_{nlp} approach. Conversely, DistilBERT demonstrates intermediate energy consumption, ranging from 3 to 20 times higher than the SVM_{nlp} approach, while occasionally exhibiting lower performance in certain cases. Finally, Generative models show very low performance but very high energy consumption values. Most likely, this depends on the fact that they are really very large models (and therefore very high consumption) developed and trained mainly to generate text (and therefore not very suitable for text classification tasks). This behavior appears to be quite recurrent in all other datasets.

EUR-LEX

The results obtained with the European Union Legislation (EUR-LEX) dataset are presented in Table 3. Consistently, the SVM_{nlp} model retains its position as the most environmentally friendly

		mF1	MF1	KWh	€	CO2
ECtHR A	**SVM$_{bow}$**	0.65	0.52	× 1.95	× 1.95	× 1.32
	SVM$_{nlp}$	0.65	0.52	**1.00**	**1.00**	**1.00**
	BERT	**0.71**	**0.64**	× 73.93	× 73.93	× 23.42
	LegalBERT	0.70	**0.64**	× 74.25	× 74.25	× 23.52
	DistilBERT	0.62	0.56	× 23.98	× 23.98	× 7.60
	GPT2 Large	0.54	0.39	× 16.52	× 16.52	× 5.23
	LLAMA2 7B	0.61	0.49	× 50.93	× 50.93	× 16.13
	LLAMA2 13B	0.69	**0.64**	× 167.21	× 167.21	× 52.97
ECtHR B	**SVM$_{bow}$**	0.75	0.65	× 1.56	× 1.56	× 1.16
	SVM$_{nlp}$	0.75	0.65	**1.00**	**1.00**	**1.00**
	BERT	**0.80**	0.73	× 62.49	× 62.49	× 21.83
	LegalBERT	**0.80**	**0.75**	× 36.56	× 36.56	× 5.00
	DistilBERT	0.71	0.61	× 3.39	× 3.39	× 1.78
	GPT2 Large	0.61	0.41	× 13.51	× 13.51	× 4.72
	LLAMA2 7B	0.70	0.58	× 42.15	× 42.15	× 14.72
	LLAMA2 13B	0.71	0.56	× 159.58	× 159.58	× 55.84

Table 2: Classification performances and the energy consumption results of different models on ECtHR datasets.

option, while maintaining highly satisfactory performance levels. Notably, the SVM$_{nlp}$ model delivers commendable performance with approximately half the power consumption compared to SVM$_{bow}$ and about three times lower energy consumption compared to BERT-based approaches. However, in this particular case, the energy savings and pollution reduction rates are relatively lower compared to the previous scenario. In this case, the classification performances returned by the generative models (GPT-2 and LLAMA2) are close to the optimal ones but at the expense of significantly higher energy consumption.

		mF1	MF1	KWh	€	CO2
EUR-LEX	**SVM$_{bow}$**	0.71	0.51	× 1.85	× 1.85	× 7.12
	SVM$_{nlp}$	0.73	0.50	**1.00**	**1.00**	**1.00**
	BERT	0.71	**0.57**	× 4.81	× 4.81	× 1.56
	LegalBERT	0.72	**0.57**	× 4.89	× 4.89	× 1.58
	DistilBERT	**0.74**	0.46	× 1.91	× 1.91	× 1.62
	GPT2 Large	0.64	0.30	× 36.28	× 36.28	× 11.75
	LLAMA2 7B	0.72	0.54	× 113.20	× 113.20	× 36.65
	LLAMA2 13B	0.72	0.56	× 291.98	× 291.98	× 94.55

Table 3: The classification performances and the energy consumption results of different models on EUR-LEX dataset.

LEDGAR

Table 4 presents the outcomes obtained from the evaluation of the Labeled Electronic Data Gathering, Analysis, and Retrieval system (LEDGAR) dataset [30]. Notably, the SVM$_{nlp}$ approach demonstrates the best performance as well as the most favorable power consumption metrics. Remarkably, the SVM$_{nlp}$ approach showcases energy savings of up to 80 times compared to fully BERT-based approaches. While DistilBERT also delivers acceptable performance, it still exhibits significantly higher energy consumption compared to the SVM$_{nlp}$ model. Similar to the previous chaos, generative models report good classification results. Even in this case, however, they require extremely high quantities of energy with considerable costs and CO_2 emissions.

		mF1	MF1	KWh	€	CO2
LEDGAR	**SVM$_{bow}$**	0.88	0.82	× 1.67	× 1.67	× 1.34
	SVM$_{nlp}$	**0.89**	**0.84**	**1.00**	**1.00**	**1.00**
	BERT	0.88	0.82	× 53.21	× 53.21	× 20.05
	LegalBERT	0.88	0.83	× 77.71	× 77.71	× 29.28
	DistilBERT	0.88	0.81	× 24.28	× 24.28	× 9.15
	GPT2 Large	0.84	0.73	× 127.20	× 127.20	× 47.93
	LLAMA2 7B	0.88	0.81	× 409.38	× 409.38	× 154.24
	LLAMA2 13B	0.85	0.75	× 1085.94	× 1085.94	× 409.15

Table 4: Classification performance and the energy consumption results of different models on LEDGAR dataset.

SCOTUS

The results obtained from the evaluation of the US Supreme Court (SCOTUS) dataset [26] are reported in Table 5. These findings align with the previous cases and reaffirm the observed trend. Moreover, in this particular case, the SVM$_{nlp}$ approach demonstrates significant superiority over other models, while simultaneously outperforming them in terms of energy consumption. Notably, the SVM$_{nlp}$ approach exhibits F1 values approximately 10 points higher than both BERT and DistilBERT, as well as 3 points higher than LegalBERT. Importantly, these performance advantages are achieved while maintaining energy savings of approximately 2 times compared to DistilBERT and 15-20 times compared to LegalBERT and BERT, respectively. In this dataset, the generative models showed extremely poor performance. The energy needs were not particularly high but these models did not prove particularly suitable for dealing with the texts in this dataset. The experiments were repeated several times to ensure the validity of the poor results obtained.

		mF1	MF1	KWh	€	CO2
SCOTUS	**SVM$_{bow}$**	0.78	0.69	× 1.33	× 1.33	**1.00**
	SVM$_{nlp}$	**0.79**	**0.70**	**1.00**	**1.00**	× 1.29
	BERT	0.68	0.58	× 19.36	× 19.36	× 6.82
	LegalBERT	0.76	0.67	× 15.10	× 15.10	× 5.31
	DistilBERT	0.68	0.57	× 1.95	× 1.95	× 1.69
	GPT2 Large	0.36	0.15	× 8.76	× 8.76	× 3.08
	LLAMA2 7B	0.34	0.10	× 9.54	× 9.54	× 3.36
	LLAMA2 13B	0.35	0.19	× 61.29	× 61.29	× 21.58

Table 5: Classification performances and the energy consumption results of different models on SCOTUS dataset.

Unfair ToS

Finally, Table 6 presents the results obtained from evaluating the Unfair Terms of Services (Unfair ToS) dataset [12]. Unfair ToS is the smallest dataset within the LexGLUE benchmark. The tests demonstrate that the SVM$_{bow}$ model achieves optimal energy savings while maintaining performance levels very close to the best models. However, noteworthy competition arises from the SVM$_{nlp}$ model, which showcases comparable performance and energy savings. Although BERT-based models deliver superior performance, concerns arise regarding their energy consumption, which averages around 30 times and 60 times higher compared to the SVM$_{nlp}$ approach and SVM$_{bow}$ model, respectively. In this dataset, on the contrary, the generative models demonstrated excellent results in text classification, returning the best performances (the same as the BERT-based models). However, even in this case, energy consumption proved to be extremely high, with factors up to thousands of times.

		mF1	MF1	KWh	€	CO2
Unfair-ToS	**SVM$_{bow}$**	0.95	0.79	**1.00**	**1.00**	**1.00**
	SVM$_{nlp}$	0.95	0.80	× 1.81	× 1.81	× 2.41
	BERT	**0.96**	0.81	× 112.33	× 112.33	× 52.62
	LegalBERT	**0.96**	0.83	× 84.15	× 84.15	× 39.42
	DistilBERT	**0.96**	0.80	× 54.36	× 54.36	× 25.46
	GPT2 Large	0.95	0.63	× 455.27	× 455.27	× 213.24
	LLAMA2 7B	**0.96**	0.82	× 1474.58	× 1474.58	× 690.67
	LLAMA2 13B	**0.96**	**0.84**	× 3858.71	× 3858.71	× 1807.35

Table 6: Classification performances and the energy consumption results of different models on Unfair-ToS dataset.

4.2 The “in production” scenario

Upon completing the research and development phase, which involves iterative model selection through training-validation-test iterations, a final solution is chosen for deployment in the production environment. The production phase represents the concluding stage of the machine learning lifecycle within the industry. The selected model is executed with high frequency for analyzing a continuous stream of documents and generating predictions. In our analysis, we specifically aimed to compare the energy requirements of different models when employed in the production step. For each model and dataset, we conducted investigations using a standardized set of documents. To represent real-world scenarios, we utilized a sample consisting of 100 documents. The resource requirements in the prediction step were evaluated by randomly selecting 100 documents from the test splits of each dataset within the LexGLUE benchmark. It is important to note that performance values, such as F1 scores, are not available in this particular analysis, as they can only be evaluated during the research and development phase.

The results are reported in Table 7 and we can see how the SVM$_{bow}$ approach exhibits the lowest energy consumption values, thereby resulting in reduced costs and lower carbon footprint. Nonetheless, the SVM$_{nlp}$ model remains an excellent alternative, demonstrating energy consumption levels that range from 2 to 25 times higher than the lightweight SVM$_{bow}$. Conversely, the BERT-based models continue to exhibit remarkably high energy consumption values, with some cases reaching up to x4000 times the energy consumption of a standard SVM$_{bow}$ model. Finally, even in this type of analysis, the generative models continue to show extremely high energy consumption values. In this case, moreover, the returned values have extremely high orders of magnitude compared to both the simple SVM-based models and the more complex BERT-based models.

4.3 Final considerations

Taking into account the evidence that emerged both in research and development (R&D) phase and the ”in-production” scenario, the SVM$_{nlp}$ approach emerges as a formidable contender, striking a perfect equilibrium between performance (with F1 scores in close proximity to BERT-based models) and sustainability (exhibiting optimal energy consumption and CO_2 emissions comparable to the baseline SVM$_{bow}$ model). The investigation indicates that in many real cases, the use of extremely complex models is not automatically reflected in an optimal choice. In fact, they report results very close to those obtained with much simpler (and in some cases inferior) models but with significantly high energy consumption (and therefore costs and CO2 emissions). These results bring into question the justification of employing significantly more energy for marginal performance improvements. Despite the growing attention of public opinion towards the issues of energy saving and emissions, the importance of eco-friendly Machine Learning (ML) has not received the recognition it deserves. The current trend of focusing on larger deep neural networks

	Model	Time	KWh	€	CO2
ECtHR A	**SVM$_{bow}$**	**~ 0.5 sec**	**1.00**	**1.00**	**1.00**
	SVM$_{nlp}$	× 20.26	× 2.70	× 2.70	× 5.77
	BERT	× 42.13	× 577.06	× 577.06	× 55.37
	LegalBERT	× 43.63	× 576.81	× 576.81	× 55.35
	DistilBERT	× 25.79	× 342.32	× 342.32	× 32.85
	GPT2 Large	× 51.22	× 712.52	× 712.52	× 68.37
	LLAMA2 7B	× 60.90	× 762.77	× 762.77	× 44.53
	LLAMA2 13B	× 108.34	× 1530.8	× 1530.8	× 146.9
ECtHR B	**SVM$_{bow}$**	**~ 0.43 sec**	**1.00**	**1.00**	**1.00**
	SVM$_{nlp}$	× 20.06	× 2.68	× 2.68	× 5.73
	BERT	× 48.46	× 665.10	× 665.10	× 63.82
	LegalBERT	× 47.50	× 649.21	× 649.21	× 62.30
	DistilBERT	× 29.48	× 394.62	× 394.62	× 37.87
	GPT2 Large	× 58.41	× 660.76	× 660.76	× 63.40
	LLAMA2 7B	× 86.03	× 1178.4	× 1178.4	× 113.07
	LLAMA2 13B	× 132.4	× 1656.7	× 1656.7	× 158.97
EUR-LEX	**SVM$_{bow}$**	**~ 0.10 sec**	**1.00**	**1.00**	**1.00**
	SVM$_{nlp}$	× 26.87	× 2.14	× 2.14	× 6.36
	BERT	× 131.95	× 483.03	× 483.03	× 64.61
	LegalBERT	× 134.61	× 533.24	× 533.24	× 71.33
	DistilBERT	× 123.23	× 337.77	× 337.77	× 45.18
	GPT2 Large	× 296.99	× 2472.90	× 2472.90	× 330.80
	LLAMA2 7B	× 473.02	× 3955.51	× 3955.51	× 529.12
	LLAMA2 13B	× 639.35	× 5343.67	× 5343.67	× 714.81
LEDGAR	**SVM$_{bow}$**	**~ 0.02 sec**	**1.00**	**1.00**	**1.00**
	SVM$_{nlp}$	× 64.88	× 5.11	× 5.11	× 15.21
	BERT	× 711.90	× 2523.67	× 2523.67	× 337.59
	LegalBERT	× 741.44	× 2640.76	× 2640.76	× 353.25
	DistilBERT	× 656.83	× 1743.25	× 1743.25	× 233.19
	GPT2 Large	× 1951.19	× 15998.4	× 15998.4	× 2140.09
	LLAMA2 7B	× 2424.72	× 20029.2	× 20029.2	× 2679.27
	LLAMA2 13B	× 3114.44	× 21267.1	× 21267.1	× 2844.86
SCOTUS	**SVM$_{bow}$**	**~ 1.43 sec**	**1.00**	**1.00**	**1.00**
	SVM$_{nlp}$	× 55.27	× 4.40	× 4.40	× 13.10
	BERT	× 8.45	× 32.71	× 32.67	× 4.38
	LegalBERT	× 9.20	× 34.72	× 34.72	× 4.64
	DistilBERT	× 7.78	× 21.45	× 21.45	× 2.87
	GPT2 Large	× 17.47	× 70.64	× 70.64	× 9.45
	LLAMA2 7B	× 27.01	× 144.29	× 144.29	× 19.30
	LLAMA2 13B	× 38.03	× 236.83	× 236.83	× 31.68
Unfair-ToS	**SVM$_{bow}$**	**~ 0.01 sec**	**1.00**	**1.00**	**1.00**
	SVM$_{nlp}$	× 91.07	× 7.66	× 7.66	× 22.79
	BERT	× 1610.41	× 3765.22	× 3765.22	× 503.67
	LegalBERT	× 1769.35	× 4112.35	× 4112.35	× 550.10
	DistilBERT	× 1549.53	× 3381.16	× 3381.16	× 452.29
	GPT2 Large	× 4443.59	× 38463.7	× 38463.7	× 5145.23
	LLAMA2 7B	× 5796.96	× 50807.2	× 50807.2	× 6796.39
	LLAMA2 13B	× 8902.85	× 78603.5	× 78603.5	× 10514.67

Table 7: Comparison of time and energy consumption of the models for each dataset in the production scenario.

should also take into account the energy consumption and ecological impact, which are vital aspects of this shift in paradigm. The findings presented in this study have the potential to motivate machine learning researchers to integrate environmental analyses as crucial elements of their research endeavors.

5. Conclusions

This paper presents a comprehensive comparative study of different text classification models in a specific domain, examining their performance (F1 scores), energy consumption (KWh), costs (€), and carbon footprint (CO_2) metrics. The chosen domain of focus is the "legal" area, with the evaluation conducted using the LexGLUE benchmark dataset, consisting of seven legal domain-specific datasets. For the investigation, three widely utilized families of models in text classification are considered: classic Support Vector Machines (SVMs), Large Language Models (LLM) and Generative Models (GenLLM).

For the SVM-based approaches, a linear SVM model is employed alongside two distinct feature representations: the classic Bag-Of-Word approach (SVM$_{bow}$), and an advanced representation enriched with linguistic and semantic features (SVM$_{nlp}$). For the NLP analysis required in the SVM$_{nlp}$ model, we employed the expert.ai hybrid natural language platform which consists in an integrated environment for deep language understanding and provides a complete natural language workflow with end-to-end support for annotation, labeling, model training, testing and workflow orchestration [15]. From the LLM-based models, three BERT-based models were selected: BERT, LegalBERT, and DistilBERT. Finally, for the generative models, we considered GPT-2 and two models from the LLAMA2 family: LLAMA2 7b and LLAMA2 13b.

The objective of this study was to examine the trade-off between performance and economic and ecological aspects of various text categorization approaches when applied in a real-world context. To accomplish this, we conducted two distinct types of investigations. Firstly, we explored a research and development (R&D) scenario where we followed a standard procedure involving training, validation, and testing phases. Secondly, we delved into the ”in production” scenario, where the selected model was deployed and continually utilized to analyze a continuous stream of documents and generate predictions.

The findings of the study reveal that adopting simple approaches can achieve performance comparable to Large Language Models (LLMs) and Generative models, over the majority of LexGLUE datasets, while simultaneously yielding substantial energy savings and reducing CO_2 emissions. These results bring into question the justification of employing significantly more energy for marginal performance improvements. Considering the outcomes from both scenarios, it becomes evident that often simpler SVM-based models offer an exceptional solution. It strikes an ideal balance between performance (with F1 scores closely rivaling those of BERT-based and Generative models) and considerations related to cost and ecological compatibility, allowing for significant energy savings and optimal resource utilization.

Despite the existence of collaborative research on this topic, as mentioned in the literature [19], the significance of eco-friendly Machine Learning (ML) has not garnered the attention it truly deserves. The prevailing trend towards larger deep neural networks should encompass considerations of energy consumption and the ecological impact, which are crucial aspects of this paradigm shift. The results showcased in this study hold the potential to inspire machine learning researchers to incorporate environmental analyses as integral components of their research activities.

Funding

The work was partially funded by:

- “enRichMyData - Enabling Data Enrichment Pipelines for AI-driven Business Products and Services”, an Horizon Europe (HE) project, grant agreement ID: 101070284 [16];
- “IbridAI - Hybrid approaches to Natural Language Understanding”, a project financed by the Regional Operational Program “FESR 2014-2020” of Emilia Romagna (Italy), resolution of the Regional Council n. 863/2021.

[15] https://www.expert.ai/products/expert-ai-platform/

[16] https://doi.org/10.3030/101070284

References

[1] Stephan Bloehdorn and Andreas Hotho. Boosting for text classification with semantic features. In *Web Mining and Web Usage Analysis*, 2004. *Cited on page(s):* 5

[2] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy, July 2019. Association for Computational Linguistics. *Cited on page(s):* 5, 8

[3] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. *Cited on page(s):* 5

[4] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics. *Cited on page(s):* 6

[5] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland, May 2022. Association for Computational Linguistics. *Cited on page(s):* 4, 5

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. *Cited on page(s):* 5

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. *Cited on page(s):* 6

[8] Leonidas Gee, Leonardo Rigutini, Marco Ernandes, and Andrea Zugarini. Multi-word tokenization for sequence compression. In Mingxuan Wang and Imed Zitouni, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 612–621, Singapore, December 2023. Association for Computational Linguistics. *Cited on page(s):* 2

[9] Leonidas Gee, Andrea Zugarini, Leonardo Rigutini, and Paolo Torroni. Fast vocabulary transfer for language model compression. In *The 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, UAE, 12 2022. *Cited on page(s):* 2

[10] Sinan Gultekin, Achille Globo, Andrea Zugarini, Marco Ernandes, and Leonardo Rigutini. An energy-based comparative analysis of common approaches to text classification in the legal domain. In Dhinaharan Nagamalai (Eds) David C. Wyld, editor, *The 4th International Conference on NLP & Text Mining (NLTM 2024)*, volume 14, Copenhagen, Denmark. *Cited on page(s):* 2, 3

[11] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *International conference on machine learning*, pages 1737–1746. PMLR, 2015. *Cited on page(s):* 2

[12] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: An automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139, 2019. *Cited on page(s):* 5, 10

[13] Kadan Lottick, Silvia Susai, Sorelle A. Friedler, and Jonathan P. Wilson. Energy usage reports: Environmental awareness as part of algorithmic accountability. *CoRR*, abs/1911.08354, 2019. *Cited on page(s):* 3, 7

[14] Sasha Luccioni, Victor Schmidt, Alexandre Lacoste, and Thomas Dandres. Quantifying the carbon emissions of machine learning. In *NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning*, 2019. *Cited on page(s):* 3

[15] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy

Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report. Technical report, 2023. *Cited on page(s):* 7

[16] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI Blog, 2018. *Cited on page(s):* 6

[17] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. *Cited on page(s):* 7

[18] Leonardo Rigutini. *Automatic Text Processing: Machine Learning Techniques*. LAP LAMBERT Academic Publishing, 07 2010. Saarbrücken, DE. isbn: 978-3-8383-7452-9 Book. *Cited on page(s):* 5

[19] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Körding, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer T. Chayes, and Yoshua Bengio. Tackling climate change with machine learning. *CoRR*, abs/1906.05433, 2019. *Cited on page(s):* 2, 13

[20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. *Cited on page(s):* 2

[21] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. *Cited on page(s):* 6

[22] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CoRR*, abs/1907.10597, 2019. *Cited on page(s):* 3

[23] Sam Scott and Stan Matwin. Feature engineering for text classification. In *International Conference on Machine Learning*, 1999. *Cited on page(s):* 5

[24] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, mar 2002. *Cited on page(s):* 5

[25] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203, 2019. *Cited on page(s):* 7

[26] Harold J. Spaeth, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2020 supreme court database, version 2020 release 1. *Cited on page(s):* 5, 10

[27] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 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, July 2019. Association for Computational Linguistics. *Cited on page(s):* 2

[28] 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. Llama: Open and efficient foundation language models. Technical report, 2023. *Cited on page(s):* 7

[29] 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. Llama 2: Open foundation and fine-tuned chat models. Technical report, 2023. *Cited on page(s):* 7

[30] Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France, May 2020. European Language Resources Association. *Cited on page(s):* 5, 9

[31] Alex K. S. Wong, John W. T. Lee, and Daniel S. Yeung. Use of linguistic features in context-sensitive text classification. In *International Conference on Machine Learning and Computing*, 2005. *Cited on page(s):* 5

[32] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. *CoRR*, abs/2104.08671, 2021. *Cited on page(s):* 5

[33] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017. *Cited on page(s):* 2