EthioLLM: Multilingual Large Language Models for Ethiopian Languages with Task Evaluation

Atnafu Lambebo Tonja[1,2,*,†], **Israel Abebe Azime**[3,*,†], **Tadesse Destaw Belay**[1,†], **Mesay Gemeda Yigezu**[1,†], **Moges Ahmed Mehamed**[4,†], **Abinew Ali Ayele**[5,6,†], **Ebrahim Chekol Jibril**[7,†], **Michael Melese Woldeyohannis**[8,†], **Olga Kolesnikova**[1], **Philipp Slusallek**[3], **Dietrich Klakow**[3], **Shengwu Xiong**[4], **Seid Muhie Yimam**[6,†]

[†] Ethio NLP, [1] Instituto Politécnico Nacional, Mexico, [2] Lelapa AI, [3] Saarland University, Germany, [4] Wuhan University of Technology, China, [5] Bahir Dar University, Ethiopia, [6] Universität Hamburg, Germany, [7] Istanbul Technical University, Turkey, [8] Addis Ababa University, Ethiopia

Abstract

Large language models (LLMs) have gained popularity recently due to their outstanding performance in various downstream Natural Language Processing (NLP) tasks. However, low-resource languages are still lagging behind current state-of-the-art (SOTA) developments in the field of NLP due to insufficient resources to train LLMs. Ethiopian languages exhibit remarkable linguistic diversity, encompassing a wide array of scripts, and are imbued with profound religious and cultural significance. This paper introduces EthioLLM – multilingual large language models for five Ethiopian languages (Amharic, Ge'ez, Afan Oromo, Somali, and Tigrinya) and English, and Ethiobenchmark – a new benchmark dataset for various downstream NLP tasks. We evaluate the performance of these models across five downstream NLP tasks. We open-source our multilingual language models, new benchmark datasets for various downstream tasks, and task-specific fine-tuned language models and discuss the performance of the models. Our dataset and models are available at the EthioNLP HuggingFace repository.

Keywords: EthioLLM, Language models, Ethiopian languages, Low resource languages

1. Introduction

Large language models (LLMs) show a significant advancement in the field of artificial intelligence (AI) (Kasneci et al., 2023). In particular, the introduction of transformer (Vaswani et al., 2017) models has sparked the creation of powerful and effective multilingual pre-trained language models such as GPT (Brown et al., 2020), XLM-RoBERTa (Conneau et al., 2019), mT5 (Xue et al., 2020), and mBERT (Devlin et al., 2018), which have attained cutting-edge performance in a variety of downstream NLP applications (Conneau et al., 2019; Devlin et al., 2018; Alabi et al., 2022; Dossou et al., 2022; Ogueji et al., 2021; Xue et al., 2020). These Pre-trained language models (PLMs) often outperform and may be tailored to a wide range of natural language processing (NLP) tasks (Kassner et al., 2021) including news classification (Adelani et al., 2023), machine translation (Wang et al., 2023a; Lyu et al., 2023), sentiment analysis (Yadav and Vishwakarma, 2020; Alsayat, 2022), named entity recognition (Pan et al., 2017), part-of-speech tagging (Chiche and Yitagesu, 2022; Nguyen and Nguyen, 2020) and fake news detection (Kong et al., 2020; Aggarwal et al., 2020). However, a substantial portion of this development has been focused on high-resource languages. African languages have received especially less attention in this area (Ogueji et al., 2021). Nevertheless, efforts are being made to address the challenges of low-resource languages, with a growing interest in developing Afro-centric models to improve NLP tasks for African languages. AfroLM (Dossou et al., 2022), AfriBERTa (Ogueji et al., 2021), AmRoBERTa (Yimam et al., 2021), and AfroXLMR (Alabi et al., 2022) aimed to bridge this gap by focusing on African languages, capturing their linguistic nuances, and improving language processing for these languages. However, those models have limitations as they did not cover most Ethiopian languages. Ethiopia has over 85 spoken languages, but only a few have been included in developing NLP tasks and tools. Among these low-resource Ethiopian languages, there is a lack of pre-trained models and resources, which limits their ability to contribute to advancing AI research (Tonja et al., 2023; Yimam et al., 2021).

In this paper, we introduce **EthioLLM** – a multilingual pre-trained large language model for five Ethiopian languages with a new benchmark dataset for various downstream NLP tasks. Our contributions are as follows: **(1)** We introduce the first multilingual language models focusing on five Ethiopian languages and English. **(2)**

[*] Equal Contribution.

We introduce **Ethiobenchmark** – new benchmark datasets for various downstream NLP tasks. We compiled new datasets by amalgamating content from multiple sources to achieve broader language coverage for our study. Data sources for creating benchmark data and details about reconstruction are mentioned in Section 4.7. All data releases were executed in consultation with the original authors if the data had not been released previously. **(3)** We evaluate our models on existing datasets of MasakhaNEWS (Adelani et al., 2023), MasakhaNER (Adelani et al., 2021), AfriSenti (Muhammad et al., 2023a,c) and new benchmark datasets. **(4)** We open-source[1] our multilingual language models, its training corpus, the new benchmark datasets, and the new task-specific fine-tuned models. We aim to promote collaboration and streamline research and development for low-resource languages, especially within the context of Ethiopian languages.

2. Related Works

Some research efforts have been dedicated to creating multilingual language models that can be applied to low-resource languages, with the aim of mitigating the inequalities between languages with ample resources and those with limited resources. Among prominent works, Conneau et al. (2019) introduced XLM-R, a multilingual masked language model trained on CommonCrawl[2] data for 100 languages, including three Ethiopian languages. Feng et al. (2020) presented a language-agnostic BERT sentence embedding (LaBSE) model supporting 109 languages, including three Ethiopian languages. Devlin et al. (2019) developed mBERT, a multilingual variant of BERT trained in 104 languages, including four African languages. Xue et al. (2020) presented mT5, a massively multilingual pre-trained text-to-text transformer using a Common Crawl-based dataset covering 101 languages.

Geographic-based multilingual pre-trained language models have also been developed to address language under-representation (Ogueji et al., 2021). Dossou et al. (2022) introduced AfroLM, a multilingual language model that employed a novel self-active learning framework entirely trained from scratch on a dataset encompassing 23 African languages, including two Ethiopian languages. Ogueji et al. (2021) presented AfriBERTa, a language model covering 11 African languages, including four Ethiopian languages. Alabi et al. (2022) built Afro-XLMR by performing multilingual adaptive fine-tuning for 17 most-resourced African languages, including three Ethiopian languages and three other high-resource languages (Arabic, French, and English) widely spoken on the African continent to encourage cross-lingual transfer learning. Pre-training approaches for encoder-only models are extended to encoder-decoder models by introducing AfriTeVa, a pre-trained on 10 African languages from scratch (Jude Ogundepo et al., 2022). How to scale these encoder-decoder models to new languages and domains is investigated by Adelani et al. (2022), a multilingual language model covering 517 African languages.

For Ethiopian-centric languages, Yimam et al. (2021) introduced AmRoBERTa, a RoBERTa model trained using Amharic corpus.

Most of these models cover 11 to 110 languages, and only a few Ethiopian languages (2 to 4 languages) are represented due to the lack of large monolingual corpora on the web. Ethiopian languages lack common benchmark datasets for various downstream NLP tasks to evaluate and use for different NLP-related research.

Our study introduces EthioLLM, a multilingual large language model that accommodates five Ethiopian languages and English. Of these, three languages (Amharic, Ge'ez, and Tigrinya) employ the distinctive Ge'ez writing script, while the remaining two use the Latin script. EthioLLM is developed through the utilization of both XLMR and mT5 architectures in their large, base, and small variants. This multilingual language model is specifically engineered to offer enhanced support for Ethiopian languages by taking into account their diverse scripts and the prevalence of popular languages within the region.

3. EthioLLM

3.1. Training Data and Languages

Even though training LMs requires a large number of datasets (Ogueji et al., 2021), the works by Alabi et al. (2022); Dossou et al. (2022) showed the possibility of training LMs for languages with a limited amount of data. We followed the same strategy to train EthioLLM as the first step towards developing language models for low-resource Ethiopian languages by collecting available monolingual datasets from different sources for five Ethiopian languages. We collected data from local news media (Fana TV[3], EBC[4], BBC news[5] and Walta[6]), the Bible, social media (Facebook and Twitter(X)), and educational textbooks.

[1] https://github.com/EthioNLP/EthioLLM
[2] https://commoncrawl.org/
[3] https://www.fanabc.com/
[4] https://www.ebc.et/
[5] https://www.bbc.com/
[6] https://waltainfo.com/

We focused on training our language models with clean data and conducted further pre-processing and cleaning. We also worked on verifying that downstream task training datasets won't end up in the language model training data. Table 1 shows the selected languages and monolingual dataset used for LMs training.

3.2. Models

3.2.1. Encoder-only models

We trained three multilingual encoder-only models (small, base, and large) with three different parameter configurations. Our encoder-only models used the same parameter setup as AfroXLMR (Alabi et al., 2022) for all the models. We trained two new tokenizers, one with a 70K vocabulary size and the other one with a 250K vocabulary size. We used a tokenizer with a 70K vocabulary size to train EthioLLM-small and the other one for EthioLLM-base and EthioLLM-large. For EthioLLM-base and EthioLLM-large, the vocabulary sizes are adopted from Alabi et al. (2022), but we wanted to experiment with a smaller vocabulary size for the smaller models to reduce the model size in addition to other hyperparameters.

We adopted the language adaptive fine-tuning (LAFT) strategy proposed by Alabi et al. (2022); Chi et al. (2021); Wang et al. (2023b) to train encoder-only models. We tested encoder-only models with different tokenizer sizes starting from 70k-250k for all model variants, but we selected 70K for small and 250K for base and large models based on our initial evaluation in Masakhane-NEWS (Adelani et al., 2023) and MaskahneNER (Adelani et al., 2021) tasks. We also experimented with training from XLMR (Conneau et al., 2019) and AfroXLMR (Alabi et al., 2022) model checkpoints. Based on our initial task evaluation using similar datasets used in tokenizer evaluation, XLMR (Conneau et al., 2019) model outperformed models trained from AfroXLMR (Alabi et al., 2022).

3.2.2. Encoder-Decoder models

To train encoder-decoder models, we adopted the work done by Jude Ogundepo et al. (2022). After sampling from each language, we created 40k vocab size tokenizers for the mt5 small variant model following the Afriteva-small configuration.

We experimented with different model starting points and observed initializing models from Xue et al. (2020) gives better results. Our models are trained for a million steps, and we experimented with different task-specific parameters for different tasks. Our initial assumption that the African-centric models could help if they were used as a starting point did not result in interesting output. We also learned that longer training steps and data cleaning help to get better performance on the small sequence models.

4. Downstream Tasks and Datasets

To evaluate our models in diverse downstream tasks, we selected news classification, machine translation, hate speech detection, named entity recognition, part of speech tagging, sentiment analysis, and question analysis tasks. We also created new benchmark datasets for Ethiopian languages (refer to Section 4.7).

4.1. News Classification

News classification is one of the text classification problems in NLP, in which news articles are categorized into different classes such as Business, Entertainment, Sports, and others (Adelani et al., 2023). To address this problem, datasets in four languages (Amharic, Oromo, Tigrinya, and Somali) were collected from publicly available sources. The MasakhaNEWS dataset (Adelani et al., 2023) includes a total of 4,512 news articles categorized into seven different classes. Additionally, a new benchmark dataset was gathered. Specifically, for the Amharic language, 24,265 news articles were obtained from Azime and Mohammed (2021), and 1,875 news articles were sourced from MasakhaNEWS, resulting in a total of 26,140 articles. Similarly, for the Tigrinya language, 2,397 news articles were obtained from the work of Yohannes and Amagasa (2022) and 1,356 news articles were sourced from MaskhaNEWS, resulting in a total of 3,753 articles.

4.2. Machine Translation (MT)

MT is a widely used NLP application that automatically translates one language to another to facilitate communication between people who speak different languages (Forcada, 2017). Many machine translation works (Biadgligne and Smaïli, 2021; Gezmu et al., 2021; Teshome et al., 2015; Abate et al., 2018; Teshome and Besacier, 2012; Ashengo et al., 2021; Ambaye and Yared, 2000) use different statistical machine translation approaches. The work by Belay et al. (2022) is done by fine-tuning an available multilingual pre-trained model (M2M100 418M) from NLLB Team et al. (2022). Most of the works use traditional approaches and cover two parallel languages, except for the work of Abate et al. (2018), which covers English and five Ethiopian languages (Amharic, Tigrinya, Afan-Oromo, Wolaytta, and Ge'ez). We combined available MT datasets from the works of Biadgligne and Smaïli (2021); Abate et al. (2018); Gezmu et al. (2021); Belay et al. (2022) and

Language	script	Family/branch	# Speakers	Explored	Data Source	# Token (M)	# Sentences
Amharic (amh)	Ge'ez	Afro-Asiatic / Ethio-Semitic	57M	yes	⋆, †, ∗	153,509,645	9,365,829
English (eng)	Latin	Indo-European / Germanic	1268M	yes	⋆, ∗	76,587,128	2,275,996
Afaan Oromo (orm)	Latin	Afro-Asiatic / Cushitic	37M	yes	†, ⋆, ∗	22,448,422	1,040,175
Ge'ez (gez)	Ge'ez	Afro-Asiatic / Ethio-Semitic	UNK	no	†	1,086,578	95,899
Somali (som)	Latin	Afro-Asiatic / Cushitic	22.3M	no	⋆, ∗	17,589,974	558,161
Tigrinya (tir)	Ge'ez	Afro-Asiatic / Ethio-Semitic	9M	yes	†, ⋆, ∗	28,290,680	1,344,586

Table 1: **Language model pre-training corpus**: including language family, number of L1 & L2 speakers (Eberhard et al., 2023), and number of tokens and sentences for each language. Data source symbols are : ∗ = news, ⋆ = Social media, and † = Spiritual (bible).

HornMT online repository[7] into one new benchmark dataset. We present the statistics of the new MT benchmark dataset in Table 2.

4.3. Hate Speech

Detecting hate speech plays a crucial role in content moderation by identifying and screening out harmful or offensive language from online platforms, thereby fostering a safer online environment (Davidson et al., 2017; Mathew et al., 2021). Detecting hate speech in low-resource languages is challenging due to sparse data, linguistic diversity, and complex cultural nuances, making it difficult to develop accurate and contextually aware models (**?**Ousidhoum et al., 2019; **?**).

Based on prior research on Amharic hate speech, we compiled a new benchmark dataset for Amharic comprising approximately 52K data entries sourced from various studies, including 5.3k from Ayele et al. (2022), 30k from the research by Tesfaye and Kakeba (2020), 15k from the investigation conducted by Ayele et al. (2023). Moreover, for the Afaan Oromo language, we utilized a dataset of 12.8k entries from Ababu and Woldeyohannis (2022).

4.4. Sentiment Analysis

Sentiment analysis constitutes a prominent domain in the field of Natural Language Processing, focusing on the automated detection of emotions or opinions expressed in digital content, including social media posts, blog articles, and reviews. This discipline leverages computational techniques to discern and classify the sentiments or viewpoints encoded in textual data sourced from the internet (Agarwal et al., 2011; Taboada et al., 2011; **?**).

The AfriSenti dataset, as meticulously curated by Muhammad et al. (2023a), is designed with a specific focus on African languages. In our research, we harnessed a total of 9,480 Amharic samples and 55,774 samples of Tigrinya. Out of the 55,774 Tigrinya samples, 2,398 were obtained from AfriSenti (Muhammad et al., 2023a), while the remaining 53,374 samples were sourced from the work of Tela et al. (2020). By amalgamating these two datasets, we created EthioNER dataset as a benchmark dataset for Tigrinya, as elaborated in Section 4.7.

4.5. Named Entity Recognition (NER)

Named Entity Recognition (NER) is a fundamental NLP task that involves the identification and classification of predefined information entities within text, which can include proper names, numerical expressions, and temporal references. In our work, we've developed a novel benchmark dataset by amalgamating existing publicly available NER datasets for Amharic, originating from the research of Gambäck and Sikdar (2017) and Jibril and Tantuğ (2023). These two datasets differ in terms of entity classes: Gambäck and Sikdar (2017) is annotated with six classes (PER, LOC, ORG, TIME, TTL, and O-other), while Jibril and Tantuğ (2023) features four classes (PER, LOC, ORG, and O-other). To harmonize the classes, we excluded the TIME and TTL categories from Gambäck and Sikdar (2017). Consequently, the new Amharic NER benchmark dataset comprises 292,367 tokens, categorized into four distinct classes. Furthermore, we have created a separate test dataset for the Ge'ez language to evaluate the zero-shot performance of our language models.

4.6. Part-of-Speech (POS) Tagging

POS tagging stands as one of the sequence labeling tasks within the realm of NLP, where each word (token) in a given sentence is assigned a part of speech tag or another philological class (Keiper et al., 2016). To assess the POS tagging capabilities of our models, we employed a publicly available Amharic POS tagging dataset comprising 33,940 sentences (440,941 words) from the research of Gashaw and Shashirekha (2020) and data from the Habit project[8] for Amharic, Tigrinya, Oromo, and Somali, with the Habit project data yet to be evaluated by researchers. We merged two Amharic datasets to create a novel benchmark

[7] https://github.com/asmelashteka/HornMT

[8] https://habit-project.eu/wiki

dataset. Additionally, we curated a new Ge'ez POS tagging test dataset to evaluate the zero-shot performance of our models. The statistics of this new benchmark dataset are presented in Table 2.

4.7. New Benchmark Dataset

We have amalgamated similar yet independently available datasets into a unified resource, thus creating the **EthioBenchmark** dataset, tailored for a range of downstream NLP tasks in various Ethiopian languages. While previous research efforts have predominantly focused on individual Ethiopian languages, there remains a dearth of comprehensive downstream task datasets spanning multiple languages of Ethiopia, thereby impeding the progress of future research (Tonja et al., 2023). The **EthioBenchmark** dataset has been developed to address this gap and facilitate forthcoming research endeavors in Ethiopian languages.

Henceforth, we will collectively refer to these new benchmark datasets as **EthioBenchmark**, and designate them as *EthioMMT*, *EthioPOS*, *EthioNEWS*, *EthioHate*, *EthioSenti*, and *EthioNER* for machine translation, POS tagging, news classification, hate speech detection, sentiment analysis, and named entity recognition, respectively. By creating this comprehensive benchmark dataset encompassing multiple Ethiopian languages, we aim to provide a foundation for generating new experimental results that can fuel future analyses in this domain.

For Tigrinya, we have amalgamated existing datasets for machine translation, POS tagging, hate speech detection, and sentiment analysis, thus creating an extensive benchmark dataset. We conducted evaluations using our models to establish baseline results. Comprehensive details regarding **EthioBenchmark** dataset, encompassing its sources, revised data splitting ratios, and pertinent statistical information, can be found in Table 2. Additionally, we curated new test dataset for Ge'ez by translating sentences from the Amharic NER and POS tagging test sets, resulting in Ge'ez test datasets comprising 1,374 and 1,022 samples for NER and POS tagging, respectively.

5. Results

We compare the performance of our model against SOTA models that include Ethiopian languages in various downstream tasks using publicly available datasets and newly curated benchmark datasets.

5.1. News Classification

Table 3 summarizes different models evaluated on the MasakhaNEWS (Adelani et al., 2023) dataset, using a weighted F1-score as the performance measure. These models are divided into several categories: general multilingual models, Afro-centric models, our encoder-only models, Afro-centric seq2seq models, and our seq2seq models. When comparing the general multilingual models (XLM-R) to the Afro-centric models (AfroXLMR-large and AfroLM), it is clear that the Afro-centric models consistently outperform the general multilingual models for all four languages. AfroXLMR-large achieves higher scores than AfroLM, indicating superior overall performance.

Our encoder-only models (EthioLLM-small, EthioLLM-base, and EthioLLM-large) demonstrate competitive performance compared to the Afro-centric models. In most languages, EthioLLM-small outperforms AfroLM, taking into account parameter size differences. Additionally, EthioLLM-base showed better performance for Amharic and Afan Oromo languages but showed lower performance for Somali and Tigrinya compared to AfroLM.

Seq2seq models (AfriTeVa-base and AfriMT5-base) performed less than all encoder-only models across all languages. Our seq2seq model (EthioMT5-small) achieves competitive results compared to the Afro-centric seq2seq models. EthioMT5-small outperforms AfriTeVa-base in all languages and outperforms AfriMT5-base in Amharic, Afaan Oromo, and Tigrinya languages.

Overall, our encoder-only models demonstrate competitive performance on the MasakhaNEWS dataset. For seq2seq models, our model outperformed the AfriTeVa-base in all tasks and showed comparative performance with the AfriMT5-base.

Table 4 presents the performance of our models in the new benchmark dataset for Amharic and Tigrinya languages. As we can see from the table, EthioLLM-large outperformed the other models for Amharic. However, it is important to note that having the highest number of parameters, as seen in EthioLLM-large, does not always guarantee the highest accuracy for both languages. Tigrinya EthioLLM-small outperformed others.

5.2. Sentiment Analysis

Table 5 summarizes the evaluation results for general, Afro-centric, and our models. For Amharic and Tigrinya we utilized the AfriSenti Muhammad et al. (2023a) dataset. Additionally in Table 6 for Tigrinya, we conducted evaluations across all our models using the new **EthioSenti** benchmark dataset.

For Amharic, XLMR-large, AfroLM-large, and EthioLLM-large exhibited similar results, achieving an F1 score of approximately 61%, while EthioLLM-base outperformed AfroLM with an F1 score of 58%. Among the sequence-to-sequence

NLP Task	# Source	Section	amh	orm	som	tir	gez
EthioMT	5	4.2	1,286,902	15,484	78,426	78,426	14,720
EthioPOS	2	4.6	22.3M	5.3M	82.4M	2.7M	1,022
EthioNEWS	2	4.1	26,140	1,615	1,463	3,753	–
EthioSenti	2	4.4	–	–	–	55,772	–
EthioNER	2	4.5	296,247	–	–	–	1,374

Table 2: **EthioBenchmark** datasets statistics for each downstream NLP task and language. Under each language category, "-" indicates that we did not compile a new benchmark dataset for that language/task.

Model(#Pram)	amh	orm	som	tir
SOTA encoder-only models (Adelani et al., 2023)				
XLM-R(550M)	93.1	88.4	76.1	62.7
AfroXLMR-l(550M)	94.4	92.1	86.9	89.5
AfroLM (264M)	90.3	83.5	72.0	83.5
Our encoder only models				
EthioLLM-s (139M)	92.55	80.84	64.01	82.22
EthioLLM-b(278M)	91.50	84.53	64.78	76.70
EthioLLM-l(550M)	94.18	90.89	77.92	84.58
SOTA seq2seq models (Adelani et al., 2023)				
AfriTeVa-b(229M)	87.0	82.9	58.0	55.2
AfriMT5-b(580M)	90.2	83.9	77.8	80.8
Our seq2seq model				
EthioMT5-s (85M)	90.04	85.96	72.44	82.23

Table 3: **Baseline results on MasakhaNEWS** . Evaluation is based on a weighted F1-score. We compared our models with general and Afro-centric models. s = small, b = base, and l = large.

Model(#Pram)	amh	tir
Our encoder only models		
EthioLLM-s (139M))	86.53	83.84
EthioLLM-b(278M)	87.28	79.51
EthioLLM-l (550M)	88.94	83.31

Table 4: **Baseline results on EthioNEWS dataset**. Evaluation is based on a weighted F1-score. We only evaluated with our multilingual models. s = small, b = base, and l = large.

Model(#Pram)	amh	tir*
SOTA encoder-only models (Muhammad et al., 2023a)		
XLMR-l (550M)	61.8	–
AfroXLMR-l (550M)	61.6	62.6
Our encoder only models		
EthioLLM-s (139M)	56.38	38.05
EthioLLM-b (278M)	58.12	35.74
EthioLLM-l (550M)	61.21	41.52
Afro-centric seq2seq LM		
AfriMT5-b (580M)	49.4	36.9
Our seq2seq model		
EthioMT5-s (85M)	51.6	29.5

Table 5: **Sentiment analysis baseline results on AfriSenti corpus**. Evaluation is based on a weighted F1-score. s = small, b = base, and l = large. *= zero-shot performance using Amharic as source language

Model(#Pram)	tir
Our encoder only models	
EthioLLM-s (139M)	91.09
EthioLLM-b (278M)	89.24
EthioLLM-l (550M)	88.86

Table 6: **Sentiment analysis baseline results on EthioSenti corpus**. Evaluation is based on a weighted F1-score. s = small, b = base, and l = large.

models, Amharic results show the EthioMT5-small model outperformed the AfriMT5-base, achieving an F1 score of 51.6%. For Tigrinya zero shot task AfriMT5-base outperformed EthioMT5-small with an F1 score of 36.9%.

For **EthioSenti** Tigrinya results, EthioLM-small outperformed all encoder-only models, attaining an F1 score of 91%, while EthioLLM-base and large demonstrated comparable results.

5.3. Hate speech

To assess our model's performance, we evaluated and compared against the SOTA model. We employed two language datasets provided by Ayele et al. (2022), Tesfaye and Kakeba (2020), Ayele et al. (2023), Abebaw et al. (2022) and Ababu and Woldeyohannis (2022) for our evaluation. We tested Afro-centric models such as AfroXLMR-large, Afro LM, and the general multilingual model XLM-R for the Amharic and Afan Oromo languages.

Table 7 summarizes the hate speech results for Amharic and Afaan Oromo. As shown in the table, EthioLLM-large outperformed other models for both languages with an F1-score of 73% and

87%, respectively, whereas EthioLLM-small and base showed comparable results.

Model(#Pram)	amh	orm
General multilingual models		
XLM-R (550M)	31.06	82.89
Afro-centric models		
AfroXLMR-l (550M)	67.73	83.87
AfroLM (264M)	61.69	81.40
Our encoder only models		
EthioLLM-s (139M)	60.90	84.68
EthioLLM-b(278M)	64.81	83.24
EthioLLM-l (550M)	**73.54**	**87.28**

Table 7: **Baseline results on EthioHate dataset**. Evaluation is based on a weighted F1-score. We compared our multilingual models with other models. s = small, b = base, and l = large.

5.4. Named Entity Recognition (NER)

We evaluated our models in the NER task using the MasakhaNER dataset (Adelani et al., 2021), which is a publicly available, high-quality dataset for NER in ten African languages, including only Amharic from Ethiopian languages. For Ge'ez language, we prepared a new NER test set. Table 8 shows the performance of our models in the Amharic NER task with SOTA models comparison. As we can see from the result, EthiLM-large outperformed all other models with an F1-score of 79%, while EthioLLM-small and base showed comparable results.

Model(#Pram)	amh
SOTA models	
XLM-R (Alabi et al., 2022) (550M)	76.18
AfroXLMR-l (550M) (Alabi et al., 2022)	78.0
AfroLM (264M)(Dossou et al., 2022)	73.84
Our encoder only models	
EthioLLM-s (139M))	68.99
EthioLLM-b(278M)	69.9
EthioLLM-l (550M)	**79.42**

Table 8: **Baseline results on our MaskhaneNER dataset**. Evaluation is based on a weighted F1-score. We compared our multilingual models with others. s = small, b = base, and l = large.

In Table 9, we evaluated our models in EthioNER datasets. We used Amharic as the source language to evaluate Ge'ez's zero-shot performance. For Amharic, EthioLLM-large outperformed base and small models with an F1-score of 78%, while EthioLLM-small and base have shown comparable results. All models have shown a promising result for Ge'ez zero-shot task, while EthioLLM-large outperformed the rest with an F1-score of 74%.

Model(#Pram)	amh	gez*
Our encoder only models		
EthioLLM-s (139M))	71.83	73.67
EthioLLM-b(278M)	73.06	73.79
EthioLLM-l (550M)	**78.02**	**74.84**

Table 9: **Baseline results on EthioNER dataset**. Evaluation is based on a weighted F1-score. * shows the zero-shot performance using Amharic as source language. s = small, b = base, and l = large.

5.5. Part of Speech Tag (POS)

Table 10 shows the results of our models. We evaluated the model on our benchmark dataset. Our EthioLLM-large model archives 90.36%, 99.98%, and 79.67% on amh, orm, and tir tasks, respectively.

Model(#Pram)	amh	orm	tir	gez*
Our encoder only models				
EthioLLM-s (139M)	86.86	99.95	78.33	35.84
EthioLLM-b(278M)	85.09	99.95	71.93	34.84
EthioLLM-l (550M)	90.36	99.98	79.67	37.63

Table 10: **Baseline results on EthiPOS tag dataset**. Evaluation is based on a weighted F1-score. * shows the zero-shot performance using Amharic as a source language. s = small, b = base, and l = large.

5.6. Machine Translation

Source	amh	eng	orm	som	tir	gez
EthioMT5-S 85M						
amh	-	17.0	0.84	0.88	0.84	5.30
eng	5.45	-	1.30	2.60	0.70	0.70
M2M100 418M						
amh	-	37.60	∗	2.90	2.86	∗
eng	13.70	-	∗	9.60	9.60	∗

Table 11: **Baseline sacreBleu results of EthioMT** on Flores-200 (NLLB Team et al., 2022) for languages except for Ge'ez. ∗ = languages not covered in M2M100. Results of M2M100 are from (NLLB Team et al., 2022) paper for *eng-xx* and *xx-eng* model, and we fine-tuned for the others.

Table 11 presents baseline results for EthioMMT dataset. We utilized the Flores-200 dataset (NLLB Team et al., 2022) for evaluation across all language pairs except for Ge'ez. For Ge'ez, we created our test split and subsequently reported the results on this custom test split. To compare the performance of our models, we used NLLB Team et al. (2022) results for the languages mentioned in the paper and finetuned for the rest.

NLLB Team et al. (2022) is the state-of-the-art in MT, but we showed the closeness we can achieve to the model with a smaller MT5 model. This model can be a good experimental platform for MT tasks with fewer trainable parameters. This lower score shown in machine translation by our MT5 models is also observed in models like Afriteva and AfriMT5. Our model also covers two previously uncovered languages in NLLB Team et al. (2022), which we found beneficial in Ethiopian languages.

6. Discussion

We compared our models with current SOTA models that include Ethiopian languages. Our models show comparable results with SOTA models. From our models, the EthioLLM-large model shows comparable results in news classification and sentiment analysis tasks and outperforms the existing SOTA model in named entity recognition and hate speech tasks. EthioLLM-small with a parameter size of 139M showed comparable results with AfroLM (Dossou et al., 2022) and outperformed XLM-R (Conneau et al., 2019) in sentiment analysis and hate speech detection.

We showed that our EthioMT5-small model performs better or is on par with the other base models on the classification tasks. This can be attributed to the longer training and data cleaning we did to train our language model. The same explanation doesn't work for tasks like machine translation, where our model fails short compared to m2m100 models. This is understandable given the smaller size of the model, but for machine translation tasks, the best approach would be to fine-tune m2m100 models directly.

Our models exhibit promising results in zero-shot evaluation for Ge'ez, suggesting that they may also perform well for low-resource languages incorporated during language pre-training. We released[9] the EthioLLM models, EthioBenchmark dataset, and our top-performing task-specific models as open-source resources, aiming to encourage further research in Ethiopian languages.

[9] https://github.com/EthioNLP/EthioLLM

7. Conclusion and Future Work

In this paper, we presented EthioLLM, the first attempt to train multilingual language models for five Ethiopian languages and English. We tested our EthiLLM with the available benchmark datasets like MasakhaNEWS, MasakhaNER, and AfriSenti. We also created **EthioBenchmark** dataset for various downstream tasks for five Ethiopian languages by combining the available corpus. Additionally, we created a new task for Ge'ez. We included a minimum of two downstream tasks for each language in the language models. Our models have outperformed some and demonstrated comparable performance with respect to the current SOTA models in different cases.

As shown in the results section, our sequence-to-sequence models were tested with only a machine translation sequence-to-sequence task. In addition to the tasks we tried, we plan to train both the base and large versions of these models and introduce several other sequence-to-sequence tasks apart from machine translation.

Limitations

In this work, we presented models and downstream task evaluation for five Ethiopian languages with a publicly available evaluation dataset and created a new benchmark dataset as one of the contributions for the languages left behind by current technology. Despite our efforts, a significant gap exists in the downstream task creation, spanning multiple languages. The primary challenge lies in developing a diverse set of tasks that encompass all languages within the language model. Another challenge we encounter is acquiring a sufficient amount of data for language model training. Due to the scarcity of corpora. There are more than 85 languages in Ethiopia but we only covered 5 of them in this study because of the scarcity of corpus.

8. Bibliographical References

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring amharic hate speech data collection and classification approaches. In *Proceedings of the 14th International Conference on RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING (RANLP 2023*, pages 59–59.

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022. The 5Js in

Ethiopia: Amharic hate speech data annotation using Toloka Crowdsourcing Platform. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 114–120, Bahir Dar, Ethiopia.

Abrhalei Tela, Abraham Woubie, and Ville Hautamaki. 2020. Transferring monolingual model to low-resource language: The case of tigrinya.

Ahmed Alsayat. 2022. Improving sentiment analysis for social media applications using an ensemble deep learning language model. *Arabian Journal for Science and Engineering*, 47(2):2499–2511.

Akshay Aggarwal, Aniruddha Chauhan, Deepika Kumar, Sharad Verma, and Mamta Mittal. 2020. Classification of fake news by fine-tuning deep bidirectional transformers based language model. *EAI Endorsed Transactions on Scalable Information Systems*, 7(27):e10–e10.

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):1–25.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021. Extended parallel corpus for Amharic-English machine translation. *arXiv preprint arXiv:2104.03543*.

Angelina McMillan-Major, Amandalynne Paullada, and Yacine Jernite. 2022. An interactive exploratory tool for the task of hate speech detection. In *Proceedings of the Second Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 11–20, Seattle, Washington. Association for Computational Linguistics.

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca J Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)*, pages 30–38.

Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023. Natural language processing in ethiopian languages: Current state, challenges, and opportunities. *arXiv preprint arXiv:2303.14406*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 14867–14875, Palo Alto, CA, USA. Association for the Advancement of Artificial Intelligence.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Named entity recognition for amharic using deep learning. In *2017 IST-Africa Week Conference (IST-Africa)*, pages 1–8. IEEE.

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*.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

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. 2022. 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, et al. 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 Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, et al. 2023. Masakhanews: News topic classification for african languages. *arXiv preprint arXiv:2304.09972*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. Ethnologue: Languages of the World. Twenty-third edition. Dallas, Texas: SIL International. Url: http://www.ethnologue.com.

Ebrahim Chekol Jibril and A Cüneyd Tantuğ. 2023. Anec: An amharic named entity corpus and transformer based recognizer. *IEEE Access*, 11:15799–15815.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. A scheme for news article classification in a low-resource language. In *International Conference on Information Integration and Web*, pages 519–530. Springer.

Ibrahim Gashaw and H L Shashirekha. 2020. Machine learning approaches for amharic parts-of-speech tagging. *arXiv preprint arXiv:2001.03324*.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Serengeti: Massively multilingual language models for africa. *arXiv preprint arXiv:2212.10785*.

Israel Abebe Azime and Nebil Mohammed. 2021. An amharic news text classification dataset. In *2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, Minneapolis, Minnesota. 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.

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.

Lena Keiper, Andrea Horbach, and Stefan Thater. 2016. Improving POS tagging of German learner language in a reading comprehension scenario. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 198–205, Portorož, Slovenia. European Language Resources Association (ELRA).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Marta R. Costa-jussà NLLB Team, 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. 2022. No language left behind: Scaling human-centered machine translation.

Mikel L Forcada. 2017. Making sense of neural machine translation. *Translation spaces*, 6(2):291–309.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023b. Nlnde at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. *arXiv preprint arXiv:2305.00090*.

Mulu Gebreegziabher Teshome and Laurent Besacier. 2012. Preliminary experiments on English-Amharic statistical machine translation. In *Spoken Language Technologies for Under-Resourced Languages*, pages 36–41, Cape Town, South Africa.

Mulu Gebreegziabher Teshome, Laurent Besacier, Girma Taye, and Dereje Teferi. 2015. Phoneme-based English-Amharic statistical machine translation. In *AFRICON 2015*, pages 1–5, Addis Ababa, Ethiopia. IEEE.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.

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.

Seid Muhie Yimam, Abinew Ali Ayele, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. 2021. Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11):275.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, et al. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *arXiv e-prints*, pages arXiv–2201.

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, et al. 2023a. Afrisenti: A twitter sentiment analysis benchmark for african languages. *arXiv preprint arXiv:2302.08956*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M Mohammad, and Meriem Beloucif. 2023b. Semeval-2023 task 12: Sentiment analysis for african languages (afrisenti-semeval). *arXiv preprint arXiv:2304.06845*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023c. SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Sheng How Kong, Li Mei Tan, Keng Hoon Gan, and Nur Hana Samsudin. 2020. Fake news detection using deep learning. In *2020 IEEE 10th symposium on computer applications & industrial electronics (ISCAIE)*, pages 102–107. IEEE.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-lingual English-Ethiopian languages statistical machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3102–3111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Surafel Getachew Tesfaye and Kula Kakeba. 2020. Automated amharic hate speech posts and comments detection model using recurrent neural network. *Preprint*. Version 1.

Tadesse Ambaye and Mekuria Yared. 2000. English to Amharic machine translation using statistical machine translation. *Master's thesis*.

Tadesse Destaw, Seid Muhie Yimam, Abinew Ayele, and Chris Biemann. 2022. Question answering classification for Amharic social media community based questions. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*.

Tadesse Destaw Belay, Atnafu Lambebo Tonja, Olga Kolesnikova, Seid Muhie Yimam, Abinew Ali Ayele, Silesh Bogale Haile, Grigori Sidorov, and Alexander Gelbukh. 2022. The effect of normalization for bi-directional amharic-english neural machine translation. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 84–89. IEEE.

Teshome Mulugeta Ababu and Michael Melese Woldeyohannis. 2022. Afaan Oromo hate speech detection and classification on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6612–6619, Marseille, France. European Language Resources Association.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, volume 11, pages 512–515, Montréal, QC, Canada. Association for Computational Linguistics.

Tilahun Abedissa, Ricardo Usbeck, and Yaregal Assabie. 2023. Amqa: Amharic question answering dataset. *arXiv preprint arXiv:2303.03290*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Yeabsira Asefa Ashengo, Rosa Tsegaye Aga, and Surafel Lemma Abebe. 2021. Context based machine translation with recurrent neural network for English–Amharic translation. *Machine Translation*, 35(1):19–36.

Yohanens Biadgligne and Kamel Smaïli. 2021. Parallel corpora preparation for English-Amharic machine translation. In *International Work-Conference on Artificial Neural Networks*, pages 443–455. Springer, Cham.

Zeleke Abebaw, Andreas Rauber, and Solomon Atnafu. 2022. Design and implementation of a multichannel convolutional neural network for hate speech detection in social networks. *Revue d'Intelligence Artificielle*, 36(2):175–183.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.