Does Corpus Quality Really Matter for Low-Resource Languages?

Mikel Artetxe[1] **Itziar Aldabe**[2] **Rodrigo Agerri**[2]
Olatz Perez-de-Viñaspre[2] **Aitor Soroa**[2]
[1]Meta AI
[2]HiTZ Center, University of the Basque Country (UPV/EHU)
artetxe@meta.com
{itziar.aldabe,rodrigo.agerri,olatz.perezdevinaspre,a.soroa}@ehu.eus

Abstract

The vast majority of non-English corpora are derived from automatically filtered versions of CommonCrawl. While prior work has identified major issues on the quality of these datasets (Kreutzer et al., 2021), it is not clear how this impacts downstream performance. Taking representation learning in Basque as a case study, we explore tailored crawling—manually identifying and scraping websites with high-quality content—as an alternative to filtering CommonCrawl. Our new corpus, called EusCrawl, is similar in size to the Basque portion of popular multilingual corpora like CC100 and mC4, yet it has a much higher quality according to native annotators. For instance, 66% of documents are rated as high-quality for EusCrawl, in contrast with $< 33\%$ for both mC4 and CC100. Nevertheless, we obtain similar results on downstream NLU tasks regardless of the corpus used for pre-training. Our work suggests that NLU performance in low-resource languages is not primarily constrained by the quality of the data, and other factors like corpus size and domain coverage can play a more important role.

1 Introduction

Large-scale pre-training has resulted in a paradigm shift in NLP (Bommasani et al., 2021). While recent progress has been primarily driven by scaling up on model size and compute, both data quantity and quality have been shown to play a critical role (Kaplan et al., 2020; Rae et al., 2022). Nevertheless, existing efforts on data curation have primarily focused on English, and recent work on multilingual pre-training has relied on automatically filtered versions of CommonCrawl. For instance, XLM-R was trained on CC100 (Conneau et al., 2020), mT5 was trained on mC4 (Xue et al., 2021), and XGLM was trained on CC100-XL (Lin et al., 2021), which were all obtained by running language identification on several CommonCrawl snapshots and filtering through language-agnostic approaches. Unfortunately, Kreutzer et al. (2021) identified major issues on the quality of such multilingual datasets, ranging from language identification errors to boilerplate and non-linguistic content. However, the practical impact of these issues has not been studied, and it is unclear the extent to which higher-quality data could lead to better performance in low-resource languages.

In this paper, we take representation learning in Basque as a case study, and explore tailored crawling (i.e., manually identifying and scraping websites with high-quality content) as an alternative to filtering CommonCrawl. We introduce EusCrawl, a new corpus for Basque comprising 12.5M documents from 33 websites with Creative Commons content. EusCrawl is similar in size to the Basque portion of CC100 and mC4, but it has substantially less issues and a higher perceived quality according to our blind audit with native annotators. However, we find that this improvement does not carry over to downstream NLU tasks, as masked language models pre-trained on either corpora obtain similar results on 5 benchmarks. Our results suggests that data quantity and domain coverage play a more important role, prompting for methods to exploit more diverse sources of data in low-resource languages.

This paper makes the following contributions: (i) we release EusCrawl, a high-quality corpus for Basque comprising 12.5M documents and 423M tokens;[1] (ii) we manually assess the quality of EusCrawl in comparison with mC4 and CC100, finding that it has substantially less issues and a higher perceived quality according to native annotators; (iii) we compare masked language models pre-trained on EusCrawl, mC4, CC100 and Wikipedia[2] on 5 NLU tasks, finding that they all perform similarly with the exception of Wikipedia; and (iv)

[1] https://www.ixa.eus/euscrawl/. Meta AI was not involved in the collection and distribution of the corpus.

[2] Models available at https://dl.fbaipublicfiles.com/euscrawl/roberta-eus-{euscrawl|mc4|cc100|wikipedia}-{base|large}.tar.gz.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7383 - 7390

	Size	Tokens	Docs	Source
mC4 (Xue et al., 2021)	4,387 MiB	1,004M	30,098k	Filtered CommonCrawl
CC100 (Conneau et al., 2020)	2,027 MiB	416M	16,761k	Filtered CommonCrawl
Wikipedia	313 MiB	66M	2,685k	Wikipedia dump
EusCrawl (ours)	2,149 MiB	423M	12,528k	Tailored crawling (see Table 2)

Table 1: Basque corpora used in our experiments. We report uncompressed text size, number of SentencePiece tokens (using a 50K vocabulary learned in each corpus), and number of documents.

	Size	Tokens	Docs	License	Domain
Tokikom†	784 MiB	153M	4,961k	CC-BY-SA	Local media
Berria	525 MiB	101M	2,193k	CC-BY-SA	National newspaper
Hitza‡	418 MiB	80M	2,257k	CC-BY-NC-ND	Regional newspapers
Wikipedia	313 MiB	68M	2,685k	CC-BY-SA	Encyclopedia
Argia	101 MiB	20M	370k	CC-BY-SA	News magazine
Bilbo Hiria irratia	7 MiB	1M	54k	CC-BY-NC-SA	Radio station
Sarean	2 MiB	0.3M	8k	CC-BY-SA	Technology blog

Table 2: Data sources used to build EusCrawl. †Tokikom is a network of local media; we include Aiaraldea, Aikor, Anboto, Tolosaldeko Ataria, Aiurri, Erran, Euskalerria Irratia, Goiena, Guaixe, Hiruka, Karkara, Maxixatzen, Plaentxia, Alea, Noaua, Txintxarri, Uztarria, Amezti, Zarauzko Hitza, Kronika and Geuria. ‡Hitza is a family of regional newspapers; we include Bidasoko Hitza, Busturialdeko Hitza, Goierriko Hitza, Irutxuloko Hitza, Lea-Artibai eta Mutrikuko Hitza, Oarsoaldeko Hitza and Urola Kostako Hitza.

we obtain state-of-the-art results on several NLU benchmarks in Basque, outperforming prior work that relied on non-public corpora.

2 Experimental setup

We next detail the corpora compared in our experiments (§2.1), and the qualitative and downstream evaluation settings (§2.2 and §2.3).

2.1 Corpora

We compare 4 Basque corpora in our experiments: mC4, CC100, Wikipedia and EusCrawl. Table 1 summarizes their details. **mC4**[3] and **CC100**[4] are, to the best of our knowledge, the two largest public corpora for Basque. They were introduced to train mT5 (Xue et al., 2021) and XLM-R (Conneau et al., 2020), respectively, and were built by filtering CommonCrawl. **Wikipedia** has been a popular source for multilingual data (Pires et al., 2019; Conneau and Lample, 2019; Artetxe et al., 2020). We extract text from a Wikipedia dump using the WikiExtractor tool.[5] **EusCrawl** is a new corpus we introduce. Instead of filtering CommonCrawl, we do tailored crawling on 33 websites with high-quality content in Basque, mostly on the news domain. We build ad-hoc scrapers to extract text from these websites, resulting in higher coverage[6] and cleaner text compared to general purpose approaches. We only use content with a Creative Commons license. Table 2 summarizes all the sources we use.

2.2 Qualitative evaluation

We manually audit the quality of EusCrawl in comparison with mC4 and CC100 by randomly sampling 100 documents from each corpus (a total of 300 documents), and asking native annotators to assess their quality.[7] We ensure that the evaluation is blind by showing the documents in a random order and not revealing what corpus they were sampled from. For each document, we ask the annotators to assess if the document has any problem in each of the following categories: **langID** (the document is not in Basque), **language variety** (the document is not written in standard and correct Basque), **coherence** (the document has gaps and/or some portions are not connected), **noise** (the document is

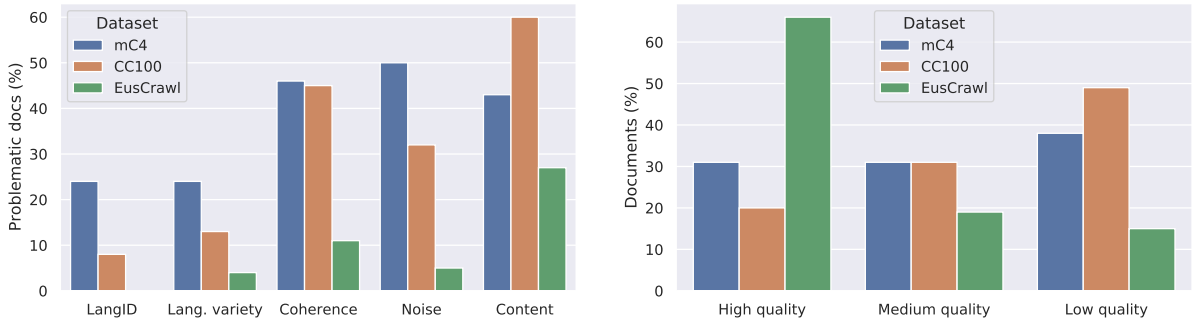
[3]We use the version released by AllenAI at https://github.com/allenai/allennlp/discussions/5265

[4]We use the version from https://data.statmt.org/cc-100/

[5]https://github.com/attardi/wikiextractor

[6]While one may expect the websites we crawl to be covered by mC4 and CC100, a large fraction of this content is missing in them. This is both because CommonCrawl is far from being a complete dump of the Internet, and the filtering applied by CC100 and mC4 is noisy, removing valid content.

[7]So as to control for the variance across annotators, we asked two additional native speakers to evaluate a random subset of 100 documents. The main findings were consistent across all the 3 annotations, so we omit results for brevity.



(a) Issues

(b) Overall quality

Figure 1: Data audit results. EusCrawl has a much higher quality than mC4 and CC100. See §2.2 for more details.

not clean) and **content** (the document seems to have been generated automatically and/or has no meat). In addition, we ask annotators to classify each document according to its **perceived quality** as high-quality (the document does not have quality issues and the annotator thinks that it would be good to have it in the corpus), medium-quality (the document has some minor issues and the annotator is unsure if it would be good to have it in the corpus), or low-quality (the document has major issues and the annotator thinks that it would be better not to have it in the corpus). Refer to Appendix A for the complete instructions given to annotators.

2.3 Downstream evaluation

In addition to the qualitative evaluation, we pre-train RoBERTa models (Liu et al., 2019) on each corpus, and evaluate fine-tuning them on the following **NLU benchmarks**: topic classification on BHTC (Agerri et al., 2020), sentiment classification on Behagune (Agerri et al., 2020), stance detection on VaxxStance (Agerri et al., 2021), Named Entity Recognition (NER) on EIEC (Alegria et al., 2006), and extractive conversational Question Answering (QA) on Elkarrizketak (Otegi et al., 2020). We provide additional details on these datasets in Appendix B.

We **pre-train** each model for 125k steps with a batch size of 2048 and a sequence length of 512, using the same hyperparameters as Liu et al. (2019). We train RoBERTa-base models for our main comparison using a learning rate of 7e-4, and further train a RoBERTa-large model on EusCrawl with a learning rate of 4e-4 to understand the effect of scaling. In all cases, we use the final checkpoint without early stopping. We use SentencePiece (Kudo and Richardson, 2018) for tokenization, using a 50k vocabulary learned in each separate corpus.

For **fine-tuning**, we use the same hyperparameters as Agerri et al. (2020). For topic classification, sentiment classification and stance detection, we use a batch size of 16, a learning rate of 2e-5 with linear decay and a warmup of 6%, and train the model for 10 epochs. For NER and QA, we use a batch size of 32, a constant learning rate of 5e-5, and train for 4 epochs. We did not perform any hyperparameter tuning or model selection, and report results on the test set. The development sets, when available, were not used.

3 Results

3.1 Qualitative evaluation

As shown in Figure 1, EusCrawl has the best quality by a large margin in all the axes that we consider. mC4 has a slightly higher perceived quality and less content-related issues than CC100, but more problematic documents in the other categories.

More concretely, we find that both mC4 and CC100 have a high proportion of documents with coherence, noise and content-related issues. In addition, mC4 has a significant number of langID and language variety problems. In contrast, EusCrawl has minimal issues in all categories but content, where it still does substantially better than mC4 and CC100. Taking a closer look, we find that most of these content-related issues in EusCrawl correspond to short, template-based Wikipedia articles (e.g., *Placosoma is a a genus of lizards in the family Gymnophthalmidae. They live in Brazil.*[8]), which should be easy to filter in future iterations. Finally, we find that the overall quality of EusCrawl documents is also much better according to native annotators, with approximately two thirds of the docu-

[8]Original text in Basque: *Placosoma Gymnophthalmidae familiako narrasti genero bat da. Brasilen bizi dira.*

		Topic class.	Sentiment	Stance det.	NER	QA	Avg
Prior best	Agerri et al. (2020)	76.8	78.1	–	87.1	–	–
	Otegi et al. (2020)	–	–	–	–	35.0	–
	Lai et al. (2021)	–	–	57.3†	–	–	–
RoBERTa-base	mC4	75.3 ±0.7	**80.4** ±1.5	59.1 ±5.2	86.0 ±1.0	35.2 ±1.8	67.2
	CC100	76.2 ±0.4	78.8 ±1.2	**63.4** ±3.5	85.2 ±1.2	35.8 ±1.1	67.9
	Wikipedia	70.0 ±0.8	72.4 ±2.3	53.2 ±4.6	71.6 ±13.1	27.4 ±0.2	58.9
	EusCrawl	76.2 ±0.6	77.7 ±1.4	57.4 ±4.7	86.8 ±0.6	34.6 ±1.8	66.5
RoBERTa-large	EusCrawl	**77.6** ±0.5	78.8 ±0.9	62.9 ±2.3	**87.2** ±0.4	**38.3** ±1.3	**69.0**

Table 3: Downstream results. We report average F1 and standard deviation across 5 runs (micro F1 in all tasks except stance detection, where we report macro F1 of the *favor* and *against* classes following common practice). †Best result among systems that rely exclusively on textual data.

ments being annotated as high-quality, compared to less than one third for both mC4 and CC100.

All in all, our qualitative evaluation provides further evidence that multilingual corpora derived from CommonCrawl have major quality issues, and shows that tailored crawling can be an effective alternative to obtain high-quality data.

3.2 Downstream tasks

We report our downstream results in Table 3.

In contrast with the qualitative evaluation, we find that there is not a clear winner among mC4, CC100 and EusCrawl. In fact, when looking at RoBERTa-base results, we find that mC4 does the best on sentiment classification, CC100 does the best on stance detection and QA, and EusCrawl does the best on NER. Wikipedia lags behind them all by a large margin. It is worth noting that the variance is high in certain tasks, which we attribute to the small size of the test sets and their unbalanced nature, but the general trends are consistent.

These results suggest that corpus quality issues in low-resource languages do not have a a major impact on NLU performance. Instead, we find evidence that it is the size and domain of the training corpus that is more important. This would explain why Wikipedia obtains the worst results, as it is substantially smaller than the other corpora and restricted to a narrow domain. Similarly, this is also consistent with EusCrawl performing worse than mC4 and CC100 on sentiment analysis and stance detection, as the domain of these benchmarks (tweets) is different from the domain of EusCrawl (primarily news, see Table 2), while CommonCrawl-derived corpora are presumably more diverse.

Finally, we find that scaling to RoBERTa-large brings consistent improvements in all tasks. Thanks to this, we are able to outperform the best published results in all the 5 benchmarks. Note that we achieve this pre-training exclusively on Creative Commons data that we release publicly, while prior work relied on private datasets.

4 Conclusions

Taking Basque as a case study, our work gives further evidence that CommonCrawl-derived corpora have major quality issues in low-resource languages. At the same time, we show that ad-hoc crawling websites with high-quality content can be an effective alternative to collect data in such languages. Our resulting corpus EusCrawl has a higher quality than mC4 and CC100 according to our manual data audit, while being similar in size. Nevertheless, this improvement in quality does not carry over to downstream performance on NLU tasks, where we find evidence that data quantity and domain coverage are more important factors.

Our work leaves important lessons for future efforts on low-resource languages. First of all, we find that, even if CommonCrawl derived multilingual corpora do have major quality issues as raised by prior work (Kreutzer et al., 2021), these issues do not have a significant impact in NLU tasks. This suggests that investing on bigger and more diverse datasets might be more fruitful than addressing such quality issues in low-resource settings. Given that the amount of written text in such languages is ultimately limited, we believe that developing effective cross-lingual transfer methods to exploit multilingual data is a promising future direction. Having said that, it should be noted that our study is limited to NLU tasks in a single language. It is possible that data quality plays a more important role in generation tasks, which we leave for future work to study. In addition, we think that

it would be valuable to conduct similar studies in other languages to corroborate our findings.

Finally, we note that prior work on Basque NLP has often relied on private resources (Agerri et al., 2020). Our work sets a new state-of-the-art on a diverse set of NLU benchmarks, and it does so using public data alone. By releasing our corpus, we hope to facilitate future work in Basque NLP, and encourage open and reproducible science using public resources.

Limitations

Our evaluation focuses on NLU tasks, and it is possible that data quality plays a different role in generation tasks. We note, however, that generation quality is harder to evaluate through automatic metrics, which is why we decided to focus on NLU tasks. Moreover, the corpora that we compare differ on various aspects other than the data quality (e.g., the domain), and it is hard to isolate the effect of quality from the rest. In any case, we believe that our main claim still holds, in that data quality has a minor impact relative to such other factors. Finally, our work builds on EusCrawl—a new high-quality corpus that we introduce for Basque—and our analysis is thus limited to this language. It would be interesting to collect high-quality corpora for other low-resource languages, and conduct a similar comparison to corroborate that our findings also apply more broadly.

Acknowledgments

We would like to thank Nikolas Vicuña for his help on expanding EusCrawl, Naman Goyal for his advice on pre-training RoBERTa, and Arantza Rico and Paloma Rodriguez-Miñambres for their help with the manual data audit. We are also grateful to all Basque media and content creators that share their work under a Creative Commons license, making a resource like EusCrawl possible.

Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de-Viñaspre and Aitor Soroa were supported by the Basque Government (excellence research group IT1343-19 and DeepText project KK-2020/00088), and by the projects: (i) DeepKnowledge (PID2021-127777OB-C21) funded by MCIN/AEI/10.13039/501100011033 and FEDER Una manera de hacer Europa; (ii) Disargue (TED2021-130810B-C21), MCIN/AEI /10.13039/501100011033 and European Union NextGenerationEU/PRTR; (iii) DeepR3 (TED2021-130295B-C31) funded by MCIN/AEI /10.13039/501100011033 and EU NextGeneration programme EU/PRTR. Rodrigo Agerri's work is also supported by the RYC-2017-23647 fellowship (MCIN/AEI /10.13039/501100011033 y por El FSE invierte en tu futuro).

References

Rodrigo Agerri, Roberto Centeno, María Espinosa, Joseba Fernandez de Landa, and Álvaro Rodrigo Yuste. 2021. Vaxxstance@iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

Iñaki Alegria, Olatz Arregi, Nerea Ezeiza, and Izaskun Fernández. 2006. Lessons from the development of a named entity recognizer for Basque. *Procesamiento del Lenguaje Natural*, 36:25–37.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

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, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark 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 Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani,

Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan 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 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.

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*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

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 Suárez, 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. 2021. Quality at a glance: An audit of web-crawled multilingual datasets.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Mirko Lai, Alessandra Teresa Cignarella, Livio Finos, and Andrea Sciandra. 2021. Wordup! at vaxxstance 2021: Combining contextual information with textual and dependency-based syntactic features for stance detection. In *IberLEF@SEPLN*.

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. 2021. Few-shot learning with multilingual language models.

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.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442, Marseille, France. European Language Resources Association.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

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.

A Annotation instructions

Table 4 reports the complete instructions used for the qualitative evaluation as given to the annotators.

B Downstream evaluation

We next provide additional details on the datasets used for downstream evaluation:

- **Topic classification**: The Basque Headlines Topic Classification (BHTC) dataset (Agerri et al., 2020) contains 12k headlines from the Argia news magazine classified into 12 thematic categories[9]. We use the standard splits containing 8662 examples for training, 1861 for development and 1860 for testing.

- **Sentiment classification**: The Behagune dataset[10] comprises 2936 tweets in Basque labeled as positive, negative or neutral. We used the same splits for train (80%), test (10%) and development (10%) as in Agerri et al. (2020).

- **Stance detection**: We used the VaxxStance dataset (Agerri et al., 2021), which offers tweets labeled as expressing an AGAINST, FAVOR or NEUTRAL stance with respect to vaccines. It contains 1070 tweets for training and 313 for testing[11].

- **Named Entity Recognition (NER)**: EIEC[12] (Alegria et al., 2006) is a Basque NER dataset composed of 44K training tokens (3817 unique entities) and 15K test tokens (931 entities).

- **Question Answering (QA)**: Elkarrizketak is an extractive conversational QA dataset (Otegi et al., 2020) that contains 377 dialogues (301 train, 38 development and 38 test) and 1,634 question/answer pairs (1,306 train, 161 development and 167 test)[13].

[9] https://hizkuntzateknologiak.elhuyar.eus/assets/files/bhtc.tgz

[10] https://hizkuntzateknologiak.elhuyar.eus/assets/files/behaguneadss2016-dataset.tgz

[11] https://vaxxstance.github.io/

[12] http://ixa2.si.ehu.eus/eiec/eiec_v1.0.tgz

[13] http://ixa.si.ehu.es/node/12934

LangID	EGOKIA: Dokumentua euskaraz dago. *CORRECT: The document is in Basque.*
	ARAZOAK: Dokumentuaren zati esanguratsu bat ez dago euskaraz. *PROBLEMATIC: A significant portion of the document is not in Basque.*
Hizkuntza *Lang. variety*	EGOKIA: Dokumentua hizkuntza estandar eta zuzenean idatzia dago. *CORRECT: The document is written in standard and correct language.*
	ARAZOAK: Dokumentua ez dago hizkuntza estandar edo zuzenean idatzia (adb. euskalkiren batean dago ala itzulpen automatikoaren bidez sortua dirudi). *PROBLEMATIC: The document is not written in standard and correct language (e.g., it is written in a dialect using non-standard Basque, or it seems to be generated through machine translation).*
Koherentzia *Coherence*	EGOKIA: Dokumentua koherentea da, eta hasieratik bukaerara unitate bat osatzen du. *CORRECT: The document is coherent, and it constitutes a single unit from the beginning to the end.*
	ARAZOAK: Dokumentua ez da koherentea: hutsuneak ditu edota atal batzuk ez dute elkarren artean loturarik (dokumentu ezberdinak dirudite). *PROBLEMATIC: The document is not coherent: it has gaps and/or some portions do not seem connected (they seem to come from separate documents).*
Garbitasuna *Noise*	EGOKIA: Dokumentuko testua garbia da. *CORRECT: The text in the document is clean.*
	ARAZOAK: Dokumentua ez da erabat garbia, eta benetako testuaz gain webguneko bestelako elementuak daude (menuetako testua, html kodea...). *PROBLEMATIC: The document is not entirely clean, and there are other elements in addition to the real content (text from menus, HTML code...).*
Edukia *Content*	EGOKIA: Dokumentua pertsona batek sortua dirudi eta gutxieneko mami bat du. *CORRECT: The document seems to have been created by a human and has some minimum meat.*
	ARAZOAK: Dokumentuak automatikoki sortua dirudi edota ez du inolako mamirik (adb futbol ligako sailkapen-taula). *The document seems to have been generated automatically and/or has no meat at all (e.g., a soccer standing table).*
Kalitate orokorra *Overall quality*	ALTUA: Dokumentua kalitatezkoa da, eta corpusean izatea komeniko litzatekeela uste dut. *HIGH: The document is of good quality, and I think that it would be good to have it in the corpus.*
	ERTAINA: Dokumentuak arazo batzuk ditu baina ez dira larriak, eta ez nago ziur ea corpusean izatea komeniko litzatekeen. *MEDIUM: The document has minor issues, and I am not sure if it would be good to have it in the corpus.*
	BAXUA: Dokumentuak arazo nabarmenak ditu. Ez dut uste corpusean izatea komeniko litzatekeenik. *LOW: The document has major issues. I think that it would be better not to have it in the corpus.*

Table 4: Annotation instructions used for the qualitative evaluation. We report the original instructions in Basque, as well as the corresponding translation into English.