Topic Classification and Headline Generation for Maltese using a Public News Corpus

Amit Kumar Chaudhary, Kurt Micallef, Claudia Borg
Department of Artificial Intelligence, University of Malta
amit.k.chaudhary.22@um.edu.mt, kurt.micallef@um.edu.mt, claudia.borg@um.edu.mt

Abstract

The development of NLP tools for low-resource languages is impeded by the lack of data. While recent unsupervised pre-training approaches ease this requirement, the need for labelled data is crucial to progress the development of such tools. Moreover, publicly available datasets for such languages typically cover low-level syntactic tasks. In this work, we introduce new semantic datasets for Maltese generated automatically using associated metadata from a corpus in the news domain. The datasets are a news tag multi-label classification and a news abstractive summarisation task by generating its title. We also present an evaluation using publicly available models as baselines. Our results show that current models are lacking the semantic knowledge required to solve such tasks, shedding light on the need to use better modelling approaches for Maltese.

Keywords: Maltese, datasets, multi-label classification, abstractive summarisation, less-resourced languages

1. Introduction

In recent years, the capabilities of language models have drastically improved. This is in part due to the pre-training paradigm, as models leverage large amounts of unlabelled text corpora thereby reducing the need for large amounts of labelled datasets. Although relatively less data is required to train such models, labelled datasets are still essential to measure the performance of language models quantitatively. Moreover, as language models get better results, the need for more challenging datasets arises, in order to continuously improve the capabilities of such models.

While high-resource languages such as English have a large body of work on datasets that have been consolidated to a standard benchmark over the years, such as GLUE (Wang et al., 2018), low-resource languages lack this. There have been recent efforts focused on building resources for a variety of languages, such as Universal Dependencies (Nivre et al., 2017), which has grown to cover more than 100 languages. Despite this, such a dataset only tests low-level syntactic knowledge of a model.

As a result, recent works have focused on building similar resources for semantic tasks such as XNLI (Conneau et al., 2018) (Natural Language Inference) and PAWS-X (Yang et al., 2019) (Paraphrase Identification). However, these resources lack language coverage compared to Universal Dependencies – XNLI covers 15 languages and PAWS-X covers 8 languages – most of which can be considered as high-resource languages.

In parallel, there have been efforts to create various evaluation datasets for particular languages. In this work, we focus on building resources for Maltese, which is considered a low-resource language (Rosner and Borg, 2022). Micallef et al. (2022) have built new language models for Maltese – BERTu and mBERTu. Although they achieved state-of-the-art results, the tasks used for the evaluation were primarily low-level morphosyntactic tagging tasks: Part-of-Speech Tagging using the MLRS POS data (Gatt and Čéplö, 2013), Dependency Parsing using the Maltese Universal Dependencies Treebank (Čéplö, 2018), and Named-Entity Recognition using WikiAnn (Pan et al., 2017). Nevertheless, they also include the Sentiment Analysis dataset from Martínez-García et al. (2021), obtaining positive results with these models.

This Sentiment Analysis dataset is a composition of two smaller datasets from Dingli and Sant (2016) and Cortis and Davis (2019). The resultant dataset comprises 851 sentences with merged binary sentiment labels. In comparison, the sentiment subtask in the GLUE benchmark (Wang et al., 2018), called SST2 (Stanford Sentiment Treebank), has 70K samples. Such small datasets can be prone to overfitting when comparing different models and cause large variances in performance based on factors like random seeds.

Hence, in this work, we build new datasets for Maltese, focusing on testing the semantic capabilities of a model. A major bottleneck in creating larger datasets is the need for involving human annotators. In this work, we sidestep the need for significant human annotation efforts, by making use of publicly available corpora and their associated metadata. Specifically, we build two new datasets from Maltese news data for topic classification and headline generation, using tags and title data from the news portals. We make these datasets publicly

LREC-COLING 2024, pages 16274–16281
20-25 May, 2024.

available.[1]

The rest of this paper is organised as follows. We present a brief overview of dataset collection approaches relevant to our work in Section 2. In Section 3, we present our approach for the data collection and labelling. Our evaluation setup on this data is presented in Section 4 followed by the results in Section 5.

2. Related Work

Many approaches use automatic machine translation to translate existing English benchmarks to specific target languages, along with partial human translation for additional quality checks. This includes the SuperGLUE benchmark (Wang et al., 2019) which was translated to languages such as Russian (Shavrina et al., 2020) and Slovene (Žagar and Robnik-Sikonja, 2022). While this approach is interesting, there is an issue that the SuperGLUE test set is private and model predictions have to be uploaded to a web interface to get the results. There is also a rate limit on the number of submissions. Additionally, these works use cloud-based services such as GoogleMT for machine translation, which makes the work in other languages costly. This translation approach has also been used by OpenAI (2023) to translate the benchmark task MMLU (Hendrycks et al., 2020) from English to evaluate the multilingual capabilities of GPT-4 against prior GPT versions on a wide range of languages.

Using a machine-translation approach makes it hard to effectively measure how much the model "understands" a language since it is given translations and not original data in the target language. Furthermore, we argue that such an approach is inadequate for our case, since translation models for a low-resource language such as Maltese results in a significant number of errors, making the results unreliable.

One approach in current literature to bypass human annotation has been to generate benchmark datasets automatically from web corpora based on associated metadata. Zhang et al. (2015) repurpose AG's corpus[2] of one million English news articles and take the top four largest categories as labels to create a news classification task based on article title. Similarly, there have also been works to create sentiment datasets based on the ratings given by users on movie and product reviews sites (Pang and Lee, 2005; Ni et al., 2019). Chalkidis et al. (2021) use document metadata to create a multi-class classification task. We use a similar approach to the latter to create our news category dataset outlined in Section 3.1.

Yang (2022) extract news article and headline pairs and use this data for abstractive summarisation. This approach is similar to that used by our news headline dataset presented in Section 3.2.

In our work, we take inspiration from the work done in other languages for automatic dataset generation to build two new semantic tasks for Maltese, doing so by leveraging publicly available corpora and its associated metadata.

3. Tasks

In this section we describe our procedure for building the different datasets. For both datasets, we make use of the *press_mt* portion of Korpus Malti v4.0 (Micallef et al., 2022), which contains articles scraped from various news portals in Malta.

Each instance contains a list of texts, corresponding to the sentences in an article. In addition, we also collect metadata available in the corpus to aid in the building of each dataset. The additional fields that we make use of are the URL, the title, and the tags of the article. In total, this contains a set of 44,823 articles, but we perform some filtering to increase the quality of the data.

First of all, we filter the dataset to only include article pages and remove pages such as category or author pages. We also normalise all the URLs to use a consistent format so that we can identify the news portal from the article URL. We drop duplicate articles from the dataset and any articles with no article content. We also drop articles whose article content is less than ten words.

During our analysis, we found that certain article contents contained CSS or JavaScript code when they were originally scraped to be included in Korpus Malti. We wrote specific regular expressions for each news portal to find those parts and removed them from the contents. We also find that certain news article contents include specific phrases repeated. For example, for the Newsbook portal *"Read in English ."* was repeated in the contents of each article. We remove such repeated phrases from the different portal contents.

After this filtering, we end up with a set of 25,403 articles, which we use to build our datasets.

3.1. News Category Classification

For this task, we build a multi-label topic classification dataset to classify news article contents into a set of defined categories. We use a semi-automated pipeline with partial human annotation.

[1] https://huggingface.co/datasets/MLRS/maltese_news_categories and https://huggingface.co/datasets/MLRS/maltese_news_headlines

[2] http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

We leverage the fact that news portals already assign specific tags on their articles and that can be used as our weak labels to define this task.

However, the categories used by the different news portals differ. In total, we collected 233 unique tags from all the portals available. Due to the lack of standardisation for tags across news portal, two of our co-authors, who are native speakers of Maltese, manually went through the list of tags as follows. Similar tags were merged into a higher-level category. For example, the “EU” label contains articles which were originally labelled as “European Union” or “Unjoni Ewropea”, amongst others. Labels that were deemed to be too specific were dropped as part of this process. This served as a basis for the rest of the automated pipeline.

Some of the articles were only tagged using generic categories (such as, “News”, “Local”, “Headlines”, “Uncategorised”, and “Archived”) and hence we dropped such articles. This reduces our working dataset size from 25,403 to 16,058 articles. We also dropped tags that had less than 100 articles in total. With this, we get a total of 21 tags. We then calculate the co-occurrence between various categories and drop categories that overlap more than 75% with another category. This procedure removes four tags, as well as some articles which were tagged exclusively with such a tag. For example the Family tag co-occured 93% with the Culture, tag, and hence it did not provide any additional signal compared to the other tags.

At the end of our pipeline, we have 15,374 articles assigned to 17 categories. A breakdown of the number of articles for each tag is shown in Table 1. Note that articles having multiple tags are counted more than once in this table (once for each tag), and hence, the total from this table is larger.

We create 70% training, 15% validation, and 15% test splits using iterative stratification from scikit-multilearn (Szymański and Kajdanowicz, 2017).

3.2. News Headline Generation

In this section we describe a dataset composed of article contents and their corresponding titles. The main purpose of this dataset is to perform headline generation based on the content of a news article, similar to a summarisation task. In addition to the article content and title, we also extract the base URL of an article so that it can be used as a signal to generate using different styles, according to the news portal.

As we do not perform any additional filtering the resultant dataset is composed of 25,403 article-title pairs. These span across 8 different news portals. We split the data as 70% training, 15% validation, and 15% test.

Tag	Articles		Tokens
	Count	Percentage	Count
Court	860	5.59	404K
Covid	1,735	11.29	458K
Culture	2,186	14.22	750K
EU	240	1.56	93K
Economy	321	2.09	113K
Education	191	1.24	56K
Entertainment	3,147	20.47	837K
Environment	147	0.96	39K
Health	290	1.89	82K
Immigration	120	0.78	30K
International	3,957	25.74	785K
Opinion	321	2.09	231K
Politics	1,294	8.42	682K
Religion	465	3.02	186K
Social	203	1.32	98K
Sports	3,066	19.94	835K
Transport	241	1.57	75K

Table 1: Article distribution by News Tag

4. Evaluation Setup

To evaluate datasets, we make use of transformer-based models available from previous works.

4.1. News Category Classification

For the tagging task (Section 3.1), we make use of encoder-only models, all using the BERT architecture (Devlin et al., 2019). We use BERTu and mBERTu, which achieve state-of-the-art results on currently available Maltese tasks (Micallef et al., 2022). Both these models were pre-trained using Korpus Malti (Micallef et al., 2022) which includes the news data used in Section 3. However, they were never exposed to the article contents in order, nor were they exposed to any of the additional metadata we make use of in the datasets.

We also use a monolingual English BERT and a multilingual model mBERT (Devlin et al., 2019) to get a baseline for models that have not been specifically pre-trained for Maltese. All models are composed of 12 transformer layers, where BERTu has 126 million parameters, BERT has 109 million parameters, and both mBERT and mBERTu have 179M parameters.

The chosen models are fine-tuned using the transformers library (Wolf et al., 2020), by adding a linear classification layer on top of the encoder. Following Micallef et al. (2022), we fine-tune each model for at most 200 epochs, with an early stopping of 20 epochs on the validation set. We use five different random seeds for each task and report the mean and standard deviation of the score in our results. We use a learning rate of 4e-5, a batch size of 32, and a dropout of 0.5.

We use macro-averaged F1 for our evaluation metric, reporting the average over 5 independent runs with different random seeds.

4.2. News Headline Generation

For the generation task presented in Section 3.2, we make use of models having decoder architectures, which enables us to generate text. We make use of multilingual models, since there is no available Maltese model with such an architecture.

Specifically we make use of BLOOMZ and mT0, which are decoder-only and encoder-decoder models, respectively (Muennighoff et al., 2023). These models are based on the BLOOM (Scao et al., 2023) and mT5 (Xue et al., 2021) models, respectively, which were pre-trained on multilingual corpora, but are fine-tuned further using instruction tuning on a variety of tasks. Instruction tuning was performed using 46 natural languages and 13 programming languages by using prompts for each task to generate the target result. While Maltese is not part of this instruction tuning data, Xue et al. (2021) report that the initial pre-training data for mT5 contains 5.2 billion Maltese tokens (0.64% of the entire corpus). However, given that this data was scraped automatically from the web, its authenticity is questionable (Kreutzer et al., 2022).

We use the 560 million parameter model for BLOOMZ and the small variant with 300 million parameters for mT0. We use the smallest available models primarily to ease memory requirements. Moreover, the small model for mT0 primarily has 6 transformer layers each for the encoder and decoder, hence being comparable in size to the BERT-based models outlined in Section 3.1.

Fine-tuning such models typically requires running examples through the model while updating the whole network, and hence, is relatively more expensive than the traditional fine-tuning approach used in Section 3.1. Thus, we rely on prompting to extract headlines. We make use of 9 prompt templates for the CNN/Daily Mail summarisation dataset (Hermann et al., 2015) which was used to fine-tune BLOOMZ and mT0 (Muennighoff et al., 2023) as this is the closest to our setup in terms of task and domain. Following Muennighoff et al. (2023), we use English prompts with the Maltese articles and corresponding titles, since they demonstrate that this gives better results than using machine-translated prompts in the target language.

We use BLEU as our main automatic evaluation metric. Since we have multiple prompts, we only extract the headlines once for each prompt. We report the average across all prompts.

Model	Macro F1
BERT	61.51 ± 0.87
mBERT	61.06 ± 1.41
mBERTu	66.65 ± 1.34
BERTu	**67.04 ± 2.02**

Table 2: Experimental results for News Category Classification. All figures shown are the mean and standard deviations over 5 runs with different random seeds. The best-performing model is **bolded**.

Model	BLEU
BLOOMZ	**0.84 ± 0.36**
mT0	0.28 ± 0.14

Table 3: Experimental results for News Headline Generation. All figures shown are the mean and standard deviations over all 9 prompt templates used. The best-performing model is **bolded**.

5. Results

We now present the results using the evaluation setup presented in Section 4.

Table 2 shows the results after fine-tuning the models on the News Category Classification task. Models pre-trained on Maltese text perform better than our baseline models as expected. On this task, BERTu performs better than mBERTu by a slight margin with a mean F1 score of 67.04. However, the gap between BERTu and mBERTu and the other models is generally smaller than that for the tasks used by Micallef et al. (2022). We also note that the results for BERTu and mBERTu are quite lower than that observed on other tasks, indicating that this task is more challenging for these models.

For the Headline Generation tasks, all models achieve BLEU scores close to 0, as shown in Table 3. The results can be partially attributed to having a hard task, since titles do not necessarily summarise, but are sometimes statements to draw attention. However, such results highlight that these models fail to generate anything meaningful. This finding is unsurprising given that both models are not fine-tuned on Maltese data, and that they have at best a few Maltese sentences during pre-training.

6. Conclusion

In this work, we presented two new semantic datasets for Maltese in the news domain. We leveraged an existing corpus comprising of news articles and associated metadata to create the datasets, with minimal human effort. Although the new datasets are relatively smaller than what is available for higher-resourced languages, these new resources are significantly larger than currently available Maltese datasets. Moreover, these datasets offer ways to test the Maltese semantic knowledge of language models as they can be used for multi-label classification and abstractive summarisation tasks.

Our evaluation on the classification task shows that existing models pre-trained on Maltese – BERTu and mBERTu (Micallef et al., 2022) – perform better than the baselines. However, the results for these models and gap with the baselines is relatively lower than what is observed for other tasks (Micallef et al., 2022). Existing multilingual models obtain terrible results on the headline generation task, highlighting that these models do not perform as well on unseen languages on higher-level sematic tasks.

Future work should enhance existing models to improve the performance on these tasks, by using language adaptation techniques (Yong et al., 2023) using Maltese corpora such as Korpus Malti (Micallef et al., 2022). In addition, while models such as BERTu and mBERTu have leveraged monolingual Maltese corpora to obtain better results, we believe that low-resource languages are still impeded by the data scarcity problem, particularly for higher-level semantic tasks. Thus, we encourage further research to improve low-resource language modelling using novel approaches.

7. Ethics Statement

The data that we used from Korpus Malti inherently has biases. In particular, several news portals may express biased or partisan opinions regarding certain topics. For this reason, we retain information related to the news portal in the dataset to allow future work to make use of this.

8. Acknowledgements

This work is based on the first author's master thesis, which was conducted under a scholarship from the Erasmus Mundus Program Language and Communication Technologies(EMLCT). We also acknowledge LT-Bridge Project (GA 952194) and DFKI for access to the Virtual Laboratory.

9. Bibliographical References

Slavomír Čéplö. 2018. *Constituent order in Maltese: A quantitative analysis*. Ph.D. thesis, Charles University, Prague.

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

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*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Keith Cortis and Brian Davis. 2019. A social opinion gold standard for the Malta government budget 2018. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 364–369, Hong Kong, China. Association for Computational Linguistics.

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

Alexiei Dingli and Nicole Jessica Sant. 2016. Sentiment analysis on maltese using machine learning.

Albert Gatt and Slavomír Čéplö. 2013. Digital Corpora and Other Electronic Resources for Maltese. In *Proceedings of the International Conference on Corpus Linguistics*, pages 96–97. UCREL, Lancaster, UK.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Antonio Martínez-García, Toni Badia, and Jeremy Barnes. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, Online. Association for Computational Linguistics.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

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 (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Mike Rosner and Claudia Borg. 2022. *Report on the Maltese Language*. Language Technology Support of Europe's Languages in 2020/2021. Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne (Series Editors). Available online at `https://european-language-equality.eu/deliverables/`.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar,

Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak,

Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

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)*, pages 4717–4726, Online. Association for Computational Linguistics.

P. Szymański and T. Kajdanowicz. 2017. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv: Arxiv-1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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*, pages 38–45, Online. Association for Computational Linguistics.

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

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Zijian Győző Yang. 2022. Neural text summarization for Hungarian. *Acta Linguistica Academica*, 69(4):474 – 500.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Aleš Žagar and Marko Robnik-Sikonja. 2022. Slovene superglue benchmark: Translation and evaluation. *International Conference On Language Resources And Evaluation*.