

Afaan Oromo Hybrid Modelling: A Case based Optimized Intelligence in Information Retrieval System's Localization

Amin Tuni¹, Sisay Tumsa², Durga Prasad Sharma³, BhupeshKumar Singh⁴, Mario A. Bochicchio⁵

¹Arba Minch University, Arba Minch Ethiopia

²Arba Minch University, Arba Minch Ethiopia

³AMUIT MOEFDRE under UNDP & Research Adviser, MAISM-RTU, Kota, India

ABSTRACT

The data and information resources available with variety and velocity over the Internet and the World Wide Web (WWW) are dramatically changing their dynamics every day. Due to such information overloading, the access or retrieval task of the desired data or information becomes complex. This new phenomenon makes it difficult for the users to recognize and retrieve the relevant information which satisfies their needs. We are living in a world where diversity of the data and information in diverse languages have become a typical challenge. In the case of Ethiopia; a linguistic localization is paramount where an enormous amount of digital data and information resources in Afaan Oromo are being generated every day. These phenomenon features of data and information resources are again challenged by archival and searching issues in these large pools of documents. The prime aim of this research study was to develop a hybrid information retrieval system for localized significance in Afaan Oromo so as to enable the users to search and retrieve their required and relevant data information efficiently. Prior research studies in Afaan Oromo linguistic domain clearly indicate that the IR systems have yet not attained any promising attention for system performance. Also, several attempts made for Afaan Oromo IR (AOIR) system using a hybrid approach but their efficiency still needs a significant improvement. To solve such problems, this research study tried to integrate different types of approaches for the IR system towards the improvement of the performance of the AOIR system. The developed prototype has basic IR subsystems both for indexing and searching. For experimental analysis, 1000 Afaan Oromo text documents were collected from print media news articles (i.e. Oromia Broadcasting Network, VOA Afaan Oromo), the Afaan Oromo Bible, websites, books, and online News. The different text operations such as; tokenization, normalization, stop-word removal, and stemming were used to identify content bearing terms and vocabularies. The tfidf term weighting scheme was applied to compute the term weight in the documents. After the experimental analysis in Python 3.7, the average result achieved was 96.6% precision, 90.0% recall, and 93.3% F-measure respectively. The system performance was still found to be affected by the problem of polysemy. Therefore, the research recommends the additional work for improving the system performance so as to advance the AOIR using different techniques.

Keywords

Afaan Oromo, AOIR, Hybrid Model, Localization, Information Retrieval, VSM, PM

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

In the current era, the data, information, and knowledge overloading over the Internet have explored multifold opportunities along with risks and challenges. The social media and the multilingual contents of the web have created additional challenges for linguistically localized tools and techniques. Information Retrieval (IR) systems are perceived as problem solving agents. These are used for the identification of pertinent documents from the large pool document repositories [1]. The research studies [2][3][4] indicate that the ability of IR systems towards retrieving the relevant documents has become a critical challenge when document collections are very large. Usually, the IR can be defined as a search for appropriate and user

specified documents from unstructured documents repositories [5]. Also, the IR system stores and manages information documents and enables users to find their desired information needs in salient languages (i.e. localized or globalized). In general more than 80 types of languages are spoken in Ethiopia. Afaan Oromo is mostly spoken in African nation Ethiopia along with other native countries like Kenya, Djibouti, and Somalia. From the year 1995, it is considered as an official language of Oromia state of Ethiopia. This language is spoken by Oromo people as their native language, which is almost 40% of the total population of Ethiopia[6]. This language has its own scripting system called Qubeetha. It is a Latin-based alphabet. The IR system is a key technique for searching the suitable documents as per the need of representation, storage, and

knowledge management. It is very important to develop the IR system for local languages to localize the systems for better information management. The information seeker tries to express their query based on their information needs. The expressed queries are matched with the document representation which is extracted during an indexing time using different types of IR matching algorithms. The indexing of documents and the searching queries are typically coordinated using different types of algorithms and functions such as cosine for vector space modeling and probabilistic modeling. Using the algorithm similar marks the documents are accessed to the users to check the suitability of document with respect to their information needs[7]. The challenge for proper representation of large collection of documents and the competitive mechanisms in accurate indexing of documents can lead the integration of Artificial Intelligence (AI) into IR for optimizing the searching and indexing through case-based optimized intelligence.

Related works

Afaan Oromo is a Cushitic language and spoken by more than 40 million people of Ethiopia, Kenya, Somalia, and Egypt [8]. In 1991 Qubee that Afaan Oromo writing script was accepted as the official script of Afaan Oromo [9]. Afaan Oromo is also used for teaching and learning at primary and junior secondary schools throughout the Oromia region. It is also used in five universities of Ethiopia like Addis Ababa, Haramaya University, Jimma University, Dilla University, and Adama University as an important language [1]. In Ethiopia; there have been many attempts to develop the Information Retrieval System for the Amharic Language [8][9][10], as the official language of Ethiopia. The research

efforts done for the Amharic language include the development of retrieval algorithms for Amharic documents which are mainly written using Ethiopic (Ge'ez script) characters. Additionally, there was an iconic attempt to develop an Amharic search engine [11]. The research works on Cross-lingual developed Information Retrieval systems for Afaan Oromo with various enhanced accuracy and precision [12][13][14]. Prior to these research works, there was no other IR system for the Afaan Oromo language. Significant efforts made by other researchers enabled the retrieval of documents in Ethiopian languages specifically English in the Afaan Oromo query. Also, research efforts are done for developing a language translation for cross-lingual IR systems in Afaan Oromo[11][12] but these efforts are still lagging behind in solving the desired issues and problems of the target users and stakeholders. The target users looking for Afaan Oromo documents with Afaan Oromo query may not get suitable tools, techniques, search, and retrieval environments to find information as per their needs. The CLIR algorithm applied was able to follow either of machine translation or parallel corpuses only. The research efforts made in prior studies were not found aligned or customized for searching in Afaan Oromo documents using Afaan Oromo queries.

In addition to this, there were other attempts on developing information retrieval systems with salient perspectives of Afaan Oromo Language. These efforts were focused on the Afaan Oromo Text retrieval system, Afaan Oromo Search engine and evaluation of performance for Afaan Oromo by cross Lingual IR systems using different approaches to translate from Afaan Oromo to English[10]. These research efforts are summarized in the table 1.

Table1: Previous work on Afaan Oromo IRS

Author, year and title	Focus area	Approach	Aim of study	Performance
Afaan Oromo Text Retrieval system[11]	Indexing and Searching	Statistic approach	To design architecture for Afaan Oromo Information Retrieval system	Recall (62.64%) and Precision (57.5%)
Afaan Oromo-English Cross-lingual IRS[12]	Translation of document to user's query	Corpus based Approach	To develop Afaan Oromo-English CLIR	Recall (31.6%) and Precision (46.8%)
Afaan Oromo Search Engine [13][15]	Web document retrieval	Semantic approach	To design best algorithm for Afaan Oromo search engine	Precision (53%) and recall (93%)

After rigorous review and analysis of the related research works; it was observed that the efforts to design and develop the IR system on Afaan Oromo are limited and need other alternative efforts like the use of Artificial Intelligence which may combine both the statistic and probability approaches (i.e. Hybrid approach). Hence, the proposed study was aimed to design and develop an IR System for Afaan Oromo using artificial intelligence (AI) with a hybrid approach (a converged version of statistical and probabilistic) for better precision and accuracy in search and retrieval of information for improved satisfaction of users.

Information Retrieval (IR) Processes

In the IR System searching processes, the users anticipate finding the relevant information for

satisfying their information needs. The most appropriate documents based on user's specified needs are called relevant documents (ReIDs). Its contrary which do not satisfy are called irrelevant (IRelDs). In general practices, the different users may posture the same query to an IR system and judge the relevance of the retrieved documents in different manners. For instance; some users may like the results, but others may dislike them. There are three basic processes that an IR system should support: 1) the representation of the content of the large collection of documents (Indexing), 2) the representation of the user's information required as a query form (Searching), and 3) the comparison of the two representations (Matching)[14][16].

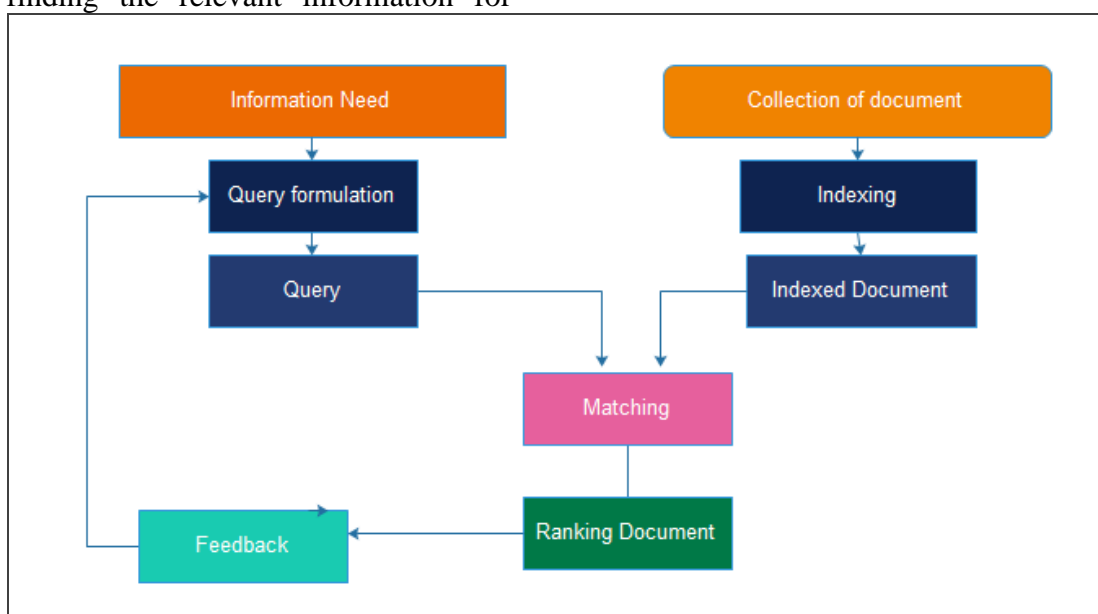


Figure 1: Information retrieval processes

The Indexing Process

Indexing is an offline process of extracting index terms from a large pool of documents collection and organizing them using indexing structure to speed up the search mechanism[17][18]. In general, it is a language-dependent process that varies from language to language.

1.1. Inverted Index

In IR system development process is an inverted index. It is very important for the creation of document index. This index enables the searching easy for users and used for representing documents for the mappings from content, such as words or numbers to its locations from the set of pooled document repositories. In this process, the construction of this indexing technique follows the four critical steps [18][19]- 1) collecting large

collection of documents to be indexed, 2) tokenization of the collected documents, turning each collected document into a list of tokens in words and sentences, 3) applying the Linguistic preprocessing, producing a list of standardized and stemmed tokens, which are the index bearing words, and 4) Creating index for the documents for each content bearing words by creating an inverted index, consisting of a directory and postings.

Tokenization: Tokenization is the process of cutting on white spaces and throwing away punctuation characters. If the original document is “Oromiyaan, magaalaa fi baadiyyaaqabdi” (Oromia has the urban and Rural) the tokens will be “Oromiyaan”, “magaalaa”, “fi”, “baadiyyaa”, “qabdi”. The indexing process contains three

straightforward steps: 1) Identifying the data source, 2) transforming into a logical view, and 3) Creating an index of the content bearing word on the logical view.

Linguistic Processing

In the IR system development process, the normalization deals with the situation folding mechanism. In this process, initially every word, phrase, and sentence occurred within the collected documents should be translated into a similar case format. In this linguistic process there are two very important techniques. These techniques are applied for removing every frequent term from the indexed documents which are poor in document contents bearing. And also, conflation of words to their base form or root after that where the morphological variants of words are stripped to single suffix entry.

1.2. Query Processing

In an IR system when the collected large collections of documents are indexed, the subsequent stage is to represent user’s query formulation into an internal form. Afterward, the system describes and transforms the formulated query into system’s internal common sense of document representation by applying query preprocessing techniques as the same with content bearing indexing. When these processes are completed, the lists of query bag of words weight is calculated and based on their weights, document index were created. The exact formula to calculate query term weighting is presented in table 2.

$$W_{qi} = (0.5 + \frac{0.5t_{fiq}}{Maxtf}) * Log(\frac{N}{ni})$$

Where:

Table 2: description of the mathematical formula for term weights

Method	Denoted
Weight of term for Query	W_{qi}
Term Frequency	t_{fiq}
Maximum Value of Term Frequency	max tf
Total collection of Document	N
Total quantity of document query term found	ni

Proposed IR System Prototype

In this research paper, the proposed Afaan Oromo IR system development process involves numerous information retrieval techniques and

methods. IR system designed involves two main components i.e. indexing and searching. The basic architecture of the information retrieval system is depicted in Figure 2.

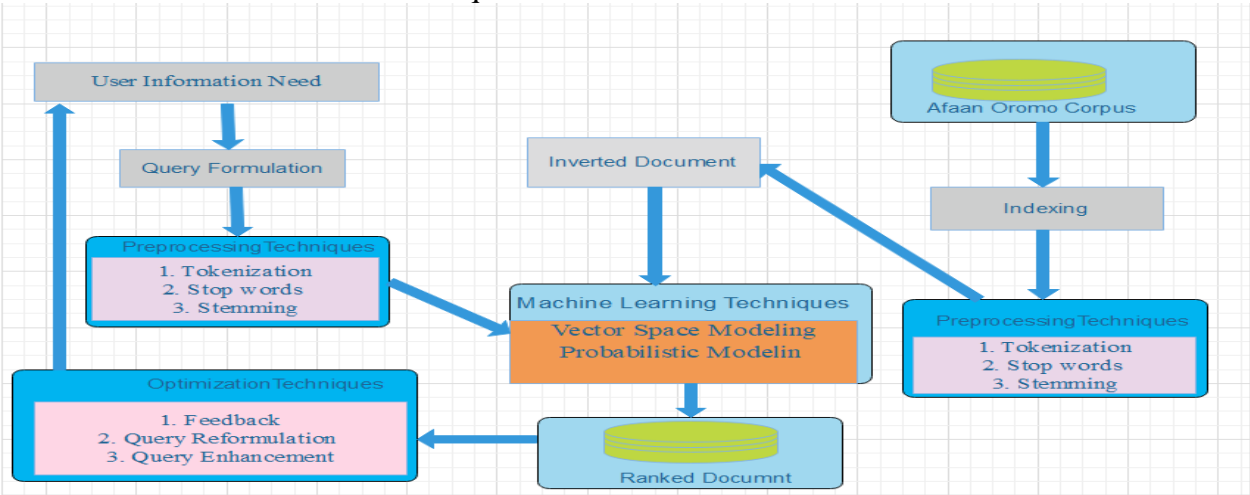


Figure 2: Afaan Oromo Hybrid Model for IR system

Data Preprocessing and Corpus Preparation

Afaan Oromo text corpus for the IR system was used to organize them by index files for enhancing the search strategy. For the indexing of the large

collection of documents, preprocessing is a very important step to make ready the documents for fast information retrieval. In searching, a similar text pre-processing technique is followed as it was practiced in the indexing part. Now, the matching

algorithm hybrid techniques (Vector space and Probability) models are used to get and rank the pertinent documents.

In this system development i.e. IR System (IRS), the preparing corpus is needed for evaluation of the system. For this purpose, documents were collected from different organizations (Afaan Oromo and Folklore department of Jimma university Oromo Cultural Center & Tourism office, and Oromia Broadcasting News) along with other online resources (Web news from VOA and Bariisa Magazine). These documents were passed through different procedures in order to index and use the retrieval system. Then different linguistic preprocessing was applied.

Algorithm for Afaan Oromo language tokenization

An IRS stemming is a very important step to remove the stop words from the indexed and searched documents. However, the stemmer algorithms which are applicable for English may not be suitable for Afaan Oromo word or sentences. For this purpose, a rule-based Afaan Oromo Stemmer algorithm as presented in Table 3 [13][15][20] was used.

Inverted Index

In IR system terminologies, a data structure that is important for effective and efficient access of the documents is indexed for the text by their words. In this inverted file, there are frequency documents and occurrences of each word in a text can also be represented. At this stage language preprocessing should be done carefully. Finally, the index file is created. The Index file includes two files i.e. vocabulary file and post file. This systematic internal document representation enabled IR system can be used to quickly find the relevant documents as per users need specifications[18][21].

Vector Space Model (VSM)

In the architecture of IR Systems, there are different IR models such as Boolean, VSM and Probabilistic. The matching algorithm plays a very important role. The VSM is a way of representing different documents in the spaces through the words that they contain as a vector. It is an algebraic model representing textual information as a vector and users queries on the plane in different angles. These vectors represent the importance of a term in the tf*idf weighting technique and even the absence or presence of the term in the document[22][23]. There are three

stages: content bearing creation, weighting the indexed terms and the last stage is the ranking of the required documents. This is as per the users query and depends on Euclidean distance and cosine function similarity measure.

Probabilistic Model (PM)

In this research study, different IR models were applied such PM and VSP to provide better search results and improves the searching result to find the relevant documents from large collection of documents to satisfy the information needs of users. In this IR models, latent semantic index (PLSI) was described by scholars[22][4] as a IR technique in which the users query is expanded and then Term Frequency Indexing (TFI) was applied to calculate the document ranking scores for relevant information retrieval. This IR model takes less time and produces better search results [4][17].

11.1 Designed and Proposed Afaan Oromo IR System Architecture

The designed and proposed Afaan Oromo IR system architecture for Case-based Optimized Intelligent IRS models using hybrid techniques for the Afaan Oromo information retrieval system focuses on two main parts of Information Retrieval System, which are indexing (Offline Process) and searching (Online Process). To improve the performance of searching in the proposed system, a large collection of Afaan Oromo documents are collected and organized after the preprocessing and Inverted Index was created. And also, for the online process i.e. searching for user's information needs the similar text preprocessing mechanism is applied to a formulated query. Then, an IR similarity measurement technique is done by applying a hybrid model (Vector space and Probabilistic) algorithm to achieve and rank relevant documents.

Dataset preparation and Document preprocessing

In this research study, dataset is prepared to create a desired corpus. It was prepared by gathering Afaan Oromo documents that are collected from different sources such as Education document, History, Gadaa System and Politics documents.

12.1 Document Pre-processing

In this research study, the document preprocessing technique was applied on collected documents which include the linguistic tasks.

In natural language processing, a tokenization process is very important. This is used for separating words and sentences from one another in the documents which may differ from one language to another language. But some of the languages are not Latin based writing script for instance Arabic, Japanese, Korean and Chinese languages. In these language scripts, the separations of words and sentence characters or space are not applicable. Afaan Oromo uses the Latin based writing script character and white spaces are used to separate word and sentences into individual words. Based on this rule, a tokenization technique was applied for this language word and sentence separation into different tokens of Afaan Oromo to create a group of tokens (based on the white space character).

For the normalization process, the writing style of some of the languages used different characters (Upper and Lower cases) to write the same word. Similarly, in Afaan Oromo, the writing script also uses an identical word with spell variation. This can be a common problem that might be appearing in the corpus which needs to be used for system performance evaluation purposes. Consequently, if those arguments are not substituted with the inconsistent group of word in the corpus, may be the system can treat as different concepts and which reduces the system performance.

In the Stop-word Removal Process, the common words cannot discriminate retrieving of the relevant information from the collected documents. This is a critical process in IR applications because, may it may cause the removal of some of the important words instead and it can mislead the information retrieved grammatically.

The Word Stemming [24] is another technique that was applied for removing words inflected by different morphological words (prefixes, infixes, and suffixes) which reduce a different word dissimilarity that have the same meaning.

As a consequence, the performance of the proposed system is considered to be an improved system. Primarily, it enhances the information retrieval evaluation matrixes that are recall and precision. Then, applying the stemming to this language document corpus and firing a query are considered major contributions in the proposed IR

system Architecture. It showcases the better performance in terms of efficiency and effectiveness. In addition, stemming techniques are language dependent tasks and therefore it may be similar to the Afaan Oromo System as well to make it a unique solution but with language reliant on limitations. Since, the proposed system was developed based on the Afaan Oromo morphological structure, and therefore each language compulsorily have their language specific stemming methods[25]. These are similar to Afaan Oromo language. This research categorically used a stemming algorithm developed by Debele[13][15] for rule base approach along with a dictionary-based algorithm. The Term Weighting is another process used in indexing and searching processes. In an indexing and searching process, assigning the weight to the term is known as “term weighting”. This process is very important to discover the suitable information as per the needs of the users for their satisfaction. Here calculating the term weight in the document collection enables to identify the important terms that signify document gratified and also used to discriminate certain specific documents from the pooled collection of the documents using term frequency. This can be used to indicate the local document (local factor) means how this term is important in specific documents and global document (global factor) as well. This means how abundant the term is important in whole collection of documents.

As explained before, the local term weighting process is to calculate the term weight contingent only on the occurrences of the word in the precise document excluding the other documents. To get the local term weighting, the several techniques are used to calculate the term importance in the documents.

The Universal term weighting was applied to provide stresses on the terms differentiating during the indexing and searching process. Accordingly, the inverse document frequency (idf) is the most popular measure of words' importance throughout the whole collection. This implies, rare terms or words have high idf and common terms or words have low idf value[29].

$$idf = \log_2(N/df_i)$$

Therefore, $tf \cdot idf$ computed as: $tf \cdot idf = tf_{ij} \cdot \log_2(N/df_i)$

In this research study, the term frequency-inverse document frequency (tf*idf) term weighting arrangement was used to compute the weight of each term since it used to normalize term weighting and popular indexing mechanisms.

In this research study, the *Inverted Index Process*[29] was created from collected documents that are corpus and applying the preprocessing techniques (Tokenization, Normalization, stop-word removal, and stemming). Since an overturned file is a technique that enables a proposed IR system for quick searching information so it has an important

influence on the performance of the complete retrieval system and therefore it was used to improve the performance of the proposed system to some level.

During this research study, three IR system performance evaluation techniques as mentioned in the table 4 were applied for measuring the performance for the planned IR System Architecture. The performance evaluation parameters are precision, recall, and F-measure. These system performance evaluation matrixes are depicted in the table 3.

Table 3: System Performance evaluation

	Relevant	Not Relevant	Precision (Pr)	Recall (Re)	F-Score
Retrieved	RR (TP)	RNR (TN)	RR/ (RR+ NRR)	RR/(RR+ NRNR)	2*Pr*Re/ (Pr + Re)
Not Retrived	NRR (FP)	NRNR (FN)			
	TP+FP	TN+FN			
Total			96.8%	90.0%	93.3%

Experimental results and Implementation

In this section, the proposed IR System architecture was transformed into a functional prototype designed based on the hybridization of the Probabilistic and the Vector space Modeling. Here the Case-based Optimized Intelligent IRS models were used in designing the Afaan Oromo Information Retrieval System. Later the evaluation was done and presented as presented in table 4.

In this section, the Index Construction was done for the proposed functional prototype. To

facilitate the information retrieval, the proposed IR system prototype has two integrated components i.e. indexing and searching.

In both the indexing and searching processes, the preprocessing techniques were applied (tokenizing, normalization, stop word removal, stemming) for creating an inverted file (Inverted document file) structure which in turn includes vocabulary file, stemmed file, query file, stemmed-query, and posting file. Figure 3 presents a fragment Python code that used to tokenize and standardizes the terms in the document during index construction.

```
In [*]: 1 import os
2 import re
3 import nltk
4 characters = ".,!#$%^&*();:\n\t\\\"'?!{}[]<>@123456789"
5 def tokenize(document):
6     terms = document.lower().split()
7     return [term.strip(characters) for term in terms]
8 s=open("C:/Users/USER/AppData/Local/Programs/Python/Python38-32/corpus/stopword.txt", 'r')
9 os.chdir('C:/Users/USER/AppData/Local/Programs/Python/Python38-32/corpus/doc')
10 stoplist=s.read()
11 s.close()
12 path = 'C:/Users/USER/AppData/Local/Programs/Python/Python38-32/corpus/indexfiles/'
13
```

Figure 3: Python Code for Tokenization and Normalization

As presented in the fragment code of figure 3, the documents are tokenized based on the space between the words and normalized by removing a character from it and finally change to the lower cases. In addition to this, the index terms selected

in this study are important terms were selected to create index which are not part of stop-words that identified in different research based on Afaan Oromo grammatical purpose only.

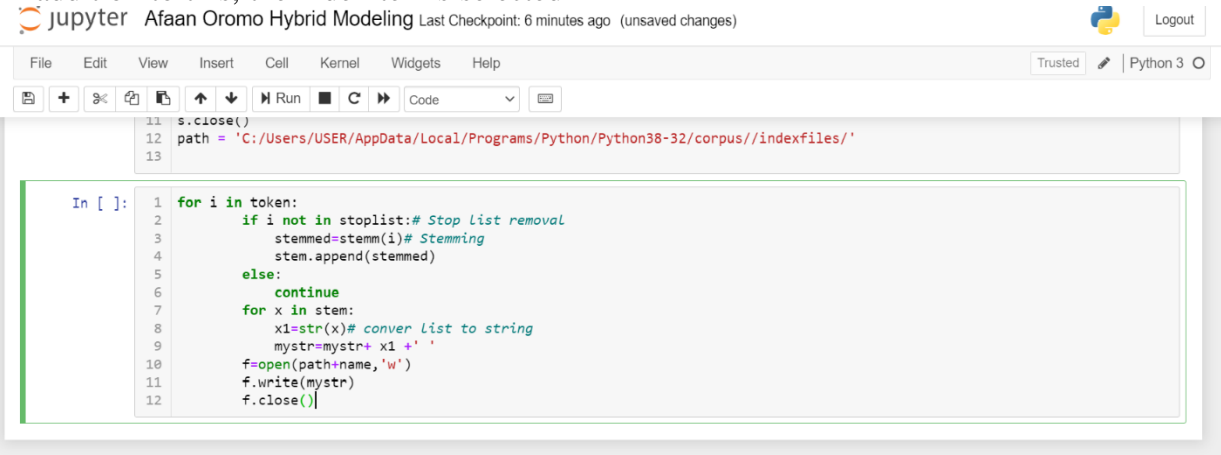


Figure 4: Python code for Stop

word removal

In this research study, an identified Afaan Oromo language stop word list (1298 words) was saved in a text separate file.. And also, different character

and numbers are removed (normalized) as presented in the figure 4 from the words then tokenized technique was applied.

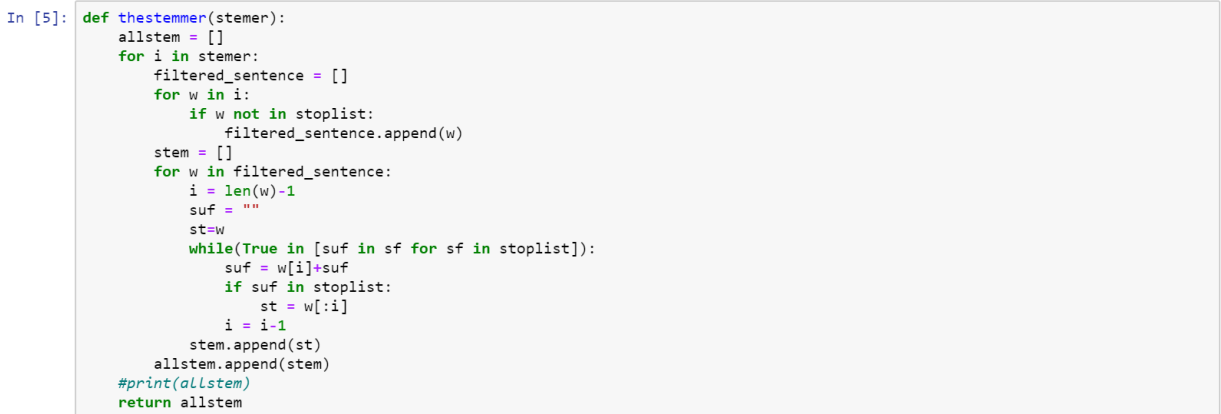


Figure 5: Afaan Oromo stemming python code

As presented in figure 5, stemmed terms were created depending on the Afaan Oromo stemmer algorithm rule. Firstly, the algorithm checked the measure of the token, then implements rules of stripping inflections of words. The inverted file

which has three separates files vocabulary, stemmed file, and posting file as presented in Table 4 and Table 5 also depicts the structure of the inverted file (vocabulary file and posting file).

Table 4: Post file

Vocabulary File		
Terms	Doc frequency	Collection frequency
AbbaaB iyyaa	2	15
Addaa	5	7

In the Information Retrieval system, the searching process is the second main component that deals with the user specified needs of information in query form and then applies the preprocessing

Table 5: Vocabulary file

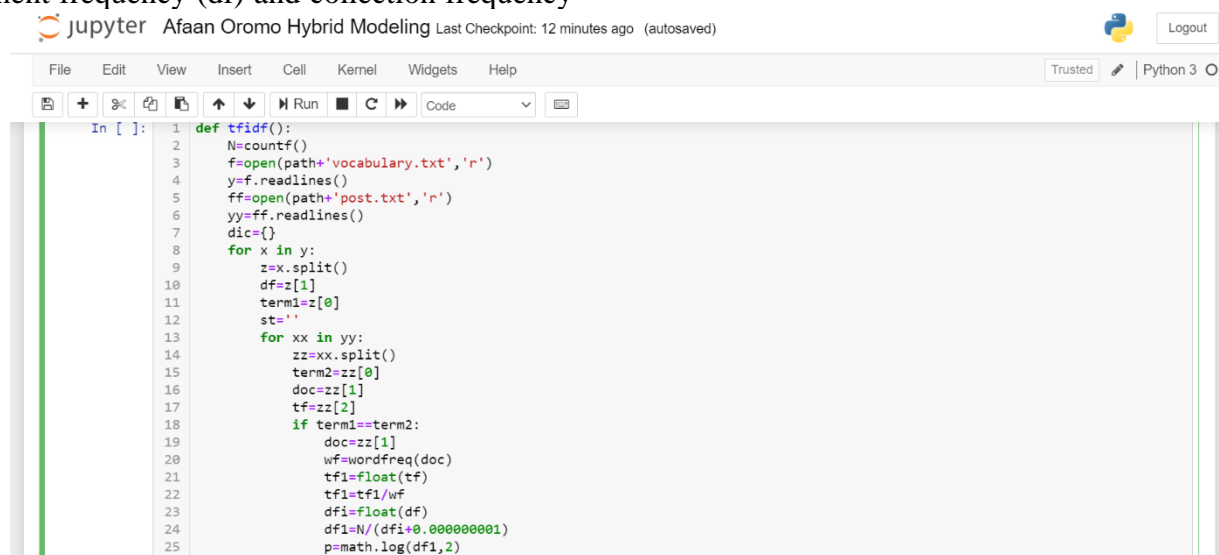
Post File			
	Doc ID	Term frequency	Location
	Doc 1	3	111,231
	Doc 10	14	214,133

techniques (tokenization, Normalization, Elimination stop word, and stemming).

2. Hybrid Model (VSM & Probabilistic) Construction

For the proposed Afaan Oromo IRS Hybrid Model (a mix of VSM and Probabilistic) a matching algorithm was applied for searching the user's information needs. In this process the created vocabulary file holds each unique free from stop-word, tokenized in the documents with its document frequency (df) and collection frequency

(CF). Both df and TF are cross-referenced into the separated file that is a posting file. Posting file holds terms with their TF, term location (positions) in the documents, and document ID that contains the term. Therefore, the search algorithm that is hybrid brings together these models in order to calculate similarity and rank documents.



```

1 def tfidf():
2     N=countf()
3     f=open(path+'vocabulary.txt','r')
4     y=f.readlines()
5     ff=open(path+'post.txt','r')
6     yy=ff.readlines()
7     dic={}
8     for x in y:
9         z=x.split()
10        df=z[1]
11        term1=z[0]
12        st=''
13        for xx in yy:
14            zz=xx.split()
15            term2=zz[0]
16            doc=zz[1]
17            tf=zz[2]
18            if term1==term2:
19                doc=zz[1]
20                wf=wordfreq(doc)
21                tf1=float(tf)
22                tf1=tf1/wf
23                df1=float(df)
24                df1=N/(df1+0.000000001)
25                p=math.log(df1,2)

```

Figure 6: Python code Integrating vocabulary and posting files for TF

The created Term Frequency and Inverted Document Frequency (TFIDF) for bearing terms (vocabulary file and posting file) which is indicated in Table 5 and Table 6 are two separate files, but both of them are used together for term weighting and similarity measurements. So, in order to implement a hybrid (VSM and Probabilistic) model, the vector of query versus document is created by using these files.

In this research, twenty (20) queries were selected for evaluating the performance of the proposed IR system for the relevance of the queries in all the documents. These queries were presented based on their term weighting and application of hybrid modeling. So, both the documents and queries were arranged as term frequency (TF), collection frequency (CF), and document frequency (DF).

Conclusion

This research paper proposed IR system architecture and its functional prototype. The hybrid approach was used for designing the functional prototype of the Afaan Oromo information retrieval (IR) system. Different IR approaches applied for the different languages of international and local levels were reviewed for

conceptual understanding of nature, structure, and pattern in comparison to the Afaan Oromo information retrieval and writing systems. The Afaan Oromo documents were collected from different sources to prepare the Corpus and different algorithms were applied to design and implement the functional prototype. The collected Afaan Oromo text documents were passed through the natural language processing such as document preprocessing methods, and term weighting calculation from indexed documents. These activities were accomplished for identifying very important terms for representation of index and searching purposes by applying the natural language processing technique. Based on the results obtained, a proposed hybrid approach was applied to order the relevant documents for the user queries. Lastly, the prepared corpus was indexed created using a file structure of the language to evaluate the proposed prototype. In this research paper, the IR system evaluation was performed over the proposed system prototype and achieved 96.8% precision, 90.0% recall, and 93.3% F-measure respectively.

Recommendation

In this research study, an attempt was made to develop a prototype using a hybrid approach for the Afaan Oromo IR system. To develop a complete IR system for Afaan Oromo, it needs a long period of time, inputs from different domain experts such as linguistics, NLP, and IT, and the large size corpus from different sources for the complete Afaan Oromo IR system. The alliance of all these different domain experts and the accessibilities of the resources created the multifold complexities to develop a fully-fledged Afaan Oromo IR system.

Therefore, the proposed IR system also needs additional performance improvement. Based on the knowledge gap obtained from the rigorous literature review and the research outcomes of this research, the following recommendations are forwarded for the future research directions.

- This research study is focused only on text documents, but other types of document like video, audio, graphics, and pictures can be included in the futures researches
- Machine learning techniques can be integrated into the video, audio, graphics, and picture documents to come up with a fully-fledged functional system.

References

- [1] A. Introduction and I. Retrieval, "Online edition (c) 2009 Cambridge UP," no. c, 2009.
- [2] J. Kekiil and I. S. Fin-, "IR evaluation methods for retrieving highly relevant documents," vol. 51, no. 2, pp. 243–250, 2017.
- [3] M. Sanderson, "Language Engineering: Christopher D . Manning , Prabhakar Raghavan , Hinrich Schütze , Introduction to Information Retrieval , Cambridge University Press . 2008 . ISBN-13 978-0-521-86571-5 , xxi + 482 pages .," no. January, pp. 100–103, 2010, doi: 10.1017/S1351324909005129.
- [4] H. Cui and J. Wen, "Query Expansion using Query," no. June 2014, 2002, doi: 10.1145/511446.511489.
- [5] F. Overview, D. Libraries, and D. Warehouses, "Introduction to Information Retrieval Systems," Inf. Storage Retr. Syst., pp. 1–25, 2005, doi: 10.1007/0-306-47031-4_1.
- [6] K. K. Tune and V. Varma, "Oromo - English Information Retrieval Experiments at CLEF 2006," pp. 25–28, 2006.
- [7] B. Saini, V. Singh, and S. Kumar, "Information retrieval models and searching methodologies: Survey," Inf. Retr. Boston., vol. 1, no. 2, p. 20, 2014.
- [8] I. Bedane, "The Origin of Afaan Oromo : Mother Language," vol. 15, no. 12, 2015.
- [9] "9 Afaan Oromo." .
- [10] L. Ballesteros and B. Croft, "Dictionary Methods for Cross-Lingual Information Retrieval."
- [11] G. G. Eggi, "Afaan Oromo Text Retrieval System," no. June, 2012.
- [12] "No Title," no. June, 2011.
- [13] D. Tesfaye, "Addis Ababa University Faculty Of Informatics Department Of Information Science Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach School Of Graduate Studies Faculty of Informatics," 2010.
- [14] R. A. Kumar, M. A. Jabbar, and Y. V. B. Reddy, "Information Retrieval systems and Web Search Engines : A Survey," no. January, pp. 1–4, 2017, doi: 10.22161/ijaers/nctet.2017.25.
- [15] T. Debela, "Addis Ababa University Department of Computer Science Afaan Oromo Search Engine November 2010.
- [16] S. Chawla, "Improving Information Retrieval Precision by Finding Related Queries with similar Information Need using Information Scent Suruchi Chawla," pp. 486–491, 2008, doi: 10.1109/ICETET.2008.23.
- [17] H. Kaur, "Indexing Process Insight and Evaluation."
- [18] H. Redwan, T. Mindaye, and S. Atnafu, "Search Engine for Amharic Web Content," no. September, pp. 1–6, 2009.
- [19] W. Tesema, "Information Technology & Software Engineering Afan Oromo Sense

- Clustering in Hierarchical and Partitional Techniques,” vol. 6, no. 5, 2016, doi: 10.4172/2165-7866.1000191.
- [20] A. Bruck and T. Tilahun, “Enhancing Amharic Information Retrieval System Based on Statistical Co-Occurrence Technique,” no. December, pp. 67–76, 2015.
- [21] A. B. Manwar, “A VECTOR SPACE MODEL FOR INFORMATION RETRIEVAL: A MATLAB APPROACH,” vol. 3, no. 2, pp. 222–229.
- [22] M. Pannu, A. James, and R. Bird, “A Comparison of Information Retrieval Models,” no. May, 2014, doi: 10.1145/2597959.2597978.
- [23] I. S. T. For, A. Society, and I. Science, “A S TEMMING P ROCEDURE AND S TOPWORD L IST General Stopword List,” 2000.
- [24] W. Ben and A. Karaa, “A NEW STEMMER TO I MPROVE I NFORMATION RETRIEVAL,” vol. 5, no. 4, pp. 143–154, 2013.
- [25] S. Zhu, “Information Retrieval using Hellinger Distance and Sqrt-cos Similarity,” no. Iccse, pp. 925–929, 2012.
- [26] O. A. S. Ibrahim, “Term frequency with average term occurrences for textual information retrieval,” *Soft Comput.*, 2015, doi: 10.1007/s00500-015-1935-7.
- [27] G. Paltoglou and M. Thelwall, “A Study of Information Retrieval Weighting Schemes for Sentiment Analysis .,” no. May 2014, 2010.
- [28] N. B. Defersha and G. Mamo, “A Two Steps Approach for Afan Oromo Nonfiction Text Categorization,” vol. 3, no. 1, pp. 107–120, 2018.