School of Information and Technology Engineering
Addis Ababa Institute of Technology
Addis Ababa University

News Classifier

Software Stream

PREPARED BY:

Fikernew Birhanu	UGR/9932/13
Firaol Ibrahim	UGR/0841/13
Fikremariam Anteneh	UGR/9301/13
Firaol Bogale	UGR/1469/13
Fedasa Bote	UGR/6761/23

Date: Jan 26, 2025
Submitted to: Dr. Fantahun Bogale

Amharic News Classification

This project focuses on classifying Amharic news articles into six distinct categories:

- ሀገር አቀፍ ዜና (National News)
- መዝናኛ (Entertainment)
- ስፖርት (Sports)
- ቢዝነስ (Business)
- አለም አቀፍ ዜና (World News)
- ፖለቲካ (Politics)

Resources

- Interactive Website for testing the models
- The News Dataset
- Finetuned Classifier Model on Hugging Face
- GitHub Repository

Endpoints

The website can be used for testing, but if specific endpoints are needed, you can make POST requests with `Content-Type: application/text` by adding the news text. Use the following endpoints:

- https://fikreanteneh-amharicnewsclassifier.hf.space/predict/xlm_roberta_finetuned
- https://fikreanteneh-amharicnewsclassifier.hf.space/predict/count_vectorizer
- https://fikreanteneh-amharicnewsclassifier.hf.space/predict/tfidf_vectorizer

Note: When testing the website the first load might take upto 50 second. Since we are on a free tier the instance is killed after 30 minute of inactivity. Thank you for understanding.

Summary

This Amharic News Classifier draws inspiration from the research paper, "An Amharic News Text Classification Dataset" by Israel Abebe Azime and Nebil Mohammed. The paper emphasizes the challenges of performing text classification in low-resource languages like Amharic due to the lack of labeled data. It introduces a dataset of over 50,000 news articles categorized into six classes, offering a baseline for classification tasks using simple models such as Naive Bayes with Count Vectorizer and TF-IDF.

We used the same dataset for our classification task and fine-tuned XLM RoBERTa, a Transformer model, for Amharic news classification. XLM RoBERTa is the only model we found that is trained on the Amharic language and performs well. The

model was pretrained on 100 different languages, including Amharic, and has over 200 million parameters. We deployed the fine-tunded model on Hugging Face and can be found here.

Unlike traditional models like Naive Bayes, which rely on simple text features such as word counts or TF-IDF, XLM RoBERTa leverages deep learning and contextual embeddings to deliver higher performance on sequence classification tasks.

We built an API and deployed it on Hugging Face Spaces. The api can be integrated into any application. We also created a small website for users to test the models without writing any code, available here.

The deployed models are:

- XLM RoBERTa (Fine-tuned for Classification on Amharic News)
- Naive Bayes with Count Vectorizer
- Naive Bayes with TF-IDF

Dataset

Dataset is Available at: Amharic News Dataset on Hugging Face

The dataset consists of 50,706 news articles in the Amharic language, categorized into six classes:

- ሀገር አቀፍ ዜና (National News): 20,564 articles
- መዝናኛ (Entertainment): 3,873 articles
- ስፖርት (Sports): 9,812 articles
- ቢዝነስ (Business): 9,307 articles
- አለም አቀፍ ዜና (World News): 6,515 articles
- ፖለቲካ (Politics): 6,635 articles

It was collected from the following sources:

- Addis Admas: 1839 articles
- Addis Maleda: 847 articles
- Al-Ain Amharic: 887 articles
- Amhara MM: 2438 articles
- BBC Amharic: 816 articles
- Ethiopian Press: 5597 articles
- Ethiopian Reporter: 6280 articles
- Fana Broadcasting: 7700 articles
- Soccer Ethiopia: 8595 articles
- VOA Amharic: 6943 articles
- Walta: 8764 articles

Sample Data

Headline	Article	Category	Date	Views	Link
የኦሊምፒክ ማጣሪያ ተሳታፊዎች ...	ብርሃን ፈይሳየኢትዮጵያ ቦክስ ፌዴሬሽን በየዓመቱ የሚያዘጋጀው የክለቦች ቻምፒዮና በአዲስ አበባ ከተማ በመካሄድ ላይ ይገኛል።...	ስፖርት	January 14, 2021	2	https://www.press.et/Ama/?p=39481
አዲስ ዘመን ድሮ...	የአዲስ ዘመን ጋዜጣ ቀደምት ዘገባዎች በእጅጉ ተነባቢ ዛሬም ላገኛቸው በእጅጉ ተነባቢ ናቸው። ...	መዝናኛ	December 28, 2020	4	https://www.press.et/Ama/?p=38334

Preprocessing

Normalization: Applied character-level normalization to correct inconsistencies in the Amharic script (e.g., ጸሀይ and ፀሐይ). These characters represent the same word, but the Amharic language does not enforce a strict rule on which character to use. Therefore, we normalized them to a consistent character representation.

Filtering: Removed rows with missing values in the category column.

Splitting: Used an 80-20 split for training and testing, ensuring a diverse representation of categories in both subsets.

Implementation Overview

1. XLM RoBERTa (Pretrained and Finetuned)

We fine-tuned the XLM-RoBERTa model to classify Amharic news articles into six distinct categories. To adapt the model to our task, we added a classification layer on top of the pre-trained XLM-RoBERTa model. The fine-tuning process was conducted on the Amharic news dataset. During training, we used a batch size of 16, a learning rate of 5e-5, a maximum sequence length of 512 tokens and the model was trained for 5 epochs.

The entire training and fine-tuning process utilized the Transformers library from Hugging Face. Once fine-tuned, the model was deployed on Hugging Face's model hub for easy access and integration into applications.

Evaluation Results

Category	Precision	Recall	F1 Score
ሀገር አቀፍ ዜና (National News)	0.9113	0.8849	0.8979
መዝናኛ (Entertainment)	0.8681	0.7315	0.7940
ስፖርት (Sports)	0.9732	0.9932	0.9831
ቢዝነስ (Business)	0.7129	0.7631	0.7372
አለም አቀፍ ዜና (World News)	0.8984	0.9192	0.9087
ፖለቲካ (Politics)	0.8378	0.8427	0.8402
Average	**0.8670**	**0.8557**	**0.8602**

2. Count Vectorizer

We used the Naive Bayes approach with Count Vectorizer as a feature extraction method. Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions between the features. The Count Vectorizer is used to convert a collection of text documents into a matrix of token counts. Then the Naive Bayes model is trained on the count matrix to classify the news articles into one of the six categories.

For model training, we used Scikit-learn's MultinomialNB model with default parameters. The model was trained on the training set and evaluated on the test set.

Evaluation Results

Category	Precision	Recall	F1 Score
ፖለቲካ (Politics)	0.67	0.58	0.62
ሀገር አቀፍ ዜና (National News)	0.88	0.48	0.62
ስፖርት (Sports)	0.96	0.94	0.95
ዓለም አቀፍ ዜና (World News)	0.45	0.89	0.60
ቢዝነስ (Business)	0.40	0.75	0.52
መዝናኛ (Entertainment)	0.26	0.80	0.39
Average	**0.60**	**0.74**	**0.62**

3. TF-IDF

The TF-IDF (Term Frequency-Inverse Document Frequency) approach is another simple yet effective method for text classification. It is based on the idea that words that occur frequently in a document but rarely in other documents are more important for classification. The TF-IDF vectorizer is used to convert a collection of text documents into a matrix of TF-IDF features. Then the Naive Bayes model is trained on the TF-IDF matrix to classify the news articles into one of the six categories.

For model training, we used Scikit-learn's MultinomialNB model with default parameters again.

Evaluation Results

Category	Precision	Recall	F1 Score
ፖለቲካ (Politics)	0.55	0.75	0.63
ሀገር አቀፍ ዜና (National News)	0.93	0.47	0.63
ስፖርት (Sports)	0.97	0.94	0.96
ዓለም አቀፍ ዜና (World News)	0.66	0.73	0.69
ቢዝነስ (Business)	0.37	0.83	0.52
መዝናኛ (Entertainment)	0.23	0.84	0.36
Average	**0.62**	**0.76**	**0.63**

Comparison of Models (Averages)

Model	Precision	Recall	F1 Score
XLM-RoBERTa	0.8670	0.8557	0.8602
Count Vectorizer + Naive Bayes	0.60	0.74	0.62
TF-IDF + Naive Bayes	0.62	0.76	0.63