An Open Dataset and Model for Language Identification

Laurie Burchell and **Alexandra Birch** and **Nikolay Bogoychev** and **Kenneth Heafield**
Institute for Language, Cognition, and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, UK
{laurie.burchell,a.birch,n.bogoych,kenneth.heafield}@ed.ac.uk

Abstract

Language identification (LID) is a fundamental step in many natural language processing pipelines. However, current LID systems are far from perfect, particularly on lower-resource languages. We present a LID model which achieves a macro-average F1 score of 0.93 and a false positive rate of 0.033% across 201 languages, outperforming previous work. We achieve this by training on a curated dataset of monolingual data, the reliability of which we ensure by auditing a sample from each source and each language manually. We make both the model and the dataset available to the research community. Finally, we carry out detailed analysis into our model's performance, both in comparison to existing open models and by language class.

1 Introduction

Language identification (LID) is a foundational step in many natural language processing (NLP) pipelines. It is used not only to select data in the relevant language but also to exclude 'noise'. For this reason, effective LID systems are key for building useful and representative NLP applications.

Despite their importance, recent work has found that existing LID algorithms perform poorly in practice compared to test performance (Caswell et al., 2020). The problem is particularly acute for low-resource languages: Kreutzer et al. (2022) found a positive Spearman rank correlation between quality of data and size of language for all of the LID-filtered multilingual datasets they studied. In addition, for a significant fraction of the language corpora they studied, less than half of the sentences were in the correct language. They point out that such low-quality data not only leads to poor performance in downstream tasks, but that it also contributes to 'representation washing', where the community is given a false view of the actual progress of low-resource NLP.

For applications such as corpus filtering, LID systems need to be fast, reliable, and cover as many languages as possible. There are several open LID models offering quick classification and high language coverage, such as CLD3 or the work of Costa-jussà et al. (2022). However, to the best of our knowledge, none of the commonly-used scalable LID systems make their training data public. This paper addresses this gap through the following contributions:

- We provide a curated and open dataset covering 201 languages. We audit a sample from each source and each language making up this dataset manually to ensure quality.
- We train a LID model on this dataset which outperforms previous open models. We make this model publicly available.[1]
- We analyse our model and use our findings to highlight open problems in LID research.

2 Background

There is a long history of research into LID using a plethora of methods (Jauhiainen et al., 2019). For high-coverage LID, Dunn (2020) presents a model covering 464 languages, whilst Brown (2014) includes as many as 1366 language varieties. Unlike our work, the training data in both cases has not been manually checked for quality. Recent work by Adebara et al. (2022) presents a LID system covering 517 African languages and varieties where the training data has been curated manually. However, as far as we are aware this data is not easily available.

Costa-jussà et al. (2022) released a substantial piece of research aiming to improve machine translation coverage for over 200 languages. As part of this, they provided several professionally-translated datasets for use as test and development sets. For

[1] github.com/laurieburchell/open-lid-dataset

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 865–879
July 9-14, 2023

this reason, we use their system as our benchmark. However, whilst they did release scripts to recreate their parallel data,[2] they did not provide—or even document—the monolingual data used to train their LID system, saying only that they use "publicly available datasets" supplemented with their own dataset NLLB-Seed. By providing an open dataset, we aim to facilitate futher research.

3 Dataset

3.1 Data sources

We wanted to be as confident as possible that our dataset had reliable language labels, so as to avoid the problems noted in existing corpora (Kreutzer et al., 2022). We therefore avoided web-crawled datasets and instead chose sources where we felt the collection methodology made it very likely that the language labels were correct.

The majority of our source datasets were derived from news sites, Wikipedia, or religious text, though some come from other domains (e.g. transcribed conversations, literature, or social media). A drawback of this approach is that most of the text is in a formal style. Further work could collect data from a wider range of domains whilst maintaining trust in the labels. We checked that each dataset was either under an open license for research purposes or described as free to use. A full list of sources is given in Appendix A, and further information including licenses is available in the code repository accompanying this paper.

3.1.1 Language selection

Our initial aim was to cover the same languages present in the FLORES-200 Evaluation Benchmark[3] so that we could use this dataset for evaluation and compare our results directly with Costa-jussà et al. (2022). However, during the curation process, we decided to exclude three languages.

Firstly, though Akan and Twi are both included as separate languages in FLORES-200, Akan is actually a macrolanguage covering a language continuum which includes Twi. Given the other languages in FLORES-200 are individual languages, we decided to exclude Akan.

Secondly, FLORES-200 includes Modern Standard Arabic (MSA) written in Latin script. It is true that Arabic dialects are often written in Latin characters in informal situations (e.g. social media). However, MSA is a form of standardised Arabic which is not usually used in informal situations. Since we could not any find naturally-occurring training data, we excluded MSA from the dataset.

Finally, we excluded Minangkabau in Arabic script because it is now rarely written this way, making it difficult to find useful training data.[4]

3.2 Manual audit process

The first step in our manual audit was to check and standardise language labels, as these are often inconsistent or idiosyncratic (Kreutzer et al., 2022). We chose to copy the language codes in Costa-jussà et al. (2022), and reassign macrolanguage or ambiguous language codes in the data sources we found to the dominant individual language. Whilst this resulted in more useful data for some languages, for other languages we had to be more conservative. For example, we originally reassigned text labelled as the macrolanguage Malay (*msa_Latn*) to Standard Malay, but this led to a large drop in performance as the former covers a very diverse set of languages.

Two of the authors then carried out a manual audit of a random sample of all data sources and languages:[5] one a native Bulgarian speaker (able to read Cyrillic and Latin scripts and Chinese characters), and the other a native English speaker (able to read Latin, Arabic and Hebrew scripts). For languages we knew, we checked the language was what we expected. For unfamiliar languages in a script we could read, we compared the sample to the Universal Declaration of Human Rights (UDHR) or failing that, to a sample of text on Wikipedia. We compared features of the text which are common in previous LID algorithms and could be identified easily by humans: similar diacritics, word lengths, common words, loan words matching the right cultural background, similar suffixes and prefixes, and vowel/consonant patterns (Jauhiainen et al., 2019, Section 5). For scripts we could not read, we checked that all lines of the sample matched the script in the UDHR.

3.3 Preprocessing

We kept preprocessing minimal so that the process was as language agnostic as possible. We used the

[2] `github.com/facebookresearch/fairseq/tree/nllb`
[3] `github.com/facebookresearch/flores/blob/main/flores200`
[4] `omniglot.com/writing/minangkabau.htm`, `ethnologue.com/language/min`
[5] Specifically, we used the following command on each file to select lines to audit: `shuf <file> | head -n 500 | less`

scripts provided with Moses (Koehn et al., 2007) to remove non-printing characters and detokenise the data where necessary. We then filtered the data so that each line contained at least one character in the expected script (as defined by Perl) to allow for borrowings. Finally, we followed Arivazhagan et al. (2019) and Costa-jussà et al. (2022) and sampled proportionally to $p_l^{0.3}$, where p_l is the fraction of lines in the dataset which are in language l. This aims to ameliorate class skew issues.

3.4 Dataset description

The final dataset contains 121 million lines of data in 201 language classes. Before sampling, the mean number of lines per language is 602,812. The smallest class contains 532 lines of data (South Azerbaijani) and the largest contains 7.5 million lines of data (English). There is a full breakdown of lines of training data by language in Appendix C.

4 Model and hardware

We used our open dataset to train a *fasttext* LID model using the command-line tool (Joulin et al., 2017). It embeds character-level n-grams from the input text, and then uses these as input to a multi-class linear classifier. We used the same hyperparameters as Costa-jussà et al. (2022) (NLLB), which we list in Appendix B. We trained our model on one Ice Lake node of the CSD3 HPC service. Each node has 76 CPUs and 256GiB of RAM. Our model takes c. 1hr 45mins to train and contains 60.5 million parameters. Inference over the 206,448 lines of the test set takes 22.4 secs (9216.4 lines/sec).

5 Evaluation

5.1 Test sets

We use the FLORES-200 benchmark provided by Costa-jussà et al. (2022) for evaluation. It consists of 842 distinct web articles sourced from English-language Wikimedia projects, with each sentence professionally translated into 204 languages. The target side is human-verified as in the right language, making it suitable for use as a LID evaluation set. For each language, 997 sentences are available for development and 1012 for dev-test (our test set).[6] We remove the three languages discussed in Section 3.1.1 from FLORES-200, leaving 201 languages in the test set: FLORES-200*.

[6] 992 sentences are withheld by Costa-jussà et al. (2022) as a hidden test set.

5.2 Other LID systems

We compare our model's performance to two other open-source LID systems: `nllb218e` (NLLB)[7] and `pycld3 0.22` (CLD3).[8] We discuss how we ensured a fair comparison below.

NLLB is a *fasttext* model. We were surprised to discover that whilst it does cover 218 languages, it only includes 193 of the 201 languages in FLORES-200*. This is despite the fact that the NLLB LID model and the original FLORES-200 evaluation set were created as part of the same work (Costa-jussà et al., 2022). Referring to the analysis in the original paper, the authors note that "Arabic languoids and Akan/Twi have been merged after linguistic analysis" (Costa-jussà et al., 2022, Table 5, p. 32). We discuss the reason to merge Akan and Twi in Section 3.1.1, but we judge Arabic dialects to be close but distinct languages. Our model performs poorly on Arabic dialects with the highest F1 score only 0.4894 (Moroccan Arabic). This is likely due to the general difficulty of distinguishing close languages combined with particularly sparse training data. We assume these poor results led to Arabic dialects (save MSA) being excluded from the NLLB LID classifier. We remove eight Arabic dialects from the test set when comparing our model and NLLB, leaving 193 languages.

CLD3 is an n-gram based neural network model for LID. It uses different language codes to the other two models, so we normalise all predictions to BCP-47 macrolanguage codes to allow fair comparison. We test on the 95 languages that all models have in common after normalisation.

6 Results

Our results are given in Table 1. We evaluate all models using F1 scores and false positive rate (FPR). We report macro-averages to avoid down-weighting low-resource languages (Kreutzer et al., 2022). Following Caswell et al. (2020), we report FPR to give a better indication of real-world performance when there is significant class skew.

We achieve an F1 score of 0.927 and a FPR of 0.033% on FLORES-200*. We also outperform both NLLB and CLD3 on the mutual subsets of FLORES-200*. Since NLLB and our model share the same architecture and the same parameters, we attribute our success to our training data selection and manual audit process.

[7] `tinyurl.com/nllblid218e`
[8] `pypi.org/project/pycld3`

System	Supported languages.	FLORES-200* 201 languages		FLORES200*∩ NLLB 193 languages		FLORES-200*∩ CLD3 95 languages	
		F1 ↑	FPR ↓	F1 ↑	FPR ↓	F1 ↑	FPR ↓
CLD3	107	-	-	-	-	0.968	0.030
NLLB	218	-	-	0.950	0.023	0.985	0.019
Our model	201	**0.927**	**0.033**	**0.959**	**0.020**	**0.989**	**0.011**

Table 1: A comparison of open-source LID systems. *Supported languages* gives the number of languages the classifier claims to support. Each column gives the classifier's performance on a test set containing the intersection of languages each classifier claims to support. We report macro-averages of F1 scores and false positive rates (FPRs).

Notably, our F1 score jumps to 0.959 and FPR falls to 0.020% when we exclude the eight Arabic dialects from the test set to compare with NLLB. The 95 languages covered by CLD3, NLLB, and our model are mostly high resource, and so it is unsurprising that we achieve the highest F1 score (0.989) and lowest FPR (0.011%) on this subset.

We notice that the Pearson correlation between the number of lines of training data and F1 score for each language is only 0.0242. This is not unexpected: some of the least resourced languages achieve perfect scores on the test set due to high domain overlap, whereas the higher-resourced languages might get lower scores on the test set but have better robustness across domains. Full results by language are available in Appendix C.

6.1 Performance by language category

Using the taxonomy and list of languages in Joshi et al. (2020), we label each of the languages in our dataset according to its level of data availability (0 = least resourced, 5 = best resourced). We leave out 5 languages missing from the taxonomy, plus the 8 Arabic dialects not covered by NLLB. Table 2 compares the mean F1 score and FPR of our model and for that of Costa-jussà et al. (2022) (NLLB). Our model has a higher or equal F1 score in every category and a lower or equal FPR in every category but one, showing our model's improved performance across languages with different amounts of available data.

We note that class zero (the least-resourced languages) shows the smallest change in performance. We speculate that this is an artifact of the curation of our training dataset. For the best-resourced languages with more sources to choose from, it is likely that there is a significant difference between our training data and that used to train the model in Costa-jussà et al. (2022). However, for the least-resourced languages, the sheer lack of resources means that overlap between our data and that used by Costa-jussà et al. (2022) is more likely. We suspect this is the reason we see little difference in performance for class zero in Table 2. Unfortunately, without access to the training data used to train NLLB, we cannot verify this assumption.

Class	Count	F1 ↑ Ours	F1 ↑ NLLB	FPR ↓ Ours	FPR ↓ NLLB
0	28	**0.900**	0.897	0.014	**0.013**
1	94	**0.981**	0.968	**0.013**	**0.013**
2	16	**0.990**	0.963	**0.009**	0.043
3	25	**0.983**	0.974	**0.007**	0.013
4	18	**0.951**	**0.951**	**0.051**	0.055
5	7	**0.897**	0.855	**0.163**	0.620

Table 2: For each language class in the taxonomy of Joshi et al. (2020), we give the count of the languages covered by the classifier in that class, mean F1 score, and mean FPR for our model and for that of Costa-jussà et al. (2022) (NLLB). 0–5 = least to best resourced.

6.2 Case study: Chinese languages

Despite our model outperforming NLLB overall, NLLB achieved a noticeably higher F1 score on Yue Chinese (0.488 vs. 0.006). Figure 1 shows the confusion matrices for our model and NLLB between the three Chinese languages. Our model performs well on Simplified and Traditional Chinese, but almost never predicts Yue Chinese, instead classifying it as Chinese (Traditional). The NLLB model is also unable to distinguish between Yue and Chinese (Traditional), but mixes the two classes instead.

We asked four native speakers to inspect our training data and the FLORES-200 test set. They noted that there was a mismatch in domain for Yue Chinese, as much of our training data was written colloquial Yue Chinese whereas the test set consisted of formal writing. Furthermore, they were unable to distinguish with high confidence between Yue and Chinese (Traditional) as the two languages are very similar when written formally. This is an example of a wider problem with LID:

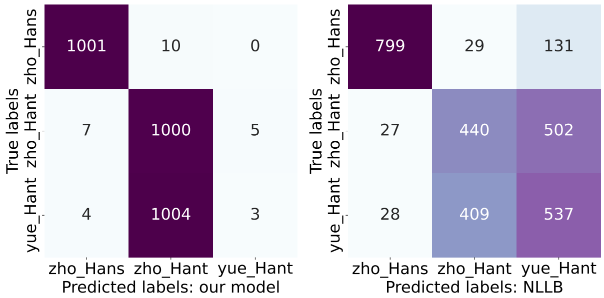

Figure 1: Confusion matrices for our model (L) and NLLB (R), showing the confusion in classification by each model on the FLORES-200 test set between Chinese (Simplified) (*zho_Hans*), Chinese (Traditional) (*zho_Hant*), and Yue Chinese (*yue_Hant*) classes.

the language covered by a particular label may vary widely, making single-label classification difficult.

7 Conclusion

We present an open dataset covering 201 languages, which we curate and audit manually to ensure high confidence in its data and language labels. We demonstrate the quality of our dataset by using it to train a high-performing and scalable LID model. Finally, we provide detailed analysis into its performance by class. We make both our model and our dataset available to the research community.

Limitations

Our dataset and model only covers 201 languages: the ones we were able to test with the FLORES-200 Evaluation Benchmark. In addition, because our test set consists of sentences from a single domain (wiki articles), performance on this test set may not reflect how well our classifier works in other domains. Future work could create a LID test set representative of web data where these classifiers are often applied. Finally, most of the data was not audited by native speakers as would be ideal. Future versions of this dataset should have more languages verified by native speakers, with a focus on the least resourced languages.

Ethics Statement

Our work aims to broaden NLP coverage by allowing practitioners to identify relevant data in more languages. However, we note that LID is inherently a normative activity that risks excluding minority dialects, scripts, or entire microlanguages from a macrolanguage. Choosing which languages to cover may reinforce power imbalances, as only some groups gain access to NLP technologies.

In addition, errors in LID can have a significant impact on downstream performance, particularly (as is often the case) when a system is used as a 'black box'. The performance of our classifier is not equal across languages which could lead to worse downstream performance for particular groups. We mitigate this by providing metrics by class.

Acknowledgements

This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology & Language Sciences.

The experiments in this paper were performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

Special thanks to Pinzhen Chen, Steven Chien, Bryan Li, Lushi Chen and Victoria Lee for their help with Chinese languages.

References

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. 2018. DART: A large dataset of dialectal Arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Ralf Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–632, Doha, Qatar. Association for Computational Linguistics.

Ralf D Brown. 2012. Finding and identifying text in 900+ languages. *Digital Investigation*, 9:S34–S43.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Jonathan Dunn. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.*, pages 3622–3627. European Language Resources Association.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2022. Can we use word embeddings for enhancing Guarani-Spanish machine translation? In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 127–132, Dublin, Ireland. Association for Computational Linguistics.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Rudali Huidrom, Yves Lepage, and Khogendra Khomdram. 2021. EM corpus: a comparable corpus for a less-resourced language pair Manipuri-English. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 60–67, Online (Virtual Mode). INCOMA Ltd.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Omid Kashefi. 2018. Mizan: A large persian-english parallel corpus. *arXiv preprint arXiv:1801.02107*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kang Kwong Luke and May LY Wong. 2015. The hong kong cantonese corpus: design and uses. *Journal of Chinese Linguistics Monograph Series*, 1(25):312–333.

Salima Medhaffar, Fethi Bougares, Yannick Estève, and Lamia Hadrich-Belguith. 2017. Sentiment analysis of Tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 55–61, Valencia, Spain. Association for Computational Linguistics.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Atul Kr Ojha. 2019. English-bhojpuri smt system: Insights from the karaka model. *arXiv preprint arXiv:1905.02239*.

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. Tep: Tehran english-persian parallel corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 68–79. Springer.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Martin Thoma. 2018. The wili benchmark dataset for written language identification. *arXiv preprint arXiv:1801.07779*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jihad Zahir. 2022. Iadd: An integrated arabic dialect identification dataset. *Data in Brief*, 40:107777.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

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, Portorož, Slovenia. European Language Resources Association (ELRA).

A Data sources

We use the following data sources to build our open dataset. We chose sources as those which were likely to have trustworthy language labels and which did not rely on other LID systems for labelling.

- Arabic Dialects Dataset (El-Haj et al., 2018)
- Bhojpuri Language Technological Resources Project (BLTR) (Ojha, 2019)
- Global Voices (Tiedemann, 2012)
- Guaraní Parallel Set (Góngora et al., 2022)
- The Hong Kong Cantonese corpus (HKCanCor) (Luke and Wong, 2015)
- Integrated dataset for Arabic Dialect Identification (IADD) (Zahir, 2022; Alsarsour et al., 2018; Abu Kwaik et al., 2018; Medhaffar et al., 2017; Meftouh et al., 2015; Zaidan and Callison-Burch, 2011)
- Leipzig Corpora Collection (Goldhahn et al., 2012)
- LTI LangID Corpus (Brown, 2012)
- MADAR 2019 Shared Task on Arabic Fine-grained Dialect Identification (Bouamor et al., 2019)
- EM corpus (Huidrom et al., 2021)
- MIZAN (Kashefi, 2018)
- MT-560 (Gowda et al., 2021; Tiedemann, 2012; Post et al., 2012; Ziemski et al., 2016; Rozis and Skadiņš, 2017; Kunchukuttan et al., 2018; Agić and Vulić, 2019; Esplà et al., 2019; Qi et al., 2018; Zhang et al., 2020; Bojar et al., 2013, 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020)
- NLLB Seed (Costa-jussà et al., 2022)
- SETIMES news corpus (Tiedemann, 2012)
- Tatoeba collection (Tiedemann, 2012)
- Tehran English-Persian Parallel (TEP) Corpus (Pilevar et al., 2011)
- Turkish Interlingua (TIL) corpus (Mirzakhalov et al., 2021)
- WiLI benchmark dataset (Thoma, 2018)
- XL-Sum summarisation dataset (Hasan et al., 2021)

B LID model hyperparameters

- Loss: softmax
- Epochs: 2
- Learning rate: 0.8
- Embedding dimension: 256
- Minimum number of word occurences: 1000
- Character n-grams: 2–5
- Word n-grams: 1
- Bucket size: 1,000,000
- Threads: 68

All other hyperparameters are set to *fasttext* defaults.

C Performance of our LID model by language

			Our model		NLLB	
Language code	**Language**	**Training data**	**F1 score ↑**	**FPR ↓**	**F1 score ↑**	**FPR ↓**
ace_Arab	Acehnese	6191	0.9679	0.0079	0.9704	0.0074
ace_Latn	Acehnese	18032	0.9980	0.0005	0.9936	0.0035
acm_Arab	Mesopotamian Arabic	4862	0.0328	0.0040	-	-
acq_Arab	Ta'izzi-Adeni Arabic	1598	0.0020	0.0000	-	-
aeb_Arab	Tunisian Arabic	18758	0.3398	0.0479	-	-
afr_Latn	Afrikaans	1045638	0.9995	0.0000	0.9985	0.0010
ajp_Arab	South Levantine Arabic	28190	0.1906	0.0158	-	-
als_Latn	Tosk Albanian	506379	1.0000	0.0000	0.9980	0.0020
amh_Ethi	Amharic	606866	0.9995	0.0005	0.9990	0.0010
apc_Arab	North Levantine Arabic	67952	0.2334	0.0983	-	-
arb_Arab	Modern Standard Arabic	7000000	0.3077	1.1280	0.1903	4.2579
ars_Arab	Najdi Arabic	23194	0.0184	0.1374	-	-
ary_Arab	Moroccan Arabic	25411	0.4894	0.7643	-	-
arz_Arab	Egyptian Arabic	52327	0.4235	1.0875	-	-
asm_Beng	Assamese	161726	1.0000	0.0000	1.0000	0.0000
ast_Latn	Asturian	35815	0.9901	0.0045	0.9902	0.0069
awa_Deva	Awadhi	4957	0.6770	0.0040	0.9611	0.0084
ayr_Latn	Central Aymara	142628	1.0000	0.0000	0.9980	0.0005
azb_Arab	South Azerbaijani	532	0.7514	0.0000	0.8805	0.0069
azj_Latn	North Azerbaijani	462672	0.9990	0.0005	0.9970	0.0030
bak_Cyrl	Bashkir	65942	1.0000	0.0000	0.9990	0.0005
bam_Latn	Bambara	9538	0.6107	0.4926	0.6194	0.4826
ban_Latn	Balinese	15404	0.9789	0.0015	0.9712	0.0030
bel_Cyrl	Belarusian	84846	1.0000	0.0000	1.0000	0.0000
bem_Latn	Bemba	383559	0.9796	0.0193	0.9739	0.0252
ben_Beng	Bengali	490226	0.9925	0.0000	0.9995	0.0005
bho_Deva	Bhojpuri	69367	0.8921	0.1136	0.9335	0.0153
bjn_Arab	Banjar	6192	0.9604	0.0257	0.9524	0.0163
bjn_Latn	Banjar	21475	0.9857	0.0064	0.8336	0.1721
bod_Tibt	Standard Tibetan	2514	0.8045	0.0000	0.9637	0.0366
bos_Latn	Bosnian	330473	0.6928	0.0939	0.5954	0.0584
bug_Latn	Buginese	7527	0.9970	0.0005	0.9765	0.0054
bul_Cyrl	Bulgarian	610545	1.0000	0.0000	0.9995	0.0000
cat_Latn	Catalan	115963	1.0000	0.0000	0.9873	0.0129
ceb_Latn	Cebuano	1002342	0.9995	0.0005	0.9995	0.0000
ces_Latn	Czech	424828	0.9975	0.0015	0.9990	0.0010
cjk_Latn	Chokwe	36244	0.9023	0.0025	0.8688	0.0089
ckb_Arab	Central Kurdish	17792	1.0000	0.0000	1.0000	0.0000
crh_Latn	Crimean Tatar	19148	0.9920	0.0005	0.9829	0.0000
cym_Latn	Welsh	98719	1.0000	0.0000	1.0000	0.0000
dan_Latn	Danish	2789406	0.9881	0.0035	0.9946	0.0020
deu_Latn	German	653914	1.0000	0.0000	0.9907	0.0094
dik_Latn	Southwestern Dinka	25911	0.9995	0.0000	0.9925	0.0000
dyu_Latn	Dyula	17351	0.0421	0.0282	0.0480	0.0228
dzo_Tibt	Dzongkha	6899	0.8585	0.1635	0.9679	0.0005
ell_Grek	Greek	3312774	1.0000	0.0000	1.0000	0.0000
eng_Latn	English	7544560	0.9941	0.0049	0.9792	0.0213
epo_Latn	Esperanto	339280	1.0000	0.0000	0.9970	0.0030
est_Latn	Estonian	3331470	0.9990	0.0005	0.9985	0.0015
eus_Latn	Basque	622029	0.9990	0.0005	0.9985	0.0015
ewe_Latn	Ewe	585267	0.9980	0.0020	0.9970	0.0030
fao_Latn	Faroese	40022	1.0000	0.0000	0.5052	0.0000
fij_Latn	Fijian	360981	0.9985	0.0005	1.0000	0.0000
fin_Latn	Finnish	2613970	0.9995	0.0005	0.9995	0.0005
fon_Latn	Fon	31875	0.9980	0.0000	0.9970	0.0000
fra_Latn	French	586938	0.9950	0.0000	0.9961	0.0035
fur_Latn	Friulian	55622	0.9985	0.0015	0.9980	0.0000
fuv_Latn	Nigerian Fulfulde	14419	0.9865	0.0005	0.9810	0.0040
gaz_Latn	West Central Oromo	335769	0.9990	0.0010	0.9995	0.0005
gla_Latn	Scottish Gaelic	52665	0.9975	0.0025	0.9985	0.0010
gle_Latn	Irish	211460	1.0000	0.0000	0.9980	0.0020
glg_Latn	Galician	42017	0.9970	0.0025	0.9931	0.0049

Table 3: For each language covered by our model, we give the number of lines of deduplicated training data in our dataset, as well as the class F1 score and class false positive rate (FPR) for our model and for the model described in Costa-jussà et al. (2022) (NLLB).

			Our model		NLLB	
Language code	**Language**	**Training data**	**F1 score ↑**	**FPR ↓**	**F1 score ↑**	**FPR ↓**
grn_Latn	Guarani	57458	0.9975	0.0025	0.9965	0.0015
guj_Gujr	Gujarati	836618	1.0000	0.0000	1.0000	0.0000
hat_Latn	Haitian Creole	299853	0.9970	0.0030	0.9985	0.0005
hau_Latn	Hausa	347741	0.9893	0.0109	0.9970	0.0025
heb_Hebr	Hebrew	944918	0.9990	0.0010	1.0000	0.0000
hin_Deva	Hindi	1089471	0.8477	0.1749	0.8722	0.1454
hne_Deva	Chhattisgarhi	52819	0.9362	0.0311	0.9300	0.0134
hrv_Latn	Croatian	832967	0.7441	0.1863	0.7335	0.2645
hun_Latn	Hungarian	2870535	1.0000	0.0000	0.9926	0.0074
hye_Armn	Armenian	368832	1.0000	0.0000	1.0000	0.0000
ibo_Latn	Igbo	491594	0.9995	0.0005	0.9995	0.0005
ilo_Latn	Ilocano	976648	0.9990	0.0010	0.9985	0.0015
ind_Latn	Indonesian	1694230	0.9279	0.0435	0.8198	0.2087
isl_Latn	Icelandic	43554	1.0000	0.0000	0.7621	0.3125
ita_Latn	Italian	479663	0.9940	0.0000	0.9721	0.0282
jav_Latn	Javanese	65595	0.9917	0.0079	0.9767	0.0218
jpn_Jpan	Japanese	876783	1.0000	0.0000	0.9808	0.0104
kab_Latn	Kabyle	52634	0.8551	0.1695	0.8579	0.1652
kac_Latn	Jingpho	11365	1.0000	0.0000	1.0000	0.0000
kam_Latn	Kamba	52674	0.9001	0.0005	0.7581	0.0010
kan_Knda	Kannada	357780	1.0000	0.0000	1.0000	0.0000
kas_Arab	Kashmiri	6203	0.9839	0.0000	0.9710	0.0000
kas_Deva	Kashmiri	6694	0.9860	0.0010	0.9840	0.0005
kat_Geor	Georgian	417604	1.0000	0.0000	1.0000	0.0000
kaz_Cyrl	Kazakh	51577	0.9995	0.0000	0.9995	0.0000
kbp_Latn	Kabiye	53275	1.0000	0.0000	1.0000	0.0000
kea_Latn	Kabuverdianu	5665	0.9652	0.0000	0.9610	0.0000
khk_Cyrl	Halh Mongolian	168540	1.0000	0.0000	1.0000	0.0000
khm_Khmr	Khmer	60513	0.9995	0.0000	0.9990	0.0000
kik_Latn	Kikuyu	96402	0.9628	0.0376	0.9636	0.0341
kin_Latn	Kinyarwanda	447057	0.8872	0.0069	0.9788	0.0119
kir_Cyrl	Kyrgyz	372399	1.0000	0.0000	1.0000	0.0000
kmb_Latn	Kimbundu	92635	0.9394	0.0534	0.9361	0.0514
kmr_Latn	Northern Kurdish	15490	0.9985	0.0010	0.9956	0.0045
knc_Arab	Central Kanuri	6196	0.7017	0.0000	0.7026	0.0000
knc_Latn	Central Kanuri	6256	0.9990	0.0005	0.9965	0.0015
kon_Latn	Kikongo	209801	0.9946	0.0045	0.9936	0.0049
kor_Hang	Korean	1772136	1.0000	0.0000	0.9961	0.0040
lao_Laoo	Lao	23529	1.0000	0.0000	0.9995	0.0000
lij_Latn	Ligurian	28641	0.9980	0.0015	0.9774	0.0025
lim_Latn	Limburgish	48151	0.9965	0.0015	0.9870	0.0010
lin_Latn	Lingala	546344	0.9990	0.0010	0.9956	0.0030
lit_Latn	Lithuanian	2663659	0.9985	0.0010	0.9990	0.0010
lmo_Latn	Lombard	35402	0.9975	0.0020	0.9696	0.0109
ltg_Latn	Latgalian	15585	0.9985	0.0000	0.9920	0.0000
ltz_Latn	Luxembourgish	37674	0.9995	0.0000	0.9995	0.0000
lua_Latn	Luba-Kasai	292972	0.9960	0.0005	0.9936	0.0035
lug_Latn	Ganda	251105	0.9941	0.0045	0.9921	0.0069
luo_Latn	Luo	138159	0.9985	0.0015	0.9975	0.0005
lus_Latn	Mizo	195262	0.9985	0.0000	0.9945	0.0005
lvs_Latn	Standard Latvian	2872096	0.9990	0.0005	0.9936	0.0064
mag_Deva	Magahi	6208	0.9620	0.0133	0.9311	0.0213
mai_Deva	Maithili	15385	0.9880	0.0010	0.9871	0.0040
mal_Mlym	Malayalam	379786	1.0000	0.0000	1.0000	0.0000
mar_Deva	Marathi	1017951	0.9990	0.0010	0.9951	0.0049
min_Latn	Minangkabau	31469	0.9931	0.0030	0.5143	0.0010
mkd_Cyrl	Macedonian	561725	0.9995	0.0005	1.0000	0.0000
mlt_Latn	Maltese	2219213	0.9985	0.0015	0.9995	0.0005
mni_Beng	Meitei	47146	0.9941	0.0059	0.9995	0.0000
mos_Latn	Mossi	197187	0.9814	0.0005	0.9684	0.0000
mri_Latn	Maori	48792	0.9995	0.0005	0.9985	0.0005
mya_Mymr	Burmese	452194	1.0000	0.0000	1.0000	0.0000
nld_Latn	Dutch	2929602	0.9970	0.0015	0.9830	0.0173
nno_Latn	Norwegian Nynorsk	101140	0.9828	0.0104	0.9697	0.0208
nob_Latn	Norwegian Bokmal	1783598	0.9719	0.0148	0.9829	0.0139

Table 3: For each language covered by our model, we give the number of lines of deduplicated training data in our dataset, as well as the class F1 score and class false positive rate (FPR) for our model and for the model described in Costa-jussà et al. (2022) (NLLB).

			Our model		NLLB	
Language code	**Language**	**Training data**	**F1 score ↑**	**FPR ↓**	**F1 score ↑**	**FPR ↓**
npi_Deva	Nepali	60345	0.9980	0.0020	0.9980	0.0020
nso_Latn	Northern Sotho	560068	0.9868	0.0119	0.9839	0.0134
nus_Latn	Nuer	6295	0.9995	0.0000	0.9980	0.0015
nya_Latn	Nyanja	789078	0.9966	0.0035	0.9460	0.0163
oci_Latn	Occitan	32683	0.9941	0.0054	0.9835	0.0163
ory_Orya	Odia	92355	1.0000	0.0000	1.0000	0.0000
pag_Latn	Pangasinan	294618	0.9990	0.0005	0.9970	0.0010
pan_Guru	Eastern Panjabi	357487	1.0000	0.0000	1.0000	0.0000
pap_Latn	Papiamento	403991	0.9768	0.0232	0.9839	0.0158
pbt_Arab	Southern Pasto	63256	0.9980	0.0015	0.9970	0.0010
pes_Arab	Western Persian	1758215	0.5570	0.5356	0.6385	0.4381
plt_Latn	Plateau Malgasy	47284	1.0000	0.0000	1.0000	0.0000
pol_Latn	Polish	3403455	0.9956	0.0045	0.9849	0.0153
por_Latn	Portuguese	3800360	0.9941	0.0040	0.9854	0.0143
prs_Arab	Dari	6662	0.5144	0.1122	0.4589	0.0608
quy_Latn	Ayacucho Quechua	154448	1.0000	0.0000	1.0000	0.0000
ron_Latn	Romanian	443200	0.9985	0.0015	0.9985	0.0015
run_Latn	Rundi	459617	0.9044	0.0973	0.9782	0.0104
rus_Cyrl	Russian	7000000	0.9990	0.0005	0.9990	0.0010
sag_Latn	Sango	255491	0.9990	0.0000	0.9970	0.0005
san_Deva	Sanskrit	39988	0.9900	0.0000	0.9885	0.0010
sat_Olck	Santali	8875	1.0000	0.0000	1.0000	0.0000
scn_Latn	Sicilian	40023	0.9956	0.0035	0.9936	0.0054
shn_Mymr	Shan	21051	1.0000	0.0000	0.9985	0.0000
sin_Sinh	Sinhala	361636	1.0000	0.0000	1.0000	0.0000
slk_Latn	Slovak	3153492	0.9970	0.0010	0.9995	0.0005
slv_Latn	Slovenian	3023266	0.9966	0.0030	0.9985	0.0015
smo_Latn	Samoan	367828	0.9985	0.0010	0.9985	0.0010
sna_Latn	Shona	764419	0.9941	0.0059	0.9941	0.0059
snd_Arab	Sindhi	26107	0.9990	0.0000	0.9980	0.0020
som_Latn	Somali	217413	0.9995	0.0005	1.0000	0.0000
sot_Latn	Southern Sotho	2030	0.9567	0.0000	0.7552	0.0000
spa_Latn	Spanish	677548	0.9921	0.0049	0.9922	0.0074
srd_Latn	Sardinian	47480	0.9961	0.0030	0.9773	0.0000
srp_Cyrl	Serbian	310259	0.9995	0.0000	1.0000	0.0000
ssw_Latn	Swati	114900	0.9911	0.0020	0.9916	0.0015
sun_Latn	Sundanese	47458	0.9926	0.0035	0.9599	0.0252
swe_Latn	Swedish	2747052	1.0000	0.0000	0.9990	0.0005
swh_Latn	Swahili	228559	0.9284	0.0771	0.8815	0.1345
szl_Latn	Silesian	34065	0.9960	0.0000	0.9875	0.0015
tam_Taml	Tamil	552180	1.0000	0.0000	1.0000	0.0000
taq_Latn	Tamasheq	10266	0.7907	0.0010	0.7916	0.0000
taq_Tfng	Tamasheq	6203	0.9505	0.0084	0.8513	0.0000
tat_Cyrl	Tatar	257828	1.0000	0.0000	0.9995	0.0000
tel_Telu	Telugu	276504	0.9990	0.0000	1.0000	0.0000
tgk_Cyrl	Tajik	135652	1.0000	0.0000	1.0000	0.0000
tgl_Latn	Tagalog	1189616	1.0000	0.0000	0.9970	0.0025
tha_Thai	Thai	734727	1.0000	0.0000	1.0000	0.0000
tir_Ethi	Tigrinya	333639	0.9995	0.0000	0.9995	0.0000
tpi_Latn	Tok Pisin	471651	1.0000	0.0000	0.9980	0.0000
tsn_Latn	Tswana	784851	0.9693	0.0311	0.8424	0.1859
tso_Latn	Tsonga	756533	0.9961	0.0035	0.9907	0.0089
tuk_Latn	Turkmen	160757	1.0000	0.0000	1.0000	0.0000
tum_Latn	Tumbuka	237138	0.9956	0.0035	0.9816	0.0183
tur_Latn	Turkish	823575	0.9936	0.0064	0.9840	0.0163
twi_Latn	Twi	545217	0.9990	0.0000	0.9420	0.0005
tzm_Tfng	Central Atlas Tamazight	8142	0.9535	0.0395	0.8854	0.1296
uig_Arab	Uyghur	57231	1.0000	0.0000	0.9995	0.0005
ukr_Cyrl	Ukrainian	1140463	0.9995	0.0005	1.0000	0.0000
umb_Latn	Umbundu	220396	0.9776	0.0079	0.9687	0.0208
urd_Arab	Urdu	412736	0.9849	0.0153	0.9735	0.0272
uzn_Latn	Northern Uzbek	1519230	0.9990	0.0010	0.9995	0.0005
vec_Latn	Venetian	43478	0.9961	0.0020	0.9916	0.0035
vie_Latn	Vietnamese	881145	0.9995	0.0005	0.9873	0.0129
war_Latn	Waray	282772	1.0000	0.0000	0.9990	0.0010

Table 3: For each language covered by our model, we give the number of lines of deduplicated training data in our dataset, as well as the class F1 score and class false positive rate (FPR) for our model and for the model described in Costa-jussà et al. (2022) (NLLB).

			Our model		NLLB	
Language code	**Language**	**Training data**	**F1 score ↑**	**FPR ↓**	**F1 score ↑**	**FPR ↓**
wol_Latn	Wolof	28784	0.9970	0.0020	0.9950	0.0010
xho_Latn	Xhosa	921590	0.9858	0.0119	0.9779	0.0148
ydd_Hebr	Eastern Yiddish	911	0.9990	0.0000	1.0000	0.0000
yor_Latn	Yoruba	531904	0.9990	0.0010	0.9956	0.0030
yue_Hant	Yue Chinese	63254	0.0059	0.0025	0.4877	0.3229
zho_Hans	Chinese (Simplified)	1046823	0.9891	0.0054	0.8559	0.0277
zho_Hant	Chinese (Traditional)	2018541	0.6605	0.5020	0.4651	0.2176
zsm_Latn	Standard Malay	404380	0.9495	0.0346	0.9351	0.0307
zul_Latn	Zulu	951688	0.9828	0.0104	0.9696	0.0267

Table 3: For each language covered by our model, we give the number of lines of deduplicated training data in our dataset, as well as the class F1 score and class false positive rate (FPR) for our model and for the model described in Costa-jussà et al. (2022) (NLLB).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
in separate limitations section at end

☑ A2. Did you discuss any potential risks of your work?
in separate ethics statement at end

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract is where you would expect; main claims are in bullets in introduction

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

B ☑ Did you use or create scientific artifacts?

section 3 describes dataset creation; section 4 describes model selection

☑ B1. Did you cite the creators of artifacts you used?
Appendix A

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
section 3.1 explains how to find full list of licenses (in repo as it is very long and subject to change)

☑ 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)?
Section 3.1 explains how all datasets are open for academic use and explains how to find the full terms on the github repo

☒ 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?
data is all in the public domain (section 3.1 explains that sources are mainly news sites and Wikipedia)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.1 gives overview of dataset domain; full information is in the repo because of length

☑ 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.
Summary statistics for training data are in section 3.4; full breakdown by class is in appendix B due to length. Description of train and dev splits is in section 5.1

C ☑ Did you run computational experiments?

section 4 describes the model, section 5 describes evaluation and results

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 4

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?
we used the same hyperparameter values as the model in No Language Left Behind as we are comparing datasets rather than models. Hyperparameters are in appendix B

☒ 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?
We do give the mean across classes but we didn't run multiple experiments because we are presenting a dataset rather than a modelling paper.

☑ 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.)?
section 3.3

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
section 3.2

☐ 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. annotation was done by the authors

☑ 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)?
section 3.2 (annotation done by the authors)

☐ 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?
section 3.2