MasakhaNEWS: News Topic Classification for African languages

David Ifeoluwa Adelani[1*], Marek Masiak[1*], Israel Abebe Azime[2], Jesujoba Oluwadara Alabi[2], Atnafu Lambebo Tonja[3,6], Christine Mwase[4], Odunayo Ogundepo[5], Bonaventure F. P. Dossou[6,7,8,9], Akintunde Oladipo[5], Doreen Nixdorf, Chris Chinenye Emezue[9,10], Sana Sabah al-azzawi[11], Blessing K. Sibanda, Davis David[12], Lolwethu Ndolela, Jonathan Mukiibi[13], Tunde O. Ajayi[14], Tatiana Moteu Ngoli[15], Brian Odhiambo, Abraham Toluwase Owodunni, Nnaemeka C. Obiefuna, Muhidin Mohamed[16], Shamsuddeen Hassan Muhammad[17], Teshome Mulugeta Ababu[18], Saheed S. Abdullahi[19], Mesay Gemeda Yigezu[3], Tajuddeen Gwadabe, Idris Abdulmumin[20], Mahlet Taye Bame, Oluwabusayo O. Awoyomi[21], Iyanuoluwa Shode[22], Tolulope Anu Adelani, Habiba Abdulganiy Kailani, Abdul-Hakeem Omotayo[23], Adetola Adeeko, Afolabi Abeeb, Anuoluwapo Aremu, Olanrewaju Samuel[24], Clemencia Siro[25], Wangari Kimotho[26], Onyekachi Raphael Ogbu, Chinedu E. Mbonu[27], Chiamaka I. Chukwuneke[27,28], Samuel Fanijo[29], Jessica Ojo, Oyinkansola F. Awosan, Tadesse Kebede Guge[30], Sakayo Toadoum Sari[26,31], Pamela Nyatsine, Freedmore Sidume[32], Oreen Yousuf, Mardiyyah Oduwole[33], Kanda P. Tshinu, Ussen Kimanuka[34], Thina Diko, Siyanda Nxakama, Sinodos G. Nugussie[18], Abdulmejid Tuni Johar, Shafie Abdi Mohamed[34], Fuad Mire Hassan[35], Moges Ahmed Mehamed[36], Evrard Ngabire[37], Jules Twagirayezu, Ivan Ssenkungu, and Pontus Stenetorp[1]

[∀] Masakhane NLP, Africa, [1]University College London, United Kingdom, [2] Saarland University, Germany, [3] Instituto Politécnico Nacional, Mexico, [4]Fudan University, China, [5] University of Waterloo, Canada, [6] Lelapa AI, [7]McGill University, Canada, [8] Mila Quebec AI Institute, Canada, [9] Lanfrica, [10] Technical University of Munich, Germany [11] Luleå University of Technology, Sweden, [12]Tanzania Data Lab, Tanzania [13] Makerere University, Uganda, [14] Insight Centre for Data Analytics, Ireland, [15]Paderborn University, Germany, [16]Aston University, UK, [17]University of Porto, Portugal, [18] Dire Dawa University, Ethiopia [19]Kaduna State University, Nigeria, [20]Ahmadu Bello University, Nigeria, [21]The College of Saint Rose, USA [22]Montclair State University, USA, [23] University of California, Davis, [24]University of Rwanda, Rwanda [25]University of Amsterdam, The Netherlands, [26]AIMS, Cameroon, [27]Nnamdi Azikiwe University, Nigeria , [28]Lancaster University, United Kingdom, [29]Iowa State University, USA, [30]Haramaya University, Ethiopia, [31]AIMS, Senegal, [32]BIUST, Botswana, [33]NOUN, Nigeria, [34] PAUSTI, Kenya, [34] Jamhuriya University, Somalia, [35]Somali National University, [36]Wuhan University of Technology, China, [37]Deutschzentrum an der Universität Burundi, Burundi

Correspondence: d.adelani@ucl.ac.uk

Abstract

Despite representing roughly a fifth of the world population, African languages are underrepresented in NLP research, in part due to a lack of datasets. While there are individual language-specific datasets for several tasks, only a handful of tasks (e.g. named entity recognition and machine translation) have datasets covering geographical and typologically-diverse African languages. In this paper, we develop MasakhaNEWS—the largest dataset for news topic classification covering 16 languages widely spoken in Africa. We provide and evaluate a set of baseline models by training classical machine learning models and fine-tuning several language models. Furthermore, we explore several alternatives to full fine-tuning of language models that are better suited for zero-shot and few-shot learning, such as: cross-lingual parameter-efficient fine-tuning (MAD-X), pattern exploiting training (PET), prompting language models (ChatGPT), and prompt-free sentence transformer fine-tuning (SetFit and the co:here embedding API). Our evaluation in a few-shot setting, shows that with as little as 10 examples per label, we achieve more than 90% (i.e. 86.0 F1 points) of the performance of fully supervised training (92.6 F1 points) leveraging the PET approach. Our work shows that existing supervised approaches work well for all African languages and that language models with only a few supervised samples can reach competitive performance, both findings which demonstrate the applicability of existing NLP techniques for African languages.

[*] Equal contribution

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
November 1–4, 2023.

1 Introduction

News topic classification is a text classification task in NLP that involves categorizing news articles into different categories like sports, business, entertainment, and politics. It has shaped the development of several machine learning algorithms over the years, such as topic modeling (Blei et al., 2001; Dieng et al., 2020) and deep learning models (Zhang et al., 2015; Joulin et al., 2017). Similarly, news topic classification is a popular downstream task for evaluating the performance of large language models (LLMs) for both fine-tuning and prompt-tuning setups (Yang et al., 2019; Sun et al., 2019; Brown et al., 2020; Liu et al., 2023).

Despite the popularity of the task in benchmarking LMs, most of the evaluation have only been performed on English and a few other high-resource languages. It is *unclear how this approach extends to pre-trained multilingual language models* for low-resource languages. For instance, BLOOM (Scao et al., 2022) was pre-trained on 46 languages, including 22 African languages (mostly from the Niger-Congo family). However, extensive evaluation on these set of African languages was not performed due to lack of evaluation datasets. In general, only a handful of NLP tasks such as machine translation (Adelani et al., 2022a; NLLB-Team et al., 2022), named entity recognition (Adelani et al., 2021, 2022b), and sentiment classification (Muhammad et al., 2023) have standardized benchmark datasets covering several geographical and typologically-diverse African languages. Another popular task that can be used for evaluating the downstream performance of language models is news topic classification, but human-annotated datasets for benchmarking topic classification using language models for African languages are *scarce*.

In this paper, we address two problems: the lack of evaluation datasets and lack of extensive evaluation of LMs for African languages. We create a large-scale **news topic classification** dataset covering 16 typologically-diverse languages widely spoken in Africa, including English and French, with the same label categories across all languages. Our dataset is also suitable for **news headline generation** task (Aralikatte et al., 2023): a special type of text summarization. We provide several baseline models using both classical machine learning approaches and fine-tuning LMs. Furthermore, we explore several alternatives to full fine-tuning of language models that are better suited for zero-shot and few-shot learning (e.g. 5-examples per label) such as cross-lingual parameter-efficient fine-tuning (MAD-X (Pfeiffer et al., 2020)), pattern exploiting training (PET) (Schick and Schütze, 2021a), prompting ChatGPT LLM, and prompt-free, sentence transformer fine-tuning (SetFit) (Tunstall et al., 2022a), and the co:here embedding API.

Our evaluation in a zero-shot setting shows the potential of prompting ChatGPT for news topic classification for low-resource African languages. We found that GPT-3.5-Turbo has impressive result for languages that make use of Latin script, but *perform poorly for non-Latin based scripts like Amharic and Tigrinya*. However, *GPT-4 was able to overcome this challenge for non-Latin script* with impressive performance matching the result of cross-lingual transfer experiments from a related African language.

In a few-shot setting, we show that with as little as 10 examples per label, we achieved more than 90% (i.e. 86.0 F1 points) of the performance of full supervised training (92.6 F1 points) leveraging the PET approach. We hope this encourages the NLP community to benchmark and evaluate LLMs on more low-resource languages. For reproducibility, we release our data and code under academic license or CC BY-NC 4.0 on Github.[1]

2 Related Work

News topic classification , an application of text classification, is a popular task in natural language processing. There are various news topic classification datasets, including BBC News (Greene and Cunningham, 2006), AG News (Zhang et al., 2015), and the multimodal N24News (Wang et al., 2022), all of which are English datasets. In addition, there is the IndicNLP News (Kunchukuttan et al., 2020) which is a multilingual dataset for Indian langauges. For African languages, only a handful of human annotated datasets exists, such as the Hausa & Yorùbá dataset (Hedderich et al., 2020) (only covering news headline), KINNEWS & KIRNEWS datasets for Kinyarwanda and Kirundi (Niyongabo et al., 2020), and Tigrinya News (Fesseha et al., 2021). Others are semi-automatically created using predefined topics from news websites like Amharic news (Azime and Mohammed, 2021) and ANTC dataset (Alabi et al., 2022)—that covered five African languages (Lingala, Somali, Naija,

[1] https://github.com/masakhane-io/masakhane-news

Malagasy, and isiZulu). These datasets, however, have limitations due to the fact that they were created with little or no human supervision and using different labeling schemes. In contrast, in this work we present news topic classification data for 16 typologically diverse African languages with a consistent labeling scheme across all languages.

Prompting Language Models using manually designed prompts to guide text generation has recently been applied to a myriad of NLP tasks, including topic classification. Models such as GPT-3 (Brown et al., 2020) and T5 (Raffel et al., 2020; Sanh et al., 2022) are able to learn more structural and semantic relationships between words and have shown impressive results even in multilingual scenarios when tuned for different tasks (Chung et al., 2022; Muennighoff et al., 2023). One approach to prompt-tuning a language model for topic classification is to design a "template" for classification and insert a sequence of text into template (Gao et al., 2021; Shin et al., 2020).

There are some other approaches to few-shot learning without prompting. One of them is SetFit (Tunstall et al., 2022a), which takes advantage of sentence transformers to generate dense representations for input sequences. These representations are then passed through a classifier to predict class labels. The sentence transformers are trained on a few examples using contrastive learning where positive and negative training pairs are sampled by in-class and out-class sampling. Another common approach is Pattern-Exploiting Training also known as PET (Schick and Schütze, 2021a). PET is a semi-supervised training approach that used restructured input sequences to condition language models to better understand a given task, while iPET (Schick and Schütze, 2021b) is an iterative variant of PET that is also shown to perform better.

3 Languages

Table 1 presents the languages covered in along with information on their language families, their primary geographic regions in Africa, and the number of speakers. Our dataset consists of a total of 16 typologically-diverse languages, and they were selected based on the availability of publicly available news corpora in each language, the availability of native-speaking annotators, geographical diversity and most importantly, because they are widely spoken in Africa. English and French are official languages in 42 African countries, Swahili is native to 12 countries, and Hausa is native to 6 countries. In terms of geographical diversity, we have four languages spoken in West Africa, seven languages spoken in East Africa, two languages spoken in Central Africa (i.e. Lingala and Kiswahili), and two spoken in Southern Africa (i.e chiShona and isiXhosa). Also, we cover four language families, Niger-Congo (8) Afro-Asiatic (5), Indo-European (2), and English Creole (1). The only English creole language is Nigerian-Pidgin, also known as Naija. Each language is spoken by at least 10 million people, according to Ehnologue (Eberhard et al., 2021).

4 MasakhaNEWS dataset

4.1 Data Source

The data used in this research study were sourced from multiple reputable news outlets. The collection process involved crawling the British Broadcasting Corporation (BBC) and Voice of America (VOA) websites. We crawled between 2k–12k articles depending on the number of articles available on the websites. Some of the websites already have some pre-defined categories, we make use of this to additionally filter articles that do not belong to categories we plan to annotate. We took *inspiration* of news categorization from **BBC English** with six (6) pre-defined and well-defined categories *("business", "entertainment", "health", "politics", "sports", and "technology")* with over 500 articles in each category. For English, we only crawled articles belonging to these categories while for the other languages, we crawled all articles. Our target is to have around **3,000** articles for annotation but three languages (Lingala, Rundi, and Somali) have less than that. Table 2 shows the news source per language and the number of articles crawled.

4.2 Data Annotation

We recruited volunteers from the Masakhane community—an African grassroots community focused on advancing NLP for African languages.[2] The annotators were asked to label 3k articles into eight categories: "*business*", "*entertainment*", "*health*", "*politics*", "*religion*", "*sports*", "*technology*", and "*uncategorized*". Six of the categories are based on BBC English major news categories, the "*religion*" label was added since many African

[2]all annotators are were included as authors of the paper.

Language	Family/branch	Region	# speakers	News Source	# articles
Amharic (amh)	Afro-Asiatic / Ethio-Semitic	East Africa	57M	BBC	8,204
English (eng)	Indo-European / Germanic	Across Africa	1268M	BBC	5,073
French (fra)	Indo-European /Romance	Across Africa	277M	BBC	5,683
Hausa (hau)	Afro-Asiatic / Chadic	West Africa	77M	BBC	6,965
Igbo (ibo)	Niger-Congo / Volta-Niger	West Africa	31M	BBC	4,628
Lingala (lin)	Niger-Congo / Bantu	Central Africa	40M	VOA	2,022
Luganda (lug)	Niger-Congo / Bantu	Central Africa	11M	Gambuuze	2,621
Naija (pcm)	English Creole	West Africa	121M	BBC	7,783
Oromo (orm)	Afro-Asiatic / Cushitic	East Africa	37M	BBC	7,782
Rundi (run)	Niger-Congo / Bantu	East Africa	11M	BBC	2,995
chiShona (sna)	Niger-Congo / Bantu	Southern Africa	11M	VOA & Kwayedza	11,146
Somali (som)	Afro-Asiatic / Cushitic	East Africa	22M	BBC	2,915
Kiswahili (swa)	Niger-Congo / Bantu	East & Central Africa	71M-106M	BBC	6,431
Tigrinya (tig)	Afro-Asiatic / Ethio-Semitic	East Africa	9M	BBC	4,372
isiXhosa (xho)	Niger-Congo / Bantu	Southern Africa	19M	Isolezwe	24,658
Yorùbá (yor)	Niger-Congo / Volta-Niger	West Africa	46M	BBC	6,974

Table 1: **Languages covered in and Data Source**: including language family, region, number of L1 & L2 speakers, and number of articles from each news source.

news websites frequently cover this topic. Other articles that do not belong to the first seven categories, are assigned to the "*uncategorized*" label.

For each language, the annotation followed two stages. In the **first stage**, we randomly shuffled the entire dataset and asked annotators to label the first 200 articles manually. In the **second stage**, we made use of active learning by combining the first 200 annotated articles with articles with pre-defined labels where available, and trained a classifier (i.e. by fine-tuning AfroXLMR-base (Alabi et al., 2022)). We ran predictions on the rest of the articles, and asked annotators to correct the mistakes of the classifier. This approach helped to speed up the annotation process.

Annotation tool We make use of an in-house annotation tool to label the articles. Appendix A shows an example of the interface of the tool. To further simplify the annotator effort, we ask annotators to label articles based on the headlines instead of the entire article. However, since some headlines are not very descriptive, we decided to concatenate the headline and the first two sentences of the news text to provide additional context to annotators.

Inter-agreement score We report Fleiss Kappa score (Fleiss et al., 1971) to measure the agreement of annotation. Table 2 shows that all languages have a moderate to perfect Fleiss Kappa score (i.e. 0.55 - 0.85), which shows a high agreement among the annotators recruited for each language. Languages with only one annotator (i.e. Luganda and Rundi) were excluded in the evaluation.

Deciding a single label per article After annotation, we assigned the final label to each article by majority voting. Each label of an article needs to be agreed by a minimum of two annotators to be assigned the label. We only had exceptions for Luganda and Rundi, since they had one annotator. Our final dataset for each language consist of a minimum of 72 articles per topic, and a maximum of 500, except for English language where the classes are roughly balanced. We excluded the infrequent labels so we do not have a highly unbalanced dataset. The choice of a minimum of 72 articles ensures a minimum of 50 articles in the training set. [3] Our target is to have at least four topics per language with a minimum of 72 articles. This approach worked smoothly except for two languages: Lingala ("politics", "health" and "sports") and chiShona ("business", "health" and "politics"), where we had only three topics with more than 72 articles. To ensure we have more articles per class, we had to resolve the conflict in annotation between Lingala annotators to ensure we have more labels for the "business" category. This approach still results in infrequent classes for chiShona. We had to crawl additional "sports" articles from a local chiShona website (*Kwayedza*), followed by manual filtering of unrelated sports news.

Data Split Table 2 provides the data split for languages. We also provide the distribution of articles by topics. We divided the annotated data into TRAIN, DEV and TEST split following 70% / 10% / 20% split ratio.

5 Baseline Experiments

We trained baseline text classification models by concatenating the news headline and news text using different approaches.

[3] since we require 50 instances per class or 50-shots for the few-shot experiments in (§6.2.2)

Language	Train/Dev/Test	# topics	Topics (number of articles per topic) # bus	# ent	# health	# pol	# rel	# sport	# tech	# Annotator	Fleiss Kappa
Amharic (amh)	1311/ 188/ 376	4	404	-	500	500	-	471	-	5	0.81
English (eng)	3309/ 472/ 948	6	799	750	746	821	-	1000	613	7	0.81
French (fra)	1476/ 211/ 422	5	500	-	500	500	-	500	109	3	0.83
Hausa (hau)	2219/ 317/ 637	7	399	500	493	500	493	497	291	5	0.85
Igbo (ibo)	1356/ 194/ 390	6	292	366	424	500	73	285	-	4	0.65
Lingala (lin)	608/ 87/ 175	4	82	-	193	500	-	95	-	2	0.56
Luganda (lug)	771/ 110/ 223	5	169	-	228	500	91	116	-	1	-
Oromo (orm)	1015/ 145/ 292	4	-	119	447	500	-	386	-	3	0.63
Naija (pcm)	1060/ 152/ 305	5	97	460	159	309	-	492	-	4	0.66
Rundi (run)	1117/ 159/ 322	6	76	158	372	500	73	419	-	1	-
chiShona (sna)	1288/ 185/ 369	4	500	-	425	500	-	417	-	3	0.63
Somali (som)	1021/ 148/ 294	7	114	139	354	500	73	148	135	3	0.55
Kiswahili (swa)	1658/ 237/ 476	7	316	98	500	500	292	500	165	4	0.72
Tigrinya (tir)	947/ 137/ 272	6	80	167	395	500	-	125	89	2	0.63
isiXhosa (xho)	1032/ 147/ 297	5	72	500	100	308	-	496	-	3	0.89
Yorùbá (yor)	1433/ 206/ 411	5	-	500	398	500	317	335	-	5	0.80

Table 2: **MasakhaNEWS dataset**. The size of the annotated data, news topics, and number of annotators. Topics are labelled by their prefixes in the table (**topics**): **bus**iness, **ent**ertainment, **health**, **pol**itics, **rel**igion, **sport**, **tech**nology.

5.1 Baseline Models

We trained three classical ML models: Naive Bayes, multi-layer perceptron, and XGBoost using the popular `sklearn` tool (Pedregosa et al., 2011). We employed the "CountVectorizer" method to represent the text data, which converts a collection of text documents to a matrix of token counts. This method allows us to convert text data into numerical feature vectors.

Furthermore, we fine-tune nine kinds of multilingual text encoders, seven of them are BERT/RoBERTa-based i.e. XLM-R (base & large) (Conneau et al., 2020), AfriBERTa-large (Ogueji et al., 2021), RemBERT (Chung et al., 2021), AfroXLMR (base & large) (Alabi et al., 2022), and AfroLM (Dossou et al., 2022), the other two are mDeBERTaV3 (He et al., 2021a), and LaBSE (Feng et al., 2022). mDeBERTaV3 pretrained a DeBERTa-style model (He et al., 2021b) with replaced token detection objective proposed in ELECTRA (Clark et al., 2020). On the other hand, LaBSE is a multilingual sentence transformer model that is popular for mining parallel corpus for machine translation.

Finally, we fine-tuned four multilingual Text-to-Text (T2T) models, mT5-base (Xue et al., 2021), Flan-T5-base (Chung et al., 2022), AfriMT5-base (Adelani et al., 2022a), AfriTeVA-base (Jude Ogundepo et al., 2022). The fine-tuning and evaluation of the multilingual text-encoders and T2T models were performed using HuggingFace Transformers (Wolf et al., 2020) and PyTorch Lightning[4]. The models were fine-tuned on Nvidia V100 GPU for 20 epochs, batch size of 32, $1e-5/5e-5$ lr, and max. sequence length of 256.

The LMs evaluated were both massively multilingual (i.e. typically trained on over 100 languages around the world) and African-centric (i.e. trained mostly on languages spoken in Africa). The African-centric multilinual text encoders are all modeled after XLM-R. AfriBERTa was pretrained from scratch on 11 African languages, AfroXLMR was adapted to African languages through fine-tuning the original XLM-R model on 17 African languages and 3 languages commonly spoken in Africa, while AfroLM was pretrained on 23 African languages utilizing active learning. Similar to the multilingual text encoders, the T2T models used in this study were pretrained on hundreds of languages, and they are all based on the T5 model (Raffel et al., 2020), which is an encoder-decoder model trained with the span-mask denoising objective. mT5 is a multilingual version of T5, and Flan-T5 was fine-tuned on multiple tasks using T5 as a base. The study also included adaptations of the original models, such as AfriMT5-base, as well as AfriTeVA-base, a T5 model pre-trained on 10 African languages.

5.2 Baseline Results

Table 3 shows the result of training several models on TRAIN split and evaluation on the TEST split for each language. Our evaluation shows that classical ML models are worse in general than fine-tuning multilingual LMs on average, however, the drop in performance is sometimes comparable to LMs if the language was not covered during the pre-training. For example, MLP, NaiveBayes and

[4] https://pypi.org/project/pytorch-lightning/

Model	size	amh	eng	fra	hau	ibo	lin	lug	orm	pcm	run	sna	som	swa	tir	xho	yor	AVG
classical ML																		
MLP	<20K	92.0	88.2	84.6	86.7	80.1	84.3	82.2	86.7	93.5	85.9	92.6	71.1	77.9	81.9	94.5	89.3	85.7
NaiveBayes	<20K	91.8	83.7	84.3	85.3	79.8	82.8	84.0	85.6	92.8	79.9	91.5	74.8	76.6	71.4	91.0	84.0	83.7
XGBoost	<20K	90.1	86.0	81.2	84.7	78.6	74.8	83.8	83.2	93.3	79.2	94.3	68.5	74.9	75.2	91.1	85.2	82.8
multilingual text encoders																		
AfriBERTa	126M	90.6	88.9	76.4	89.2	87.3	87.0	85.1	89.4	98.1	91.3	89.3	83.9	83.3	87.0	86.9	90.3	87.8
XLM-R-base	270M	90.9	90.6	90.4	88.4	82.5	87.9	65.3	82.2	97.8	85.9	88.9	73.8	85.6	54.6	78.6	84.5	83.0
AfroXLMR-base	270M	94.2	92.2	**92.5**	91.0	90.7	93.0	89.4	92.1	98.2	91.4	95.4	85.2	88.2	86.5	94.7	93.0	91.7
AfroLM	270M	90.3	87.7	77.5	88.3	85.4	85.7	88.0	83.5	95.9	86.8	92.5	72.0	83.2	83.5	91.4	86.5	86.1
mDeBERTa	276M	91.7	90.8	89.2	88.6	88.3	81.6	65.7	84.7	96.8	89.4	93.9	72.0	84.6	78.7	90.5	89.3	86.0
LABSE	471M	92.5	91.6	90.9	90.0	91.6	89.6	86.8	86.7	98.4	91.1	94.6	82.1	87.6	83.8	94.7	92.1	90.3
XLM-R-large	550M	93.1	92.2	91.4	90.6	84.2	91.8	73.9	88.4	98.4	87.0	88.9	76.1	85.6	62.7	89.2	84.5	86.1
AfroXLMR-large	550M	**94.4**	**93.1**	91.1	**92.2**	**93.4**	**93.7**	**89.9**	**92.1**	**98.8**	**92.7**	**95.4**	**86.9**	**87.7**	**89.5**	**97.3**	**94.0**	**92.6**
RemBERT	559M	92.4	92.4	90.8	90.5	91.1	91.5	86.7	88.7	98.2	90.6	93.9	75.9	86.7	69.9	92.5	93.0	89.1
multilingual text-to-text LMs																		
AfriTeVa-base	229M	87.0	80.3	71.9	85.8	79.9	82.8	60.2	82.9	95.2	80.0	84.4	58.0	80.7	55.2	69.4	86.4	77.5
mT5-base	580M	78.2	89.8	59.0	82.7	76.8	80.8	75.0	79.2	96.1	85.7	90.4	75.0	76.1	65.1	71.8	86.2	80.0
Flan-T5-base	580M	54.5	92.4	88.9	84.5	86.6	90.6	84.1	85.8	97.8	87.3	90.6	76.0	79.0	41.5	90.8	88.0	82.4
AfriMT5-base	580M	90.2	90.3	87.4	87.9	88.0	88.6	84.8	83.9	96.6	91.0	91.5	77.8	84.4	80.8	91.6	88.8	87.7

Table 3: **Baseline results on** . We compare several ML approaches using both classical ML and LMs. Average is over 5 runs. Evaluation is based on weighted F1-score. Africa-centric models are in gray color

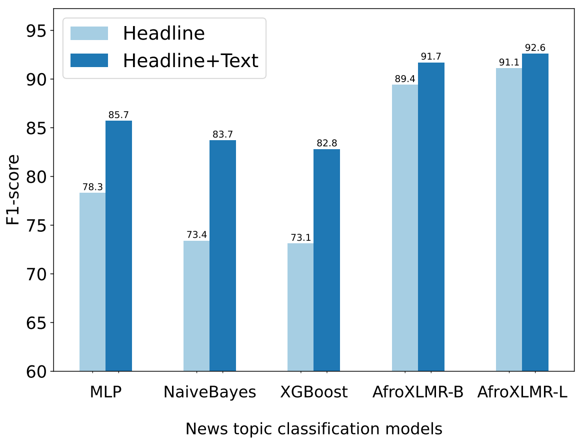

Figure 1: **Comparison of article content type used for training news topic classification models**. We report the average across all languages when either `headline` or `headline+text` is used

XGBoost have better performance than AfriBERTa on `fra` and `sna` since they were not seen during pre-training of the LM. Similarly, AfroLM had worse result for `fra` for the same reason. On average, XLM-R-base, AfroLM, mDeBERTaV3, XLM-R-large gave 83.0 F1, 86.1 F1, 86.0 F1, and 86.1 F1 respectively, with worse performance compared to the other LMs ($87.8 - 92.6$ F1) because they do not cover some of the African languages during pre-training (see Table 6) or they have been pre-trained on a small data (e.g. AfroLM pretrained on less than 0.8GB despite seeing 23 African languages during pre-training). Larger models such as LABSE and RemBERT that cover more languages performed better than the smaller models, for example, LABSE achieved over of 2.5 F1 points over AfriBERTa.

The best result achieved is by AfroXLMR-base/large with over 4.0 F1 improvement over AfriBERTa. The larger variant gave the overall best result due to the size. AfroXLMR models benefited from being pre-trained on most of the languages we evaluated on. We also tried multilingual T2T models, but none of the models reach the performance of AfroXLMR-large despite their larger sizes. We observe the same trend that the adapted mT5 model (i.e. AfriMT5) gave better result compared to mT5 similar to how AfroXLMR gave better result than XLM-R. We found *FlanT5-base to be competitive to AfriMT5* despite seeing few African languages, however, *the performance was very low for languages that uses the Ge'ez script* like `amh` and `tir` since the model do not support Ge'ez.

Headline-only training We compare our results using `headline+text` (as shown in Table 3) with training on the article `headline`—with shorter content, we find out that fine-tuned LMs gave impressive performance with only headlines while classical ML methods struggle due to shorter content. Figure 1 shows the result of our comparison. AfroXLMR-base and AfroXLMR-large both improve by (2.3) and (1.5) F1 points respectively if we use `headline+text` instead of `headline`. Classical ML models improve the most when we make use of `headline+text` instead of `headline`; MLP, NaiveBayes and XGBoost improve by large F1 points (i.e. $7.4 - 9.7$). Thus, for the remainder of this paper, we make use of `headline+text`. Appendix B provides the breakdown of the result by languages for the comparison of `headline` and `headline+text`.

SRC LANG	amh	eng	fra	hau	ibo	lin	lug	orm	pcm	run	sna	som	swa	tir	xho	yor	AVG	AVG^{src}
Fine-tune (AfroXLMR-base)																		
hau	81.8	78.8	72.9	91.5	83.2	74.4	57.5	63.3	93.2	81.6	85.5	63.3	80.7	73.2	77.4	80.4	77.4	76.2
swa	89.5	82.4	86.7	80.8	81.5	74.5	66.5	63.8	92.7	86.2	83.6	74.7	87.3	71.8	72.6	80.4	79.7	79.1
MAD-X																		
hau	81.0	79.5	72.2	90.3	87.4	82.6	84.4	80.2	91.2	76.0	89.9	66.5	81.2	72.6	82.8	87.4	81.6	81.0
swa	91.0	80.9	86.1	81.2	83.0	85.0	75.1	82.6	94.2	86.9	90.1	74.6	88.4	77.6	80.7	88.8	**84.1**	**84.0**
PET																		
None	67.2	53.3	51.7	42.1	50.4	28.6	27.0	43.9	63.1	57.9	62.2	39.2	53.8	45.2	56.0	49.7	49.5	49.7
SETFIT																		
None	75.8	61.6	60.1	53.3	53.1	59.6	40.1	38.9	72.0	55.1	66.6	49.4	55.2	37.8	49.3	63.7	55.7	55.9
ChatGPT (GPT 3.5 Turbo) - Mar 23 version																		
None	33.3	79.3	67.6	59.4	65.0	62.3	59.4	62.9	93.2	73.6	73.0	62.0	69.3	41.4	73.9	80.1	66.0	66.2
ChatGPT (GPT 3.5 Turbo) - May 24 version																		
None	36.1	79.5	69.6	70.1	78.3	75.1	64.7	72.0	93.1	82.2	84.5	72.3	75.9	45.0	78.0	81.7	72.4	72.3
GPT 4 – May 24 version																		
None	88.5	79.1	77.3	76.5	84.0	82.6	77.9	70.0	**96.2**	**88.6**	**90.8**	**77.3**	75.0	76.7	**83.1**	83.7	81.7	82.5

Table 4: **Zero-shot learning on** . We compare several approaches such as using MAD-X, PET and SetFit. We excluded the source languages hau and swa from the average (AVG^{src}).

6 Zero-shot and Few-shot transfer

6.1 Methods

Here, we compare different zero-shot and few-shot methods:

Fine-tune (Fine-tune on a *source language*, and evaluate on a *target language*) using AfroXLMR-base. This is only used in the **zero-shot setting**.

MAD-X (Pfeiffer et al., 2020, 2021) - a parameter efficient approach for cross-lingual transfer leveraging the modularity, and portability of adapters (Houlsby et al., 2019). We followed the same **zero-shot** setup as Alabi et al. (2022), however, we make use of hau and swa as source languages since they cover all the news topics used by all languages. The setup is as follows: (1) We train language adapters using monolingual news corpora of our focus languages. We perform language adaptation on the *news* corpus to match the domain of our dataset, similar to (Alabi et al., 2022). (2) We train a task adapter on the source language labelled data using source language adapter. (3) We substitute the source language adapter with the target language to run prediction on the target language test set, while retaining the task adapter.

PET/iPET (Schick and Schütze, 2021a,b), also known as (**I**terative) **P**attern **E**xploiting **T**raining is a semi-supervised approach that makes use of few labelled examples and a prompt/pattern to a LM for few-shot learning. It involves three steps. (1) designing of a prompt/pattern and a verbalizer (that maps each label to a word from LM vocabulary). (2) train an LM on each pattern based on few labelled examples (3) distill the knowledge of the LM on unlabelled data. Therefore, PET leverages unlabelled examples to improve few-shot learning. iPET on the other hand, repeats step 2 and 3 iteratively. We make use of the same set of patterns used for AGNEWS English dataset (Zhang et al., 2015) provided by the PET/iPET authors. The patterns are (1) $P_1(x) = \_\_\_\_ : a, b$ (2) $P_2(x) = a(\_\_\_\_)b$ (3) $P_3(x) = \_\_\_\_ - ab$ (4) $P_4(x) = ab(\_\_\_\_)$ (5) $P_5(x) = \_\_\_\_News : ab$ (6) $P_6x) = [Category : \_\_\_\_]ab$, where a is the news headline and b is the news text. In evaluation, we take average over all patterns.

SetFit (Tunstall et al., 2022b) is a few-shot learning framework based on sentence transformer models (Reimers and Gurevych, 2019) like LaBSE following two steps. **Step 1** fine-tunes the sentence transformer model using a few labelled examples with contrastive learning—where positive examples, are K-examples from a class c, and negative examples pairs are labelled examples with random labels from other classes. Contrastive learning approach enlarges the size of training data in few-shot scenarios. In **Step 2**, the fine-tuned sentence transformer model is used to extract rich sentence representation for each labelled example, followed by logistic regression for classification. The advantage of this approach is that it is faster and requires no prompt unlike PET. We use this in both **zero- and few-shot setting**. For the zero-shot setting, SetFit creates dummy example N-times (we set $N = 8$, similar to the SetFit paper) like **"this sentence is {}"** where { } can be any news topic like "sports".

Co:here multilingual sentence transformer co:here introduced a multilingual embedding model *multilingual-22-12* [5], which supports over a hundred languages, including most of the languages included in . This is only for the few-shot setting.

OpenAI ChatGPT API[6] is an LLM trained on a large chunk of texts to predict the next word like GPT-3 (Brown et al., 2020), followed by a set of instructions in a prompt based on human feedback. It leverages Reinforcement Learning from Human Feedback (RLHF), similar to InstructGPT (Ouyang et al., 2022) to make the LLM to interact in a conversational way. We prompt the OpenAI API based on GPT-3.5 Turbo and GPT-4 to categorize articles into news topics. For the prompting, we make use of a simple template from Sanh et al. (2022): *'Is this a piece of news regarding {{"business, entertainment, health, politics, religion, sports or technology"}}? {{INPUT}}'*. We make use of the first 100 tokens of `headline+text` as {{INPUT}}. The completion of the LLM can be a single word, a sentence, or multiple sentences. We check if a descriptive word relating to any of the news topics has been predicted. For example, "economy", "economic", "finance" is mapped to "business" news. We provide more details on the ChatGPT evaluation in Appendix C.

For all few-shot settings, we tried K samples/shots per class where $K = 5, 10, 20, 50$. We make use of LaBSE as the sentence transformer for SetFit, and AfroXLMR-large as the LM for PET.

6.2 Results

6.2.1 Zero-shot evaluation

GPT-3.5-Turbo performs poorly on non-Latin scripts Table 4 shows the result of zero-shot evaluation using FINE-TUNE, MAD-X, PET, SETFIT and GPT-3.5-TURBO (March 2023 version). Our result shows that cross-lingual zero-shot transfer from a source language with same domain and task (i.e FINE-TUNE & MAD-X), gives superior result (+11 F1) than PET, SetFit, and GPT-3.5-TURBO. GPT-3.5-TURBO gave better results with over +9.0 F1 point better than SETFIT and PET showing that capabilities of instruction-tuned LLMs over smaller LMs. However, the results of CHATGPT were poor (< 42.0) for non-Latin based languages like Amharic and Tigrinya which makes use of the Ge'ez script. The languages that make use of Latin script have over 59.0%. Surprisingly, some results of GPT-3.5-TURBO are comparable to the FINE-TUNE approach for some languages (English, Luganda, Oromo, Naija, Somali, isiXhosa, and Yorùbá), without leveraging any additional technique apart from prompting the LLM.

GPT-3.5-Turbo evaluation improves with newer versions We repeated GPT-3.5-TURBO evaluation using a newer version (May 23, 2023 version), our results suggest a significant improvement of the result for 14 (out of 16) languages in our evaluation. This implies that the newer version of the model seems to be better than older versions for the news topic classification task.

GPT-4 overcomes the limited non-Latin capabilities of GPT-3.5-Turbo We also evaluated on GPT-4 on the 16 languages in zero-shot setting. Our results shows a significant improvement in performance over GPT-3.5-TURBO by over +9 points. Surprinsingly, GPT-4 was able to overcome the limitation of GPT-3.5-TURBO for languages with non-Latin script (i.e Amharic and Tigrinya) with impressive performance, matching the performance of cross-lingual transfer experiment from a related African language (i.e. FINE-TUNE `hau/swa`$\rightarrow$ `xx` and MAX-X `hau`$\rightarrow$ `xx`).

The large performance gap between GPT-3.5-Turbo and GPT-4 may be due to either the former being a distilled version of a more powerful model created to reduce inference cost, which also significantly affected its performance on non-Latin scripts.[7] [8] Alternatively, GPT-4 may just be a bigger and better model with more multilingual and non-Latin capabilities.

Leveraging labelled data from other languages is more effective In general, it may be advantageous to consider leveraging knowledge from other languages with available training data when no labelled data is available for the target language. Also, we observe that Swahili (`swa`) achieves better result as a source language than Hausa (`hau`) especially when transferring to `fra` (+13.8), `lug` (+9.0), and `eng` (+3.6). The reason for the impressive performance from Swahili to Luganda might be due to both languages belonging to the same Greater Lake Bantu language sub-group, but it is

[5] https://docs.cohere.ai/docs/text-classification-with-classify

[6] https://openai.com/blog/chatgpt

[7] https://arstechnica.com/information-technology/2023/07/is-chatgpt-getting-worse-over-time-study-claims-yes-but-others-arent-sure/

[8] https://platform.openai.com/docs/models/gpt-3-5

Model	amh	eng	fra	hau	ibo	lin	lug	orm	pcm	run	sna	som	swa	tir	xho	yor	AVG
Fine-tune (AfroXLMR-large)																	
5-shots	68.4	55.1	58.0	35.8	71.3	52.7	29.2	39.2	92.5	71.2	70.2	18.1	42.5	30.2	46.5	62.7	52.7
10-shots	75.5	75.2	65.9	64.6	86.1	72.6	31.3	56.8	95.8	87.3	80.8	38.9	73.8	36.3	61.7	69.4	67.0
20-shots	88.5	85.6	78.3	85.2	90.4	80.8	48.4	41.1	97.4	90.0	92.3	63.6	82.9	67.3	83.1	84.3	78.7
50-shots	91.4	87.5	86.9	88.8	87.3	91.0	75.2	71.3	96.4	89.8	95.5	85.3	86.6	86.2	94.1	90.2	87.7
Fine-tune (LaBSE)																	
5-shots	71.6	67.4	61.3	60.7	63.6	65.9	59.5	43.3	86.5	65.6	83.1	25.4	49.1	36.1	46.0	71.2	59.7
10-shots	79.0	77.1	76.8	79.7	77.1	70.2	68.3	58.5	94.5	81.9	84.8	44.8	77.2	51.8	69.9	79.8	73.2
20-shots	90.3	84.7	83.1	85.1	82.0	82.2	70.4	72.3	95.5	86.0	90.6	66.6	84.3	69.0	80.5	86.0	81.8
50-shots	89.6	86.3	85.6	87.1	86.4	88.4	80.6	77.8	96.7	87.9	93.0	80.1	85.3	79.6	87.4	88.6	86.3
PET																	
5-shots	89.9	80.8	72.3	82.6	85.0	82.9	79.0	89.2	94.5	87.7	88.9	69.5	79.6	59.7	84.3	84.0	81.9
10-shots	91.1	81.7	83.3	86.6	86.1	87.6	84.0	91.8	96.6	90.8	91.4	74.9	81.1	69.2	88.9	90.5	86.0
20-shots	92.7	86.4	82.8	89.1	88.6	89.2	83.8	94.9	96.7	88.7	93.3	81.6	83.5	72.4	91.5	91.0	87.9
50-shots	92.9	89.2	89.1	90.9	90.6	89.6	86.7	96.0	97.2	90.9	94.8	84.2	84.2	76.4	93.5	92.4	89.9
SetFit																	
5-shots	68.3	69.6	64.3	76.0	78.9	48.3	28.9	38.8	91.2	74.8	85.8	68.9	76.8	73.1	84.0	60.2	68.0
10-shots	84.8	82.0	80.5	79.4	71.4	77.8	49.5	57.3	92.8	83.8	89.2	65.1	81.2	64.9	83.6	76.5	76.2
20-shots	87.9	78.5	83.9	83.3	81.8	86.6	71.7	61.0	97.4	87.0	83.2	69.4	79.2	64.9	78.4	85.0	80.0
50-shots	88.6	76.6	83.8	83.0	77.3	81.9	60.8	63.6	93.6	85.6	90.6	67.9	76.5	69.8	83.8	86.0	79.3
Cohere sentence embedding API																	
5-shots	66.0	65.9	60.2	74.2	72.0	69.8	50.2	50.0	74.0	61.2	78.1	52.8	67.7	60.1	68.3	71.9	65.2
10-shots	80.1	72.5	71.4	80.4	75.7	78.4	65.5	57.2	84.9	78.2	85.0	60.4	73.8	59.8	83.2	80.1	74.2
20-shots	87.6	78.0	78.4	82.9	77.7	86.9	70.2	63.9	88.7	82.7	86.6	65.3	79.0	64.8	88.2	83.9	79.1
50-shots	90.2	80.9	83.2	85.6	81.9	87.7	78.0	70.6	94.9	84.1	90.5	68.9	77.6	72.8	90.4	88.4	82.9

Table 5: **Few-shot learning on** . We compare several few-shot learning approaches: PET, SetFit and Cohere Embedding API.

unclear why Hausa gave worse results than Swahili when adapting to English or French. However, with few examples, PET and SetFit methods are powerful without leveraging training data and models from other languages.

6.2.2 Few-shot evaluation

Table 5 shows the result of the few-shot learning approaches. With only 5-shots, we find all the few-shot approaches to be better than the usual FINE-TUNE baselines for most languages. However, as the number of shots increases, they have comparable results with SETFIT and CO:HERE API especially for $K = 20, 50$ shots. However, we found that PET achieved very impressive results with 5-shots (81.9 on average), matching the performance of SETFIT/CO:HERE API with 50-shots. The results are even better with more shots i.e ($k = 10$, 86.0 F1), ($k = 20$, 87.9 F1), and ($k = 50$, 89.9 F1). Surprisingly, with 50-shots, PET gave competitive result to the full-supervised setting (i.e. fine-tuning all TRAIN data) that achieved (92.6 F1) (see Table 3). It's important to note that PET make use of additional unlabelled data while SetFit and Cohere API do not. In general, our result highlight the importance of getting few labelled examples for a new language we are adapting to, even if it is as little as 10 examples per class—which is typically not time-consuming to annotate (Lauscher et al., 2020; Hedderich et al., 2020).

7 Conclusion

In this paper, we created the largest news topic classification dataset for 16 typologically diverse languages spoken in Africa. We provide an extensive evaluation using both full-supervised and few-shot learning settings. Furthermore, we study different techniques of adapting prompt-based tuning and non-prompt methods of LMs to African languages. Our experimental results shows that prompting LLMs like ChatGPT perform poorly on the simple task of text classification for several under-resourced African languages especially for non-Latin based scripts. Furthermore, we showed the potential of prompt-based few-shot learning approaches like PET (based on smaller LMs) for African languages. Our work shows that existing supervised approaches work well for all African languages and that language models with only a few supervised samples can reach competitive performance, both findings which demonstrate the applicability of existing NLP techniques for African languages.

In the future, we plan to extend this dataset to more African languages, include the evaluation of other multilingual LLMs like BLOOM, mT0 (Muennighoff et al., 2022) and XGLM (Lin et al., 2022), and extend analysis to other text classification tasks like sentiment classification (Shode et al., 2022, 2023; Muhammad et al., 2023).

8 Limitations

One major limitation of our work is that we did not evaluate extensively the performance of ChatGPT LLM on several African languages and tasks such as question answering, and text generation tasks. Our evaluation is only limited to text classification and may not generalize to many tasks. However, we feel that if it perform poorly on text classification, the result may even be worse on more difficult NLP tasks. Also, there is a challenge that our result may not be fully reproducible since we use the ChatGPT API where the underlining LLM are often updated or improved with time. It might be that the support for non-Latin based script may improve significantly in few months. This limitation also applied to the co:here embedding API.

9 Ethics Statement

Our work aims to provide benchmark dataset for African languages, we do not see any potential harms when using our news topic classification datasets and models to train ML models, the annotated dataset is based on the news domain, and the articles are publicly available, and we believe the dataset and news topic annotation is unlikely to cause unintended harm. Also, we do not see any privacy risks in using our dataset and models because it is based on news domain.

Acknowledgments

We would like to thank Yuxiang Wu for the suggestions on the few-shot experiments. We are grateful for the feedback from the anonymous reviewers of AfricaNLP and IJCNLP-AACL that helped improved this draft. David Adelani acknowledges the support of DeepMind Academic Fellowship programme. This work was supported in part by Oracle Cloud credits and related resources provided by Oracle. Finally, we are grateful to OpenAI for providing API credits through their Researcher Access API programme to Masakhane for the evaluation of GPT-3.5 and GPT-4 large language models.

References

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 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, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. 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 M. Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, 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.

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.

Rahul Aralikatte, Ziling Cheng, Sumanth Doddapaneni, and Jackie Chi Kit Cheung. 2023. Vārta: A large-scale headline-generation dataset for indic languages. *ArXiv*, abs/2305.05858.

Israel Abebe Azime and Nebil Mohammed. 2021. An amharic news text classification dataset. *CoRR*, abs/2103.05639.

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Bonaventure FP Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. 2022. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages. *arXiv preprint arXiv:2211.03263*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. Ethnologue: Languages of the world. twenty-third edition.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Awet Fesseha, Shengwu Xiong, Eshete Derb Emiru, Moussa Diallo, and Abdelghani Dahou. 2021. Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. *Information*, 12(2).

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 377–384, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*.

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.

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid. Association for Computational Linguistics.

Anoop Kunchukuttan, Divyanshu Kakwani, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages. *arXiv preprint arXiv:2005.00085*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

NLLB-Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo G. Ponferrada, Efrat Levkovizh, ..., Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Iyanuoluwa Shode, David Ifeoluwa Adelani, and Anna Feldman. 2022. yosm: A new yoruba sentiment corpus for movie reviews.

Iyanuoluwa Shode, David Ifeoluwa Adelani, JIng Peng, and Anna Feldman. 2023. NollySenti: Leveraging transfer learning and machine translation for Nigerian movie sentiment classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 986–998, Toronto, Canada. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022a. Efficient few-shot learning without prompts. *ArXiv*, abs/2209.11055.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022b. Efficient few-shot learning without prompts.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. 2022. N24News: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6768–6775, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Annotation Tool

Figure 2 provides an example of the interface of our in-house annotation tool.

B Comparing different article content types

Table 7 provides the comparison between using only news `headline` and `headline+text` for training. We find significantly improvement on average when we make use of `headline+text` for training across all models and languages especially for classical ML methods (MLP, NaiveBayes, and XGBoost).

C ChatGPT Evaluation

We prompted ChatGPT for news topic classification using the following template: *'Is this a piece of news regarding {{"business, entertainment, health, politics, religion, sports or technology"}}? {{INPUT}}'*. The completion may take different forms e.g. a single word, sentence or multiple sentences. Examples of such predictions are:

1. sports
2. This is a piece of news regarding sports.
3. This is a piece of sports news regarding the CHAN 2021 football tournament in Cameroon. It reports that the Mali national football team has advanced to the semi-finals after defeating the Congo national team in a match that ended in a penalty shootout.
4. This is a piece of news regarding sports. It talks about the recent match between Tunisia and Angola in the African Cup of Nations. Both teams scored a goal, and the article mentions some of the details of the game, such as the penalty and missed chances.
5. I'm sorry, but I'm having trouble understanding this piece of news as it appears to be in a language I don't recognize. Can you please provide me with news in English so I can assist you better?

To extract the right category, we make use of a simple verbalizer that maps the news topic to several indicative words (capitalization ignored) for the category like:

(a) 'business': {'business', 'finance', 'economy'. 'economics' }

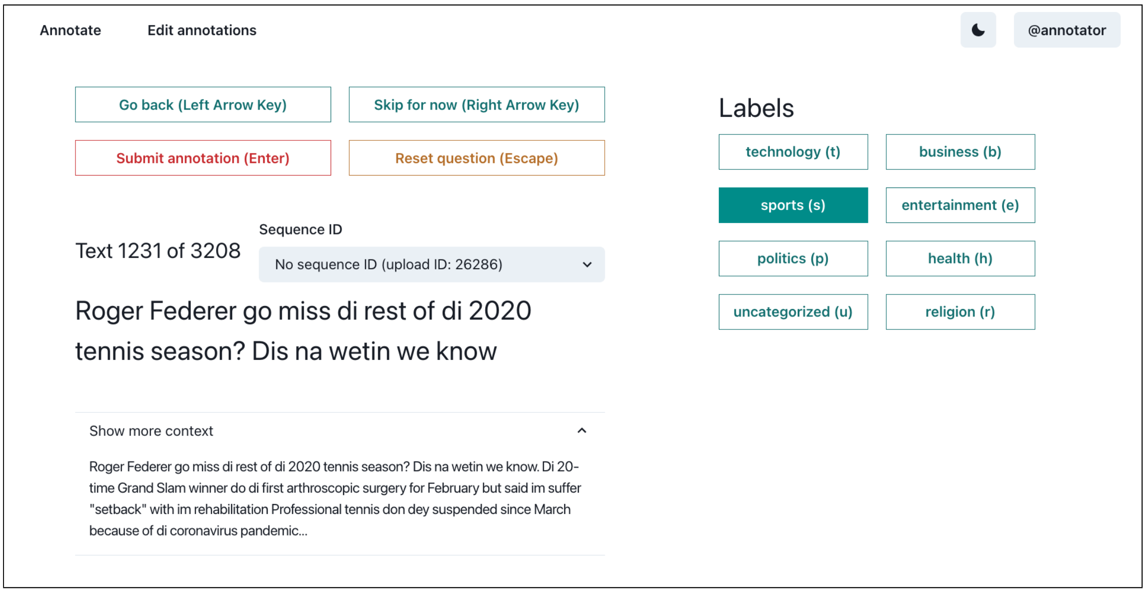

Figure 2: **Interface of our in-house Annotation tool**. Annotators can correct the pre-defined category assigned and also edit their annotation

(b) 'entertainment': {'entertainment' , 'music' }

(c) 'health': {'health' }

(d) 'politics': {'politics', 'political' }

(e) 'religion': {'religion' }

(f) 'sports': {'sports', 'sport' }

(g) 'technology': {'technology' }

When the right category is not obvious, like (5 : "I'm sorry, but I'm having trouble understanding this piece of news as it appears to be in a language I don't recognize. "), we choose a random category before computing F1-score.

LLM	LLM size	# Lang.	# African Lang.	Focus languages covered
XLM-R-base/large	270M/550M	100	8	amh, eng, fra, hau, orm, som, swa, xho
AfriBERTa-large	126M	11	11	amh, hau, ibo, orm, pcm, run, swa, tir, yor
mDeBERTa	276M	110	8	amh, eng, fra, hau, orm, swa, xho
RemBERT	575M	110	12	amh, eng, fra, hau, ibo, sna, swa, xho, yor
AfriTeVa-base	229M	11	11	amh, run, hau, ibo, orm, pcm, swa, tir, yor
AfroXLMR-base/large	270M/550M	20	17	amh, eng, fra, hau, ibo, orm, pcm, run, sna, swa, xho, yor
AfriMT5-base	580M	20	17	amh, eng, fra, hau, ibo, orm, pcm, run, sna, swa, xho, yor
FlanT5-base	580M	60	5	eng, fra, ibo, swa, yor

Table 6: Languages covered by different multilingual Models and their sizes

Model	size	amh	eng	fra	hau	ibo	lin	lug	orm	pcm	run	sna	som	swa	tir	xho	yor	AVG
Headline																		
MLP	<20K	86.7	72.6	69.8	80.4	77.8	79.4	74.6	81.9	87.5	73.8	84.9	71.4	69.3	80.7	79.1	83.0	78.3
NaiveBayes	<20K	88.8	71.6	70.0	76.6	75.8	74.0	74.6	74.2	82.6	64.3	79.5	61.7	60.6	66.0	72.5	81.4	73.4
XGBoost	<20K	83.6	71.3	67.8	77.4	71.3	76.7	68.7	77.7	80.8	71.3	84.6	63.4	66.4	62.1	69.4	77.5	73.1
AfroXLMR-base	270M	91.8	87.0	92.0	89.2	87.8	89.0	87.4	87.4	97.4	87.8	94.5	85.9	85.0	85.7	93.5	88.6	89.4
AfroXLMR-large	550M	93.0	89.3	91.8	91.0	90.7	91.4	87.7	90.9	98.2	89.3	95.9	**87.1**	86.6	88.5	96.2	90.3	91.1
Headline+Text																		
MLP	<20K	92.0	88.2	84.6	86.7	80.1	84.3	82.2	86.7	93.5	85.9	92.6	71.1	77.9	81.9	94.5	89.3	85.7
NaiveBayes	<20K	91.8	83.7	84.3	85.3	79.8	82.8	84.0	85.6	92.8	79.9	91.5	74.8	76.6	71.4	91.0	84.0	83.7
XGBoost	<20K	90.1	86.0	81.2	84.7	78.6	74.8	83.8	83.2	93.3	79.2	94.3	68.5	74.9	75.2	91.1	85.2	82.8
AfroXLMR-base	270M	94.2	92.2	**92.5**	91.0	90.7	93.0	89.4	92.1	98.2	91.4	95.4	85.2	88.2	86.5	94.7	93.0	91.7
AfroXLMR-large	550M	**94.4**	**93.1**	91.1	**92.2**	**93.4**	**93.7**	**89.9**	**92.1**	**98.8**	**92.7**	**95.4**	86.9	**87.7**	**89.5**	**97.3**	**94.0**	**92.6**

Table 7: **Baseline results on** . We compare different article content types (i.e `headline` and `headline+text`) used to train news topic classification models. Average is over 5 runs. Evaluation is based on weighted F1-score.