Amharic News Topic Classification: Dataset and Transformer-Based Model Benchmarks

Dagnachew Mekonnen Marilign
HiLCoE
School of Computer Science and Technology
Ethiopia
dagnachewmm@hilcoeschool.com

Eyob Nigussie Alemu
Addis Ababa University
Ethiopia
eyob.alemu@aau.edu.et

Abstract

News classification is a downstream task in Natural Language Processing (NLP) that involves the automatic categorization of news articles into predefined thematic categories. Although notable advancements have been made for high-resource languages, low-resource languages such as Amharic continue to encounter significant challenges, largely due to the scarcity of annotated corpora and the limited availability of language-specific, state-of-the-art model adaptations. To address these limitations, this study significantly expands an existing Amharic news dataset, increasing its size from 50,000 to 144,000 articles, thus enriching the linguistic and topical diversity available for the model training and evaluation. Using this expanded dataset, we systematically evaluated the performance of five transformer-based models: mBERT, XLM-R, DistilBERT, AfriBERTa, and AfroXLM in the context of Amharic news classification. Among these, AfriBERTa and XLM-R achieved the highest F1-scores of 90.25% and 90.11%, respectively, establishing a new performance baseline for the task. These findings underscore the efficacy of advanced multilingual and Africa-centric transformer architectures when applied to under-resourced languages, and further emphasize the critical importance of large-scale, high-quality datasets in enabling robust model generalization. This study offers a robust empirical foundation for advancing NLP research in low-resource languages, which remain underrepresented in current NLP resources and methodologies.

1 Introduction

Amharic, Ethiopia's official working language, is spoken by millions and is the second most widely used Semitic language after Arabic.As internet penetration expands, there is a significant increase in the consumption of online content, particularly in digital news. The shift from traditional to digital media has amplified the volume of unstructured text data, presenting opportunities for advanced NLP applications, such as text classification(TC).

Topic classification (TC) plays a vital role in organizing unstructured textual data, enabling automated news categorization, and supporting recommendation systems. Although transformer-based models and large language models (LLMs) have significantly advanced TC in high-resource languages, Amharic remains notably underrepresented in this domain. This gap is largely due to the scarcity of large-scale, labeled datasets and the limited application of modern NLP techniques specifically tailored to the linguistic and contextual characteristics of Amharic.

Earlier work, such as (Azime and Mohammed, 2021), introduced a foundational Amharic news dataset, but regular updates are required. Other studies relied on smaller datasets and lacked reproducibility due to the inaccessibility of resources (Kelemework, 2013; Endalie and Haile, 2021), highlighting the need for more robust approaches.

This paper presents an in-depth evaluation of transformer-based models for the classification of Amharic news. The contributions of this study are twofold:

- Expansion of the Amharic News Dataset: We significantly enhance the existing Amharic news dataset by expanding it from 50,000 to 144,000 articles, nearly tripling its size. This allows for more robust model training and evaluation, ensuring better performance and generalizability in real-world applications.

- Evaluation of Transformer Models and Benchmarking: Using the expanded dataset, we fine-tune and evaluate five popular transformer-based models: mBERT, XLM-R, DistilBERT, AfriBERTa, and AfroXLM. These models are trained to classify news articles into six categories, which are Local News, International

Proceedings of the 9th Widening NLP Workshop, pages 130–135

News, Politics, Sports, Business, and Entertainment. Through this evaluation, we establish benchmark results by conducting a comparative performance analysis of these models.

The findings of this study offer valuable insight into the application of state-of-the-art transformer models for low-resource languages, such as Amharic. The results support the development of more accurate and efficient news classification systems, with potential applications in content aggregation, personalized recommendations, and automated news filtering.

2 Related Work

The availability of organized and machine-readable data in high-resource languages has privileged them in NLP research, while low-resource languages in Africa and Asia remain underrepresented.

MasakhaNEWS (Adelani et al., 2023) explored multilingual transformers, such as mBERT and XLM-R, to classify news topics in 16 African languages, including Amharic. Using transfer learning, the study achieved promising results but faced challenges in generalization due to the limited availability of annotated data and linguistic complexity. Similarly, MasakhaNER introduced NER datasets for 10 languages, which later expanded to 20 (Adelani et al., 2021, 2022), but excluding major Ethiopian languages such as Afaan Oromo, Tigrigna, and Somali, revealing a research gap.

AfriSenti (Muhammad et al., 2023) advanced sentiment analysis in 14 African languages, including Amharic, Afaan Oromo, and Tigrigna. Although models such as AfriBERTa perform well, they still struggle with language-specific issues such as imbalance and structural variation. AfriBERTa (Ogueji et al., 2021), trained in less than 1GB of text in 11 African languages, outperformed mBERT and XLM-R in some tasks, but could not fully address concerns about data quality or diversity.

Other efforts include monolingual models such as PuoBERTa for Setswana and transformer-based TC work on Ewe, Swahili, and Kinyarwanda, demonstrating that language-specific models often outperform general-purpose models.

Among Semitic languages, Arabic leads the development of NLP research, with dedicated models such as AraBERT, MARBERT, and ArabicBERT (Abdul-Mageed et al., 2021; Alammary, 2022). Hebrew has also benefited from monolingual models such as AlephBERT and HebBERT (Seker et al., 2021; Chriqui and Yahav, 2022), which surpass multilingual baselines.

Ethiopia has more than 85 languages; however, NLP research remains limited to Amharic, the most studied language. A review by (Tonja et al., 2023) emphasized the fragmented state of NLP for Ethiopian languages, with a lack of datasets, benchmarks, and transformer-based models in particular. Even major projects such as MasakhaNEWS and AfriSenti rarely go beyond Amharic, neglecting other widely spoken languages such as Afaan Oromo and Tigrigna.

The reviewed studies have several key limitations. Although Amharic is Ethiopia's most widely spoken and official language, there has been limited progress in developing robust NLP resources. Many studies rely heavily on multilingual models that often struggle to capture the unique linguistic and contextual features of Amharic. In addition, there is a lack of language-specific standardized benchmarks for text classification in Amharic and other Ethiopian languages.

This study focuses on classifying Amharic news using five transformer-based models: mBERT, XLM-R, DistilBERT, AfriBERTa, and AfroXLM. Its main contributions include fine-tuning and evaluating these models on a significantly expanded and original Amharic news dataset, establishing benchmark results through comparative performance analysis , and providing enriched data and insights that support future NLP research in Amharic and can also encourage similar studies in other underrepresented languages.

3 Experimentation

The Expanded Amharic News Dataset developed for this study comprises 144,201 articles published between 2011 and late 2024. It integrates 92,792 newly collected articles with an existing, manually filtered set of 51,409 articles from the original dataset (reduced from 51,471 after excluding entries with incomplete or ambiguous content).

Articles were collected from 12 major Amharic news outlets, using BeautifulSoup. Data scraping followed ethical and responsible practices, involving only publicly accessible content, excluding paywalled material, and adhering to the terms of service of each source.

Category	Original Dataset	Expanded Dataset	Relative Increase (%)
Local	20,654	62,994	+205.1%
Entertainment	632	1,138	+80.1%
Sport	10,397	25,228	+142.6%
Business	3,887	16,671	+328.8%
International	6,530	13,345	+104.4%
Politics	9,309	24,825	+166.7%
Total	**51,409**	**144,201**	**+180.4%**

Table 1: Class Distribution Statistics for Original and Expanded Amharic News Datasets

The dataset was constructed through a semi-automatic pipeline, preserving editorial category labels provided by the sources (e.g., Local News, Entertainment, Sports, Business, International, Politics). A manual quality assurance step was applied to remove records with missing or invalid data. Each entry includes a headline, article body, category label, and source URL link. When available, the publication date and view count of the news article are also included; missing metadata is marked as NA. Records lacking the headline or body of the article were excluded to ensure data quality. Compared to previous Amharic news datasets (Azime and Mohammed, 2021), this expanded corpus significantly increases scale and metadata richness, offering improved support for news topic classification and other low-resource NLP tasks.

The Preprocessing steps included text cleaning, metadata curation, and stratified partitioning into training (70%), validation (10%), and test (20%) sets. The dataset not only advances Amharic text classification but also enables exploration of imbalance-handling methods and finer-grained categorization in future work. Tokenization used a pre-trained tokenizer with padding/truncation to 512 tokens, and the category labels were encoded using Scikit-learn's LabelEncoder.

This study fine-tuned and evaluated five transformer-based models: mBERT, DistilBERT, XLM-R, AfroXLM, and AfriBERTafor Amharic news classification. BERT and its variants, such as mBERT and DistilBERT, leverage masked language modeling and prediction of the next sentence to generate deep contextual representations (Devlin et al., 2019; Pires et al., 2019; Sanh, 2019). XLM-R, which is trained with a large multilingual corpus, offers strong cross-lingual performance (Conneau et al., 2020). AfroXLM and AfriBERTa, trained with African languages, improve generalization for underrepresented and morphologically complex languages, such as Amharic (Alabi et al., 2022; Ogueji et al., 2021).

Fine-tuning was performed using Hugging Face's Trainer API with batch sizes of 16 and 32, gradient accumulation steps of 4, a learning rate of 5×10^{-5}, weight decay of 0.1, and mixed-precision training (fp16). Models were trained for five epochs, evaluated using F1 score, and implemented in a Linux Kaggle environment (Tesla P100 GPU, Python 3.10.14). Pre-trained models were tokenized using AutoTokenizer and padded via DataCollatorWithPadding. The input text combined cleaned headlines and content, tokenized with truncation, and the maximum length per model. The cross-entropy loss and AdamW optimizer were used. The evaluation metrics (accuracy, precision, recall, and F1-score) were logged using the W&B. The dataset was normalized, label-encoded, and split using stratified sampling (70% train, 10% validation, 20% test) across six news categories. We used a confusion matrix to assess the model's generalization.

4 Results and Discussion

To assess the effect of batch size on model performance, each transformer-based architecture was fine-tuned using batch sizes of 16 and 32. Although the performance differences between the configurations were relatively modest, the final reported results corresponded to the setting that yielded the highest macro F1-score on the test set. Macro F1 was selected over weighted F1 as the primary evaluation metric to provide a balanced assessment across all classes, particularly considering the inherent class imbalance in the dataset. This choice ensured that the evaluation did not disproportionately favor majority classes, thereby supporting a

Model	Dataset	F1 (Macro)	F1 (Weighted)	Accuracy	Precision	Recall
mBERT	Expanded	0.57	0.6337	0.6441	0.6334	0.6441
	Original	0.50	0.5874	0.6174	0.6396	0.6174
XLM-R	Expanded	0.88	0.9011	0.9013	0.9013	0.9013
	Original	0.85	0.880	0.8790	0.8811	0.0.879
DistilBERT	Expanded	0.57	0.6350	0.6447	0.6349	0.6447
	Original	0.60	0.6720	0.6745	0.6719	0.6745
AfriBERTa	Expanded	0.89	0.9025	0.9029	0.9025	0.9029
	Original	0.87	0.8783	0.8785	0.8781	0.8785
AfroXLMR	Expanded	0.88	0.8965	0.8961	0.8965	0.2950
	Original	0.84	0.8704	0.8705	0.8707	0.8705

Table 2: Comparison of Model Performance on Original and Expanded Amharic News Datasets

Model	Dataset	Local	Entertainment	Sport	Business	International	Politics
mBERT	Expanded	0.71	0.51	0.75	0.43	0.45	0.56
	Original	0.68	0.45	0.71	0.34	0.19	0.53
XLM-R	Expanded	0.91	0.8	0.99	0.82	0.93	0.84
	Original	0.88	0.78	0.96	0.71	0.89	0.81
DistilBERT	Expanded	0.71	0.51	0.76	0.44	0.45	0.56
	Original	0.75	0.49	0.71	0.37	0.40	0.53
AfriBERTa	Expanded	0.91	0.83	0.99	0.81	0.92	0.85
	Original	0.89	0.80	0.98	0.72	0.90	0.81
AfroXLMR	Expanded	0.9	0.82	0.99	0.8	0.92	0.83
	Original	0.86	0.79	0.95	0.71	0.87	0.80

Table 3: Per-Class F1 Scores on Original and Expanded Datasets

more equitable comparison of model performance across both frequent and underrepresented categories.

The evaluation results show that the expanded Amharic news dataset significantly improved model performance across the board, especially for models with larger parameter capacities. As shown in Table 2, AfriBERTa and XLM-R achieved the highest scores across macro F1, weighted F1, and accuracy, highlighting their strong generalization when they were trained on a larger and more diverse dataset. For instance, AfriBERTa's macro F1 improved from 0.87 on the original dataset to 0.89 on the expanded version, whereas XLM-R increased from 0.85 to 0.88.

Per-class F1 analysis Table 3 further illustrates the benefits of dataset expansion. Notable gains were observed in previously underrepresented categories such as Business and International News. For example, mBERT's F1-score for International improved from 0.19 to 0.45, and Business from 0.34 to 0.43, indicating a meaningful reduction in class imbalance and better coverage of low-resource categories. Although categories such as entertainment remain challenging due to limited examples and semantic overlap, the overall classification balance was markedly improved.

AfroXLMR also maintained strong and stable performance across both datasets, while lightweight models such as mBERT and DistilBERT, despite lower overall accuracy, still benefited from the data expansion.

The expanded dataset substantially enhanced classification performance by providing more representative training samples, particularly benefiting large-scale transformer models. These results reaffirm the importance of domain-specific and linguistically aligned data for advancing NLP in low-

resource languages such as Amharic.

5 Conclusion and Recommendation

This study assessed the performance of transformer-based models for Amharic news topic classification using an expanded dataset comprising over 144,000 articles. The results demonstrated that AfriBERTa and XLM-R consistently delivered superior performance in both accuracy and F1 scores, underscoring the effectiveness of language-specific or regionally pretrained models for low-resource languages. AfroXLMR also achieved strong results, reinforcing the value of pretraining strategies that incorporate African linguistic features. In contrast, general-purpose models such as mBERT and DistilBERT struggled to capture the linguistic complexity of Amharic, particularly in terms of morphology and syntax.

Building on the findings of this study, future work should consider adopting more sophisticated classification strategies, such as multi-label and hierarchical models, to better capture topic overlap commonly found in news content. Incorporating cross-lingual transfer learning and few-shot learning techniques could also enhance model adaptability across other under-resourced African languages. Given the reliance of the datasets on editorially assigned labels, future research should investigate possible labeling inconsistencies or bias, which can impact classification performance. Introducing human-in-the-loop validation can further improve data quality and support the development of a gold-standard benchmark subset. Additionally, the rich metadata structure of the dataset opens opportunities for broader NLP applications, including news summarization, headline generation, and temporal topic modeling. To enable deployment in real-world and low-resource settings, future efforts should focus on compressing large-scale models, such as developing distilled versions of AfriBERTa without significantly compromising performance. Finally, ongoing attention to the linguistic characteristics of Amharic, including its complex morphology, orthographic variations, and context sensitivity, will be essential to build more robust and generalizable language technologies.

Limitations

Despite these promising results, this study has some limitations. The scarcity of high-quality labeled Amharic news data and the use of multilingual models not tailored for Amharic reduced performance. Limited computational resources also constrain model tuning. Additionally, reliance on static data affects the generalization the model.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. 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 7088–7105, Online. 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, and 42 others. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, and 46 others. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

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, and 26 others. 2022. 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 Oluwadara Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Multilingual language model adaptive fine-tuning: A study

on african languages. In *3rd Workshop on African Natural Language Processing*.

Ali Saleh Alammary. 2022. Bert models for arabic text classification: a systematic review. *Applied Sciences*, 12(11):5720.

Israel Abebe Azime and Nebil Mohammed. 2021. An amharic news text classification dataset. *arXiv preprint arXiv:2103.05639*.

Avihay Chriqui and Inbal Yahav. 2022. Hebert and hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *INFORMS Journal on Data Science*, 1(1):81–95.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Demeke Endalie and Getamesay Haile. 2021. Automated amharic news categorization using deep learning models. *Computational Intelligence and Neuroscience*, 2021(1):3774607.

Worku Kelemework. 2013. Automatic amharic text news classification: Aneural networks approach. *Ethiopian Journal of Science and Technology*, 6(2):127–137.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alipio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

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.

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

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.

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. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 126–139, Dubrovnik, Croatia. Association for Computational Linguistics.