Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Improving Amharic Information Retrieval Using Transformer Based Language Models

Andargachew Alayu Eshete

contactandargachew@dbu.edu.et

Debre Berhan University

Anbesaw Belete Mekonnen

werabe university

Shambel Kefelegn Bereded

Debre Berhan University

Dereje Chinkl Kassie

Debre Berhan University

Research Article

Keywords: Information Retrieval, Amharic Language, Se-mantic Search, Transformer Models, BERT, XLM-RoBERTa, Natural Language Processing.

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11114849/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Improving Amharic Information Retrieval Using Transformer Based Language Models

Andargachew Alayu Eshete[1*], Anbesaw Belete Mekonnen[2], Shambel Kefelegn Bereded[3], Dereje Chinkl Kassie[1] [1]Department of Information Systems,Debre Berhan University, Debre Berhan, Ethiopia

[2]Department of Information Systems, Werabe University, Werabe, Ethiopia

[3]Department of Computer Science,Debre Berhan University, Debre Berhan, Ethiopia

[3]Department of Information Systems,Debre Berhan University, Debre Berhan, Ethiopia

[*]Corresponding author: contactandargachew@dbu.edu.et

Abstract

Information retrieval systems are essential for obtaining relevant information from large-scale digital resources. However, current IR systems are language-dependent and have challenges in under-standing the semantic meaning of queries. This study aims to develop a transformer-based semantic information retrieval model for Amharic documents. The model was developed by collecting, cleaning and normalizing 4,372,180 Amharic documents using the design science research methodology. The evaluation was per-formed on three multilingual transformer models, XLM-RoBERTa, mBERT-Cased, and mBERT-Uncased, in different training and testing configurations in Python 3.7. The experimental results show that XLM-RoBERTa achieved the best performance with an F-measure of 92.09% (97.09% precision and 87.64% recall) for an 80/20 train-test split and 91.52% F-measure (94% precision and 88.9% recall) for a 70/30 split. The results demonstrate that transformer-based models are capable of capturing semantic relation-ships in Amharic queries and improving the performance of information retrieval. Further research can investigate advanced transformer architectures and pre-training methods to improve Amharic semantic retrieval systems.

Keywords

Information Retrieval, Amharic Language, Se-mantic Search, Transformer Models, BERT, XLM-RoBERTa, Natural Language Processing.

1. Introduction

Amharic is the most official working language in Ethiopia, used extensively for communication across the various regions of the country. According to the 2007 Ethiopian census of population [1]. Amharic is the mother tongue of some 22 mil-lion and the second language of another 7 mil-lion Ethiopians. The last census in 1994 revealed that there were close to 617 million native speakers and more than 40% of the Ethiopian population could understand the language [2]. More recent estimates place the total number of Amharic speakers combining both first- and second-language users at well over 56 million, with around 32 mil-lion native speakers and more than 25 mil-lion speaking it as a second language [3]. Since there are many people who speak Amharic either as their native language or as an additional language, building Information Retrieval Systems (IRS) capable of working with Amharic is crucial in order to allow Ethiopians to have quick and reliable access to information. The latter is essential for communicating, educating, and providing public services throughout the country, like Ethiopia [4].

Information retrieval is the process of searching and obtaining certain information resources that are relevant to a specific information need from vast document [2]. These systems underlie web search, digital libraries, electronic health records, question-answering systems, expert systems, and document classification. [5],[6]. As digital technology has spread, Amharic content online has grown rapidly. That growth has pushed demand for IR systems that can deliver relevant Amharic documents at any time [6],[7]. A good Amharic IR system should help users locate what they need without sifting through irrelevant results [8]. The retrieval process starts when a user submits a query. The system then runs an indexing and matching algorithm against its document collection to find the best matches [9]. The quality of an IR system comes down to the retrieval algorithm and how well it determines semantic relevance be-tween queries and documents [8]. IR research has advanced, but many systems still fail to meet actual user needs because they rely on lexical or keyword matching [2]. Traditional models like Bag-of-Words (BoW) and TF-IDF score documents by exact word overlap with the query. Yet people express the same need in different words, depending on their knowledge, vocabulary, or writing style[10]. For example, an Amharic document may contain the sentence አትክልት መመገብ እወዳለሁthat a user search for ሙዝ መብላት እወዳለሁ. The meanings are nearly identical, but standard techniques will likely fail because the query words do not appear in the document. ሙዝ is a specific fruit, whereas አትክልት is a broader category. መመገብ and መብላት are synonyms in different forms. Word-vector methods only match words, not meanings, so they miss these associations. This makes these methods less applicable for morphologically rich languages such as Amharic. The problem gets worse because Amharic has so many synonymous and polysemous words. Synonyms have different forms but similar meanings in context; polysemous words share the same form but carry different meanings depending on usage [1], [11]. A user might ask, የኢትዮጵያ ጠቅላይ ሚኒስትር ማን ነው or they might አሁን ያለው የኢትዮጵያ መሪ ማን ነው both seek the same answer. Similarly, በአዲስ አበባ ያሉ የቱሪዝም ቦታዎችን አሳዩኝ and አዲስ አበባ ውስጥ የሚጎበኙ ቦታዎች ምንድናቸው are different phrasings of the same request. Mean-while, polysemous words can shift meaning entirely with context, so accurate retrieval requires contextual

understanding. Deep learning (DL) has improved modern IR by moving systems beyond keyword matching toward contextual understanding [12]. As such, the potential benefits of utilizing DL methods in education, research, communication, and services for the number of communities that communicate in Amharic are immense. In order to develop an information retrieval system for Amharic users, which utilizes the semantics and meanings be-hind each query, it will be necessary to use a deep learning method [9]. Among DL architectures, Transformer-based models have shown great performance in semantic IR. Traditional word embeddings consider each word in isolation. Transformers use self-attention to process all words in a sentence at the same time, capturing their contextual relationships. This enabled the model to differentiate between two meanings of the same word and to learn that different phrases have the same meaning. Thus, most of the state-of-the-art IR systems are based on transformer-based models. Several studies have attempted Amharic IR, but cur-rent systems still failure semantic and con-textual meaning. Early work used statistical term-frequency approaches where document ranking depended heavily on query term frequency [2], [9]. These methods frequently preferred longer documents and ignored semantic relations between words. Later studies added Word2Vec embeddings with cosine similarity to find related documents. How-ever Word2Vec is limited by a fixed vocabulary and cannot represent unseen or rare words effectively. [9]. Other researchers tried stem-based and root-based CBOW and skip-gram models to im-prove semantic representation. But these captured semantics only at the word level and failed to model phrase-level or sentence-level contexts. So, as per the knowledge of the researcher, there is no previously developed Amharic-aware information retrieval system that effectively handles semantic or context-level language issues. Therefore, this study aims to develop a transformer-based Amharic information retrieval model that takes the semantics or context of the information seeker's query by using transformer models.

2. Related Work

Transformer-based Amharic information retrieval has gained significance due to its potential to enhance users' ability to retrieve relevant information from large corpus data. [13] conducted a study on information retrieval related to information regarding Covid-19 using transformer architecture. The study mainly focuses on testing the information retrieval model with the BERT model using a transformer architecture. The BERT architecture used is a transformer framework using an encoder for the process of matching documents with queries. This work uses two encoders, namely bi encoder and cross-encoder. The results of document encoding are then ranked using BM25. The scores from the two encoders are then evaluated using the MRR, mAP and NDCG methods to see which encoder is more efficient for searching data using queries. The results of data encoding using three queries using 6222 data points.

The first ranking score for the query" covid19" using a bi-encoder showed a value of 0.511 and of the 5 highest data points, there was data point that was not relevant to the query. The first ranking score for the query" covid19" using a cross-encoder shows a value of 0.920% and of the 5 highest data points, there is no data that is not relevant to the query. The first ranking score for the query" virus" using a bi-encoder shows a value of 0.582% and of the 5 highest data points, there is 1 data point that is not relevant to the query. The

first ranking score for the query" virus" using a cross-encoder shows a value of 0.707% and of the 5 highest data points, there is no data that is not relevant to the query. The first ranking score for the query" covid" using a bi-encoder shows a value of 0.426% and of the 5 highest data points, there are 4 data points that are not relevant to the query. The first ranking score for the query" covid" using a cross-encoder shows a value of 0.630, and of the 5 highest data points, there are 2 that are not relevant to the query. [14] focused on transformer-based model for Arabic language understanding. Transformer-based model has proven to be very efficient at language understanding provided they are pre-trained on a very large corpus. These models were able to set new standards and achieve state-of-the-art results for most NLP tasks.

This study uses pre-trained BERT specifically for the Arabic language in the pursuit of achieving the same success that BERT did for the English language. The performance of AraBERT is compared to multilingual BERT from Google and other state-of-the-art approaches. Performance of AraBERT 96.2% and mBERT 93% compared to the result shows AraBERT were better performance than mBERT. For Ethiopian/Amharic information retrieval, [2] conducted a study to design an Amharic information retrieval system based on the statistical co-occurrence technique to enhance the effectiveness or performance of the model that has been done in the previous research. Hence, the statistical co-occurrence technique has been deployed to increase its performance. The statistical approach that has been used for the Amharic information retrieval system performs the retrieval task based on frequently appearing terms with the query term, which ignores the meaning existing in the query term. However, in this research, long documents may have an unfair advantage in being retrieved since short documents will not be considered relevant documents.

[10] has developed Amharic information retrieval using semantic vocabulary. Automatically construct a semantic vocabulary from an Amharic text corpus using a distributional semantics approach for information retrieval. In this research, the researcher collects different document collections; semantic vocabulary has been built. The user query will match to semantic vocabulary using cosine similarities, and the query has been expanded and forwarded to the document collection to retrieve relevant documents as per the user query. However, the researcher doesn't include the phrasal words, and the performance of the system is low because of query optimization. In addition, utilized a statistical co-occurrence technique for Amharic information retrieval. However, this approach primarily focused on frequently appearing terms without considering the semantic meaning of the query term. As a result, long documents may have had an unfair advantage in retrieval, while short documents were overlooked.

[1] focused in this study on the development of a semantic information retrieval (IR) system for the Amharic language. The primary objective is to accept user queries in the form of words and retrieve semantically relevant terms from a document collection. To build the model, a corpus was collected from the 2AIRTC dataset, followed by the application of various preprocessing techniques. The research introduces a semantic IR model that integrates query expansion techniques based on deep neural learning and Amharic WordNet. The proposed approach put forward centers on the representation of Amharic texts through the root forms of words. To a large extent, this technique is used on stopword removal, indexing, term matching, and query

expansion. On the other hand, the usual stem-based representation is also experimented with, and a comparative study is done. The findings show that root-based representation yields better results for both resource construction and IR system development. To capture the statistical properties of words in documents and queries, the study uses Word2Vec and the Continuous Bag-of-Words (CBOW) model. Besides, root-based query expansion is carried out to make user queries more comprehensive, which results in better retrieval effectiveness. In other words, it is a way of enabling the system to better grasp the intent behind a query and thus return more accurate results. Lastly, a comparative analysis between stem-based and root-based approaches is conducted with a view to deciding which retrieval model is more effective. The measurement of performance is done through well-known IR evaluation metrics that clearly show the advantage of the root-based method in enhancing Amharic semantic information retrieval.

3. Data Preparation

Datasets are required to train and test the effectiveness of the IR system. However, there is no standardized corpus prepared for the Amharic language. Therefore, the study pre-pared the dataset by collecting the text documents written in the Amharic language from different sources that can represent the language in the general aspects of education, culture, arts, health, and politics. Such sources are from online free data portals of different domains like Ethiopian Reporter, Fana Broad-casting Corporation, Amharic Bible, and z-habesha.com. The collected documents are filtered manually, and text preprocessing operations have been applied on the collected dataset, which are language-dependent processes such as tokenization and normalization to make it ready for use in the IR system. For prototype evaluation, a total of 4,372,180 Amharic documents were collected and stored in the UTF-8 encoding format, which is suitable to support languages.

No	Data source	Number of Documents	TOTAL	Scenario One: 80:20 Percentage Split Test Option		Scenario two: 70:30 Percentage Split Test Option
1	Ethiopian Reporter	1098045	4,372,180	Training	3,497,744	3,060,526
2	Fana-Broadcasting Corporate	1097050		Validation	437,218	655,827
3	Amharic Bible	1089040		Testing	437,218	655,827
4	Z-Habesha.Com	1088045		TOTAL	4,372,180	4,372,180

Table 1: Dataset Data Sources and Train-Validation-Test Percentage Splits

3.1. Data Preprocessing

When you get datasets in the Amharic language, cleaning up things before usage is al-most a necessity. Otherwise, you get a lot of random stuff scattered throughout English words, numbers, hashtags, emojis,

weird symbols, and even HTML tags. example: እንኳን በደህና መጣህ123# አማርኛ Welcome, @user It has a bit of everything in it. If you actually care about Amharic text, all the English parts, random mentions, special characters, etc. just get in the way. So, in order for this data to be useful for any kind of Amharic-focused NLP work, you have to strip out anything that is not really Amharic. So that means extracting the English bits, re-moving spammy punctuation, removing undesired symbols, and doing things like HTML markup. At the same time, you don't want to be too aggressive and remove real Amharic characters or the periods and question marks that mark where sentences begin and end. The script itself can be tricky, so any cleaning you do will have to understand Unicode and the idiosyncrasies of the Amharic alpha-bet. Finally, you want to clean up the spaces too: no double spaces and no messy whites-pace. All this work will pay off eventually. You have a much cleaner dataset and your search or information retrieval models actually work better.

3.1.1. Tokenization

The first Amharic corpus had particular processing issues. The raw text was noisy with complex punctuation, numbers, special char-acters and control characters but words are separated by white spaces and the sentence termination is marked by the Arat Netsib (::). To prepare this stream of characters for robust NLP. The text must be divided into sub-word token units. Using the XLMRobertaTokenizer, the study converted the documents into this fine-grained format. For instance, the statement "ከመንግስት መስሪያ ቤቶች በተጨማሪ ህዝባዊና የልማት ድርጅቶችም ጭምር" is tokenized into sequence:"<s>", "ከመ", "ንግስት", "መስ", "ሪያ", "ቤቶች", "በ", "ተጨማሪ", "ህዝባ", "ዊና", "የልማት", "ድርጅቶች", ጭምር" and"<s>". This subword tokenization process was suitable across the entire corpus ensuring all documents were converted into the sequence-based structure required by the XLM-RoBERTa model.

3.1.2. Normalization

When different Amharic letters share the same sound but look different, we group them into a single, most likely character for consistency. Take ሀ, ሐ, ኀ, and ሃ—these all get mapped to ሀ in every word across the corpus. The Amharic language is unlike many other languages; it has plenty of characters that look different but mean the same thing, like አ, ኣ, ዐ, ዓ, ሀ, ሃ, ሐ, ሓ, ኀ, ኃ, ሰ, ሠ, ፀ, and ጸ. Even though there are language rules for using each character, people tend to write as they want, which becomes a challenge for researchers. So, words like ኀይል, ኃይል, ሀይል, ሃይል, ሐይል, and all their variations get standardized to just ሀይል. There are many homophones in Amharic let-ters. There are many homophones and characters in Amharic which include ሐ, ሑ, ሒ, ሓ, ሔ, ሕ, ሖ, ኀ, ኍ, ኂ, ኃ, ኄ, ኅ, ኆ, ሠ, ሡ, ሢ, ሣ, ሤ, ሥ, ሦ, ዐ, ዑ, ዒ, ዓ, ዔ, ዕ, ዖ, ጸ, ጹ, ጺ, ጻ, ጼ, ጽ, ጾ, ጿ, ፀ, ፁ, ፃ, ፄ, ፅ, ፆ. To avoid the spelling variation problem of Amharic, the characters must be mapped into one and common writing alpha-bet. Not only spell inconsistencies but also abbreviations like "ዓ.ም" or " ዓ/ም" ,"አ. አ" or "አ/ አ" " ዶ . ር " or " ዶ / ር" ," መ.ር" or "መ/ር" "ወ.ሮ" or "ወ/ ሮ" "አ.ቶ" or "አ/ቶ" "ዳ.ን" or "ዳ/ን" and "ት.ቤት" or "ት/ቤት" must be written to their corresponding long forms.

4. Proposed Models

The study evaluated three pre-trained trans-former models for Amharic information retrieval: XLM-RoBERTa, mBERT-Cased, and mBERT-Uncased. XLM-RoBERTa uses 12 layers, 768 hidden units, and 12 attention heads, with approximately 125 million parameters, while both mBERT variants use 12 layers and 768 hidden units with ap-proximately 110 million parameters. XLM-RoBERTa and mBERT-Cased preserve capitalization, whereas mBERT-Uncased converts input text to lowercase.

4.1. Ethics statement

This study was developed from the author's master's thesis. The internal and external examiners evaluated the thesis as part of the university's academic assessment and degree-awarding process. They did not participate as research subjects or provide annotations or relevance judgments for the experiments reported in this study.

4.2. Training Setup

We fine-tuned the models for document retrieval. For a given query, the system computes cosine similarity between the query embedding and document embeddings (derived from the [CLS] token). Documents were ranked in descending similarity.

4.3. Hyperparameter Settings

Hyperparameter	Value
Encoder Layers	12
Decoder Layers	12
Learning Rate	0.01
Optimizer	Adam
Attention Heads	12
Epochs	50
Hidden Size	768
Batch Size	32
Dropout	0.1

Table 2: Hyperparameter Configuration Used for Model Training

4.4. The Proposed Architecture

The suggested Amharic information retrieval architecture is illustrated in the figures and in-volves training, validation and testing phases. The system pre-processes the Amharic corpus, extracts features, and creates word embeddings using a fine-tuned pre-trained multi-lingual model.

We next analyze the retrieval model using precision, recall, and F1-score and verify its performance with experts. To further strengthen the study, the prototype was also presented to domain experts (examiners) for feedback and validation. Figure 3.1 illustrates the proposed architecture, which guided the design, implementation, and evaluation of the proposed Amharic information retrieval model.

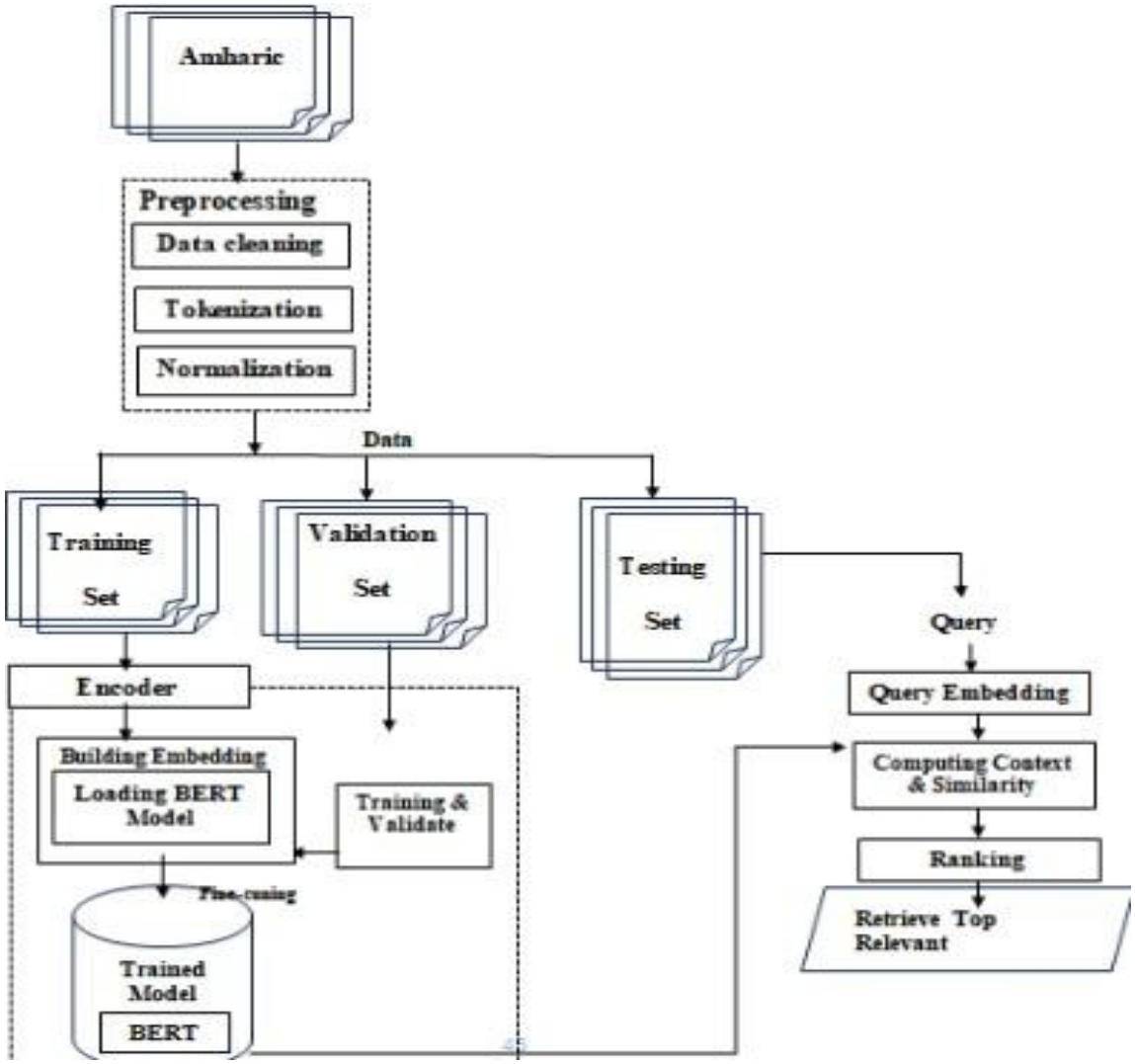


Figure 1: Overall architecture of the proposed system

XML -RoBERTa

Scenario One: 80:20 Percentage Split Test Option In this scenario, the Amharic information retrieval model was developed and evaluated by dividing the entire collected corpus of 4,372,180 into three parts: 3,497,744 documents (80%) for training and the remaining 437,218 documents (10%) for validation and 437,218

documents (10%) for testing. Hence, the detailed experimental analysis of each query's retrieval performance is presented below in a table.

No	Query	Performance Evaluation Metrics					
		Xlm-Roberta 80:20 Test Option					
		Rel.	Ret.	Ret.Rel.	Prec.	Rec.	F1
1	የኢትዮጵያ ዋና ከተማ	28697	25584	24832	97.06	86.53	91.49
2	ብዙ ቋንቋዎች	28292	26831	25932	96.64	91.65	94.07
3	የማቄዶንያ መሠረቶች	29871	27586	26938	97.65	90.18	93.73
4	እርሻ ልማት	28795	26385	25524	96.73	89.00	92.70
5	አማርኛ ቋንቋ	28675	26841	25543	95.16	89.07	92.01
6	አየር ጥራት	28993	25562	24989	97.75	86.17	91.59
7	ጤና በኢትዮጵያውያ	29872	26568	25993	97.83	87.01	92.10
8	አበበው በቀለ	28770	24842	23934	96.34	83.19	89.28
9	ጤናማ አስተሳሰብ	29891	27381	26935	98.37	90.00	94.00
10	በህይወት እኖራለሁ	28869	24617	23872	97.00	82.69	89.27
11	አባይ ባንክ	28982	24684	23992	97.19	82.78	89.40
12	የኢትዮጵያ ግዙፍ ታሪክ	29880	26361	25972	98.52	89.92	94.02
13	ከበታችነት ስሜት የሚመነጭ	29775	26657	25726	96.50	86.40	91.17
14	የንግድ ኢንዱስትሪ	28992	26992	25993	96.29	90.00	93.03
15	የአዲስ አበባ መስጊዶች	28864	26671	25972	97.37	90.00	93.54

Table 3: Performance Evaluation Metrics for the XLM-RoBERTa 80:20 Test Option

Scenario Two XLM-RoBERTa: 70:30 Percentage Split Test Option

For this purpose, the corpus consisting of 4,372,180 documents was divided into three parts: 3,060,526 documents (70%) for training, 655,827 documents (15%) for validation, and 655,827 documents (15%) for testing.

No	Query	Performance Evaluation Metrics					
		Xlm-Roberta 70:30 Test Option					
		Rel.	Ret.	Ret.Rel.	Prec.	Rec.	F1
1	የኢትዮጵያ ዋና ከተማ	43295	42105	40582	96.38	93.73	95.03
2	ብዙ ቋንቋዎች	43878	42683	41979	98.35	95.67	97.00
3	የማቄዶንያ መሠረቶች	43767	40894	39732	97.15	90.78	93.85
4	እርሻ ልማት	43858	40987	39798	97.06	90.74	93.79
5	አማርኛ ቋንቋ	42897	39867	36974	92.74	86.19	89.34
6	አየር ጥራት	43939	40983	38867	94.83	88.45	91.52
7	ጤና በኢትዮጵያውያ	43889	41972	39769	94.75	90.61	92.63
8	አበበው በቀለ	44768	42987	40983	95.33	91.54	93.39

9	ጤናማ አስተሳሰብ	43889	40973	38947	95.05	88.73	91.78
10	በህይወት እኖራለሁ	43898	40892	38782	94.84	88.34	91.47
11	አባይ ባንክ	43189	40867	38973	95.36	90.23	92.72
12	የኢትዮጵያ ግዙፍ ታሪክ	43697	40689	33768	83.00	77.23	80.01
13	ከበታችነት ስሜት የሚመነጭ	43687	40689	38634	94.9	88.43	91.55
14	የንግድ ኢንዱስትሪ	43789	40672	38672	95.08	88.3	91.56
15	የአዲስ አበባ መስጊዶች	43387	40679	36677	90.00	84.53	87.17
Total Corpus		655827			94.32	88.9	91.52

Table 4: Performance Evaluation Metrics for Xlm-Roberta 70:30 Test Option

Scenario Three, BERT-Base Multilingual Cased: 80:20 Percentage Split Test Option

The experiment is conducted by splitting the corpus of 4,372,180 documents into 80% for training and 437,218 documents (10%) for validation and 437218 documents (10%) testing. The effectiveness of the Amharic information retrieval model was evaluated using precision, recall, and F1-score.

No	Query	Performance Evaluation Metrics					
		mBERT (Cased), 80:20 Test Option					
		Rel.	Ret.	Ret.Rel.	Prec.	Rec.	F1
1	የኢትዮጵያ ዋና ከተማ	28697	25371	20348	80.20	70.9	75.26
2	ብዙ ቋንቋዎች	28292	25617	23105	90.19	81.66	85.71
3	የማቄዶንያ መሠረቶች	29871	26627	23457	88.09	78.52	87.80
4	እርሻ ልማት	28795	25361	23180	91.40	80.50	85.60
5	አማርኛ ቋንቋ	28675	27209	25341	93.13	88.37	90.68
6	አየር ጥራት	28993	25891	22657	87.50	78.14	82.55
7	ጤና በኢትዮጵያውያ	29872	26861	23271	86.63	77.90	82.03
8	አበበው በቀለ	28770	26732	23866	89.27	82.95	85.99
9	ጤናማ አስተሳሰብ	29891	27309	24537	89.84	82.08	85.78
10	በህይወት እኖራለሁ	28869	25864	21867	84.54	75.74	79.89
11	አባይ ባንክ	28982	26197	23678	90.38	81.69	85.81
12	የኢትዮጵያ ግዙፍ ታሪክ	29880	26493	23591	89.04	78.95	83.69
13	ከበታችነት ስሜት የሚመነጭ	29775	26768	23954	89.48	80.45	84.72
14	የንግድ ኢንዱስትሪ	28992	26381	24638	93.39	84.9	88.94
15	የአዲስ አበባ መስጊዶች	28864	26731	23921	89.48	82.87	86.04
Total Corpus		437218			88.84	80.37	84.7

Table 5: Performance Evaluation Metrics for mBERT (Cased), 80:20 Test Option

Scenario for BERT-Base Multilingual Cased: 70:30 Percentage Split Test Option

The model was also evaluated by splitting the collected corpus into 70% for training, 15% for validation, and the remaining 15% for testing. A detailed analysis of the developed retrieval model using the BERT-

multilingual-cased pre-trained embedding is presented below.

No	Query	Performance Evaluation Metrics					
		mBERT (Cased), 70:30 Test Option					
		Rel.	Ret.	Ret.Rel.	Prec.	Rec.	F1
1	የኢትዮጵያ ዋና ከተማ	43295	42382	33725	79.57	77.89	78.72
2	ብዙ ቋንቋዎች	43878	39786	34892	87.69	79.52	83.40
3	የማቄዶንያ መሠረቶች	43767	40782	35768	87.70	81.72	84.60
4	እርሻ ልማት	43858	40352	36751	91.07	83.79	87.30
5	አማርኛ ቋንቋ	42897	38678	34972	90.41	81.52	85.73
6	አየር ጥራት	43939	39784	35768	90.00	81.40	85.48
7	ጤና በኢትዮጵያውያ	43889	40682	32834	80.70	74.81	77.64
8	አበበው በቀለ	44768	41092	36926	89.86	82.48	86.01
9	ጤናማ አስተሳሰብ	43889	40105	32367	80.70	73.74	77.06
10	በህይወት እኖራለሁ	43898	39762	34762	87.42	79.18	83.09
11	አባይ ባንክ	43189	39642	34531	87.10	80.00	83.39
12	የኢትዮጵያ ግዙፍ ታሪክ	43697	39867	32968	82.69	75.44	78.89
13	ከበታችነት ስሜት የሚመነጭ	43687	39897	33862	84.87	77.51	81.02
14	የንግድ ኢንዱስትሪ	43789	38976	33811	86.74	77.2	81.69
15	የአዲስ አበባ መስጊዶች	43387	39713	34867	87.79	80.36	83.91
Total Corpus		655827			86.29	79.1	82.53

Table 6: Performance Evaluation Metrics for mBERT (Cased), 70:30 Test Option

Scenario five, BERT Base Multilingual-Uncased 80:20 Test option

The experiment is conducted by splitting the corpus of 4,372,180 documents into 80%training, 10% validation, and 10% testing. The effectiveness of the Amharic information retrieval model was evaluated using precision, recall, and F1-score, with the results summarized.

No	Query	Performance Evaluation Metrics					
		mBERT (Uncased), 80:20 Test Option					
		Rel.	Ret.	Ret.Rel.	Prec.	Rec.	F1
1	የኢትዮጵያ ዋና ከተማ	28697	26972	25986	96.34	90.55	93.35
2	ብዙ ቋንቋዎች	28292	26972	24998	92.68	88.35	90.46
3	የማቄዶንያ መሠረቶች	29871	26983	22986	85.18	77.00	80.88
4	እርሻ ልማት	28795	26837	21981	82.00	76.33	79.06
5	አማርኛ ቋንቋ	28675	27192	21923	80.62	76.45	78.47
6	አየር ጥራት	28993	27867	26162	94.00	90.23	90.11
7	ጤና በኢትዮጵያውያ	29872	28734	22634	78.77	75.56	77.13

8	አበበው በቀለ	28770	27864	24987	89.67	86.85	88.23
9	ጤናማ አስተሳሰብ	29891	28679	23989	83.64	80.25	81.90
10	በህይወት እኖራለሁ	28869	27891	22992	82.43	79.64	81.01
11	አባይ ባንክ	28982	28197	24986	88.61	86.21	87.39
12	የኢትዮጵያ ግዙፍ ታሪክ	29880	28972	22993	79.36	77.00	78.16
13	ከበታችነት ስሜት የሚመነጭ	29775	27986	22983	88.12	77.18	82.28
14	የንግድ ኢንዱስትሪ	28992	27979	25979	92.85	89.60	91.19
15	የአዲስ አበባ መስጊዶች	28864	27682	26524	95.81	92.00	93.86
Total Corpus		437218			87.34	82.88	84.9

Table 7: Performance Evaluation Metrics for mBERT (Uncased), 80:20 Test Option

Scenario six, BERT-Base Multilingual-Uncased: 70:30 Percentage Split Test Option

The experiments were conducted by splitting the corpus of 4,372,180 documents into 70% training, 15% validation, and 15% testing. The effectiveness of the Amharic information retrieval model was evaluated using precision, recall, and F1-score, with the results summarized.

No	Query	Performance Evaluation Metrics					
		mBERT (UnCased), 70:30 Test Option					
		Rel.	Ret.	Ret.Rel.	Prec.	Rec.	F1
1	የኢትዮጵያ ዋና ከተማ	43295	40521	34571	85.31	79.84	82.48
2	ብዙ ቋንቋዎች	43878	39572	31248	78.96	71.21	74.88
3	የማቄዶንያ መሠረቶች	43767	38142	34257	89.81	78.27	83.64
4	እርሻ ልማት	43858	36751	32628	88.78	74.39	80.95
5	አማርኛ ቋንቋ	42897	39592	34672	87.57	80.82	84.05
6	አየር ጥራት	43939	40268	34627	86.00	78.80	82.24
7	ጤና በኢትዮጵያውያ	43889	37105	32964	88.83	75.10	81.39
8	አበበው በቀለ	44768	37931	33761	89.00	75.41	81.64
9	ጤናማ አስተሳሰብ	43889	41861	36934	88.23	84.15	86.15
10	በህይወት እኖራለሁ	43898	40683	34281	84.26	78.09	81.05
11	አባይ ባንክ	43189	39531	34367	87.00	79.57	83.11
12	የኢትዮጵያ ግዙፍ ታሪክ	43697	37192	31127	83.69	71.23	76.95
13	ከበታችነት ስሜት የሚመነጭ	43687	38972	35427	90.9	81.09	85.71
14	የንግድ ኢንዱስትሪ	43789	34861	31751	91.07	72.50	80.73
15	የአዲስ አበባ መስጊዶች	43387	40105	36922	92.06	85.09	88.43
Total Corpus		655827			87.43	77.7	82.23

Table 8: Performance Evaluation Metrics for mBERT (UnCased), 70:30 Test Option

4.5. Performance Comparison

The main aim of this study is to accept user queries in the Amharic language and retrieve the most relevant documents. To achieve this, pre-trained embed-ding models are employed and fine-tuned, including XLM-RoBERTa, BERT-base multilingual cased, and uncased. During the experiment, different scenarios are also explored, such as dividing the Amharic corpus into training, validation, and testing configurations and analyzing their effectiveness using multiple performance evaluation metrics. In this section, a comparison of the models is presented, which is illustrated in figure The figure illustrates the performance comparison

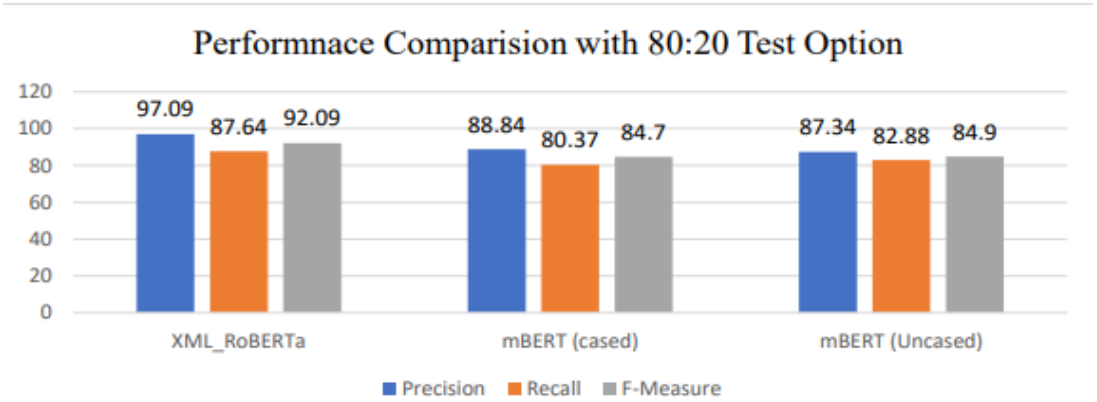


Figure 2: 80:20 test option Model Comparison parison of all pre-trained models used in this study.

Specifically, it shows the detailed results for XLM-RoBERTa, mBERT (cased), and mBERT (uncased) when the corpus is split into 80% for training, 10% for validation and 10% for testing. The results indicate that XLM-RoBERTa outperforms the other pre-trained models under this configuration. Furthermore, figure 4.4 shows performance comparison of pretrained models using corpus split ratio of 70% for training, 15% for validation and 15% for testing. Furthermore, figure shows performance comparison of pretrained models using corpus split ratio of 70% for training, 15% for validation and 15% for testing.

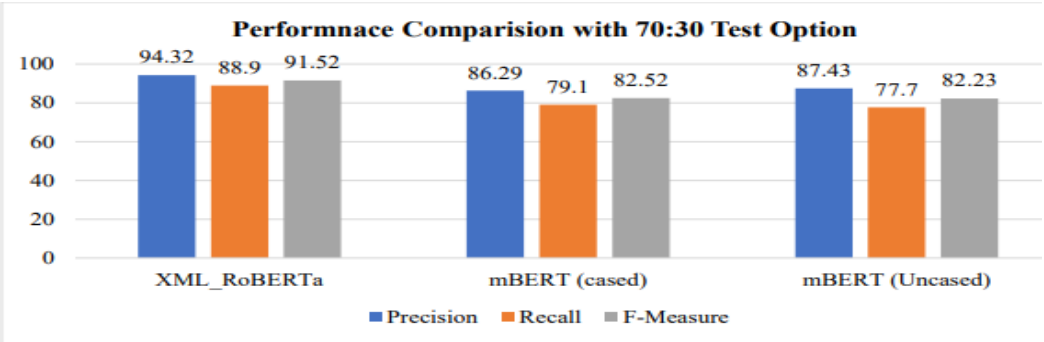


Figure 3: 70:30 test option Model Comparison

4.6. All Model Performance Comparison

BERT Embedding Models	Performance Measurement Metrics Pretrained Models		
	Precision	Recall	F-Measure

XML-RoBERTa	80:10:10 Test Option	97.09	87.64	92.09
	70:15:15 Test Options	94.32	88.9	91.52
mBERT-cased	80:10:10 Test Option	88.84	80.37	84.7
	70:15:15 Test Options	86.29	79.1	82.52
mBERT-Uncased	80:10:10 Test Option	87.34	82.88	89.9
	70:15:15 Test Options	87.43	77.7	82.23

Table 9: Comparison of Performance Measurement Metrics across Pretrained BERT Embedding Models

No	Author (Year)	Research Title	Method	Recall	Precision	F-score
1.	Bruck and Tilahun (2015) [2]	Developing Amharic IR System	Statistical Co-Occurrence Approach	73%	83%	84.5%
2.	Getnet Akalu (2019) [4]	Developing Amharic IR Using Semantic Vocabulary	Distributional Semantics Approach	84.23%	77%	72.5%
3.	Tilahun Yeshambel (2022) [1]	Developing a Semantic IR System for Amharic	Deep Neural Learning Model and Wordnet	71.23%	53%	63%
			proposed IR Method	**97.09%**	**87.64**	**92.09**

4.7. Comparison of This Study to Other Research Works

Table 10: Comparison of Literature Review Results with Proposed IR Method

4.8. Conclusion

In today's digital landscape, information is available in all types (text, audio, image, and video), as well as the ease with which users can access that information in this increasingly digitally connected world. We are seeing a massive shift in how we both receive and consume information. Technology has certainly made it easier for users to quickly access large amounts of data. But, despite these technological advances, there is still an enormous gap in being able to obtain meaningful and useful information in multiple languages (especially low-resource languages such as Amharic). Existing in-formation retrieval (IR) systems are predominantly language-dependent and often fail to support multi-lingual or non-English queries effectively. This study was to address this gap by addressing a transformer-based Amharic information retrieval (IR) model that can accept user queries in text or writ-ten Amharic and retrieve relevant Amharic documents. The research applied state-of-the-art Natural Language Processing (NLP) techniques, specifically leveraging pre-trained transformer models such as BERT, XLM-RoBERTa, and mBERT (both cased and uncased variants). These models have been fine-tuned and adapted to the Amharic language using custom training

data and employed hyperparameter settings. Three different pre-trained language models were tested: XLM-RoBERTa, mBERT-cased, and mBERT-uncased. The XLM-RoBERTa model with a hidden size of 768 and 12 transformer layers and also 12 attention heads, a feed-forward size of 3072, and a vocabulary size of 250002 outperformed the others in all evaluation metrics. The multilingual BERT model, though architecturally similar, had a smaller vocabulary size 119547 and produced 79 comparatively lower performance metrics. The data was split using multiple per-centage combinations, 80%:20%, 70%:30%, to ensure robustness and generalization. The experimental results showed that the XLM-RoBERTa model achieved the highest performance, with precision, recall, and F1-score values of 97.09%, 87.64%, and 92.09%, respectively. These results indicate that the model is highly effective in understanding and retrieving relevant Amharic documents based on user queries.

4.9. Future Work

Since Amharic and other Ethiopian languages are low-resource, building larger and high-quality corpora is essential to improve NLP performance. Current machine translation and NLP models perform poorly due to the limited availability of parallel and annotated datasets. Future work could focus on collecting, curating, and sharing publicly accessible corpora for Amharic and other low-resource Ethiopian languages. Exploring and comparing various text preprocessing techniques, such as stop word removal, stemming, and lemmatization for text, as well as investigating transformer-based and deep learning methods to enhance the performance of the Amharic information retrieval system. Appreciated direction for future research is to develop a context-aware image retrieval system. This enhancement would enable the system to use not only visual characteristics but also the semantic and contextual meaning linked to each image during the retrieval process. In this study, the researcher developed an Amharic information retrieval system focused solely on text-based document retrieval. For future research, expanding the system to support multimedia retrieval from image, audio, and video would significantly enhance its usability and applicability.

4.10. Declarations

Data Availability Statement: The dataset analyzed during the current study is publicly available through the Hugging Face Datasets repository at [15]

Author Contributions: All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Consent to participate: Not applicable.

Consent to Publish: Not applicable.

Conflict of Interest: The authors declare no conflict interests

Acknowledgments: The authors thank the editor and anonymous reviewers for their helpful comments and suggestions.

References

[1] T. Yeshambel, J. Mothe, and Y. Assabie, “Learned Text Representation for Amharic Information Retrieval and Natural Language Processing,” *Inf.*, vol. 14, no. 3, 2023, doi: 10.3390/info14030195.

[2] A. Bruck and T. Tilahun, “Enhancing Amharic Information Retrieval System Based on Statistical Co-Occurrence Technique,” *J. Comput. Commun.*, vol. 03, no. 12, pp. 67–76, 2015, doi: 10.4236/jcc.2015.312006.

[3] K. Vishwakarma, “Bilingual Hate Speech Detection on Social Media: Amharic and Afaan Oromo,” *JANOLI Int. Journals Big Data ISSN(online)*, vol. 1, no. 1, Jan. 2025, [Online]. Available: https://jijbd.janolijournals.org/

[4] B. Getnet Akalu, “Amharic Information Retrieval Using Semantic Vocabulary,” M.S. thesis College of Natural Sciences, Addis Ababa Univ., Addis Ababa, Ethiopia, 2019. [Online]. Available: http://etd.aau.edu.et/handle/123456789/22078

[5] I. Keleş, “Formalization of Natural Language Text using Syntactic Dependency,M.S. thesis, Dept. of Computer Science,” Eindhoven University of Technology, Netherlands, 2021.

[6] Z. Huang, M. Liang, J. Qin, S. Zhong, and L. Lin, “Understanding Self-attention Mechanism via Dynamical System Perspective,” *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1412–1422, 2023, doi: 10.1109/ICCV51070.2023.00136.

[7] D. N. Mhawi, H. W. Oleiwi, N. H. Saeed, and H. L. Al-Taie, “An Efficient Information Retrieval System Using Evolutionary Algorithms,” *Netw. MDPI*, vol. 2, no. 4, pp. 583–605, 2022, doi: 10.3390/network2040034.

[8] W. B. Demilie, “Implemented Stemming Algorithms for Six Ethiopian Languages,” *Int. J. Adv. Sci. Technol.*, vol. 29, no. 7, pp. 2532–2536, 2020.

[9] Y. A. Tilahun, Josiane Mothe, “Amharic Semantic Information Retrieval System,” *Commun. Comput. Inf. Sci.*, vol. 1608 CCIS, pp. 22–44, Oct. 2022, doi: 10.1007/978-3-031-14602-2_2.

[10] B. Getnet and Y. Assabie, “Amharic Information Retrieval Based on Query Expansion Using Semantic Vocabulary,” 2021, pp. 407–416. doi: 10.1007/978-3-030-80621-7_29.

[11] K. A. Mekonnen, Y. W. Alemneh, and M. de Rijke, “Optimized Text Embedding Models

and Benchmarks for Amharic Passage Retrieval," pp. 10428–10445, 2025, doi: 10.18653/v1/2025.findings-acl.543.

[12] A. Mamun, R. Uddin, T. Ahmed, and M. Islam, "A New Compression Based Index Structure for Efficient Information Retrieval," *J. Biomed. Inform.*, vol. 1, p. 5, 2012, doi: org/10.48550/arXiv.1209.5448.

[13] R. Yunitarini, E. V. Y. M. Stefany, A. Rachmad, E. K. A. Mala, S. Rochman, and Y. P. Asmara, "Implementation of Long Short-Term Memory for Traditional Medicine Information Retrieval System," vol. 25, no. September, pp. 496–502, 2026, doi: 10.47839/ijc.25.3.4845.

[14] H. Abdelazim, M. Tharwat, and A. Mohamed, "Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG)," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 11, pp. 1328–1334, 2023, doi: 10.14569/IJACSA.2023.01411135.

[15] "rasyosef/amharic-sentences-corpus · Datasets at Hugging Face." Accessed: Oct. 11, 2026. [Online]. Available: https://huggingface.co/datasets/rasyosef/amharic-sentences-corpus