bgGLUE: A Bulgarian General Language Understanding Evaluation Benchmark

Momchil Hardalov[*]
AWS AI Labs

Pepa Atanasova
University of Copenhagen, DIKU

Todor Mihaylov
Meta AI[†]

Galia Angelova **Kiril Simov** **Petya Osenova**
IICT, Bulgarian Academy of Sciences

Ves Stoyanov
Meta AI

Ivan Koychev
Sofia University

Preslav Nakov
MBZUAI

Dragomir Radev
Yale University

Abstract

We present bgGLUE (Bulgarian General Language Understanding Evaluation), a benchmark for evaluating language models on Natural Language Understanding (NLU) tasks in Bulgarian. Our benchmark includes NLU tasks targeting a variety of NLP problems (e.g., natural language inference, fact-checking, named entity recognition, sentiment analysis, question answering, etc.) and machine learning tasks (sequence labeling, document-level classification, and regression). We run the first systematic evaluation of pre-trained language models for Bulgarian, comparing and contrasting results across the nine tasks in the benchmark. The evaluation results show strong performance on sequence labeling tasks, but there is a lot of room for improvement for tasks that require more complex reasoning. We make bgGLUE publicly available together with the fine-tuning and the evaluation code, as well as a public leaderboard at https://bgglue.github.io, and we hope that it will enable further advancements in developing NLU models for Bulgarian.

1 Introduction

Natural Language Understanding (NLU) benchmarks, such as GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a), were designed for a rigorous evaluation of language models on a diverse set of natural language understanding (NLU) tasks. The wide adoption of such benchmarks has driven the rapid development of models that perform well on the tasks that are part of these benchmarks, but also beyond (Devlin et al., 2019; Liu et al., 2019). However, until recently, the focus of such benchmarks has been on English, with little interest in other languages (Bender, 2011; Ponti et al., 2019).

To address this, recent work has designed benchmarks to test models on NLU tasks on non-English languages (Le et al., 2020; Rodriguez-Penagos et al., 2021; Shavrina et al., 2020) or on multiple languages (Liang et al., 2020; Hu et al., 2020).

Here, we aim to improve the diversity of the languages represented in NLU benchmarks by proposing *bgGLUE*, a benchmark for Bulgarian that consists of nine NLU tasks, including token classification, regression, and classification. Thus far, only individual datasets in Bulgarian have been used for model development and evaluation. Small subsets of up to three downstream tasks in Bulgarian have also been included in some multilingual benchmarks (Liang et al., 2020; Hu et al., 2020). Additionally, there are existing benchmarks focusing on other Balto-Slavic languages, such as the Russian SuperGLUE (Shavrina et al., 2020) and the Slovene SuperGLUE (Žagar and Robnik-Šikonja, 2022). However, there are no comprehensive benchmarks for representatives of the Eastern South-Slavic language subgroup, and for Bulgarian in particular. We aim to address these limitations with bgGLUE.

bgGLUE unifies and facilitates access to existing datasets and tasks for Bulgarian. By including more challenging tasks such as natural language inference, fact-checking, and question answering, we ensure that it comprises a rigorous test set for NLP models developed for Bulgarian. We also provide access to the benchmark through the HuggingFace Hub (Lhoest et al., 2021) to allow for ease of use and we encourage model sharing for Bulgarian. Moreover, we fine-tune and run the first systematic evaluation of existing language models for Bulgarian, comparing and contrasting results across all tasks in the benchmark. Our evaluation results show that larger and more robustly pre-trained models yield better performance on all tasks, but also that efficiently distilled models are a strong competitor to their larger counterparts.

[*] Work done while Momchil was in the Sofia University, prior to joining Amazon.

[†] Meta AI is not involved in the creation, release, or hosting of the datasets in the benchmark.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8733–8759
July 9-14, 2023 ©2023 Association for Computational Linguistics

#	Corpus	|Train|	|Dev|	|Test|	Splits	Task	Metrics	Domain
				Token Classification				
1	**BSNLP**	724	301	301	🔄	Named Entities	Macro F1	Misc.
2	**PAN-X**	16,237	7,029	7,263	🔍	Named Entities	Macro F1	Wikipedia
3	**U.Dep**	8,907	1,115	1,116		POS Tagging	Macro F1	Misc.
				Regression / Ranking				
4	**Cinexio**	8,155	811	861	🔄🔍	Sentiment	Pear./Spear. Corr.	Movies
5	**CT21.T1**	2,995	350	357		Check-Worthiness	Avg. Precision	Tweets
				Classification Tasks				
6	**Cred.-N**	19,227	5,949	17,887	🔄✚	Humor Detection	Binary F1	News
7	**Fake-N**	1,990	221	701	🔄	Fake News	Binary F1	News
8	**XNLI**	392,702	5,010	2,490		NLI	Accuracy	Misc.
9	**EXAMS**	1,512	365	1,472	✚	Multi-Choice QA	Accuracy	*HS* Exams

Table 1: Summary of the tasks included in the *bgGLUE* benchmark. The numbers in the train, development, and test columns are in terms of examples. The following columns define the structure of the tasks. The domain is based on the source of the texts. The *EXAMS* dataset is collected from high school (HS) examinations. Splits: 🔄 new splits; 🔍 removed duplicates; ✚ new examples added/collected.

The models show strong performance on part-of-speech tagging and named entity recognition, but struggle on tasks that require more complex reasoning such as solving matriculation exams, or evaluating the credibility and the veracity of news articles. Our contributions are as follows:

- We propose the first benchmark for evaluating the capabilities of language models on NLU in Bulgarian, bgGLUE, which includes nine diverse and challenging downstream tasks.[1]
- While creating the benchmark, we curated the datasets and created standard splits, where those have not been previously available in the original publications. This facilitates the principled evaluation of all datasets in bgGLUE.
- We train and share 36 models for Bulgarian and provide the first comparative evaluation of existing models on all tasks in bgGLUE.

2 Tasks

Table 1 shows the nine datasets that are included in the bgGLUE benchmark. Table 2 shows examples from each dataset and their corresponding labels (translations are available in Table 17 in the Appendix). We present additional details such as word overlaps, domain, topic, label distributions, etc. about each dataset in Appendix C.

2.1 Token Classification

BSNLP The dataset is released as part of the Balto-Slavic NLP workshop series (Piskorski et al., 2017, 2019, 2021). The task focuses on cross-lingual document-level extraction of named entities: the systems should recognize, classify, extract, normalize, and make a cross-lingual linking of all named entity mentions in a document; detecting the position of each named entity mention is not required. The target tags are person (PER), organization (ORG), location (LOC), product (PRO), and event (EVT).

PAN-X (WikiANN) The PAN-X dataset (Pan et al., 2017) has Named Entity Recognition (NER) annotations for persons (PER), organizations (ORG), and locations (LOC). It has been constructed using the linked entities in Wikipedia pages for 282 different languages.

Universal Dependencies (U. Dep) Universal Dependencies (UD) (Nivre et al., 2020) is a framework for consistent annotation of grammar (part of speech, morphological features, and syntactic dependencies) across different human languages. UD is an open community effort with over 300 contributors producing more than 200 treebanks in over 100 languages. The dataset was collected, annotated, and later transferred into the required UD format as part of the Bulgarian treebank (BTB-BulTreeBank) project (Osenova and Simov, 2015).

[1] The bgGLUE code, data, and models are available at https://github.com/bgGLUE/bgglue.

BSNLP	**Document**: ... Канцлерът на {Германия}^{LOC} {Ангела Меркел}^{PER} и президентът на {Русия}^{LOC} {Владимир Путин}^{PER} са обсъдили по телефона реализацията на проекта "{Северен поток - 2}^{PRO}" ... По - рано компанията "{Норд стрим}^{ORG}" , която води строителството ... **Possible Tags**: Person (PER), Organization (ORG), Location (LOC), Product (PRO), Event (EVT)
Cinexio	**User Review**: Пет звезди са му малко - заслужава поне още толкова :) **Rating**: 5.0
Cred.-N	**Body:** Днес изтича срокът, в който българите, живеещи в чужбина, могат да подадат заявление за разкриване на изборна секция за предстоящия на 27 януари референдум. Според решение на Централната избирателна комисия (ЦИК) за допитването секции могат да се откриват в посолствата и консулствата на страната. За целта обаче са нужни поне 20 заявления на желаещи... **Title**: Днес изтича срокът за подаване на заявления за разкриване на секции в чужбина за референдума **Correct Label**: Credible
CT21.T1	**Tweet:** Според изследване, #COVID19 оцелява до 3 часа в аерозоли във въздуха, до 24 часа на хартиена и около 2-3 дни на стоманена или пластмасова повърхност. [URL] **Check-worthy**: Yes
EXAMS	**Paragraph:** През есента на 917 година той изпраща армия ... за да нападнат Сърбия и да накажат Гойникович за предателството му. Българският владетел отново изпраща Теодор Сигрица и Мармаис, но този път те претърпяват поражение... което принуждава Симеон да сключи примирие с Византия... **Subject**: *History* **Question:** Кои пълководци оглавяват наказателния поход на Симеон срещу възникналата сръбска опасност през 917 г.? **Candidate answers:** *(A)* Теодор Сигрица и Мармаис, *(B)* Кракра и Алусиан, *(C)* Ивац и Никулица, *(D)* Книн, Имник и Ицвоклий
Fake.-N	**Body:** Изследователят на българските пророци Христо Радев разкрива предсказания на феномена Слава Севрюкова в интервю за „България днес" „В края на 80-те години Слава Севрюкова казва, че в България изневиделица ще се появи човек, в който е преroден духът на ярък библейски герой. Има предвид Давид. Според ясновидката този българин ще изпълни много важна роля в бъдещето на страната. Дано този президент да е въпросният човек! Румен Радев изскочи от нищото, също като библейския Давид... **Title**: Petel.bg - новини - „България днес": Изкопаха изгубеното пророчество на Слава Севрюкова за България! То се сбъдва пред очите ни **Correct Label**: Fake
PAN-X	**Sentence:** Видът е разпространен в {Бурунди}^{LOC}, {Демократична република Конго}^{LOC}, {Замбия}^{LOC} и {Танзания}^{LOC}. **Possible Tags**: Person (PER), Organization (ORG), Location (LOC)
U.Dep	**Sentence:** В^{ADP} дискусията^{NOUN} ,^{PUNCT} предполагам^{VERB} ,^{PUNCT} ще^{AUX} се^{PRON} засегнат^{VERB} важни^{ADJ} въпроси^{NOUN} .^{PUNCT} **Possible Tags**: NOUN, PUNCT, ADP, VERB, ADJ, PRON, AUX, PROPN, ADV, CCONJ, DET, NUM, PART, SCONJ, INTJ
XNLI	**Text:** И той каза: Мамо, у дома съм. Той се обади на майка си веднага щом училищният автобус го е оставил. **Hypothesis:** Той се обади на майка си веднага щом училищният автобус го е оставил. **Entailment:** Neutral

Table 2: Examples from our bgGLUE benchmark. For each task, the different parts of the example are shown in **Bold**. Underlined text shows the label for that example (or the set of possible labels). The precise model's inputs and the expected outputs (labels) are shown in Table 6 in the Appendix. Translations for each examples are shown in Table 17.

2.2 Natural Language Inference

XNLI This dataset (Conneau et al., 2018) is a subset of a few thousand examples from MNLI, which has been translated into 14 languages. As with MNLI, the goal is to predict textual entailment: does sentence A imply/contradict/neither sentence B? This is a classification task: given two sentences, predict one of the three labels.

2.3 Sentiment Analysis

Cinexio The Cinexio dataset (Kapukaranov and Nakov, 2015) focuses on fine-grained sentiment analysis of movie reviews. It was automatically collected to contain movie reviews in Bulgarian from the Cinexio ticket-booking website (which is not available anymore).

2.4 News Credibility / Fact-Checking

CLEF-2021 CheckThat!, Task 1A (CT21.T1) Check-Worthiness Estimation dataset is part of the 2021 CheckThat! Lab on Detecting Check-Worthy Claims, previously Fact-Checked Claims, and Fake News (Task 1) (Shaar et al., 2021). The aim of the task is to determine whether a piece of text is worth fact-checking. More precisely, given a tweet, one has to produce a ranked list of tweets, ordered by their check-worthiness.

Credible News (Cred.-N) The *Credible News* (Hardalov et al., 2016) dataset focuses on the problem of automatically distinguishing credible from fake and humorous news. The examples are articles collected from six Bulgarian news websites. The articles cover various topics including politics (both local and global), sports, lifestyle, and pop culture. The original dataset contained news from four websites. As part of the bgGLUE initiative, we collected 6,550 new articles (5K credible and 1.5K humorous) from two new websites, and we release more than 30K ones that were not publicly available.

Fake News (Fake-N) This dataset (Society, 2017; Karadzhov et al., 2017) contains Bulgarian news articles over a fixed period of time, whose factuality was questioned. These news articles come from 377 different sources from various domains, including politics, interesting facts, and tips&tricks. The dataset was prepared for and used in the *Hack the Fake News* hackathon in 2017. We found and removed instances that were duplicated across the splits and we further randomly allocated 10% of the training instances for a development dataset, which was not available in the original version of the dataset.

2.5 Question Answering

High School Examinations (EXAMS) EXAMS (Hardalov et al., 2019, 2020) is a benchmark dataset for cross-lingual and multilingual question answering for high school examinations. It contains more than 24,000 high-quality exam questions in 26 languages, covering eight language families and 24 school subjects from the Natural Sciences and Social Sciences, among others. EXAMS offers a fine-grained evaluation framework across multiple languages and subjects, which allows a precise analysis and comparison of various models.

2.6 Scoring

In bgGLUE, we opt for the simple approach of weighing each task equally, and for tasks with multiple metrics, first averaging those scores to get a task score.

3 Data Preparation

Here, we describe the pre-processing steps we took to prepare the datasets before including them in the bgGLUE benchmark. Our main goal was to ensure that the setup evaluated the language understanding abilities of the models in a principled way and in a diverse set of domains. Since all of the datasets were publicly available, we preserved the original setup as much as possible. Nevertheless, we found that some datasets contained duplicate examples across their train/dev/test splits, or that all of the splits came from the same domain, which may overestimate the model's performance. Hereby, *we removed data leaks* and *proposed new topic-based or temporal-based (i.e., timestamp-based)* data splits where needed. We deduplicated the examples based on a complete word overlap in two pairs of normalized texts, i.e., lowercased, and excluding all stop words.

For the *BSNLP* task, we combined the data from three consecutive editions of the NER shared task. We selected all Bulgarian examples, which encompassed six different topics. We used the latest two topics for testing, and split the rest randomly at a 4:1 ratio for training and validation.

The *Credible News* dataset contained data from six sources and various news topics. We split the dataset both by topic and by source. In particular, we included all news articles from the same topic in the same split. For training and validation, we used documents from the largest sources from the two categories. Moreover, we extracted the same topics from the two and grouped them together within the splits, keeping the class ratio at 1:10. In the test dataset, we used data points from all six data sources.[2] We note that our test set contains all data points from the two new sources, which are more recent, making them even more challenging. Finally, we cleaned the texts from duplicates and removed all keywords indicating the source: the names of the authors, the media source, the URLs, etc. More details about the collection process of the news articles are given in Appendix C.

[2] We did not have overlapping topics from the same source in the different splits.

We prepared an entirely new split for *Cinexio*, as there was no standard one. First, we removed all duplicate comments and we sorted the remaining ones by publication time. Next, we split the comments using half for training and the rest equally for validation and testing. Finally, we removed from the training set the comments for the same movie that appeared in the other two splits, which changed the distribution by slightly increasing the proportion of the training set.

For *EXAMS*, we kept the original validation and test splits (Hardalov et al., 2020). We added all additional questions from Hardalov et al. (2019) to the training set, i.e., the category *history online quizzes*.

The *Fake News* dataset was released as part of a shared task and it was already split into training and testing sets. After manual inspection, we found that some of the articles had only their titles reworded and had the same content. Therefore, we removed duplicates based only on exact matches in the article's body. In addition, we designated 10% of the training examples for validation.

PAN-X contained short sentences from Wikipedia automatically annotated based on the available Wiki entities. The dataset consists of 20K examples for training and 10K for validation and testing. We kept the proposed splits, but we checked for duplicates by converting each sentence to lowercase and removing the punctuation. This resulted in the removal of 10K sentences overall.

Our analysis did not find any issues with *CT21.T1*, *U.Dep*, and *XNLI*, and thus we kept the splits as provided originally.

4 Experiments

In this section, we first describe the baseline systems we experiment with and then we present the evaluation results.

4.1 Baselines

Majority and Random Baselines The majority class baseline is calculated from the distributions of the labels in each test set. In the random baseline, each test instance is assigned a target label at random with equal probability.

Fine-tuned models Our baselines include several prominent multilingual encoder-only pretrained Transformer models. We divide them, based on their pre-training objective as follows:

I *Masked language modeling*:

- **mBERT** (Devlin et al., 2019) We use the *base* cased version, trained on 104 languages, including Bulgarian. The pre-training task is done on a Wikipedia dump for each language.
- **XLM-R** (Conneau et al., 2020) We evaluate the *Base* and the *Large* versions of the model. They are trained on filtered CommonCrawl data in 100 languages, including Bulgarian.

II *Knowledge distillation*:

- **Distil-mBERT** (Sanh et al., 2019) The model is distilled using mBERT as the teacher.
- **MiniLM$_{\mathbf{L12}}$** (Wang et al., 2020) The model is distilled using XLM-R$_{\text{Base}}$ as the teacher, on the same pre-training corpora as the latter.

III *Bulgarian downstream task*:

- **SlavicBERT** (Piskorski et al., 2021) The model is based on mBERT that is additionally pre-trained with four Slavic languages: Bulgarian, Czech, Polish, and Russian, using a stratified dataset of Russian news and Wiki articles for the other languages. Finally, the model is fine-tuned on all the languages from the BSNLP shared task.

For the token classification tasks (*BSNLP*, *U.Dep*, *PAN-X*), we predict the tag for each word based on the tag of the first sub-token. For the sentence classification tasks, we obtain the predictions based on the first special token (e.g., *[CLS]*). For *Cinexio*, we optimize a mean squared error loss. For *CT21.T1*, *Cred.-N*, *Fake-N*, and *XNLI*, we optimize the cross entropy loss. Finally, for *EXAMS*, we optimize a binary cross entropy for each candidate answer. More details about the experimental setup, the values of the model hyperparameters, and other training details can be found in Appendix A. For a description of the inputs and the outputs, we refer the reader to Appendix B.

4.2 Experimental Results

Table 3 shows the results for the baseline models fine-tuned on the *bgGLUE* tasks. Each model is trained on one task at a time. First, we see that the random and the majority baselines achieve below 20 points *bgGLUE score*, and all fine-tuned models outperform them on all tasks by a sizable margin.

Figure 1 shows the correlation between the model size and the bgGLUE score: we can see that scaling the model size brings additional performance improvements (Devlin et al., 2019; Conneau et al., 2020; Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022; Scao et al., 2022).

#	Model Name	bgGLUE Avg. →	BSNLP F1$_{\text{macro}}$	Cinexio P/S Corr.	CT21.T1 Avg. P	Cred.-N F1$_{\text{binary}}$	EXAMS Acc.	Fake-N F1$_{\text{binary}}$	U.Dep F1$_{\text{macro}}$	PAN-X F1$_{\text{macro}}$	XNLI Acc.
	Random Baselines										
-	Majority	18.52	0.00	0.00	28.41	34.14	25.68	45.08	0.01	0.00	33.33
-	Random	17.59	0.75	0.00	25.06	30.14	25.54	35.65	6.31	0.94	33.33
	Fine-tuning Baselines										
1	XLM-R$_{\text{large}}$	**75.82**	63.81	**85.69**	**69.45**	**79.73**	**36.41**	**70.31**	**99.30**	**92.96**	**84.71**
2	XLM-R$_{\text{base}}$	73.04	62.47	84.40	63.91	75.74	33.42	66.82	99.23	91.18	80.22
3	SlavicBERT	72.12	‡**65.28**	81.71	62.70	72.01	31.86	67.28	99.06	92.36	76.79
4	mBERT$_{\text{base}}$	71.08	56.13	82.07	64.79	69.17	35.39	65.65	98.99	92.11	75.39
5	MiniLM$_{\text{L12}}$	70.96	59.70	80.63	57.37	75.41	35.26	64.33	98.91	90.26	76.81
6	Distil-mBERT	69.58	52.82	80.32	65.15	67.05	34.31	65.66	98.58	90.82	71.50

Table 3: Baseline results on the bgGLUE benchmark. We show the best results in **bold** and we underline the second best result. The scores for each model are the highest ones achieved during hyper-parameter search by selecting the best model checkpoint on each task's development set. We calculate the *bgGLUE* score on the raw scores (before rounding) and then we round it to two digits. Following the notation of previous benchmarks, we multiply the results by 100. ‡*SlavicBERT* is pre-trained on all languages from the *BSNLP* NER task (not using our splits), therefore its score on that task is unrealistically high.

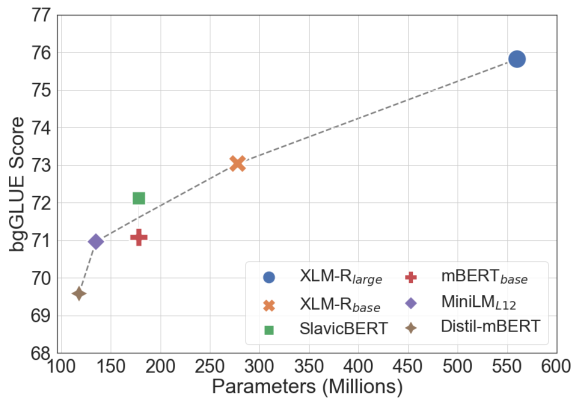


Figure 1: Scaling curve of the bgGLUE score based on the model size (in million parameters). The dotted line illustrates the average increase.

Table 5 in the Appendix summarizes the number of parameters for each model we experimented with. We can see in that there is a linear correlation between the number of parameters in the model and its performance, which holds for all models except for Distil-mBERT, where the slope is more steep. Nonetheless, it is clear that the model size is not the only factor that affects the downstream task performance. Other important factors include the pre-training dataset that was used (Liu et al., 2019; Raffel et al., 2020), the number of tokens in that training set (Hoffmann et al., 2022), whether the model is distilled or is full-size (Sanh et al., 2019; Wang et al., 2020), whether it is monolingual or multilingual (Conneau et al., 2020), etc.

As expected, the largest model, XLM-R$_{\text{Large}}$, outperformed its smaller version XLM-R$_{\text{Base}}$ by 2.78 points absolute. Moreover, the more robustly pre-trained XLM-R model on average outperformed by 2 points a similar-sized mBERT, scoring at 73.0 bgGLUE score. The highest differences between BERT-based and XLM-R-based models (including their distilled versions) are on *BSNLP*, *Cinexio*, *Cred.-N*, and *XNLI*, in favor of XLM; on *CT21.T1* and *PAN-X* this is in favor of BERT.

The gap between XLM-R and mBERT is reduced by 1 point absolute by the SlavicBERT's additional fine-tuning on downstream tasks. Although the largest improvement is observed for NER tasks,[3] we see an increase by 1-2 points also on *Cred.-N*, *Fake-N*, and *XNLI*, compared to mBERT. However, we also see that the downstream pre-fine-tuning is not beneficial for all tasks (Poth et al., 2021), and we see a drop in performance for ranking (*CT21.T1*) and question answering[4] (*EXAMS*) tasks.

Our evaluation of knowledge-distilled models shows that they are a competitive alternative to their teacher models for Bulgarian. Although they are ranked last in terms of performance, their results are 1.5–2.0 points of bgGLUE score behind the best results we obtained, and thus we believe they are a viable alternative to the full models.

[3] SlavicBERT is pre-trained on data from the 2019 edition of the BSNLP competition using different splits.

[4] Hardalov et al. (2019) observed the same when fine-tuning SlavicBERT for multiple-choice QA.

In order to measure the trade-off between model size and model performance, we compare mBERT to DistilBERT and XLM-R to MiniLM. In the case of DistilBERT, we have 30% (178M → 135M) fewer parameters, which results in a 2.2% drop in performance. In turn, MiniLM has less than half of the parameters of its teacher, i.e., 135% fewer parameters (278M → 118M), which leads to only a 2.9% relative drop in performance. The most challenging tasks for the distilled models are NER (*BSNLP* and *PAN-X*), NLI (*Cinexio* and *XNLI*), where we see a sizable gap between non-distilled models. Finally, we note that MiniLM has the worst performance on *CT21.T1*, which is also so for XLM-R. We hypothesize that this is due to the source of the dataset being Twitter. For more detailed results, we refer the reader to Appendix D.

5 Discussion

Software Tools As part of building the benchmark, we developed a set of software tools that facilitate the training and the evaluation of new models. The toolkit is implemented in PyTorch (Paszke et al., 2019), using the *transformers* library (Wolf et al., 2020). Moreover, we integrate and release publicly all datasets (in accordance with their licenses; see the next paragraph) from *bgGLUE* in the HuggingFace datasets repository (Lhoest et al., 2021).[5]

Dataset Licenses We keep the licenses as provided by their authors for all datasets included in the *bgGLUE* benchmark. Table 4 summarizes the information about each dataset and gives links to the external websites and code repositories provided by the authors. All datasets are available to use for research purposes. Some of them come with a non-commercial license, i.e., *Cinexio*, *Cred.-N*, *U.Dep*, and *XNLI*. *Cred.-N* requires signing an agreement form before obtaining the dataset.

Modeling Considerations Previous work has shown that a model's scale (Kaplan et al., 2020; Hoffmann et al., 2022) is an important factor for its performance (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; Conneau et al., 2020; Soltan et al., 2022; Zhang et al., 2022), especially in zero-shot or few-shot settings (Brown et al., 2020; Wei et al., 2021; Chowdhery et al., 2022; Ouyang et al., 2022).

Task	Public	License	Website	Code
BSNLP	✓	✗	🔗	-
Cinexio	✓	🎓	🔗	-
CT21.T1	✓	🎓	🔗	</>
Cred.-N	✗	🎓	-	</>
EXAMS	✓	🅭	-	</>
Fake-N	✓	🔓	-	</>
U.Dep	✓	🎓	🔗	</>
PAN-X	✓	✗	-	</>
XNLI	✓	🎓	🔗	</>

Table 4: Dataset licensing information. The *Cred.-N* dataset is not public and is distributed after filling up an agreement form. All other datasets are publicly distributed under different licenses: ✗ no specific license, 🎓 non commercial use only, 🅭 creative commons license open for commercial use, 🔓 MIT License. 🔗 the dataset has a website. </> the authors offer a repository with code.

Nevertheless, these models are often pre-trained on high-resource languages such as English. A few noteworthy alternatives for Bulgarian include XLM-RoBERTa (Goyal et al., 2021), multilingual T5 (Xue et al., 2021), XGLM (Lin et al., 2022), mGPT (Shliazhko et al., 2022), and the extended version of BLOOM (Yong et al., 2022). These models represent different variants of the Transformer architecture, i.e., encoder/decoder-only or sequence-to-sequence. In this work, we only included encoder-only Transformer models with less than one billion parameters (see Table 5 in the Appendix). We leave the rest to be explored by future participants in the bgGLUE benchmark and we note that some of the tasks, such as ranking or regression, require additional steps to make the different architectures work.

Although a model's scale is important for its performance, it comes with additional efficiency and computational costs, among other considerations (Bommasani et al., 2021). Fine-tuning large pre-trained language models is usually time-consuming and expensive, and it also requires a large number of manually annotated examples. A possible direction that alleviates these requirements is to use adapter-based models (Rebuffi et al., 2017; Pfeiffer et al., 2021) and other techniques for efficient training (Lester et al., 2021; Hu et al., 2021; Ben Zaken et al., 2022).

[5] https://huggingface.co/bgglue

Promising results were shown both in multilingual (Pfeiffer et al., 2020) and in cross-lingual settings (Muennighoff et al., 2022).

Another line of work, which we leave for future research, is zero-shot and few-shot learning. Recently, different techniques have been developed such as learning from demonstrations Brown et al. (2020), patterns (Schick and Schütze, 2021a,b), instructions (Mishra et al., 2022; Wang et al., 2022; Chung et al., 2022; Iyer et al., 2022), or multi-task fine-tuning Raffel et al. (2020); Wei et al. (2021); Chowdhery et al. (2022). These models require fewer examples due to their extensive fine-tuning, but they still showed Chowdhery et al. (2022) that a model's size is of crucial importance for their performance.

Finally, there is a trade-off between performance and using monolingual vs. multilingual models (Devlin et al., 2019; Conneau et al., 2020; Pyysalo et al., 2021). Extensively pre-trained monolingual models on language-specific corpora often achieve better performance compared to multilingual ones (Kuratov and Arkhipov, 2019; Canete et al., 2020; Masala et al., 2020; Delobelle et al., 2020; Chan et al., 2020; Martin et al., 2020; Cui et al., 2021; Pyysalo et al., 2021; Barry et al., 2022). However, there is no open-source large scale pre-trained monolingual Bulgarian model with extensive pre-training: most of the existing checkpoints are based on multilingual ones, and they are fine-tuned on small corpora.[6]

Leaderboard We develop our leaderboard in accordance with existing ones, e.g., the (Super)GLUE (Wang et al., 2019b,a): the participants are provided with all the training, validation, and test examples without the gold test labels. They submit an archive with their predictions for each task, and then our system automatically evaluates their predictions.

The intended use of our leaderboard is to provide a standardized way to compare the performance of different models on specific tasks, thus allowing researchers and practitioners to assess the current state of the art and to identify areas where improvements can be made. We urge against making improperly supported claims about general language understanding based on the performance on our leaderboard, and on NLP leaderboards in general (Ethayarajh and Jurafsky, 2020; Raji et al., 2021; Blasi et al., 2022).

We believe that the bgGLUE leaderboard will incentivize model and resource creation in two ways: (*i*) the participants are required to share details about their submissions, and are encouraged to release their models; we cannot force the latter, but we can ensure that the methods are reproducible to some extent; (*ii*) practice shows that the results on such leaderboards tend to saturate in several years, which will likely happen with this benchmark as well. We plan to open our platform and to work with interested researchers, first to design new leaderboards (Ma et al., 2021), second to include their datasets into bgGLUE, and third to collaborate to build new (including human-and-model-in-the-loop (Kiela et al., 2021)) and refining exciting language resources for Bulgarian.

6 Related Work

Language Understanding The release of the code and English corpora as part of the General Language Understanding Evaluation (GLUE Wang et al. (2019b)) was a push towards the development of models with improved performance on a diverse set of downstream tasks. The GLUE benchmark includes 11 NLU tasks, such as semantic textual similarity, natural language inference, and other classification tasks. Later, the benchmark was extended with additional and more sophisticated tasks in its SuperGLUE (Wang et al., 2019a) variant.

While GLUE and SuperGLUE have been established as the de-facto standard for evaluating machine learning models, they are limited to English. To foster the evaluation and the development of machine learning models for other languages, several benchmarks in other languages have been released. They can be grouped based on their language family as follows: *Romance* – French (Le et al., 2020), Catalan (Rodriguez-Penagos et al., 2021), *Balto-Slavic* – Russian (Shavrina et al., 2020), Slovenian (Žagar and Robnik-Šikonja, 2022), *Iranian* – Persian (Khashabi et al., 2021), *Altic* – Korean (Park et al., 2021), *Sino-Tibetan* – both CLUE (Xu et al., 2020), and CUGE (Yao et al., 2021) focus on Chinese, *Indic* – Kakwani et al. (2020) evaluated fine-tuned pre-trained models on multiple Indic languages, while Doddapaneni et al. (2022) focused on their zero-shot capabilities, and *Malayic* – Indonesian (Koto et al., 2020). Khanuja et al. (2020) provides further resources for code-switched languages (English with Spanish or Hindi).

[6] Reference: https://huggingface.co/models?language=bg

While Shavrina et al. (2020) provided resources for Balto-Slavic languages, there are *no existing benchmarks for languages in the South Eastern-Slavic subgroup or for Bulgarian in particular*. We address this deficiency by developing *bgGLUE*, a benchmark for Bulgarian, which is part of the South-Slavic subgroup. Hristova (2021) published a survey of the language resources available for Bulgarian, which we also include in the current benchmark, extended with more recent datasets.

The aforementioned benchmarks focus on a single language or on a single language family. Other studies looked at multiple languages. Liang et al. (2020) proposed XGLUE, a benchmark for 19 languages that covers NLU problems and language generation tasks. Hu et al. (2020) collected a cross-lingual evaluation dataset in 40 languages, later extended with 10 additional (Ruder et al., 2021), including tasks similar to the original (Super)GLUE setup such as token classification, question answering, textual similarity, natural language inference, etc. Both benchmarks include Bulgarian, but are limited to three tasks: part-of-speech (POS) tagging (Universal Dependencies Nivre et al. (2020)), named entity recognition (PAN-X/WikiAnn Pan et al. (2017)), and natural language inference (XNLI Conneau et al. (2018)). These tasks are also part of *bgGLUE*, but we extend them with additional NLU tasks, including question answering, fake news detection, sentiment analysis, etc.

More recently, a large-scale initiative for providing open access to large language models trained to perform new tasks based on few demonstrations or natural language instructions was launched as part of the BLOOM workshop (Scao et al., 2022). This led to the release of a new corpus, comprising sources in 46 natural and 13 programming languages, and a multilingual decoder-only Transformer language model pre-trained on that data. However, BLOOM was not pre-trained on Slavic languages, and it was only later that zero-shot support for Bulgarian was added Yong et al. (2022).

BIG-Bench (Srivastava et al., 2023) is another such initiative that incorporates more than 200 tasks (some not related to NLP) to test the capabilities of language models. The task topics are diverse, drawing problems from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, and beyond. Currently, there are a few non-English tasks included, but none of them is for Bulgarian.

The *low number of Bulgarian resources* that are part of these initiatives is yet another reason why *more publicly available Bulgarian resources and open-access models are needed.*

Other Modalities Existing work also quantifies the abilities of state-of-the-art models in multi-modal settings. CodeX GLUE (Lu et al., 2021) is a benchmark for program understanding and generation. Conneau et al. (2022) proposed XTREME-S that focuses on speech tasks, including speech recognition, classification, speech-to-text translation, and retrieval. Finally, IGLUE (Bugliarello et al., 2022) fills the gap in image and text evaluation, including tasks such as visual question answering, cross-modal retrieval, and grounded reasoning. Bulgarian is included as part of both XTREME-S and IGLUE. However, here we focus only on NLP tasks on text and currently, we do not include tasks with multiple modalities in the present benchmark.

7 Conclusion and Future Work

We presented *bgGLUE* – the first holistic benchmark for evaluating NLU systems in Bulgarian. It includes nine challenging tasks that cover token classification, regression/ranking, and text classification. We fine-tuned and evaluated six different pre-trained state-of-the-art language models. Our extensive evaluation showed that bgGLUE contains challenging tasks that are far from being solved. Finally, we open-sourced the cleaned versions of the datasets, including the new, more challenging splits, and the source code for training and evaluation, and we released 36 fine-tuned models (one for every task and model combination). All the released artifacts are also integrated into the HugginFace Hub. We believe that bgGLUE is a rich and challenging testbed that will cultivate prospective work on Bulgarian language understanding.

In future work, we plan to add more tasks for Bulgarian, e.g., toxicity detection (Dinkov et al., 2019). We also want to use monolingual Bulgarian datasets for pretraining, beyond Wikipedia and CommonCrawl, e.g., (Simov et al., 2002, 2004; Koeva et al., 2004, 2012, 2020), using which will require a thorough assessment in order to prevent introducing unwanted biases and hazardous behavior in the models trained on them (Bender et al., 2021; Liang et al., 2021). Finally, we plan to try recent multilingual models such as mDeBERTaV3 (He et al., 2021), mT0 and BLOOMz (Muennighoff et al., 2023).

In Memory of Professor Dragomir Radev

We dedicate this work to the memory of Dragomir Radev, who is a co-author of this paper. Drago had a tremendous impact on our community, and his legacy will live on through the countless students and colleagues whose lives he touched. Drago was not only an exceptional computer scientist, but one of the kindest and most humble people many of us have ever known. He deeply cared about Bulgarian NLP and Bulgarian NLP researchers. He was also the one who gave the idea and who remained the main driving force behind the Bulgarian GLUE project. Drago will be greatly missed...

Acknowledgements

We thank the anonymous reviewers for their helpful questions and comments, which have helped us improve the quality of the paper. We also want to thank the MSc students from Sofia University "St. Kliment Ohridski" Mihail Mihaylov and Kiril Georgiev for their help with the initial parsing and analysis. We further thank Quỳnh Mihaylova for designing the bgGLUE logo and for her priceless feedback on the design of the leaderboard.

This work is partially supported by Project UNITe BG05M2OP001-1.001-0004 funded by the Bulgarian OP "Science and Education for Smart Growth", co-funded by the EU via the ESI Funds.

The work on BulTreeBank-UD and the Balto-Slavic datasets for Bulgarian is partially supported by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, grant number DO1-301/17.12.21.

Limitations

Tasks in bgGLUE The bgGLUE benchmark is comprised of nine challenging NLU tasks, including three token classification tasks, one ranking task and five text classification tasks. While we cover three different types of tasks in the benchmark, we are restricted by the available resources for Bulgarian, and thus we could not include some other NLP tasks, such as language generation. We also consider only NLP tasks and we do not include tasks with other/multiple modalities. Finally, some of the tasks are of similar nature, e.g., we include two datasets for NER and two for credibility/fake news classification (see Section 2).

Domains in bgGLUE The tasks included in bgGLUE span over multiple domains such as social media posts, Wikipedia, and news articles and can test both for short and long document understanding. However, each task is limited to one domain and the topics within the domain do not necessarily have full coverage of all possible topics. Moreover, some of the tasks have overlapping domains, e.g., the documents in both Cred.-N and Fake-N are news articles.

Baseline Models As described in Section 5, the baseline models provided for bgGLUE include fairly small encoder-only Transformer architectures. We leave for future work other modeling architectures and modeling techniques that are known for improving the efficiency and the computational requirements of the used models, e.g., few-shot and zero-shot in-context learning and instruction-based evaluation, multi-task learning, etc.

Model Biases In this work, we did not explore whether the datasets in bgGLUE contain unwanted biases, which could also lead to potentially hazardous behavior of the baselines we trained in our experiments with the bgGLUE benchmark.

Ethics and Broader Impact

Dataset Collection

In bgGLUE, we include only datasets that are publicly available, with a license that allows at a minimum free use for academic research. We have also referenced the original work where the corresponding resources were first proposed. We encourage the users of bgGLUE to refer to the original work for licensing details.

Additionally, we carefully examined and removed the instances of the dataset that were duplicated across the training/development/test splits. Whenever development or other dataset splits were not available, we also provide new dataset splits as well. Section 2 points where such changes of the corresponding original resources were required, and the code used to filter or to produce the new splits is available in bgGLUE's code repository. We believe that the selection of publicly available datasets and the adopted dataset curation steps will foster the development and the rigorous evaluation of language models for Bulgarian.

Biases and Misuse Potential

The datasets included in bgGLUE were annotated by human annotators, who could be subject to potential biases in their annotation process. Hence, the datasets in bgGLUE could potentially be misused to develop models that make predictions that are unfair to individuals or groups. Therefore, we ask users of bgGLUE to be aware of such potential biases and risks of misuse. We note that any biases that might exist in the original resources gathered in this benchmark are unintentional and do not aim to cause harm.

Intended Use

The bgGLUE benchmark is intended to promote the development and the rigorous evaluation of language models for Bulgarian. We further believe that the benchmark will serve to examine the capabilities and the limitations of existing and emerging models on the challenging natural language understanding tasks in Bulgarian. Ideally, this could also lead to raising awareness of the potential risks associated with the use of such models developed for downstream tasks in Bulgarian.

Environmental Impact

While bgGLUE can stimulate the development of new machine learning models, it is worth noting that such models could require large computational resources for training, which contributes to global warming (Strubell et al., 2019). On the other hand, bgGLUE is intended mainly for fine-tuning pretrained large language models, which requires considerably smaller computations. Additionally, we release the benchmark and the models on the HuggingFace Hub, which further reduces the environmental impact, as fine-tuning again is computationally costly, especially for larger models.

References

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, and Jennifer Foster. 2022. gaBERT — an Irish Language Model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, LREC '22, pages 4774–4788, Marseille, France.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics)*, ACL '22, pages 1–9, Dublin, Ireland.

Emily M Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, Virtual Event, Canada.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with The Natural Language Toolkit*. O'Reilly Media, Inc.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 5486–5505, Dublin, Ireland.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021.

On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Proceedings of the Thirty-Fourth Conference on Neural Information Processing Systems*, NeurIPS '20, pages 1877–1901, Virtual.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *ICML '22*, pages 2370–2392.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *Proceedings of the Practical Machine Learning for Developing Countries Workshop*, PML4DC '20, pages 1–10, Addis Ababa, Ethiopia (Online).

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, COLING '22, pages 6788–6796, Barcelona, Spain (Online).

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, Daan van Esch, Vera Axelrod, Simran Khanuja, Jonathan Clark, Orhan Firat, Michael Auli, Sebastian Ruder, Jason Riesa, and Melvin Johnson. 2022. XTREME-S: Evaluating Cross-lingual Speech Representations. In *23rd Annual Conference of the International Speech Communication Association*, Interspeech '22, pages 3248–3252, Incheon, Korea.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 8440–8451, Online.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 2475–2485, Brussels, Belgium.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Findings 20, pages 3255–3265, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA.

Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2019. Detecting toxicity in news articles: Application to Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 247–258, Varna, Bulgaria. INCOMA Ltd.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. IndicXTREME: A multi-task benchmark for evaluating indic languages. *arXiv preprint arXiv:2212.05409*.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 4846–4853, Online.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-Scale Transformers for Multilingual Masked Language Modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP*, RepL4NLP '21, pages 29–33, Online.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2016. In Search of Credible News. In *Artificial Intelligence: Methodology, Systems, and Applications*, AIMSA '16, pages 172–180, Varna, Bulgaria.

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019. Beyond English-Only Reading Comprehension: Experiments in Zero-shot Multilingual Transfer for Bulgarian. In *Proceedings of the 2019 International Conference on Recent Advances in Natural*

Language Processing, RANLP '19, pages 447–459, Varna, Bulgaria.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 5427–5444, Online.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Proceedings of the Thirty-Sixth Conference on Neural Information Processing Systems*, NeurIPS '22, New Orleans, Louisiana, USA.

Gloria Hristova. 2021. Text Analytics in Bulgarian: An Overview and Future Directions. *Cybernetics and Information Technologies*, 21(3):3–23.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML '20*, pages 4411–4421, Virtual Event.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. OPT-IML: Scaling Language Model Instruction Meta Learning through the Lens of Generalization. *arXiv preprint arXiv:2212.12017*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.

Borislav Kapukaranov and Preslav Nakov. 2015. Fine-Grained Sentiment Analysis for Movie Reviews in Bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '15, pages 266–274, Hissar, Bulgaria.

Georgi Karadzhov, Pepa Gencheva, Preslav Nakov, and Ivan Koychev. 2017. We built a fake news / click bait filter: What happened next will blow your mind! In *Proceedings of the 2017 International Conference Recent Advances in Natural Language Processing*, RANLP '17, pages 334–343, Varna, Bulgaria.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An Evaluation Benchmark for Code-Switched NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 3575–3585, Online.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A Suite of Language Understanding Challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '21, pages 4110–4124, Online.

Svetla Koeva, Stoyan Mihov, and Tinko Tinchev. 2004. Bulgarian Wordnet–Structure and Validation. *Romanian Journal of Information Science and Technology*, 7(1-2):61–78.

Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, LREC '20, pages 6988–6994, Marseille, France.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and practice in corpus design. *Journal of Language Modelling*, 0(1):65–110.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, COLING '20, pages 757–770, Barcelona, Spain (Online).

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. In *Proceedings of the 2019 International Conference of Computational Linguistics and Intellectual Technologies*, Dialogue 2019, Moscow, Russia.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, LREC '20, pages 2479–2490, Marseille, France.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 3045–3059, Online and Punta Cana, Dominican Republic.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '21, pages 175–184, Online and Punta Cana, Dominican Republic.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards Understanding and Mitigating Social Biases in Language Models. In *Proceedings of the 38th International Conference on Machine Learning*, ICML '21, pages 6565–6576, Virtual Event.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A New Benchmark Datasetfor Cross-lingual Pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 6008–6018, Online.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP '22, pages 9019–9052, Abu Dhabi, United Arab Emirates.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations*, ICLR '19, New Orleans, Louisiana, USA.

Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin Clement, Dawn Drain, Daxin Jiang, Duyu Tang, Ge Li, Lidong Zhou, Linjun Shou, Long Zhou, Michele Tufano, MING GONG, Ming Zhou, Nan Duan, Neel Sundaresan, Shao Kun Deng, Shengyu Fu, and Shujie LIU. 2021. CodeXGLUE: A Machine Learning Benchmark Dataset for Code Understanding and Generation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, NeurIPS '21, Virtual.

Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking. In *Thirty-fifth Conference on Neural Information Processing Systems*, NeurIPS '21, pages 10351–10367, Virtual.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ACL '20, pages 7203–7219, Online.

Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. RoBERT – A Romanian BERT Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, COLING '20, pages 6626–6637, Barcelona, Spain (Online).

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL '22, pages 3470–3487, Dublin, Ireland.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual Generalization through Multitask Finetuning. *arXiv preprint arXiv:2211.01786*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, LREC '20, pages 4034–4043, Marseille, France.

Petya Osenova and Kiril Simov. 2015. Universalizing BulTreeBank: a Linguistic Tale about Glocalization. In *the 5th Workshop on Balto-Slavic Natural Language Processing*, BSNLP '15, pages 81–89, Hissar, Bulgaria.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Thirty-sixth Conference on Neural Information Processing Systems*, NeurIPS '22, pages 27730–27744, New Orleans, Louisiana, USA.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL '17, pages 1946–1958, Vancouver, Canada.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean Language Understanding Evaluation. *arXiv preprint arXiv:2105.09680*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An Imperative Style, high-performance deep learning library. In *Thirty-fourth Conference on Neural Information Processing Systems*, NeurIPS '19, Vancouver, Canada.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '21, pages 487–503, Online.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 7654–7673, Online.

Jakub Piskorski, Bogdan Babych, Zara Kancheva, Olga Kanishcheva, Maria Lebedeva, Michał Marcińczuk, Preslav Nakov, Petya Osenova, Lidia Pivovarova, Senja Pollak, Pavel Přibáň, Ivaylo Radev, Marko Robnik-Sikonja, Vasyl Starko, Josef Steinberger, and Roman Yangarber. 2021. Slav-NER: the 3rd Cross-lingual Challenge on Recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 122–133, Kiyv, Ukraine.

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 2019. The Second Cross-Lingual Challenge on Recognition, normalization, classification, and linking of named entities across Slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, BSNLP '19, pages 63–74, Florence, Italy.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The First Cross-Lingual Challenge on Recognition, normalization, and matching of named entities in Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, BSNLP '17, pages 76–85, Valencia, Spain.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to Pre-Train on? Efficient

Intermediate Task Selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 10585–10605, Online and Punta Cana, Dominican Republic.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. WikiBERT Models: Deep Transfer Learning for Many Languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics*, NoDaLiDa '21, pages 1–10, Reykjavik, Iceland (Online).

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, NeurIPS '21, Virtual.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Thirty-first Conference on Neural Information Processing Systems*, NeurIPS '17, Long Beach, California, USA.

Carlos Rodriguez-Penagos, Carme Armentano-Oller, Marta Villegas, Maite Melero, Aitor Gonzalez, Ona de Gibert Bonet, and Casimiro Carrino Pio. 2021. The Catalan Language CLUB. *arXiv preprint arXiv:2112.01894*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 10215–10245, Online and Punta Cana, Dominican Republic.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

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

Timo Schick and Hinrich Schütze. 2021a. Few-Shot Text Generation with Natural Language Instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, EMNLP '21, pages 390–402, Online and Punta Cana, Dominican Republic.

Timo Schick and Hinrich Schütze. 2021b. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '21, pages 2339–2352, Online.

Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Mucahid Kutlu Alex Nikolov, Firoj Alam Yavuz Selim Kartal, Giovanni Da San Martino, Alberto Barrón-Cedeño, Rubén Míguez, Tamer Elsayed, and Preslav Nakov. 2021. Overview of the CLEF-2021 CheckThat! Lab Task 1 on Check-Worthiness Estimation in Tweets and Political Debates. In *Working Notes of the 2021 Conference and Labs of the Evaluation Forum*, CLEF '2021, Bucharest, Romania (online).

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 4717–4726, Online.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mGPT: Few-Shot Learners Go Multilingual. *arXiv preprint arXiv:2204.07580*.

Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, and Dimitar Doikoff. 2004. A Language Resources Infrastructure for Bulgarian. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, LREC '04, Lisbon, Portugal.

Kiril Simov, Gergana Popova, and Petya Osenova. 2002. HPSG-based Syntactic Treebank of Bulgarian (BulTreeBank). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, pages 135–142.

Data Science Society. 2017. Hack the News Datathon Case – Propaganda Detection. Online; accessed 15 November 2022.

Saleh Soltan, Shankar Ananthakrishnan, Jack FitzGerald, Rahul Gupta, Wael Hamza, Haidar Khan, Charith Peris, Stephen Rawls, Andy Rosenbaum, Anna Rumshisky, et al. 2022. AlexaTM 20B: Few-Shot Learning Using a Large-Scale Multilingual Seq2Seq Model. *arXiv preprint arXiv:2208.01448*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 3645–3650, Florence, Italy.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Thirty-third Conference on Neural Information Processing Systems*, NeurIPS '19, pages 3261–3275, Vancouver, Canada.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations*, ICLR '19, New Orleans, Louisiana, USA.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Thirty-fourth Conference on Neural Information Processing Systems*, NeurIPS '20, pages 5776–5788, Virtual.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP '22, pages 5085–5109, Abu Dhabi, United Arab Emirates.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned Language Models Are Zero-Shot Learners. *arXiv preprint arXiv:2109.01652*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '20, pages 38–45, Online.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, COLING '20, pages 4762–4772, Barcelona, Spain (Online).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '21, pages 483–498, Online.

Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, Zheni Zeng, Yuxian Gu, Kun Zhou, Xuancheng Huang, Wenhao Li, Shuhuai Ren, Jinliang Lu, Chengqiang Xu, Huadong Wang, Guoyang Zeng, Zile Zhou, Jiajun Zhang, Juanzi Li, Minlie Huang, Rui Yan, Xiaodong He, Xiaojun Wan, Xin Zhao, Xu Sun, Yang Liu, Zhiyuan Liu, Xianpei Han, Erhong Yang, Zhifang Sui, and Maosong Sun. 2021. CUGE: A Chinese Language Understanding and Generation Evaluation Benchmark. *arXiv preprint arXiv:2112.13610*.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, et al. 2022. BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting. *arXiv preprint arXiv:2212.09535*.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene SuperGLUE Benchmark: Translation and Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, LREC '22, pages 2058–2065, Marseille, France.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open Pre-Trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*.

Appendix

A Model Hyper-Parameters and Training

Below, we first describe the values of some parameters that are across all models we experiment with, and then we discuss the values of some model-specific parameters:

- All our models use the AdamW (Loshchilov and Hutter, 2019) optimizer with a weight decay of 1e-8, β_1 0.9, β_2 0.999, ϵ 1e-08, and are trained for five epochs with a batch size of 16 (gradient accumulation is applied when needed), and a maximum length of 512 tokens.

- We truncate longer input sequences token by token, if the input is formed from multiple sequences (see Section B), i.e., pairs, we start from the longest one.

- All models use a warmup ratio of 0.06 from the training data. We experiment with learning rate values {2–5}e-04 for base and distilled models, and {1–3}e-04 for XLM-R_{Large}.

- The values of the hyper-parameters (including the number of training epochs) and the best checkpoints were selected on the development set. We use the target metric for each task as a checkpoint selection criterion.

- We trained our models on 5x Tesla T4 GPUs. Depending on the dataset size, the experiments took between 10 minutes, for the smaller datasets and models, and up to 2 hours, for larger datasets. Training the XNLI model took 10 hours with base models, and 20 hours for large models.

- All models were trained with half-precision (fp16) using the default PyTorch implementation.

- Table 5 shows the models' size in terms of the number of parameters.

- When evaluating the *Token Classification Tasks* if the predicted sequence was shorter than the target one (i.e., not all inputs fit into 512 tokens), we added empty tags ('*O*') until the target length was reached.

Model Name	#Params
XLM-R_{large}	560M
XLM-R_{base}	278M
SlavicBERT	178
mBERT$_{\text{base}}$	178M
Distil-mBERT	135
MiniLM$_{\text{L12}}$	118M

Table 5: Number of parameters in millions for each baseline pre-trained model included in the evaluation.

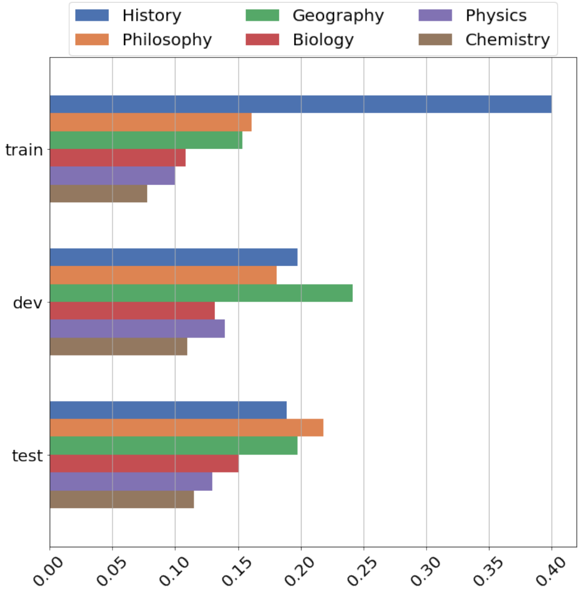


Figure 2: Subject distribution in the *EXAMS* dataset.

B Model Input, Output and Loss

Table 6 shows the inputs and the outputs for each model. We selected the formats based on previous work (Devlin et al., 2019; Liu et al., 2019) and the proposed formats on the (Super)GLUE benchmark (Wang et al., 2019b,a). For all tasks we introduce a projection layer on top of the pre-trained language model's representations. For classification tasks, the output maps to the number of classes, for regression, we project it to a single continuous value, for ranking, we obtain a probability distribution over two classes, for question answering, we rank each answer based on the log probability score, and finally, for token classification tasks, we apply the classification head on top of each token's representation. It is important to note that we use the BIO encoding for the NER tasks. We chose the loss function based on the target value. Finally, we replaced the special tokens with the corresponding ones from the baseline model.

Task	Input	Output	Loss
BSNLP	[CLS] *Document* [SEP]	BIO Tag	Per Token Cross Entropy
Cinexo	[CLS] *User Comment* [SEP]	Rating (1–5)	Mean Squared Error
CT21.T1	[CLS] *Tweet* [SEP]	Normal / Check-worthy	Binary Cross Entropy
Cred.-N	[CLS] *Title* [SEP] *News Article* [SEP]	Credible / Humorous	Binary Cross Entropy
Exams	[CLS] *Context* [SEP] *Question + Option* [SEP]	Option Ranking	Binary Cross Entropy
Fake-N	[CLS] *Title* [SEP] *News Article* [SEP]	Credible / Fake	Binary Cross Entropy
PAN-X	[CLS] *Wikipedia sentence* [SEP]	BIO Tag	Per Token Cross Entropy
U.Dep	[CLS] *Document* [SEP]	POS Tag	Per Token Cross Entropy
XNLI	[CLS] *Hypothesis* [SEP] *Premise* [SEP]	Entailment (3-way)	Cross Entropy

Table 6: Input format for each task, the special tokens are replaced with the corresponding ones from the baseline model. Expected output, e.g., tag name, class, ranking, rating, etc. Finally, the optimization loss used for training.

Topic	**Examples**
Brexit	598
Covid19	151
USElection2020	150
NordStream	130
AsiaBibi	94
Ryanair	84
Total	1,207

Table 7: Topic distribution in the *BSNLP* dataset.

Subset	**#Unique Movies**
Train	257
Dev	25
Test	47
Total	329

Table 8: Number of unique movies in each subset in the *Cinexio* dataset that the users comment about.

Subset	**#Choices**
Train	3.88
Dev	4.00
Test	4.00

Table 9: Number of options per question in the *EXAMS* dataset.

	P/S Corr.	Pearson	Spearman
XLM-R$_{\text{Large}}$	85.69	89.66	81.73
XLM-R$_{\text{Base}}$	84.40	87.91	80.90
SlavicBERT	81.71	84.76	78.66
mBERT$_{\text{Base}}$	82.07	85.22	78.92
MiniLM$_{\text{L12}}$	80.63	85.05	76.21
DistilBERT	80.32	83.55	77.09

Table 10: Fine-grained results for **Cinexio**.

C Additional Task Details

In this section, we summarize some additional characteristics for each task in the bgGLUE benchmark.

Figure 3 shows some statistics about the word overlap between subsets. To calculate the statistics, we split the texts into words using the NLTK Bird et al. (2009) tokenizer. After that, we take the number of unique words in each subset and we take the union of all common words between the first and the second subset, we then compare and divide them by the size of the superset obtained by combining the two. We see that most of the datasets have high overlap between the training and the development/testing set. This is expected as the training sets are often significantly larger compared to the other subsets, and also as we did not filter out the stop words, which cover a big part of the word tokens.

Interestingly, the only exception is the PAN-X dataset. We attribute this to the text snippets being short, designed to contain named entities, and being extracted from different Wikipedia articles.

Figure 4 shows the per task label distribution. We see that most of the tasks maintain similar distributions across labels, except for BSNLP, where we have fewer ORG and PRO tags and more PER.

BSNLP We can see in Table 7 that the most represented topic is Brexit with 600 examples (4x compared to the second topic), followed by COVID-19, US Elections 2020, and Nord Stream, each covering well above 100 examples. The other two topics, AsiaBibi and Ryanair, have less than 100 examples.

	Avg. P	P@1	P@3	P@5	P@10	P@20	P@50	R-Precision
XLM-R$_{\text{Large}}$	69.45	100.00	100.00	100.00	90.00	90.00	70.00	59.21
XLM-R$_{\text{Base}}$	63.91	100.00	100.00	100.00	90.00	75.00	68.00	57.89
SlavicBERT	62.70	100.00	100.00	100.00	80.00	70.00	64.00	61.84
mBERT$_{\text{Base}}$	64.79	100.00	100.00	80.00	90.00	85.00	72.00	60.53
MiniLM$_{\text{L12}}$	57.37	100.00	66.67	60.00	70.00	75.00	66.00	59.21
DistilBERT	65.15	100.00	100.00	80.00	90.00	90.00	68.00	56.58

Table 11: Fine-grained results for **CLEF-2021 CheckThat!, Task 1A (CT21.T1)**.

	Humorous		
	F1	P	R
XLM-R$_{\text{Large}}$	79.73	86.52	73.93
XLM-R$_{\text{Base}}$	75.74	77.53	74.04
SlavicBERT	72.01	88.97	60.48
mBERT$_{\text{Base}}$	69.17	88.91	56.60
MiniLM$_{\text{L12}}$	75.41	76.55	74.31
DistilBERT	67.05	83.27	56.11

Table 12: Fine-grained results for **Credible News (Cred.-N)**.

	Fake		
	F1	P	R
XLM-R$_{\text{Large}}$	70.31	68.20	72.55
XLM-R$_{\text{Base}}$	66.82	61.22	73.53
SlavicBERT	67.28	63.09	72.06
mBERT$_{\text{Base}}$	65.65	68.25	63.24
MiniLM$_{\text{L12}}$	64.33	58.10	72.06
DistilBERT	65.66	67.18	64.22

Table 13: Fine-grained results for **Fake News (Fake-N)**.

Cinexio Table 8 shows the number of movies in the Cinexio dataset. Each movie received 29.9 comments on average.

Cred.-N. We used a custom crawler, *BeautifulSoup* to parse the HTML, and per-site CSS selectors to extract the articles' text. The crawler was based on simple rules that collect and follow the links to articles on each starting page we pass. Our starting points are pages that contain all articles sorted by their publication date and paginated. Finally, we remove all HTML tags, images and information about the authors and the sources, retaining only the plain text. We annotated the articles as credible or humorous based on the label for their website. More details about the dataset and the pre-processing are given in Sections 2 and 3.

High School Examinations (EXAMS) Figure 2 shows the average number of options per question for each subset in the datasets. Both the *Dev* and *Test* subset have four options, but *Train* contains questions with three answers coming from online history exams collected from Hardalov et al. (2019). These examples also affect the subject distribution for the training set.

Fake News (Fake-N.) Here, we report the number of unique and common domains:

- Train vs Dev
  - #Common Domains: 106
  - Only in train: 239
  - Only in dev: 13
- Train vs. Test
  - #Common Domains: 162
  - Only in train: 183
  - Only in test: 46
- Dev vs. Test
  - #Common Domains: 90
  - Only in dev: 29
  - Only in test: 118

D Fine-Grained Results

Here, we present the fine-grained results per task. Grouped by the task types from Table 1, we include the following tables: (*i*) *Regression / Ranking* – In Table 10, we present the Spearman and the Pearson correlation values for the *Cinexio* task. Table 11 shows the metrics for *Ct21.T1*, including P@K and R-Precision; (*ii*) *Classification Tasks* – for the binary classification tasks *Cred.-N* and *Fake-N* we include the *Precision* and *Recall* for the target class, i.e., *Humorous*, and *Fake* respectively, and finally (*iii*) *Token Classification* – Tables 14, 15, and 16 include per token type P, R, and F1.

We did not include tables for *EXAMS* and *XNLI* as their target evaluation measure is *Accuracy*, and thus they are only coarse-grained.

		XLM-R$_{\text{Large}}$	XLM-R$_{\text{Base}}$	SlavicBERT	mBERT$_{\text{Base}}$	MiniLM$_{\text{L12}}$	DistilBERT
Overall	F1	63.81	62.47	65.28	56.13	59.70	52.82
	P	87.22	85.69	84.36	83.73	81.55	76.70
	R	50.30	49.15	53.24	42.21	47.08	40.28
EVT	F1	4.52	5.73	10.56	3.16	0.20	1.36
	P	62.16	45.45	49.15	51.61	10.00	13.46
	R	2.34	3.06	5.91	1.63	0.10	0.71
LOC	F1	74.95	73.89	79.25	68.46	71.45	64.94
	P	95.06	92.81	92.55	92.02	91.00	86.73
	R	61.86	61.38	69.29	54.50	58.82	51.90
ORG	F1	55.50	53.05	57.22	49.03	50.02	42.59
	P	74.00	71.31	71.31	68.28	59.06	56.41
	R	44.40	42.23	47.77	38.25	43.37	34.22
PER	F1	72.83	71.58	72.55	62.84	68.83	60.92
	P	97.80	97.15	97.57	97.27	94.05	91.60
	R	58.02	56.66	57.74	46.41	54.28	45.64
PRO	F1	40.91	38.56	36.42	34.36	35.39	32.55
	P	40.91	40.30	34.84	35.84	40.83	33.54
	R	40.91	36.96	38.14	33.00	31.23	31.62

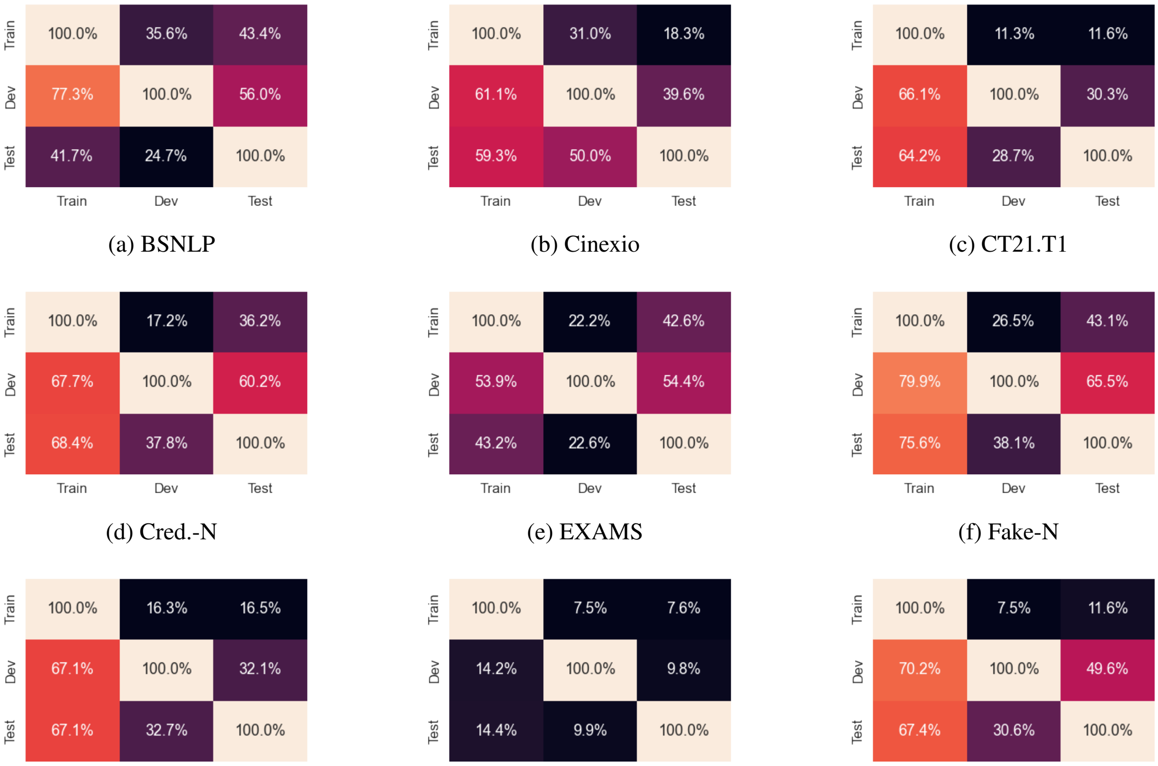
Table 14: Fine-grained results for **BSNLP**.

		XLM-R$_{\text{Large}}$	XLM-R$_{\text{Base}}$	SlavicBERT	mBERT$_{\text{Base}}$	MiniLM$_{\text{L12}}$	DistilBERT
Overall	F1	92.96	91.18	92.36	92.11	90.26	90.82
	P	92.37	90.77	91.85	91.70	89.63	90.40
	R	93.55	91.59	92.88	92.52	90.91	91.24
LOC	F1	95.21	93.66	94.97	94.43	93.37	93.88
	P	95.03	92.85	94.53	93.80	93.07	93.57
	R	95.39	94.50	95.41	95.08	93.67	94.19
ORG	F1	86.33	83.50	84.75	84.81	81.82	82.80
	P	85.35	84.15	84.80	84.81	80.86	82.57
	R	87.34	82.85	84.70	84.81	82.81	83.04
PER	F1	95.07	93.67	94.69	94.60	92.61	92.82
	P	94.25	92.92	93.56	94.17	91.80	92.06
	R	95.90	94.43	95.84	95.03	93.43	93.59

Table 15: Fine-grained results for **PAN-X (WikiAnn)**.

		XLM-R$_{\text{Large}}$	XLM-R$_{\text{Base}}$	SlavicBERT	mBERT$_{\text{Base}}$	MiniLM$_{\text{L12}}$	DistilBERT
Overall	F1	99.30	99.23	99.06	98.99	98.91	98.58
	P	99.32	99.24	99.08	99.00	98.92	98.59
	R	99.29	99.22	99.04	98.98	98.91	98.57
ART	F1	99.22	99.23	98.22	97.97	97.46	96.46
	P	99.48	98.48	96.98	96.50	96.00	94.55
	R	98.97	100.00	99.48	99.48	98.97	98.45
CONJ	F1	99.76	99.68	99.68	99.51	99.11	99.27
	P	99.68	99.68	99.52	99.51	99.19	99.35
	R	99.84	99.68	99.84	99.51	99.03	99.19
DJ	F1	99.19	98.77	98.69	98.42	98.23	97.10
	P	99.08	98.54	98.47	98.61	97.93	97.44
	R	99.31	99.00	98.92	98.23	98.54	96.77
DP	F1	99.96	99.91	99.89	99.93	99.93	99.82
	P	99.96	99.96	99.91	99.96	99.96	99.87
	R	99.96	99.87	99.87	99.91	99.91	99.78
DV	F1	99.35	98.52	97.94	98.93	97.77	97.39
	P	99.18	99.01	98.83	99.18	98.67	97.23
	R	99.51	98.04	97.05	98.69	96.89	97.55
ERB	F1	99.17	99.14	98.81	98.66	98.72	98.24
	P	99.52	99.46	99.28	98.93	99.10	98.51
	R	98.81	98.81	98.34	98.40	98.34	97.98
ET	F1	97.75	97.74	96.81	96.42	96.64	95.67
	P	98.49	99.23	97.73	97.71	97.00	96.95
	R	97.03	96.28	95.91	95.17	96.28	94.42
NTJ	F1	96.97	96.97	96.97	100.00	80.00	90.32
	P	100.00	100.00	100.00	100.00	77.78	100.00
	R	94.12	94.12	94.12	100.00	82.35	82.35
OUN	F1	99.39	99.42	99.42	99.16	99.26	98.85
	P	99.29	99.61	99.41	99.29	99.23	98.85
	R	99.50	99.23	99.44	99.02	99.29	98.85
RON	F1	99.56	99.46	98.86	98.86	98.91	98.54
	P	99.56	99.14	98.80	98.49	99.02	98.06
	R	99.56	99.78	98.91	99.24	98.80	99.02
ROPN	F1	97.52	97.87	97.55	97.62	97.31	97.31
	P	97.83	96.69	96.95	96.96	96.79	96.94
	R	97.22	99.07	98.15	98.30	97.84	97.69
UM	F1	96.76	96.54	96.30	95.61	96.06	96.04
	P	95.87	95.43	95.41	94.52	95.39	95.81
	R	97.66	97.66	97.20	96.73	96.73	96.26
UNCT	F1	99.98	99.98	99.98	99.98	99.98	99.98
	P	99.95	99.95	99.95	99.95	99.95	99.95
	R	100.00	100.00	100.00	100.00	100.00	100.00
UX	F1	97.80	97.80	97.55	97.56	97.61	97.19
	P	97.80	97.68	97.55	97.44	97.56	97.07
	R	97.80	97.92	97.55	97.67	97.67	97.31

Table 16: Fine-grained results for **Universal Dependencies (U. Dep).**



(a) BSNLP

(b) Cinexio

(c) CT21.T1

(d) Cred.-N

(e) EXAMS

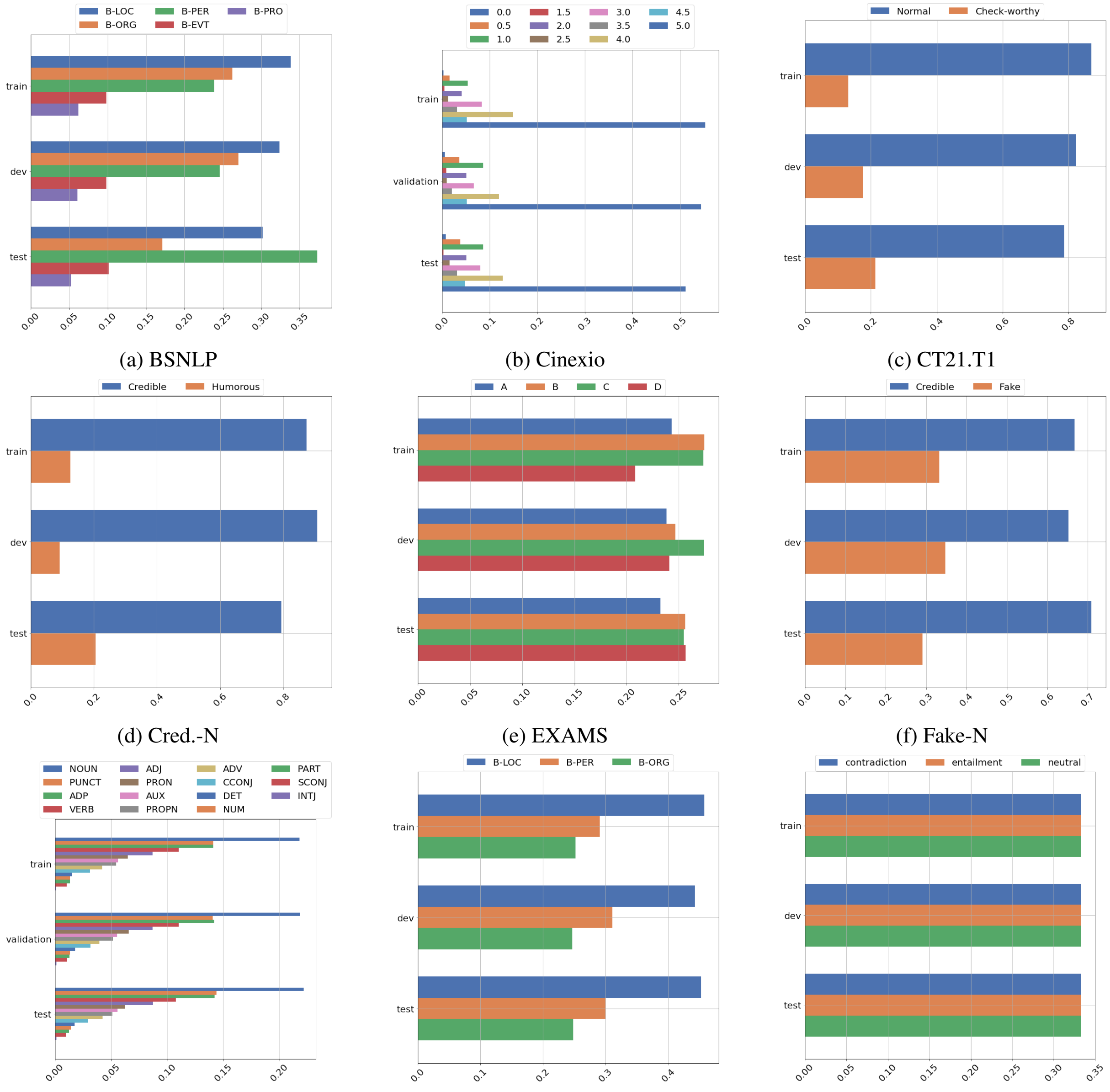
(f) Fake-N

(g) U.Dep

(h) PAN-X/WikiANN

(i) XNLI

Figure 3: Per-task vocabulary overlap. Calculated as the number of common words in the row and the column divided by the total number of unique words in the row.



(a) BSNLP

(b) Cinexio

(c) CT21.T1

(d) Cred.-N

(e) EXAMS

(f) Fake-N

(g) U.Dep

(h) PAN-X/WikiANN

(i) XNLI

Figure 4: Per-task label distribution. We remove the empty tags when we plot for the NER tasks.

BSNLP	**Document**: ... *The chancellor of {Germany}^{LOC} {Angela Merkel}^{PER} and the president of {Russia}^{LOC} {Vladimir Putin}^{PER} discussed over the phone the implementation of the project "{Nord Stream - 2}^{PRO}" ... Earlier the company "{Nord Stream}^{ORG}" which leads the construction ...* **Possible Tags**: Person (PER), Organization (ORG), Location (LOC), Product (PRO), Event (EVT)
Cinexio	**User Review**: *Five stars are not enough - it deserves at least that many :)* **Rating**: 5.0
Cred.-N	**Body:** *Today is the deadline for Bulgarians living abroad to submit an application for opening a polling station for the upcoming referendum on January 27. According to a decision of the Central Election Commission (CEC), polling stations can be opened in the embassies and consulates of the country. However, for this purpose, at least 20 applications are needed from those who wish...* **Title**: *Today is the deadline for submitting applications to open sections abroad for the referendum* **Correct Label**: Credible
CT21.T1	**Tweet:** *According to research, #COVID19 survives up to 3 hours in aerosols in the air, up to 24 hours on paper and about 2-3 days on a steel or plastic surface. [URL]* **Check-worthy**: Yes
EXAMS	**Paragraph:** *In the autumn of 917 he sent an army ... to invade Serbia and punish Gojniković for his treachery. The Bulgarian ruler again sends Theodore Sigritsa and Marmais, but this time they are defeated... which forces Simeon to conclude a truce with Byzantium...* **Subject**: *History* **Question:** *Which generals led Simeon's punitive campaign against the emerging Serbian danger in 917?* **Candidate answers:** (A) *Theodor Sigritsa and Marmais*, (B) *Cracra and Alusian*, (C) *Ivac and Nikulitsa*, (D) *Knin, Imnicus and Izvoklius*
Fake.-N	**Body:** *The researcher of Bulgarian prophets Hristo Radev reveals predictions of the Slava Sevryukova phenomenon in an interview for "Bulgaria Today" a person in whom the spirit of a bright biblical hero has been reborn. He means David. According to the clairvoyant, this Bulgarian will play a very important role in the future of the country. I hope this president is the person in question! Rumen Radev jumped out of nowhere, just like the biblical David...* **Title**: *Petel.bg - news - "Bulgaria today": Slava Sevryukova's lost prophecy about Bulgaria was dug up! It is coming true before our eyes* **Correct Label**: Fake
PAN-X	**Sentence:** *The species is distributed in {Burundi}^{LOC}, {Democratic Republic of Congo}^{LOC}, {Zambia}^{LOC} and { Tanzania}^{LOC}.* **Possible Tags**: Person (PER), Organization (ORG), Location (LOC)
U.Dep	**Sentence:** *In the discussion, I guess, important questions will be discussed.* **Possible Tags**: NOUN, PUNCT, ADP, VERB, ADJ, PRON, AUX, PROPN, ADV, CCONJ, DET, NUM, PART, SCONJ, INTJ
XNLI	**Text:** *And he said, Mother, I am at home. He called his mother as soon as the school bus dropped him off.* **Hypothesis:** *He called his mother as soon as the school bus dropped him off.* **Entailment:** Neutral

Table 17: English translations of the examples shown in Table 2.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section "Limitations"

☑ A2. Did you discuss any potential risks of your work?
In Sections "Ethics Statement" and "Limitations".

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In "Abstract" and Section 1. "Introduction".

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 2 "Tasks" and Section 4 "Experiments"

☑ B1. Did you cite the creators of artifacts you used?
Sections 2 "Tasks"

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 5. "Discussions"

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Sections 2 "Tasks" and "Ethics Statement".

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
In Sections 2 "Tasks" and "Ethics Statement".

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Sections 2 "Tasks", Section 3 "Data Prepartion" and the Appendix.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2 "Tasks", and the Appendix.

C ☑ Did you run computational experiments?

In Section 4. "Experiments" and Appendix.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.