Building Resources for Emakhuwa: Machine Translation and News Classification Benchmarks

Felermino D. M. A. Ali[1,2,3,5], Henrique Lopes Cardoso[1,2], Rui Sousa-Silva[3,4]
[1]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC / LASI)
[2]Faculdade de Engenharia da Universidade do Porto,
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[3]Centro de Linguística da Universidade do Porto (CLUP)
[4]Faculdade de Letras da Universidade do Porto,Via Panorâmica,4150-564 Porto,Portugal
[5]Faculdade de Engenharia da Universidade Lúrio, Pemba 3203, Mozambique
{up202100778, hlc}@fe.up.pt, rssilva@letras.up.pt

Abstract

This paper introduces a comprehensive collection of NLP resources for Emakhuwa, Mozambique's most widely spoken language. The resources include the first manually translated news bitext corpus between Portuguese and Emakhuwa, news topic classification datasets, and monolingual data. We detail the process and challenges of acquiring this data and present benchmark results for machine translation and news topic classification tasks. Our evaluation examines the impact of different data types—originally clean text, post-corrected OCR, and back-translated data—and the effects of fine-tuning from pre-trained models, including those focused on African languages. Our benchmarks demonstrate good performance in news topic classification and promising results in machine translation. We fine-tuned multilingual encoder-decoder models using real and synthetic data and evaluated them on our test set and the FLORES evaluation sets. The results highlight the importance of incorporating more data and potential for future improvements. All models, code, and datasets are available in the https://huggingface.co/LIACC repository under the CC BY 4.0 license.

1 Introduction

Natural Language Processing (NLP) has witnessed significant advances in recent years. However, addressing low-resource languages, including those spoken in Africa, is a persistent challenge. These languages often lack the extensive datasets and evaluation sets readily available for high-resource languages. This scarcity significantly hinders NLP progress in these languages.

Emakhuwa, spoken in northern and central Mozambique, is no exception. While it is the most spoken language in the country, it has received limited attention in NLP research. This lack of resources hinders the development of NLP applications that could benefit Emakhuwa speakers.

We present a valuable collection of Emakhuwa resources for the NLP community to address this gap. Our contributions include:

- Training and Evaluation sets: (1) the first parallel news corpora between Portuguese and Emakhuwa for Machine Translation and labeled data for News Topic Classification. (2) Optical Character Recognition post-correction dataset in Emakhuwa and Portuguese. (3) Monolingual data in Emakhuwa.

- Benchmarks: We provide benchmarks for machine translation and news topic classification tasks. Additionally, we explore how different data configurations—such as originally clean text, post-corrected OCR, and back-translated data—affect the performance of fine-tuned pre-trained models for machine translation tasks.

2 Emakhuwa

Emakhuwa (also known as Makua, Macua, or Makhuwa) is a Bantu language spoken in northern and central Mozambique, i.e., Niassa, Cabo Delgado, and Nampula, as well as in some parts of the Zambezia province. It is estimated that approximately 25% of the country's population of 30 million people make use of the language daily as an alternative to Portuguese (Ronaldo Rodrigues de Paula, 2016). The language is spoken in neighbouring countries such as Tanzania and Malawi, with relatively few speakers.

There are eight variants of Emakhuwa, with Emakhuwa-Central (ISO-639 code *vmw*) being the standard variety (Ngunga and Faquir, 2014). Emakhuwa is an SVO (Subject-Verb-Object) language written in Latin script. Unlike many other languages, it has no grammatical gender. Like other Bantu languages, it is complex and rich, featuring agglutinative morphology and tonal attributes.

3 Related Works

Evaluation datasets and benchmarks are crucial for developing NLP models, especially for low-resource languages such as those spoken in Africa. The scarcity of evaluation datasets and benchmarks for low-resource African languages has significantly hindered the progress of NLP in this region. Without these essential tools, it is difficult to assess the performance of NLP models for these languages, making it harder to identify areas for improvement and track progress. Despite these challenges, a growing movement within the research community aims to address this gap. Researchers and organizations are actively working to create and disseminate evaluation datasets and benchmarks for low-resource African languages. One prominent group in this effort is the Masakhane Research community. This community has established a robust network through which many resources have been developed and made available to the research community. Some key resources include MasakhaNEWS (Adelani et al., 2023), MAFAND-MT (Adelani et al., 2022a), MasakhaNER (Adelani et al., 2021, 2022b), or AfriSent (Muhammad et al., 2023), among others. The impact of these efforts is becoming increasingly evident. For example, in a study by Adebara et al. (2023), researchers demonstrated that their pre-trained models surpassed the benchmarks for news topic classification established by Adelani et al. (2023). Similarly, the study by Adebara et al. (2024) showed that their language models improved machine translation across multiple language pairs, thus outperforming the benchmarks set by Adelani et al. (2022a). Nevertheless, the existing evaluation datasets and benchmarks for African languages still cover a limited number of languages. For instance, although Emakhuwa is an African language that has seen some development in NLP, the data and evaluation tools are not widely available in the public domain, highlighting the need for further resource development and accessibility. For example, Ali et al. (2021) compiled what appears to be the first parallel dataset for machine translation between Emakhuwa and Portuguese. Despite their efforts in dataset construction, the reproducibility of their results is compromised due to the non-reproducible nature of their evaluation sets.

4 Data Collection

Our data collection includes (1) the first manually translated parallel news data, (2) data extracted from printed books, and (3) monolingual data. A notable aspect of our collected data is its diverse range of sources, which introduces significant variations in writing styles. This diversity includes different genres, contexts, and periods, providing a multifaceted linguistic resource. For example, the manually translated parallel data reflects the contemporary language, while the data extracted from printed books captures more of traditional narratives, dialogues, and idiomatic expressions. Monolingual data, on the other hand, spans content from modern publications and digital platforms and adds further variety. This heterogeneity is crucial for creating NLP models that can handle the complexity and richness of Emakhuwa in real-world applications. Below, we describe in detail the process of creating each of the types of data collected.

4.1 The first News *bitext* data for Emakhuwa

We created the first news translation corpus between Portuguese and Emakhuwa. To achieve this, we hired professional translators to translate a subset of news articles from the MOT (Palen-Michel et al., 2022) dataset. The articles were sourced from the VOA (Voice of America) platform and published between 2001 and 2021.

Data preparation: We selected Portuguese news articles across seven categories: politics, economy, culture, sports, health, society, and world news. We chose articles that encouraged lexical diversity within each category and maintained contextual relevance for Mozambique. Additionally, we created a glossary and guidelines to ensure consistency. The glossary was constructed by digitizing several existing bilingual dictionaries of Portuguese-Emakhuwa. Furthermore, we added News domain terminologies from the Glossaries of Political, Sports, and Social Concepts from Radio of Mozambique (Moçambique E.P., 2016).

The guidelines, however, emphasized that the translated text should convey the same meaning as the source and adhere to the latest orthography standards. Personal names were not to be translated. Additionally, loanwords adapted into Emakhuwa were to be annotated in the comments section of each segment. This was particularly useful for identifying discrepancies in loanword adaptation,

which we addressed through discussions in dedicated workshops.

Translators recruitment: The translators were selected based on their proficiency in both Portuguese and the Central variant (*vmw*) of Emakhuwa and their proven experience in Emakhuwa translation. In total, we worked with 10 translators: two are final-year bachelor's students majoring in Bantu Linguistics at the Faculty of Arts and Social Sciences (*Faculdade de Letras e Ciências Sociais da Universidade Mondlane* - FLCS), seven hold degrees in Bantu Linguistics and work as professional translators, and one is a Radio Mozambique host with vast experience in translating and broadcasting contents in Emakhuwa. The translation/revision work was paid at competitive rates and budgeted based on the word count of the source text.

Translation Workflow: We managed the translation workflow using MateCat[1]. The process involved iterative cycles with the following phases:

- **Translation**: Translators were assigned text segments to translate into Emakhuwa.
- **Revision**: Translated documents were automatically checked using our custom-developed spelling tool[2]. The revision reports were then sent back to the translators so they could revise their work and resubmit after completing the revisions.

Workshops: We conducted workshops that all translators should attend. During these workshops, we discussed potential new glossary entries and collaboratively resolved some major issues that led to translation disagreements.

As a result of the translation process, we successfully translated 1,897 news articles and segmented them into 18,540 parallel sentences spanning various topics, as detailed in Table 1.

Topic	Docs	Sent.
-politics	288	2,084
-economy	287	2,274
-culture	329	3,822
-sports	345	3,148
-health	204	2,423
-society	198	2,099
-world news	246	2,681
Total	1,897	18,540

Table 1: Summary of parallel sentences collected during translation process

4.2 OCRed Data

A significant quantity of text in Emakhuwa exists solely in printed book format, making it hardly accessible for machine processing and NLP applications. Similar to Ali et al. (2021), we have collected this data type by digitizing printed documents using Google Vision Optical Character Recognition (OCR) software (Google, 2024). This includes documents with bilingual content in Emakhuwa and Portuguese. To collect the data, we follow a similar methodology as described in (Rijhwani et al., 2020): we collected bitext data extracted from the *Método Macua* book (Centis, 2000), which is rich in cultural narratives, with tales intrinsic to Emakhuwa culture. We extracted parallel sentences from 367 pages by following the process described below:

- **Data Extraction**: First, we scanned the pages as images. Given that the Emakhuwa text and its corresponding Portuguese translation were often presented in a two-column layout, we developed an algorithm to automatically split the images into two mirror halves (i.e. crops)—one containing the Emakhuwa content and the other containing the corresponding Portuguese translation.
- **Post-Correction**: After OCR, we manually corrected any errors with the help of two volunteers, who reviewed the text in the images and corrected potential OCR errors. Post-correction was facilitated using the free tier of Label Studio[3].

In total, from this process, we have collected 1,911 parallel sentences from 369 cropped images (Table 2).

Topic	Crops	Sent.
-tales	369	1,911

Table 2: Summary of parallel sentences extracted from printed book

[1] https://www.matecat.com
[2] anonymous
[3] https://labelstud.io/

4.3 Monolingual data

A considerable amount of monolingual data is also available exclusively in Emakhuwa. Therefore, we also collected monolingual data as it holds practical utility for various NLP tasks. Below, we describe two sources from which we have collected the monolingual data:

- ***Wòniherya* News**: An online news magazine *Wòniherya* News magazine[4] with monthly publications. To date, they have published 32 editions of the magazine, all exclusively written in Emakhuwa.

- **Wikimedia data**: The Wikimedia Incubator[5], currently with a total of 1005 articles written in Emakhuwa.

In both sources, we observed that some of the text contained mixed words from other languages. Therefore, we performed a preprocessing to clean the data, we followed these steps:

1. **Segmenting into smaller sentences**: To achieve this, we used the NLTK[6] sentence tokenizer. The Portuguese NLTK sentence tokenizer was chosen due to the structural similarities between Portuguese and Emakhuwa at the sentence segmentation level.

2. **Filtering Emakhuwa sentences**: To ensure that only Emakhuwa sentences were included in the final dataset, we applied GlotLID, a language identification system (Kargaran et al., 2023), to each sentence. Sentences with fewer than three tokens or identified as languages other than Emakhuwa were excluded.

As a result, we compiled 5,021 sentences from the Wòniherya News and 9,442 from Wikimedia as summarised in Table 3.

Topic	Source	Docs	Sent.
news and miscellaneous	Wòniherya	32	5,021
miscellaneous	Wikimedia	1,005	9,442
Total		1,037	14,463

Table 3: Summary of monolingual data

[4]https://emakhuwa.org.mz/
[5]https://incubator.wikimedia.org/wiki/Category:Wp/vmw
[6]https://www.nltk.org/

5 Datasets

We gathered approximately 18,540 aligned parallel sentences, 1,897 labeled news articles, 1,911 post-corrected OCR bitext sentences, and 14,463 monolingual sentences. This data was meticulously cleaned and prepared for Machine Translation and News Topic Classification. While the dataset holds potential for other NLP tasks, such as OCR post-correction, loanword detection, and language modeling, these applications fall outside the scope of the benchmarks presented in this study.

Once cleaned, the data was divided into training (TRAIN), validation (DEV), and test (TEST) sets. For the Machine Translation task, we allocated 17k sentences for training. The dev and test sets were meticulously constructed by extracting sentences from the compiled parallel corpus. To ensure high quality and wide domain coverage in the dev and test set, we also included a small set of sentence pairs from the Ali et al. (2021) test set. The following steps were then applied to select sentences for the test and dev sets: (1) Source sentences had to contain between 5 and 150 tokens; (2) Each sentence needed to start with a capital letter and end with appropriate punctuation; (3) Sentence pairs were chosen if the length ratio between the source and target language sentences fell between 0.66 and 1.5, as recommended by (Kudugunta et al., 2023). This helps reduce potential issues caused by significant discrepancies in sentence lengths between the two languages. This ultimately resulted in 964 sentences for validation and 993 for testing.

For the News Classification dataset, the split included 1,337 articles for training, 185 for validation, and 375 for testing corresponding to a 70%, 10%, and 20% split ratio respectively.

Table 4 provides a summary of both datasets, detailing the distribution of articles across these splits and the coverage of various topics.

6 Benchmarks and Experiments

To promote reproducibility, we provide detailed descriptions of the experiments and benchmarks conducted, particularly focusing on Machine Translation and News Classification tasks.

6.1 Machine Translation

Table 5 provides a summary of the training data used to build the machine translation models in Section. In addition to the training data created during this study (shown in Table 4), we also gathered

	News Classification			Machine Translation		
	# docs			# Sents.		
Topics	TRAIN	DEV	TEST	TRAIN	DEV	TEST
#News politics	203	28	57	1,950	67	67
#News economy	202	28	57	2,119	65	90
#News culture	232	32	65	3,611	108	103
#News sports	243	34	68	3,030	63	55
#News health	144	20	40	2,256	83	84
#New society	140	19	39	1,922	90	87
#News world	173	24	49	2,515	89	77
#Religion	-	-	-	-	290	307
#Tales	-	-	-	-	20	36
#Wikipedia	-	-	-	-	28	22
#Legal	-	-	-	-	24	16
#Miscellaneous	-	-	-	-	37	49
Total	1337	185	375	17,403	964	993

Table 4: Datasets splits and distribution of examples across different topics

data from Ali et al. (2021) and from synthetically generated data. The synthetic data was produced by back-translating monolingual data in Section 4.3. The model used for back-translation (referred to as System 2 in Section 7) was trained using a combination of the real data.

	Ali-2021	News	OCR-ed	Synthetic
Topic	# Sent.	# Sent.	# Sent.	# Sent.
religious	45,983	-	-	-
news	-	-	-	-
-politics	-	2,084	-	-
-economy	-	2,274	-	-
-culture	-	3,822	-	-
-sports	-	3,148	-	-
-health	-	2,423	-	-
-society	-	2,099	-	-
-world news	-	2,681	-	-
tales	204	-	1,911	-
legal	917	-	-	-
history	93	-	-	-
miscellaneous	80	-	-	14,463
Sentences	47,273	17,403	1,911	14,463
Tokens				
vmw	951,520	541,598	12,283	27,2559
pt	1,104,279	596,066	15,887	30,8389
Vocabs				
vmw	70,825	82,196	3,447	54,836
pt	37,726	34,931	2,342	21,239

Table 5: Machine Translation training set, Ali-2021 corresponds to Ali et al. (2021) training subset. News and OCR-ed, refers to sentence pairs obtained through manual translation and OCR-extracted sentences (see Table 4 and Table 2); and Synthetic, consists of back-translated monolingual data.

6.1.1 Evaluation

To evaluate the performance of our Machine Translation systems, we used two metrics: the SacreBLEU toolkit to compute BLEU (Papineni et al., 2002) and ChrF scores (Popović, 2015). In addition to our own test set, we also evaluated the systems using the FLORES benchmark proposed by Ali et al. (2024), which consists of manually translated multi-way data from Wikipedia. The FLORES evaluation includes two splits: dev with 997 sentences and devtest with 1,012 sentences.

6.1.2 Setup

We trained bilingual models to translate in both directions, i.e., Portuguese-Emakhuwa and vice-versa, and we considered the following setups and models.

Transformer baseline We adopt a transformer architecture (Vaswani et al., 2017), implemented through the OpenNMT toolkit (Klein et al., 2017). The model architecture features an encoder and decoder, each with 6 layers, 8 attention heads, and 512 hidden units in the feed-forward network. Both source and target word embeddings are represented in 512 dimensions, and training was performed using a batch size of 32. The sentence length is set to 150 tokens, incorporating 0.1 label smoothing. We applied layer normalization and added dropout with a 0.1 probability to the embedding and transformer layers. Additionally, the Adam optimizer (Kingma and Ba, 2014) was used, and a learning rate of 0.0002. The checkpoints were saved every 1000 updates. We preprocess the input, applying the Byte Pair Encoding subword segmentation.

Multilingual For our experiments, we fine-tuned the following language models: mT5 (Xue et al., 2021), byT5 (Xue et al., 2022), and the multilin-

gual translation models M2M-100 (Fan et al., 2021) and NLLB (NLLBTeam et al., 2024). Specifically, we use mT5-base (580M parameters), byT5-base (580M parameters), M2M-100 (418M parameters), and NLLB-200's distilled variant (600M parameters). Additionally, we included African-centric language models such as AfribyT5 and AfrimT5 by Adelani et al., 2022a.

6.2 News Topic Classification

We trained two sets of models, classical Machine Learning (ML) models and multilingual text encoders based on BERT/RoBERTa (Devlin et al., 2019; Zhuang et al., 2021). For training we used the label data in Table 4. Following (Adelani et al., 2023), for classical ML models, we considered the Naive Bayes, Multi-Layer Perceptron (MLP), and XGBoost models. Conversely, for multilingual text encoders, we fine-tuned the following pre-trained models: XLM-R-large (Conneau et al., 2020), AfriBERTa-large (Ogueji et al., 2021), RemBERT (Chung et al., 2021), AfroXLMR-large (Alabi et al., 2022), AfroLM (Dossou et al., 2022), mDeBERTaV3 (He et al., 2021), SERENGETI (Adebara et al., 2023). The models were fine-tuned using an Nvidia A10 GPU over 20 epochs, with a batch size 16, a learning rate of 1×10^{-5}, and a maximum sequence length of 256 tokens.

6.2.1 Evaluation

We evaluate the models using the average Precision (P), Recall (R), and F1-score (F1), where each model's performance was assessed based on the average results from five runs on our test set.

7 Results

In this section, we discuss the benchmark results of both Machine Translation and News topic classification.

7.1 Machine Translation

Table 6 presents BLEU and ChrF scores for various machine translation models trained and fine-tuned using the [*Ali-2021 + News (Table 4)*] and evaluated on our own and the FLORES test sets. The results highlight the advantages of fine-tuning multilingual language models in low-resource settings.

Among the evaluated models presented in Table 6, the byT5-based, M2M-100, and NLLB models emerged as the top performers, achieving the highest scores in both translation directions. These results highlight the significance of fine-tuning models with extensive language coverage. Notably, the byT5-based model outperformed other text-to-text models, such as mT5-based and mT0. We attribute this advantage to the tokenization-free approach of the byT5 model, which is better suited for handling the morphological richness and orthographic variations characteristic of Emakhuwa.

Nevertheless, all models in Table 6 struggle to achieve a higher BLEU score in the *pt→vmw* direction. This outcome was anticipated, as BLEU relies on exact word n-gram matches. Emakhuwa is an agglutinative language with non-standard spelling, which poses a significant challenge to BLEU's evaluation method. Agglutination leads to a high variability in word forms. This variability, combined with non-standardized spelling, results in fewer exact n-gram matches, which BLEU heavily depends on. Further analysis of the ChrF results provides additional insights. Unlike BLEU, ChrF evaluates based on character *n*-grams, making it more resilient to agglutination and spelling inconsistencies.

How does adding more data impact performance? Here, we examine the impact of incorporating additional data on the performance of MT models. Specifically, we evaluate models fine-tuned with the additional manually curated OCR bitext and synthetic data. For this analysis, we focus on fine-tuning the NLLB-200 model in Table 6, demonstrating the best performance among the models on bootstrap significance tests on our test set (refer to Appendix A).

As shown in Table 7, we created distinct training datasets by aggregating different sources of data:

To assess the impact of different data sources on fine-tuning performance, we compare the following training data configurations:

1. **System 1** - [*Ali-2021*]: Training exclusively on Ali et al. (2021) training data.

2. **System 2** -[*Ali-2021+News*]: increase the training set by adding the training set in Table 4.

3. **System 3** -[*Ali-2021+News+OCRed*]: Building on the previous setup, this configuration adds bilingual text data extracted via OCR processing.

4. **System 4** -[*Ali-2021+News+Synthetic*]: This setup extends further the previous System 2 setup with the addition of back-translated data derived from monolingual texts.

	Our test set		FLORES dev set		FLORES devtest set	
Model	*pt→vmw*	*vmw→pt*	*pt→vmw*	*vmw→pt*	*pt→vmw*	*vmw→pt*
baseline	5.94 / 32.20	10.03 / 34.14	3.7 / 30.67	4.36 / 25.48	3.27 / 29.23	2.93 / 23.96
byT5	11.32 / 43.49	20.04 / 43.31	10.66 / 42.37	22.24 /47.01	7.49 / 36.33	14.1 / 37.75
afri-byT5	10.13 / 41.13	19.86 / 43.01	10.32 / 41.88	22.45 / 47.31	7.03 / 35.87	13.47 / 37.78
mT5	4.47 / 32.48	9.83 / 40.07	6.76 / 34.09	15.42 / 37.58	5.67 / 31.67	9.65 / 32.22
mT0	5.99 / 33.61	14.61 / 36.50	5.52 / 30.33	17.46 / 38.92	4.69 / 27.89	10.63 / 32.69
afri-mT5	6.23 / 36.83	14.92 / 39.07	5.66 / 35.37	12.12 / 38.18	4.7 / 32.7	7.39 / 32.92
M2M-100	10.92 / 44.23	20.62 / 44.11	8.25 / 39.22	21.08 / 45.31	6.92 / 36.33	13.67 / 37.46
NLLB	11.58 / 45.62	22.90 / 46.65	8.19 / 41.44	17.41 / 42.88	5.88 / 36.13	10.35 / 35.05

Table 6: Evaluation scores on the test set (shown as <BLEU> / <ChrF>) for the MT models

	System 1	**System 2**	**System 3**	**System 4**	**System 5**
# Training Sents.	~47k	~63k	~65k	~78k	~80k
Our test set					
pt→vmw	8.14/39.36	11.58 / 45.62	29.65 / 66.05	40.90 / 74.04	31.97 / 68.35
vmw→pt	14.20 / 37.33	22.90 / 46.65	23.15 / 46.47	22.42 / 45.92	22.22 / 46.25
FLORES dev set					
pt→vmw	4.67 / 33.94	8.19 / 41.44	8.83 / 40.43	9.49 / 41.89	9.23 / 41.94
vmw→pt	12.38 / 36.21	17.41 / 42.88	23.19 / 47.62	23.03 / 47.41	23.00 / 47.46
FLORES devtest set					
pt→vmw	4.17 / 32.70	5.88 / 36.13	6.93 / 37.02	7.77 / 38.42	7.62 / 38.54
vmw→pt	9.05 / 32.99	10.35 / 35.05	15.00 / 39.67	14.65 / 39.01	14.33 / 39.27

Table 7: Performance of the NLLB model across different training data configurations, evaluated on our test set as well as the FLORES development and test sets (reported as <BLEU> / <ChrF>). System 1 was fine-tuned exclusively on the dataset from Ali et al. (2021). System 2, on the other hand, was fine-tuned using a combination of Ali et al. (2021) dataset and parallel news data created in this study. System 3 further expanded the System 2 training data by adding OCR-extracted sentence pairs, while System 4 expanded System 2 training data with synthetically generated data. Finally, System 5 combined all data sources into a single training set.

5. **System 5** -All: Combine all real and synthetic training data above.

Notably, adding more data resulted in performance gain for both BLEU and ChrF in our test set, particularly in the *pt→vmw* direction. Interestingly, even with a relatively small amount of data, from the OCR-ed data (approximately 2k), System 3 depicts a remarkable performance gain in BLEU (+18.07) and ChrF (+20.40) in *pt→vmw*.

Certainly! Here is the revised text:

However, surprisingly, when we combined the OCR-ed data with other data, we observed a decline in performance. We believe that this decline may be due to the diversity present in the OCR-ed data. This data contains tales, narratives, dialogues, and idiomatic expressions, which differ from the news and religious texts found in the rest of the dataset. A future study of interest would involve conducting a deeper analysis of domain effects and measuring the impact of domain shifts on translation performance.

Based on the results in Table 7, among the training data configurations discussed above, System 4, which was fine-tuned by combining [*Ali-2021+News+Synthetic*] proved to be the most promising configuration.

Out-of-distribution test set As shown in Table 7, the improvements reported above were also reflected in the FLORES dev and test sets, demonstrating the NLLB model' generalizability capabilities.

7.2 News Topic Classification

The results of our news topic classification evaluation, summarized in Table 8, show the performance of various models when trained on the TRAIN split and evaluated on the TEST split, both for the task of classifying based on headlines ("Headline only") as well as using both headline and text ("Headline + Text") inputs. Classical machine learning and fine-

tuned multilingual language models (LMs) exhibit distinct performance patterns.

Headline only Classical ML models (NaiveBayes and MLP) outperformed several fine-tuned LMs. NaiveBayes achieved the highest F1 score (54.12), followed closely by MLP (53.52). These models excelled AfroLM, AfroXLMR-large, and SERENGETI, which likely suffered due to limited exposure to Emakhuwa during their pre-training.

Headline + Text incorporating the full text significantly improved performance for both classical ML models and fine-tuned LMs, emphasizing the importance of contextual information. NaiveBayes maintained its lead with the highest precision, recall, and F1-score (75.78, 72.43, and 72.83, respectively). Among the LMs, AfriBERTa demonstrated the best performance, with an F1-score of 70.33.

These findings underscore that existing African LMs still lack comprehensive language coverage in their pre-training data. Classical ML models like NaiveBayes thus remain competitive. However, the potential of fine-tuned African LMs is evident when provided with sufficient context, suggesting further investment in their development and pre-training on diverse African languages.

Further analysis of the confusion matrix in Figure 1 reveals several key insights into the performance of the NaiveBayes model. The model demonstrates high accuracy in the "sports" and "economy" categories, with most instances correctly classified. This indicates that these categories have distinctive features that the model effectively captures. However, there are notable misclassifications across the "politics" and "economy" categories. This suggests an overlap in topics, likely due to the geopolitical nature of certain news articles that blur the lines between political and economic content. The "society" and "world news" categories also exhibit many misclassifications. This could be attributed to these categories' broad and diverse nature, encompassing a wide range of topics. As a result, the model struggles to discriminate between them, leading to misclassifications.

8 Conclusion

This paper presents a valuable collection of NLP resources for the Emakhuwa language, including parallel text corpora, news topic classification datasets, and monolingual data. We also establish benchmark results for machine translation and news topic classification tasks, demonstrating the potential of these resources to advance NLP research in Emakhuwa.

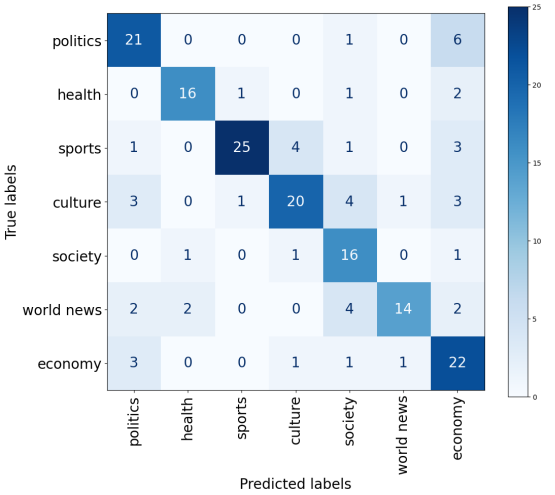

Figure 1: Confusion Matrix for the NaiveBayes model on Headline + Text

Our benchmarks for the respective tasks are quite challenging. For news topic classification, classical machine learning approaches yield better results, achieving an F1 score of 72.83. We observe relatively strong performance in machine translation, particularly when translating from Portuguese to Emakhuwa. Our model, fine-tuned from NLLB using a combination of real and synthetic data generated through back-translation, achieves a BLEU score of 40.90 and a ChrF score of 74.04.

By publicly making these resources available, we aim to foster further research and development in NLP for African languages, ultimately benefiting the Emakhuwa-speaking community.

9 Limitations

During data collection, we encountered challenges that underscored existing structural problems in Emakhuwa and Mozambican languages in general. One of the main issues we faced was the high disagreement regarding spelling forms. Despite our working with trained professionals, there were frequent discrepancies in how words were spelled. We believe that these inconsistencies are influenced by multiple factors

- Translator's native language variants: Although we previously agreed to use the standard variant, the translators came from different regions and backgrounds, and they typi-

		Headline only			Headline + Text		
	Model	P	R	F1	P	R	F1
classical ML	MLP	54.95	52.97	53.52	70.31	70.27	69.06
	NaiveBayes	**58.49**	**54.59**	**54.12**	**75.78**	**72.43**	**72.83**
	XGBoost	51.76	50.27	49.65	73.48	72.43	72.48
LMs	XLM-R-large	45.86	45.60	44.91	69.43	67.33	67.90
	AfroXLMR-large	48.17	46.93	44.64	68.90	68.53	68.25
	AfroLM	41.54	42.93	41.26	69.23	68.53	68.12
	mDeBERTa	**54.63**	**51.20**	**52.11**	69.87	69.60	69.69
	AfriBERTa	51.87	50.13	50.46	70.34	**70.40**	**70.33**
	RemBERT	46.51	38.13	39.99	**70.56**	69.33	69.59
	SERENGETI	35.77	36.53	31.65	70.29	70.13	69.76

Table 8: Performance of various Classical ML and LM models on news topic classification, on P-Precision, R-Recall, and F1 mean.

cally had disagreements during workshop discussions.

- Underdeveloped Spelling Standards: The Emakhuwa spelling system is still in its infancy, with a long journey ahead to achieve largely accepted standardized orthography. The lack of established spelling conventions for Emakhuwa leads to significant inconsistencies in written communication.

- Tonal marking: There are no clear guidelines on how to mark tones in Emakhuwa, which affects the uniformity of written text. For example, the word "Mozambique" (*Moçambique* in Portuguese) was translated using several different spellings, "*mosampikhi*", "*mosampiiki*", "*mocampiiki*", "*mosapiiki*".

- Loanword adaptation: The integration of loanwords into Emakhuwa lacks standardized adaptation rules, which causes further discrepancies. For example, the word "government" (*governo* in Portuguese) was subject to several different spellings, "*kuvernuru*", "*kuveeerunu*", "*kveerunu*", "*kuveeruni*", "*kuveernu*", "*kuveeruunu*", "*kuveru*". We also observed "lazy" translation, where translators would adapt words into Emakhuwa instead of finding the actual Emakhuwa equivalents to complete their tasks more quickly. To discourage this practice, we instructed translators to annotate all loanword adaptations during translation and revision, which added significant extra work. While this approach effectively promoted better translations, our findings still indicated a high prevalence of loanwords in contemporary Emakhuwa, with 3 out of 15 words in our data being loanwords.

Another limitation of our work is that we collected data exclusively for the Emakhuwa-central variant. Consequently, the benchmark is only applicable to this specific variant.

Acknowledgements

This work was financially supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC) as well as supported by the Base (UIDB/00022/2020) and Programmatic (UIDP/00022/2020) projects of the Centre for Linguistics of the University of Porto. Felermino Ali is supported by a PhD studentship (with reference SFRH/BD/151435/2021), funded by Fundação para a Ciência e a Tecnologia (FCT).

We extend our deepest gratitude to the Lacuna Fund and the German Research Center for Artificial Intelligence for their generous sponsorship, which enabled the creation of this dataset, and for supplying the necessary computing resources to carry out some experiments. We also wish to give special thanks to the Translation Team for their unwavering dedication, as well as to the many translators and linguists whose invaluable contributions made this project a reality.

References

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language

generation for 517 african languages. *Preprint*, arXiv:2401.01053.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022b. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of*

the 29th International Conference on Computational Linguistics, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. 2021. Towards a parallel corpus of Portuguese and the Bantu language Emakhuwa of Mozambique. *arXiv preprint*.

Felermino D. M. Antonio Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Expanding flores+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation. *Preprint*, arXiv:2408.11457.

Gino Centis. 2000. *Método Macua*, 5ª edição edition. Missionários Combonianos - Roma, Centro Catequético Paulo VI, Anchilo - Nampula, Moçambique.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

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.

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

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Google. 2024. Vision ai: Image & visual ai tools. Accessed: 2024-09-28.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

R. de Moçambique E.P. 2016. Glossários de conceitos políticos, desportivos e sociais (português-línguas moçambicanas). Retrieved from `http://197.249.65.29/moodle/file.php/1/Glosario_RMe.pdf`.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. SemEval-2023 task 12: Sentiment analysis for African languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2319–2337, Toronto, Canada. Association for Computational Linguistics.

Armindo Ngunga and Osvaldo Faquir. 2014. *Padronização da Ortografia de Línguas Moçambicanas: Relatório do VI Seminário*. Centro de Estudos das Línguas Moçambicanas.

NLLBTeam, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers,

Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chester Palen-Michel, June Kim, and Constantine Lignos. 2022. Multilingual open text release 1: Public domain news in 44 languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2080–2089, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Fábio Bonfim Duarte Ronaldo Rodrigues de Paula. 2016. Diversidade linguística em moçambique.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

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.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Bootstrap significance tests

To futher our analysis, we conducted bootstrap significance tests on the test set using the models presented in Table 6. We set the number of bootstrap samples to 1,000 and evaluated the models using the BLEU score metric.

The heatmaps in Figure 2 illustrate the pairwise significance results between the models. Each cell represents the p-value of the test, where values close to zero indicate a statistically significant difference between the corresponding models. NLLB-200 models show statistically significant superiority. Furthermore, other models show a range of performance comparisons, but none consistently outperforms the NLLB-200 models across all translation directions.

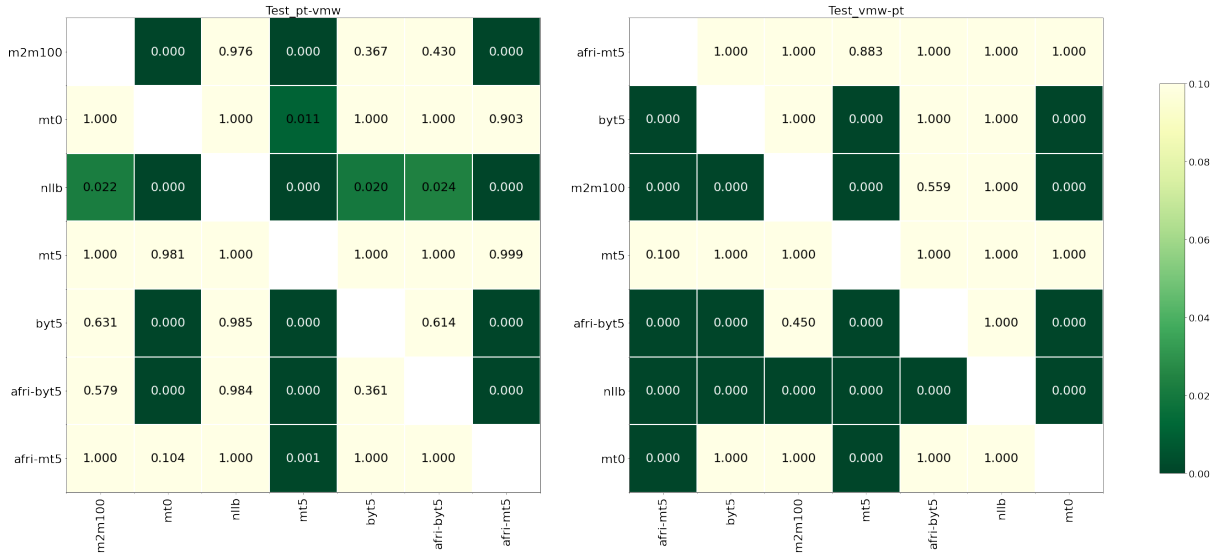

Figure 2: The Automatic metric pairwise randomized significance test results for Portuguese-to-Emakhuwa and Emakhuwa-to-Portuguese systems. The heat map cells highlight where the performance of the system in the row is significantly greater than that of the system in the column ($p < 0.05$) . The value inside each sell corresponds to the p-value. This means that, for each pair of systems, statistical significance was established if the score for the *row system* was meaningfully higher than the score for the *column system*, based on our test set.

vmw	**Atthu ootheene opooma wo Vatikaanu anatiini a ekirixitawu ya katolika.**.
en	*All citizens of Vatican City are Roman Catholics.*
pt	Todos os cidadãos da cidade do Vaticano são católicos romanos.
baseline	Todos na cidade do Vaticano apela a terra de <unk>.
afri-byT5	Toda a população na cidade do Vaticano realiza a religião católica.
afri-mT5	Todos os cidadãos em Vaticano são religiosos da igreja católica.
byT5	Toda a população na cidade do Vaticano é religiosa da cristã católica.
mT0	Todos os cidadãos da cidade de Vaticane são cristãos da igreja católica.
mT5	Todos os cidadãos na cidade de Vaticano são cristãos católicos.
M2M100	Todos os cidadãos do Vaticano são cristãos católicos.
NLLB	Todos na cidade do Vaticano são religiosos católicos.

Table 9: Example of Emakhuwa to Portuguese translations

pt	**A camada é mais fina debaixo dos mares e mais espessa abaixo das montanhas**.
en	*It is thinner under the maria and thicker under the highlands.*
vmw	Mpattapatthaaya tiwoyeva vathi wa mphareya ni yowoneya vathi wa miyaako.
baseline	Khalai atthu yahikhotta vathi-va, khukelela vasulu vaya.
afri-byT5	Okhala wira okathi wa okathi ole ti wootepexa ottuli wa iphareya ni otepexa ottuli wa miyako.
afri-mT5	Nthowa nenlo ninkhala ntoko nsuwa ntoko nsuwa ni ninkhala ntoko nsuwa ni ninkhala ntoko nsuwa ni ninkhala ntoko nsuwa.
byT5	Ekamada eyo yootepa omalela vathi va iphareya ni yootepa omalela vathi va miyaako.
mT0	Okhala wira ematta eyo enniphwanyaneya ottuli wa maasi, nto ematta eyo enniphwanyaneya ottuli wa maasi.
mT5	Ekatana eyo ti yootepa otthuneya ovikana maasi ni yootepa otthuneya ovikana maasi.
M2M100	Ekaaxa ele ti yootepa otthuneya vathi va ephareya ni yootepa otthuneya vathi va mwaako.
NLLB	Mukattelo ti woorekama vathi vathi wa ophareya ni wootepa maasi vathi wa miyaako.

Table 10: Example of Portuguese to Emakhuwa translations

Headline	Content	Label
Asuweli ahoonenela yoophattuwa oniwaakiha anamwane a Waafirika wa ettekuxa.	Asuweli ahoonenela yoophattuwa oniwaakiha anamwane a Waafirika wa ettekuxa. Anamasuwela anihimya wirra yoooneleliwa enrowa othupereriha owanana eretta, eniiriha okhwa wa anamwane 500 wa khula mwaakha. Yoosomiwa yooniheriwe enamararu ela, mahiku 30, onihimya wiira asuweli aasuwanyeyiha mithinto sa mayarelaniyo oovirikanasa siniwaakiha anamwane akina a Waafirika wa wunnuwa wa mithinto sinceene sa ettekuxa. Asuweli aahimya wiira wooneleliwa onrowa othupereriha owanana eretta, enikuxeeriha okhwa w'anamwane 500 wa khula mwaakha. Wa yoosomiwa wuulupale opakiwe, anamathokosa ahaakhulela wiira osuwanyeehiwa wa soovirikanasa sa oyareriwa wa nipuro noosuweliwa mwa erutthu enikhaliyerya ohimya mwaha wa anamwane akina aninnuwaaya axikokho aye oovelaveliha ya ettekuxa ni akinaku khaninnuwa, mmuttettheni mme atthu akina anilummwaaya ni pwilimwithi oniruuhela eretta. Mikwahai ikina, aahimya asuweli, opwanyaneya wa enamuna ya mayarelaniwo yoovirika ennivukula mpakha waattamela eriyari erixiku ya anamwane okhwa mwaha wa eretta. "Ninniwerya naanaano ohimya, voohivonya, wiira mithinto sooyareliwa muttetthe owo waacenoma ya pinaatamu onoouhela waakiha woolipe wa ettekuxa yootepexa ni ihaali sa moolumenkuni yeekeekhayi, okhalaka oviikana ankha mwaana onookhala wala onookhwa", oolavula Dominic Kwiatkowski, Purusoora a Esenturu para Ceneetika Aapinaatamu ni Inxitituutu Sanger ya Fundasawu Wellcome ni mmosa anamathokosa anihoolela epuruceetu. Muteko waavariwa ni MalariaGEN, ereete ya olumwenku ya anamathokosa a Waafiika, Aasiya ni mittetthe sikina soowaattela owereya ehasara ya ettekuxa, vanceenexa ekhaliheriwe ni Efuntasawu Wellcome. Ettekuxa yoowiiva atthu oophiyeryaka ikonto 584 eyaakha ya 2013, sintoko itaatu sa Mutthenkeso Woolumenku wa Ekumi. Ophiyerya 90 puru sento ya atthu akwiiye anamwane ohiphiya iyaakha ithanu a Waafirika Supusahariyaana. Sintoko yoosomiwa, anamathokosa awehawehale itaatu sikhumale Opurukina Faaso, Okamaronxi, Okaana, Okheeniya, Omalawi, Omali, Okampiya ni Otanzaniya ni anilikaniha ni DNA a anamwane 5.633 ni ettekuxa yootepa ni DNA ya anamwane 5.919 aniwereya ettekuxa yoohitepa. Elaleyaka muteko aya mureviixitani siyentiifika Nature, anamathokosa aaleliherya wiira locus musyaa woonineliwe onipwanyaneya vakhiviru va makhuru ya genes ni ekootiku yoolattana ni kilikoforina, makhura ololowanne ni enamuna ntoko mwaaxiitthu a ettekuxa onivoluwaawe iseelula sooxeerya sa mphomeni.	health

Table 11: Examples of News classification in Emakhuwa

Model	Size	Hyperparameters
byT5-base / afri-byT5-base	580M	• Max source length: 200 • Max target length: 200 • Batch size: 8 • Beams: 4
mT5-base / afri-mT5-base	580M	• Max source length: 200 • Max target length: 200 • Batch size: 8 • Beams: 4
mT0	580M	• Max source length: 200 • Max target length: 200 • Batch size: 8 • Beams: 4
NLLB-200-distilled-600M	600M	• Max steps: 60000
M2M100	418M	• Max tokens: 1200 • Layers: 12 • Dropout: 0.3 • Attention dropout: 0.1 • Learning rate: 3e-05 • Max update: 40000 • Emakhuwa was mapped to Swahili (sw)

Table 12: Machine Translation Models Configurations