The availability of compute and data to train larger and larger language models increases the demand for robust methods of benchmarking the true progress of LM training. Recent years witnessed significant progress in standardized benchmarking for English. Benchmarks such as GLUE, SuperGLUE, or KILT have become a de facto standard tools to compare large language models. Following the trend to replicate GLUE for other languages, the KLEJ benchmark has been released for Polish. In this paper we evaluate the progress in the field of benchmarking for under-resourced languages. We note that only a handful of languages have such comprehensive benchmarks. We also note the gap in the number of tasks being evaluated by benchmarks for resource-rich English/Chinese and the rest of the world.

In this paper we introduce LEPISZCZE, a new, comprehensive benchmark for Polish NLP with a large variety of tasks and high-quality operationalization of the benchmark. We design LEPISZCZE with flexibility in mind. The inclusion of new models, datasets, and tasks is as simple as possible, while still offering data versioning and model tracking. In the first run of the benchmark, we test 13 experiments (task and dataset pairs) based on the five most recent LMs for Polish. We use five datasets from the Polish benchmark and add eight novel datasets. As the main contribution of the paper, apart from LEPISZCZE, we provide insights and experiences learned while creating the benchmark for Polish as the blueprint to design similar benchmarks for other under-resourced languages.

---

1 klej is the word for glue in Polish
2 lepiszcze is the Polish word for glew, the Middle English predecessor of glue
1 Introduction

Lack of reproducibility is an infuriating problem in machine learning practice. The inability to reproduce evaluation results and conduct reliable model comparisons is usually related to poor code quality, unclear and cryptic selection of hyper-parameter values, the random introduction of multiple factors affecting classification performance, and lack of a well-defined evaluation protocol [Pineau et al. (2021)]. These problems can be circumvented by encouraging people to use standardized and peer-reviewed evaluation environments. The rapid development of diverse language technology has increased the need for reliable evaluation environments.

The reproducibility issues are intensifying even stronger as more novel language models emerge each year. We have seen a remarkable progress on many language understanding tasks, from language modeling [Brown et al. (2020); Rae et al. (2021); Hoffmann et al. (2022)], Named Entity Recognition [Li et al. (2020); Ye et al. (2022)], Q&A [Lan et al. (2020); Yang et al. (2019)], or various text classification tasks [Peters et al. (2018); Bingyu and Arefyev (2022)] in recent years. Moreover, in the last decade, data-centric models have become a major direction in solving most problems in the NLP area. Researchers and industry experts focus more on curated datasets and their maintenance processes. Hence, benchmarking models based on many datasets and their various and constantly changing versions is a great challenge.

![Figure 1: Tasks in large NLU benchmarks vs number of speakers, with Indian languages grouped together for readability.](image)

As shown in Figure 1, most NLP benchmarks are written for well-resourced languages such as English and Chinese. This is understandable because many datasets exist in these languages, and many research teams are working on these problems. Besides English and Chinese, languages thoroughly covered with benchmarks include Indian, Spanish, French, and Portuguese. These languages are one of the most commonly used languages in the world; hence their position in the ranking is not surprising. However, we can also find Arabic and Japanese, which are widely spoken languages, but surprisingly few tasks are covered in benchmarks for these languages. Finally, we have languages with some benchmarks, such as Romanian, Persian, Dutch or Polish, but they only cover the most basic NLP tasks.

In this work, we focus on Polish and aim to provide datasets and tools to facilitate research on Polish NLP tasks. We designed the benchmarking process so that it could be easily applied to other languages. Thus, preparing and adding benchmarks for other low-resourced languages should become much less laborious.

The Polish benchmarking tradition has a relatively short history. One of the few platforms for evaluating and comparing modern language models for Polish is the KLEJ benchmark [Rybak et al. (2020)], a single-metric benchmark defined over a limited dataset. This traditional practice for evaluating a model’s performance no longer works. Current recommendations for the comparative evaluation of LMs advocate the inclusion of diversified tasks, challenges, and tests. Hence, we wanted to rethink and design a benchmark and environment to assess models so that they can still serve as valuable progress indicators.
Our main contributions are as follows:

• We propose LEPISZCZE, a new, extensive benchmark for Polish NLP with a large variety of tasks,
• We design the benchmark and its maintenance using the best practices found in the literature, we also investigate some of the most problematic aspects of creating benchmarks,
• We extended previous Polish benchmark KLEJ with eight new datasets, published as a unified modern API,
• We share the lessons we have learned while building a benchmark, especially for other low-resource languages,
• We present the summary of training and evaluation of more than 6000 different models for LEPISZCZE, storing all information about code, dataset versions, parameters, metrics, predictions, or even information about their experimental environment.

2 Related Work

We used Google Scholar to review available NLP benchmarks for languages with at least 10 million speakers, without going deeper into dialects (i.e. German includes all German dialects without dividing into Standard or Bavarian German). We searched for <language name> NLP|NLG benchmark. Furthermore, we dismissed benchmarks consisting of a single data set, as a result we found 38 benchmarks, created by Seelawi et al. (2021); Zhang et al. (2021); Xu et al. (2020); Yao et al. (2021); McCann et al. (2018); Kiela et al. (2021); Chen et al. (2022); Xu et al. (2021); Le et al. (2020); Gehrman et al. (2021); Liu et al. (2021); Wang et al. (2018); Canete et al. (2022); Kakwani et al. (2020); Koto et al. (2020); Cahyawijaya et al. (2021); Wilie et al. (2020); Petroni et al. (2021); Rybak et al. (2020); Park et al. (2021); Kim et al. (2022); Dumitrescu et al. (2021); Guan et al. (2022); Safaya et al. (2022); Khashabi et al. (2021); Farahani et al. (2021); Gomez (2020); Blinov et al. (2022); Shavrina et al. (2020); Zagar and Robnik-Šikonja (2022); Conneau and Kiela (2018); Wang et al. (2019); Liang et al. (2020); Hu et al. (2020); Wang et al. (2022) (in alphabetical order of benchmark name), included a total of 66 different tasks. Out of those, only 34 appeared in one language, these can be divided into two groups: specialized tasks which required a larger effort to build a good data set (like diagnosis normalization, see Wang et al. (2020)), or misdefined tasks such as Named Entity Recognition in Polish KLEJ benchmark, which was not a span labeling task, but rather a text fragment classification task to detect if it contains an entity, without providing the span. We provide more detailed results of our survey in the supplementary materials.

When it comes to language coverage, only 31 languages have an existing NLP or NLG benchmark, out of 91 available in the 2022 edition of Ethnologue? Arabic, Assamese, Bengali, Chinese, Dutch, English, French, German, Gujarati, Hindi, Indonesian, Italian, Japanese, Kannada, Korean, Malayalam, Marathi, Odia, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Spanish, Swahili, Tamil, Telugu, Turkish, Urdu. The 74 tasks were not equally distributed per language, per Figure 1 Benchmarks for the two most commonly spoken languages: English and Chinese would cover around 40 tasks, while the languages with both lowest number of tasks available in benchmarks and lower numbers of native speakers were Romanian and Polish (around 10 tasks). The results of our analysis are attached in Appendix.

The situation where Polish has a disproportionate small number of tasks given nearly 40 million native speakers in its main NLP benchmarks would in itself be a good reason to expand the benchmark. However, once we take a deeper look at how tasks are formulated in KLEJ, we must acknowledge that, the number of tasks formulated in a manner established in a given NLP subfield is even smaller. The already mentioned NER task in KLEJ is not a sequence tagging, but a document classification task. Summarization in KLEJ is evaluated based on classifying pairs of text and summary, the task is to predict whether the summary summarizes the text, most benchmarks would define this as an NLG task where the model is expected to generate the summary and would be evaluated with ROGUE or BLEU measures. A similar situation is happening in the Q&A task. We thus consider KLEJ to provide only 9 tasks, marking Polish the least task-covered European language with respect to modern NLP and NLG benchmarked tasks.
3 LEPISZCZE

LEPISZCZE (IPA: [lɛpˈʂɛʃɛ]) is an open-source benchmark and a continuous-submission leaderboard, concentrating public Polish datasets (existing and new) in specific tasks. The integration of datasets and tasks with model performance and efficiency allows both academia and industry to quickly gauge performance on tasks of interest. Finally, it intends to foster a constructive competition and innovation by bringing together and promoting previously disparate resources.

Our benchmark is structured into datasets, tasks, and models. We designed the benchmark to be easy extendable and to be possible in future to build a leaderboards for various subsets of dataset, task, model triplets. Hence, we can evaluate for example what is the best static word embedding representation for SVM or XGBoost model for sequence tagging task for subset of datasets.

3.1 Lessons learned for the process of creating the LEPISZCZE benchmark

As the prerequisite for LEPISZCZE creation, we gathered some of the most important and repeated in the NLP community aspects for benchmarking.

3.1.1 Datasets for benchmark

The benchmark is as good as the selection of datasets and tasks is. If an unrepresentative collection of data and tasks is used to create a benchmark, the evaluation is of limited informative value for further development of language models. If a benchmark consists only of closely related datasets, we can evaluate only a narrow part of the model’s capabilities. Hence, one of the first and the most critical task for us was to gather a large number of diverse datasets for Polish. Benchmarking involves running many experiments and tracking their performance; Therefore, we unified all datasets into one, accessible and easy to process data format. We uploaded datasets to the HuggingFace Datasets repository and ran all experiments using the HuggingFace hub. We wanted to cover also many sources of text data in our benchmarking environment. The model’s performance could be different for books, social media, and other domain texts. Thus, having a representative collection of text data allows to evaluate the models in terms of their in-domain and out-of-domain generalization abilities.

Integrating benchmark libraries with HuggingFace Datasets platform opens new possibilities to evaluate language models in a multilingual zero-shot setting for any low-resource language. We believe using a unified dataset inventory will contribute to sustainable development of reliable evaluation data. Technically, our benchmark allows to choose any dataset or collect utterly new data, prepare data loading scripts compatible with HuggingFace platform and evaluate models in the target language.

We noticed an interesting problem when extending our collection of datasets covered in the benchmark. Should we add a dataset that is not free and publicly available? We added to the benchmark Dialogue Acts dataset — DiaBiz.Kom (more in Section 3.2) that is available only for internal usage of CLARIN-PL-Biz associates. Still, the dataset covers a significant collection of infrequent domain data for Polish, that is targeted on spoken language understanding. The results of modern language models for this dataset present much room for improvement — see Table 3. The restricted access to this dataset means that we must conduct as many experiments as possible using this dataset. However, the question is whether we should add more non publicly open datasets to the benchmark in the future.

Metrics matter Every benchmark has to provide task specific evaluation metrics. Even though we can focus on a single metric for a specific dataset and task in most cases, it could not be enough for many scenarios. Imagine a classification task, then accuracy or F-score metrics look like an obvious choice, but the classification of strong negative and negative is not as important as mixing strong negative with strong positive. The same could be with using F-score for sequence tagging problems such as named entity recognition. Do we check count positive prediction only when it is 100% overlap between prediction and gold label or lower coverage also could be correct?

Interestingly, we may want to use the benchmark to guess what model would be a good choice for our new internally created dataset. We wish to investigate not a single metric but many of them or weight them. We designed the LEPISZCZE so that many different metrics are calculated. We allow to sort https://huggingface.co/datasets
submissions based on various metrics. We also store raw predictions if we would like to add more metrics in the future without recomputing whole models. Finally, we track all python environment information and times of completion of all, ran by us experiments for the benchmark. Hence, we could build a custom leaderboard mixing model and environment metrics in the future. As Ethayarajh and Jurafsky (2020) said, a highly inefficient model would provide less utility to practitioners but not to a leaderboard since it is a cost that only the former must bear.

3.1.2 Standard splits problems

Many benchmarks such as GLUE and derived work do not revile test sets the benchmarking platform calculates the final results. Using static data splits leads to over-fitting and results in quick benchmark saturation. Another way worth to consider is based on multiple splits Gorman and Bedrick (2020) allowing to evaluate model’s performance based on many different data partitions. In our evaluation pipelines we follow the same idea and we use not only a single train, dev and test split but also potentially many different ones. In our benchmark, we decided to implement a new experimentation procedure to evaluate other splits in the next benchmark version.

3.1.3 Continuous benchmarking

The disadvantage of multi-task benchmarks such as GLUE, SuperGLUE, KLEJ, etc., is their lack of dynamics. The static benchmarks become quickly outdated, and therefore useless from a practical perspective. As part of the CLARIN-PL-Biz we plan to add more datasets, tasks, models and maintain LEPISZCZE benchmark continuously. We encourage other associated researchers to publish datasets and models in our benchmark. Many of those added by us to the LEPISZCZE datasets have been added together with their source and author contribution. Moreover, we track all model parameters and version of the dataset. Hence, there is a possibility to create a leaderboard for a specific version of the dataset in our benchmark.

3.2 Datasets in the benchmark

We started work on the LEPISZCZE benchmark by gathering datasets. Some initial datasets were published on the KLEJ benchmark. Many datasets have been described in their own research papers, but they were still quite hard to obtain, and of course, they were in different formats. We started gathering them and unifying them to fit the same form. We also decided to upload them to HuggingFace Datasets hub and make them as easy as it is possible to reuse and to version them.

**PAC — Polish Abusive Clauses Dataset** "I have read and agree to the terms and conditions" is one of the biggest lies on the Internet. Consumers rarely read the contracts they are required to accept. We conclude agreements over the Internet daily. But do we know the content of these agreements? Do we check potential unfair statements? On the Internet, we probably skip most of the Terms and Conditions. However, we must remember that we have concluded many more contracts. Imagine that we want to buy a house, a car, send our kids to the nursery, open a bank account, or many more. In all these situations, you will need to conclude the contract, but there is a high probability that you will not read the entire agreement with proper understanding. European consumer law aims to prevent businesses from using so-called "unfair contractual terms" in their unilaterally drafted contracts, requiring consumers to accept. The PAC aims to treat "unfair contractual term" as the equivalent of an abusive clause. The dataset has been created with the Office of Competition and Consumer Protection. This dataset used more than 700 contracts and gathered 4,129 examples of abusive clauses and 5,127 non-abusive contract fragments.

**AspectEmo Corpus** Kocoen et al. (2021) is an extended version of the publicly available PolEmo 2.0 corpus. The AspectEmo corpus consists of 1,465 online customer reviews from the following domains: school, medicine, hotels, and products. All documents are annotated at the aspect level with six sentiment categories: strong negative, weak negative, neutral, weakly positive, and strong positive.

5 https://ukik.gov.pl/home.php
6 https://huggingface.co/datasets/laugustyniak/abusive-clauses-pl
7 https://huggingface.co/datasets/clarin-pl/aspectemo
Compositional Distributional Semantics Corpus (CDSC-E) is an entailment classification task. It consists of 1000 pairs of sentences and human-annotated entailment labels for each pair. There are three possible classes: entailment, contradiction, and neutral.

Dialogue Acts — Diabiz.Kom It consists of 1,277,962 tokens in 1,104 transcribed call center phone conversations spanning eight domains. Each example is annotated by three linguists (in a 2+1 system, with an inter-annotator agreement of 0.78 for annotation borders and categories and 0.86 for annotation borders) with dialogue acts in compliance with ISO 64217-2:2012 standard with a layer of information concerning communicative functions. Within the benchmark, we consider the task of dialogue act classification, where each utterance is provided with its role in the dialogue. Sample and a detailed description can be obtained from [https://clarin-pl.eu/dspace/handle/11321/886](https://clarin-pl.eu/dspace/handle/11321/886) and corresponding recordings from [https://clarin-pl.eu/dspace/handle/11321/887](https://clarin-pl.eu/dspace/handle/11321/887). Diabiz.Kom is annotation layer on top of DiaBiz Pézik et al. (2022) — corpus of polish call center dialogs.

DYK Did You Know (pol. Czy wiesz?) is a dataset that consists of 4,721 human-annotated question-answer pairs. It was simplified by Rybak et al. (2020) to binary classification to label denoting if the answer contained in Wikipedia article is factually correct in light of stated question.

KPWr-NER is a part the Polish Corpus of Wrocław University of Technology (Korpus Języka Polskiego Politechniki Wrocławskiej) Broda et al. (2012) is a named entity recognition dataset focusing on fine-grained categories of entities (82 classes) using BIO notation. It contains 13,959 training and 4,323 test human annotated sentences, originating from texts covering large variety of domains, genres and sources.

NKJP-POS is a part the National Corpus of Polish (Narodowy Korpus Języka Polskiego) Przepiórkowski et al. (2012). Its objective is the part-of-speech tagging task. The dataset contains 85,663 sentences tagged with 35 tags. During the creation of corpus, texts were annotated by humans from various sources, covering many domains and genres.

PolEmo 2.0 Kocoń et al. (2019) is a dataset of online consumer reviews from four domains: medicine, hotels, products, and university. It consists of 8,216 reviews having 57,466 sentences. The aim is to predict one of the sentiment classes: positive, negative, neutral, or ambiguous. During the development of KLEJ benchmark Rybak et al. (2020) two tasks that differ in context used during evaluation have been created: in-domain and out-domain. In contrast, we preserved original data split and utilised all of the domains.

Political Advertising dataset Augustyniak et al. (2020) aims for detecting specific text chunks and categories of political advertising in the Polish language. It contains 1,705 human-annotated tweets tagged with nine categories, constituting campaigning under Polish electoral law. The authors achieved 0.65 inter-annotator agreement (Cohen’s kappa score) for the sequence tagging task, and they used an additional annotator to resolve the mismatches between the first two annotators, improving the final consistency of annotations.

PSC Polish Summaries Corpus Ogrodniczuk and Kopeć (2014) consist of 569 news summaries done by human annotators. It was simplified by Rybak et al. (2020) for the purpose of the KLEJ development. They formulated a text-similarity task by matching positive and negative pairs using procedure detailed in publication.

[https://huggingface.co/datasets/allegro/klej-cdsc-e](https://huggingface.co/datasets/allegro/klej-cdsc-e)
[https://huggingface.co/datasets/allegro/klej-dyk](https://huggingface.co/datasets/allegro/klej-dyk)
[https://huggingface.co/datasets/clarin-pl/kpwr-ner](https://huggingface.co/datasets/clarin-pl/kpwr-ner)
[https://huggingface.co/datasets/clarin-pl/nkjp-pos](https://huggingface.co/datasets/clarin-pl/nkjp-pos)
[https://huggingface.co/datasets/clarin-pl/pol emo2-official](https://huggingface.co/datasets/clarin-pl/pol emo2-official)
[https://huggingface.co/datasets/laugustyniak/political-advertising-pl](https://huggingface.co/datasets/laugustyniak/political-advertising-pl)
[https://huggingface.co/datasets/allegro/klej-psc](https://huggingface.co/datasets/allegro/klej-psc)
Punctuation Restoration is a crowdsourced text and audio data set of Polish Wikipedia pages read out loud by Polish lectors. The base dataset is divided into conversational (WikiTalks) and information (WikiNews) parts. Then the texts were read by hundred people, which resulted in 36 hours of transcription. Whole dataset contain 32 thousands texts. Punctuation restoration includes only 1000 texts - 800 train and 200 test examples. This dataset is part of PolEval 2021 Competition.

Table 1: Datasets available in the LEPISZCZE benchmark with sizes of train, dev and test sets. The datasets that were previously incorporated into KLEJ benchmark were marked with * symbol.

<table>
<thead>
<tr>
<th>Name</th>
<th>Domain</th>
<th>Task</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>#Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDSC-E*</td>
<td>image captions</td>
<td>Entailment Classification</td>
<td>8000</td>
<td>1000</td>
<td>1000</td>
<td>3</td>
</tr>
<tr>
<td>DYK*</td>
<td>Wikipedia</td>
<td>Q&amp;A Classification</td>
<td>4154</td>
<td>0</td>
<td>1029</td>
<td>2</td>
</tr>
<tr>
<td>PolEmo 2.0 (In-Domain)*</td>
<td>online reviews</td>
<td>Sentiment Analysis</td>
<td>5783</td>
<td>723</td>
<td>722</td>
<td>4</td>
</tr>
<tr>
<td>PolEmo 2.0 (Out-Domain)*</td>
<td>online reviews</td>
<td>Sentiment Analysis</td>
<td>5783</td>
<td>494</td>
<td>494</td>
<td>4</td>
</tr>
<tr>
<td>PSC*</td>
<td>online reviews</td>
<td>Paraphrase Classification</td>
<td>4302</td>
<td>0</td>
<td>1078</td>
<td>4</td>
</tr>
<tr>
<td>Abusive Clauses</td>
<td>legal texts</td>
<td>Text Classification</td>
<td>4284</td>
<td>1519</td>
<td>3453</td>
<td>2</td>
</tr>
<tr>
<td>AspectEmo</td>
<td>online reviews</td>
<td>Aspect-based Sentiment Analysis</td>
<td>1173</td>
<td>0</td>
<td>292</td>
<td>7</td>
</tr>
<tr>
<td>KPWr NER</td>
<td>misc.</td>
<td>NER</td>
<td>13959</td>
<td>0</td>
<td>4323</td>
<td>82</td>
</tr>
<tr>
<td>NKJP POS</td>
<td>misc.</td>
<td>POS Tagging</td>
<td>78219</td>
<td>0</td>
<td>7444</td>
<td>35</td>
</tr>
<tr>
<td>PolEmo 2.0</td>
<td>online reviews</td>
<td>Sentiment Analysis</td>
<td>6573</td>
<td>823</td>
<td>820</td>
<td>4</td>
</tr>
<tr>
<td>Political Advertising</td>
<td>social media</td>
<td>Sequence Tagging</td>
<td>4284</td>
<td>1519</td>
<td>3453</td>
<td>4</td>
</tr>
<tr>
<td>Punctuation Restoration</td>
<td>Wikipedia Talk, Wikinews</td>
<td>Punctuation Restoration</td>
<td>800</td>
<td>0</td>
<td>200</td>
<td>8</td>
</tr>
<tr>
<td>Diabiz.Kom</td>
<td>call center phone conversations</td>
<td>Dialogue Acts Classification</td>
<td>70454</td>
<td>8807</td>
<td>8807</td>
<td>54</td>
</tr>
</tbody>
</table>

4 Experiments

We conducted experiments on 13 datasets and five language models. Each language model was fine-tuned for a given dataset and was evaluated separately. Experiments were performed with our developed library clarin1-embeddings, which provides predefined pipelines for text classification, text pairs classification, and sequence labeling. To ensure reproducibility of our experiments, we utilized MLOps tools: Data Version Control (DVC) for pipelines and tracking of datasets and models; Weight&Biases for experiments summaries and metrics tracking. We share the code of our experiments on GitHub repository and model tracking dashboard.

4.1 Initial models for benchmark

We picked four recent transformer-based language models for Polish publicly available in the HuggingFace hub, along with one multilingual XLM-RoBERTa model. We present those models with their total number of parameters and repository location in Table 2. For fine-tuning, we utilize sequence and token classification models from the transformers library (Wolf et al., 2020), consisting of a single linear classification layer with dropout. For the initial datasets evaluation we chose both cased and uncased versions of the language models.

Table 2: Language Models used for experiments. All models can be accessed via HuggingFace repository (huggingface.co).

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params</th>
<th>HuggingFace Repository Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>PolBERT (base, cased), Kłeczek (2020)</td>
<td>132M</td>
<td>dkleczek/bert-base-polish-cased-v1</td>
</tr>
<tr>
<td>PolBERT (base, uncased), Kłeczek (2020)</td>
<td>132M</td>
<td>dkleczek/bert-base-polish-uncased-v1</td>
</tr>
<tr>
<td>HerBERT (base, cased)</td>
<td>124M</td>
<td>allegro/herbert-base-cased</td>
</tr>
<tr>
<td>HerBERT (large, cased)</td>
<td>355M</td>
<td>allegro/herbert-large-cased</td>
</tr>
<tr>
<td>Mroczkowski et al. (2021)</td>
<td>124M</td>
<td>allegro/herbert-large-cased</td>
</tr>
<tr>
<td>XLM-RoBERTa (paraphrase)</td>
<td>278M</td>
<td>sentence-transformers/paraphrase-xlm-r-multilingual-v1</td>
</tr>
</tbody>
</table>

4.2 Hyperparameter search (HPS)

To fairly compare different transformer models across various tasks, we performed a hyperparameter search (HPS) to obtain the best configuration for fine-tuning the language model to a particular task.

We performed a hyperparameter search separately for each combination of tasks and language models (which we restricted to 100 iterations). Under the hood, we utilized the Optuna framework wrapper from the clarinpl-embeddings library. We also logged each run in the hyper-parameter search via Weights&Biases PyTorch Lightning logger.

4.3 Evaluation

Experiments were computed using a server with five Titan RTX GPU cards. We logged over 6000 runs in the Weight&Biases dashboard, which took over 2000 hours to complete. We reported metrics, hyperparameters, dataset information, and package versions in each run.

HPS We used macro averaged F1 measure as the metric for the objective function. Evaluation of models in the hyper-parameter search stage was performed on the validation subset of the dataset. In case whether validation subset was missing, we randomly sampled 10% of the training subset. After obtaining the best hyper-parameter configuration, we no longer need such a subset, so we use original subsets for the final model evaluation.

Model Evaluation For each dataset and language model pair, we choose the best configuration from the HPS process in terms of the best f1-macro score on the validation subset. We retrain models five times and calculate various metrics on test sets such as accuracy, precision, recall, and f1 with different averaging (micro, macro, weighted) and class or tags metrics (accuracy, precision, recall, and f1).

4.4 Results of an initial set of trained models

Evaluation results are presented in Table 3, where we report macro averaged F1 metric for each dataset. As we can observe the performance of models above 80% in text classification datasets (except out-domain dataset), these models perform poorly considering most sequence tagging tasks. Even the best performing model (HerBERT (large), which is also considered as best in the KLEJ benchmark, shows F1-macro around 39% for AspectEmo and 46% in Punctuation Restoration. Considering those results, we can state that we still need better models that can cope with complex and underrepresented tasks. Multilingual model XLM-RoBERTa with HerBERT (large) is comparable only for one dataset (around 2 percentage points difference in means in the Abusive Clauses dataset). However, for other tasks, the gap is much bigger, and it fails even more in mentioned sequence tagging tasks with a difference of up to 70%.

21 Due to the extensive size of Dialogue Acts dataset and the fact that work on this dataset is still in progress, the HPS on this task was ran only for 10 iterations and excluding the HerBERT (large, cased) language model which is the most computationally demanding.  
22 https://www.pytorchlightning.ai
We also reported preliminary results for Dialogue Acts classification tasks. For tested language model we achieved comparable performance about 50%. Considering limited computational resources and fine-tuning scheme, these results may be updated in our future work.

Table 3: F1 Macro performance of evaluated models on the test subsets. We present values as the mean and standard deviations over 5 model retrains. Values marked with Bold present the best results for a single dataset. Additionally, we indicate datasets previously appeared in the KLEJ benchmark with *. WIP denotes the dataset for which we present preliminary results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HerBERT (base, uncased)</th>
<th>HerBERT (large, cased)</th>
<th>PolBERT (base, cased)</th>
<th>PolBERT (base, uncased)</th>
<th>XLM-RoBERTa (paraphrase)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDSC-E*</td>
<td>90.96 ± 0.73</td>
<td>90.48 ± 0.20</td>
<td>88.95 ± 0.31</td>
<td>90.62 ± 0.27</td>
<td>82.62 ± 0.88</td>
</tr>
<tr>
<td>DYK*</td>
<td>82.39 ± 1.43</td>
<td>79.58 ± 0.59</td>
<td>75.87 ± 0.98</td>
<td>74.41 ± 1.15</td>
<td>58.93 ± 7.98</td>
</tr>
<tr>
<td>PolEmo 2.0 In-Domain*</td>
<td>88.10 ± 0.36</td>
<td>88.34 ± 0.63</td>
<td>85.32 ± 0.45</td>
<td>85.71 ± 0.40</td>
<td>83.75 ± 0.45</td>
</tr>
<tr>
<td>PolEmo 2.0 Out-Domain*</td>
<td>57.31 ± 2.93</td>
<td>57.08 ± 2.03</td>
<td>54.10 ± 3.82</td>
<td>54.29 ± 1.83</td>
<td>45.12 ± 3.40</td>
</tr>
<tr>
<td>PSC*</td>
<td>97.90 ± 0.24</td>
<td>98.33 ± 0.69</td>
<td>98.95 ± 0.13</td>
<td>98.87 ± 0.10</td>
<td>58.85 ± 1.49</td>
</tr>
<tr>
<td>Abusive Clauses</td>
<td>85.66 ± 0.58</td>
<td>86.57 ± 0.91</td>
<td>85.93 ± 0.66</td>
<td>85.74 ± 0.86</td>
<td>84.32 ± 0.71</td>
</tr>
<tr>
<td>AspectEmo</td>
<td>37.28 ± 0.71</td>
<td>39.44 ± 1.74</td>
<td>30.01 ± 0.58</td>
<td>31.48 ± 1.06</td>
<td>18.42 ± 0.98</td>
</tr>
<tr>
<td>KPWr NER</td>
<td>54.22 ± 0.76</td>
<td>52.68 ± 1.39</td>
<td>48.01 ± 0.76</td>
<td>40.21 ± 0.50</td>
<td>36.13 ± 0.44</td>
</tr>
<tr>
<td>NKJP POS</td>
<td>94.59 ± 0.56</td>
<td>96.14 ± 0.38</td>
<td>94.34 ± 0.61</td>
<td>94.54 ± 0.19</td>
<td>90.29 ± 0.51</td>
</tr>
<tr>
<td>PolEmo 2.0</td>
<td>86.78 ± 0.79</td>
<td>89.33 ± 0.49</td>
<td>85.89 ± 1.25</td>
<td>85.83 ± 0.47</td>
<td>84.12 ± 0.47</td>
</tr>
<tr>
<td>Political Advertising</td>
<td>61.42 ± 1.38</td>
<td>62.16 ± 0.14</td>
<td>58.94 ± 1.92</td>
<td>62.52 ± 1.23</td>
<td>56.68 ± 0.94</td>
</tr>
<tr>
<td>Punctuation Restoration</td>
<td>45.59 ± 0.38</td>
<td>46.68 ± 0.61</td>
<td>38.89 ± 0.91</td>
<td>41.31 ± 0.59</td>
<td>14.33 ± 1.94</td>
</tr>
<tr>
<td>Dialogue Acts (WIP)</td>
<td>49.54 ± 0.74</td>
<td>×</td>
<td>50.20 ± 1.32</td>
<td>48.87 ± 0.90</td>
<td>49.05 ± 0.39</td>
</tr>
</tbody>
</table>

5 Conclusions and Future Work

In this paper we have introduced LEPISZCZE , a new comprehensive benchmark for Polish NLP. LEPISZCZE is characterized by the large variety of NLP tasks and high-quality operationalization of the benchmark. The benchmarking approach is designed to maximize the flexibility and portability to other under-resourced languages. Adding new models, datasets, or NLP tasks, is simple and intuitive.

The benchmark internally supports data versioning and model tracking for improved reproducibility. In the first run of the benchmark we tested 13 experiments (task and dataset pairs) based on five most recent LMs for Polish, to prove the usability and usefulness of LEPISZCZE . An important added value of the paper is sharing of our experiences collected during the work on the benchmark. We hope that NLP researchers working on other under-resourced languages will find our comments and suggestions useful.

We plan to add more Natural Language Understanding and Spoken Language Understanding tasks to the benchmark. We want to use our clarinpl-embeddings library to evaluate language models, other contextual embeddings, and static word representations with simpler than transformer-based models. The critical direction in the benchmark is to run experiments not only with the predefined data splits [Gorman and Bedrick, 2020] but also to use other splits to properly check the model’s robustness. We hope that this benchmark will encourage the scientific community to work in transparent and reproducible environments, leading to a rapid improvement of the current language technology for Polish language.

Acknowledgments and Disclosure of Funding

The work was partially supported by (1) the Polish Ministry of Education and Science, CLARIN-PL; (2) the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational
Programme, CLARIN – Common Language Resources and Technology Infrastructure, and (3) the Department of Artificial Intelligence at Wroclaw University of Science.

References


Checklist
1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See for examples Sections 3.1 and 5.
(c) Did you discuss any potential negative societal impacts of your work? [Yes] We checked the datasets for examples if they contain PII breaches. The most important was the Polish Abusive Clauses dataset; hence it is based on real contracts.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A]

(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4. We even provided DVC stages to reproduce experiments.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Table 3.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We talked with all authors of datasets used in the benchmark. We even started creating data sheets for some of the datasets to unify the knowledge about the dataset. We plan to publish them in the future for each benchmarked dataset.

(b) Did you mention the license of the assets? [Yes] We agreed with the authors of the datasets to add the license information to HuggingFace repositories. It will be easier to maintain it in one place.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We agreed with the authors of the datasets to add them to HuggingFace repositories. It will be easier to maintain it in one place.

(d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes] We talked with the authors. We are in constant communication with them.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] For example, the Polish Abusive Clauses dataset has been checked by the Office of Competition and Consumer Protection employees to see if it contains PII. A couple of such examples have been removed before publication.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]