Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 525–545.

Tarek Naous, Wissam Antoun, Reem Mahmoud, and Hazem Hajj. 2021. Empathetic BERT2BERT conversational model: Learning Arabic language generation with little data. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 164–172, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Tarek Naous, Christian Hokayem, and Hazem Hajj. 2020. Empathy-driven Arabic conversational chatbot. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 58–68.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *arXiv preprint arXiv:1608.07836*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Quora.com. 2017. Quora question pairs. https://www.kaggle.com/competitions/quora-question-pairs.

Simin Rao, Hua Zheng, and Sujian Li. 2021. Cross-lingual leveled reading based on language-invariant features. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2677–2682.

Ankit Rathi. 2020. Deep learning apporach for image captioning in Hindi language. In *2020 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, pages 1–8. IEEE.

Biswarup Ray, Avishek Garain, and Ram Sarkar. 2021. An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing*, 98:106935.

Shigehiko Schamoni, Julian Hitschler, and Stefan Riezler. 2018. A dataset and reranking method for multimodal mt of user-generated image captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 140–153.

Alok Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2022. Attention based video captioning framework for Hindi. *Multimedia Systems*, 28(1):195–207.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Sergey Smetanin. 2022. Rusentitweet: A sentiment analysis dataset of general domain tweets in russian. *PeerJ Computer Science*, 8:e1039.

Sergey Smetanin and Michail Komarov. 2019. Sentiment analysis of product reviews in russian using convolutional neural networks. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, pages 482–486.

Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories.

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, volume 17, pages 4096–4102.

Jeniya Tabassum, Mounica Maddela, Wei Xu, and Alan Ritter. 2020. Code and named entity recognition in StackOverflow. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

TripAdvisor. Topic modelling on Trip Advisor dataset - Kaggle. https://www.kaggle.com/code/imnoob/topic-modelling-lda-on-trip-advisor-dataset/notebook.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 163–173.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497.

Mengting Wan, Rishabh Misra, Ndapandula Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2605–2610.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. 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 3621–3625.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2019. Text readability assessment for second language learners. *arXiv preprint arXiv:1906.07580*.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

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.

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023. MultiCapCLIP: Auto-encoding prompts for zero-shot multilingual visual captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11908–11922. Association for Computational Linguistics.

Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. 2022. MDIA: A benchmark for multilingual dialogue generation in 46 languages. *arXiv preprint arXiv:2208.13078*.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multilingual misinformation. *arXiv preprint arXiv:2210.15954*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

A More details about README++

A.1 Textual Domains

This section provides a description of how sentences were collected from each domain of README++. Table 15 shows statistics of the corpus and Table 16 summarizes the sources from which data was collected for each domain in each language, including publicly available web resources or open-source datasets.

- WIKIPEDIA: Wikipedia is an attractive source of multilingual text since most articles are available in a large number of languages. Further, articles belong to a variety of topics where writing style and technicality differ significantly. We select 9 Wikipedia topics and, from each, randomly sample 5 different articles that discuss a certain sub-topic within that topic. For example, an article on *"Information Theory"* belongs to the *"Technology"* topic. We scrape the Arabic, English, French Hindi, and Russian versions of each article.

- NEWS ARTICLES: We leverage resources used for news category classification research, which we find publicly available datasets for in Arabic (Alfonse and Gawich, 2022) and English (Misra, 2022). No similar public resource was found for the other languages.

- RESEARCH: We collect text from medical, law, politics, and economics research papers in each language if available. We use open-access research archives such as arxiv[1] or HAL[2]. We also search for open-access research articles published under a Creative Commons license on Google Scholar using the same keyword in each language. We notice that research papers from natural sciences or technology are much less frequent in non-English languages as most researchers in those areas publish their work in English.

- LITERATURE: We collect sentences from different types of literature *(Novels, History, Biographies, Children's Stories)* using books that are in the public domain. For English, French, and Russian, we use Project Gutenberg[3] that archives old books for which U.S. copyright has expired. For Arabic, we use Hindawi Books[4] which provide free Arabic books in many genres and topics. For Hindi, the law in India states that the copyright terms of books end 60 years after the death of an author and comes under the public domain[5]. Similar laws for most countries of the world are present with varying number of years[6]. We thus manually search for books in Hindi whose copyrights have expired according to these lengths. For example, we used Hindi novels by Premchand, Sarat Chandra Chattopadhyay, Rabindranath Tagore and Devaki Nandan Khatri.

- TEXTBOOKS: Textbooks are obtained from the Open Textbook Library[7] for English and Hindawi Books for Arabic which provide openly licensed textbooks. For Hindi textbooks, we use publicly available school textbooks from the National Council of Educational Research and Training in India [8] which provides books at various high-school levels and in different subjects. No similar openly available resource was found for French and Russian.

- LEGAL: We identify multiple governmental type of documents that we group under the "legal" domain, which include:

  Constitutions: We sample sentences from the U.S. constitution for English, the Lebanese constitution for Arabic, the Indian constitution for Hindi, the French constitution for French, and the Russian constitution for Russian.

  Judicial Rulings: We used recent public decisions by law courts, such as the Supreme Court in the US [9], to collect sentences from judicial rulings, in addition to using legal datasets with such content (Kapoor et al., 2022).

  United Nations Parliament: We collect samples from the United Nations (UN) Parallel Corpus (Ziemski et al., 2016) which contains official records and parliamentary documents of the UN.

[1] arxiv.org
[2] hal.science
[3] gutenberg.org
[4] hindawi.org
[5] https://copyright.gov.in/Documents/handbook.html
[6] en.wikipedia.org/wiki/List_of_countries%27_copyright_lengths
[7] open.umn.edu/opentextbooks/books
[8] ncert.nic.in/
[9] law.cornell.edu/supremecourt/text