International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 4 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195414

Language Independent and Multilingual Language Identification using Infinity Ngram Approach

Kidst Ergetie Andargie, Tsegay Mullu Kassa
Department of Information and Technology, Wachemo University, Hossana, Ethiopia

ABSTRACT

Now days it is possible to get massive amount of multilingual digital information that are generated, propagated, exchanged, stored and accessed through the web each day across the world. Such accumulation of multilingual digital data becomes an obstacle for information acquisition. In order to tackling such difficulty language identification is the first step among many steps that are used for information acquisition. Language identification is the process of labeling given text content into corresponding language category.

In past decades research works have been done in the area of language identification. However, there are issues which are not solved until: multilingual language identification, discriminating language category of very closely related languages documents and labelling the language category for very short texts like words or phrases.

In this investigation, we propose an approach which able to eradicate unsolved issues of language identification (i.e. multilingual and very short texts language identification) without language barrier. In order to attain this we adopt an approach of that uses all character ngram features of given text unit (i.e. word, phrase or etc).

Moreover, the proposed approach has a capability of identify the language of a text at any text unit (i.e. word, phrase, sentence or document) in both monolingual and multilingual setting. The reason behind this capability of proposed approach is due to adopting word level features, in which every words need to be classify with regard to its language category. The infinity ngram approach uses all character ngrams of text unit together in order to label the language category of given text per word level.

In order to observe the effectiveness of the proposed approach four experimental techniques are conducted: pure infinity character ngram, infinity ngram with location feature and infinity ngram with sentence and document level reformulation. The experimental result indicates that an infinity ngram with location feature and along with sentence and document level reformulation achieves a promising result, which is an average F-measure of 100% at word, phrase, sentence, document level in monolingual setting. As well, for multilingual setting also attains an average F-measure of 100% for both sentence and document level, but for phrase level achieves 84.33%, 88.95% and 90.19% For Amharic, Geeze and Tigrigna respectively. Beside this, at word level achieves 83.16%, 80.96% and 85.85% for Amharic, Geeze, and Tigrigna respectively.

Keywords : Language Identification, Multilingual, Infinity Character Ngram, Ngram Location, Language Independent

I. INTRODUCTION

Now days, the multilingual textual data are getting more and more available on the glob network. In order to use such textual data one should know the language in which it is written or it has to translate to local language or mother tongue of an individual. To achieve this language translator is required, and such kinds of language processing tools are language dependent. Hence, there is demand of an automated tools and technique which identify the language of the written text and then select the required tools for further processing of text based on language of written text.

The solution to such problem is language identification, which is process of identifying a language in which text document is written. The problem of language identification is familiar, since one of the characteristics of being human is the ability to communicate complex and sophisticated thoughts and ideas. This is only possible through use of a common language. The research in language identification aims to attain human ability of language recognition in automatic way.

To achieve the identification of language without human intervention, in past decades a number of computational approaches have been developed through use of different algorithms and data structures. To detect the language for given text document is written effectively in automatic manner is enabling technology that increases accessibility of data.

Moreover, in order to apply natural language processing (NLP) techniques to real world data language identification is typically the first step to ensure that only documents in relevant languages are subjected for further processing. Similarly, in information storage and retrieval it is common to index documents in a multilingual collection by the language that are written in. Hence, language identification is necessary for document collection where the language of documents is not known such as data crawled from World Wide Web.

On the other hand, machine translation also needs detection of the language of a text for routing a suitable translator, since a translator to be effective first need to know language of given text. Collection of text documents on the web which gives for language identification an input may be written in only one or multiple language. When compare the two situations, processing of monolingual documents is fairly simple compared to multilingual documents, since knowing one language and knowing several languages are quiet difference.

According to [1], the main challenges of multilingual language identification is according to the following reasons:

a. Segmentation of documents: - identify the regions of a document in different languages is a problem to processing multilingual documents. In multilingual language identifier knowing language switching is big challenge, this specifies how frequently or where a shift from one language to another can occur in a document. Once the region is identified, the language of the content in that region can be easily identified and used for further processing.
b. Common words: - in very similar languages, certain words are used commonly in all languages and this makes language identification is difficult. Since, in similar languages share a great deal of lexical and grammatical features.

The task of labeling documents to a unique language, which is called monolingual language identification is solved problem [2]. Hence, now a day's multilingual language identification is a hot research area. As far as researchers knowledge there are limited number of works done on this issue. However, there are issues not solved until now on this area (i.e. identify any level of text in both monolingual and multilingual setting) [1].

In order to solve such problem we propose an approach that able to detect language of text at any level in both monolingual and multilingual setting using infinity character ngram with different features (i.e. location, sentence and document level reformulation). To examine effectiveness of proposed approach we aims to adopt on top of Ethiopian Semitic language (i.e. Amharic, Geeze and Tigrigna language), which are very similar natural languages that share same lexical and grammatical features.

The remaining part of this paper organized as follows. Section II discusses related works. In Section III, we present the proposed approach of multilingual language identification. Experimental result and discussion stated in Section IV. Section V presents conclusion and future works.

II. RELATED WORK

In this section, the most related researches and approaches in language identification are summarized. Due to long research history of this area it is difficult to give a complete review of all existing works. Hence we try to discuss on multilingual related research works in the state-of-the-art.

Prager [3] proposed model that supports multilingual language identification for textual documents. The researcher adopt vector space model in order to model the test document and language and also use cosine similarity to find a best match between a feature vector test document and language domain. The feature vectors for both parties are represented by frequency counts over byte n-grams ($2 \leq n \leq 5$) and words. This work shows how to construct vectors representative of particular combinations of languages independent of the relative proportions, and proposes a method for choosing combinations of languages to consider for any given document. One weakness of this approach is that for exhaustive coverage, this method is factorial in the number of languages, and as such intractable for a large set of languages. Furthermore, calculating the parameters for the virtual mixed languages becomes unfeasibly complex for mixtures of more than 3 languages.

Teahan [4], proposed an approach of language identification suitable for multilingual documents setting. The approach first segmenting the text into monolingual blocks and detecting the language using an eight word window size. The labelling of each blocks is done based cross entropy calculated using a fixed order character based markov model and experimental result indicates the segmenting of given text and language labeling in this work achieves an accuracy of over 99%. This experimental result shows that the method are significant improvement over existing literature works related to this area.

Rehurek and Kolkus [5], proposed a framework that segments input text into monolingual blocks and perform language segmentation by computing a relevance score between terms and languages smoothing across adjoining terms and finally identifying points of transition between high and low relevance, which are considered as boundaries between languages. The researchers try to overcome some issues of language identification like on very short texts and multilingualism issues. The experimental result of this finding indicates the proposed approach achieves a better result on addressing these language identification limitations.

Yamaguchi and Tanaka-Ishii [6] use a minimum description length approach, embedding a compressive model to compute the description length of text segments in each language. They present a linear-time dynamic programming solution to optimize the location of segment boundaries and language labels. Their data was artificially created by randomly sampling and concatenating text segments (40-160 characters) from monolingual texts. The experimental result of this study achieves an F-scores for language identification and with 40 to 120 characters segmentation were 0.98 and 0.94 respectively. This approach concurrently detects multilingual documents and segments them by language, but the approach is computationally very expensive.

Kind and Abney [7] proposed a word level language identification model through tokenizing the text into words and classify the language to the corresponding language category. In this approach, each words in the document is labelled with a specific language. To achieve this researchers used a conditional random fields and introduce a technique to estimate the parameters using only monolingual data , an important consideration as there is no readily-available collection of manually-labeled multilingual documents with word-level annotations. The experimental result indicates a conditional random field model trained with generalized expectation criteria was the most accurate and performed consistently as the amount of training data was varied. Finally, researchers specifically mention the need for an automatic method to examine multilingual language identification with high accuracy.

Nguyen and Dogruoz [8] introduce a two-pass approach to processing Turkish-Dutch bilingual documents. To do so first each word of document label to language category independently and the second pass uses the local context of a word to further refine the predictions for better effectiveness. The experimental result indicates the approach achieves a promising result, an accuracy of 98%. This result reveal that language models are more robust than dictionaries and adding context improves the performance. They evaluate their methods from different perspectives based on how language identification at word level can be used to analyze multilingual data.

III. PROPOSED SYSTEM

A. System Architecture

The main goal of the proposed model is to identify a language label of each word in a text. To do so, this study develop framework which identify the language of document written in one language (i.e. monolingual documents) and textual document written in more than one language (i.e. multilingual documents). As shown in Figure 1, the proposed architecture for language identification for multilingual setting is structured into two main phases, training and testing phase.

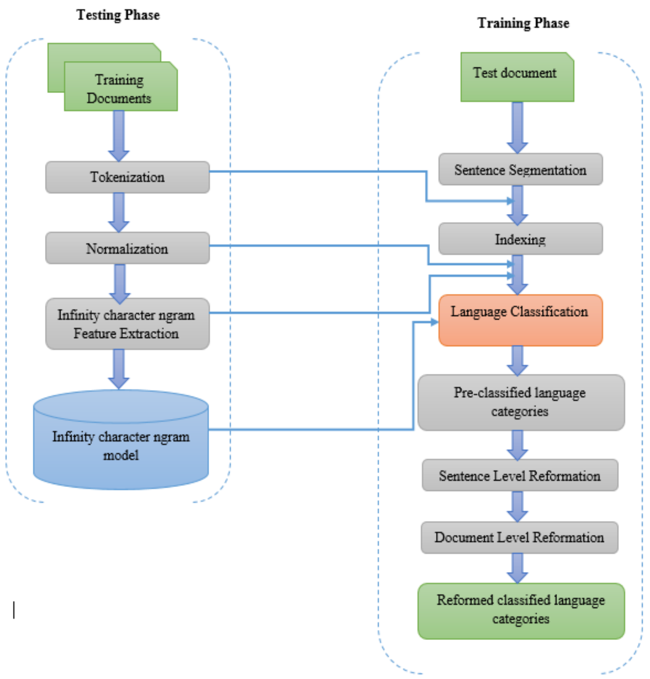

Testing Phase
Training Documents
Tokenization
Normalization
Infinity character ngram Feature Extraction
Infinity character ngram model
Training Phase
Test document
Sentence Segmentation
Indexing
Language Classification
Pre-classified language categories
Sentence Level Reformation
Document Level Reformation
Reformed classified language categories

Figure 1: General Architecture of Proposed Multilingual Language Identifier

A. Sentence Segmentation

This module is a preliminary step and responsible to divide a string of input text into meaningful units called sentences. As shown in Figure 1. This module is fundamental particularly intended for testing phase, since identifying and indexing of sentence for the given text document is relevant for sentence level reformation module. To do so, an Ethiopian sematic language sentence boundary markers, which are ‘።’, ‘?’, ’!’, ‘¡’ are used. In this approach a text document with single word or phrase, meaningful components of sentence and has not end of sentence markers considered as a single sentence to make suitable sentence indexing operation.

B. Tokenization

This module is deals with sentence segmentation into words, units that are meaningful for distinguishing between languages. Since, in this investigation language identification is done for each individual words independently, each bag of sentence provided by previous module tokenized into bag of words. As shown in Figure 1, this module is used for both training and testing phase. In order to attain we substitute all occurrences of multiple white spaces with a single white space , and then split it by word separator i.e. white space character.

C. Indexing

This module is responsible to provide index information at both sentence and word level of given test document. During sentence segmentation index is assigned for each sentence in text document and this information is very important to perform sentence level reformation, which is language category improvement at sentence level. The index is assigned for each sentence of text sequentially with integer in ascending manner (i.e. 1 … n) and this is used as unique identification of each sentence in document. Beside this, the index information is also provided to tokens of test document after tokenization module is done. But, this index information contains an actual position of each words in text and later used to identify the language switching point during language labelling.

D. Normalization

As shown in Figure 1, this module is shared for both training and testing phase and it is concerned to normalize the homophone characters in Ethiopia Sematic language. In Geez writing system there are characters having same pronunciation but different symbols and consideration of these characters as different reduces effectiveness of our language identification task. Hence, in this study this normalization process replace those homophone characters through representative and common character symbols. Beside this, particularly for training phase this module is also responsible to clean all unnecessary characters (i.e. all special characters, digits) to build language profiles. Since these characters are a member of all Ethiopian sematic languages but not unit of particular language. However, nothing is removed during testing phase. Since all parts of text document including these unnecessary characters are expected to be label with one of the language category.

E. Infinity Character Ngram Extraction

An extracted ngram features for a given word are more in number or rich a reliability of language identification at word level is effective. Hence, to make the proposed approach effective for multilingual language detection at word level extraction of rich ngram features for a word is required. To do so, we used an approach that used a combination of all character ngram size features of a word in once called infinity character ngram. This approach have been introduced for text classification [9], which extracts all character ngrams of a string as feature for document classification task. This researchers uses this novel approach, since tokenized words are not enough for determining class of a document, ultimately through experiment learning a classifier by using all character ngrams achieves a better result [9].

Therefore, in order to get such benefit of infinity character ngram, we adopt this novel approach for multilingual language identification at word level. During this approach the size of ngram used to extract character sequence for both training and testing is not fixed rather it depends on the word length. The size of character ngram (n) in infinity ngram vary from word to word and it depends on a word length (w), maximum at n = |w|. So, all ngram types range from 2 to |w| are extracted to represent a given word with length w maximum of ngram. This novel approach produces a considerably rich character ngram feature set in comparison with fixed length character ngram representation and this is enhances effectiveness of language identification at word level.

Furthermore, in order to capture the entire word for these short words with infinity ngram, we pad each word with one special character to denote the beginning and end of a word, and use infinity character ngrams extracted from these modified words. For example, from the Amharic word ና, we derive the infinity character ngrams *$ና*, *$ና#*, and *ና#*, with *$ና#* indicating that the entire word is represented. To clarify capability of infinity ngram for extraction of rich character ngram features per word level take Amharic word “እንዳይከሰት” and for this word it is possible to extract 28 numbers of character ngram features as shown below in Figure 2.

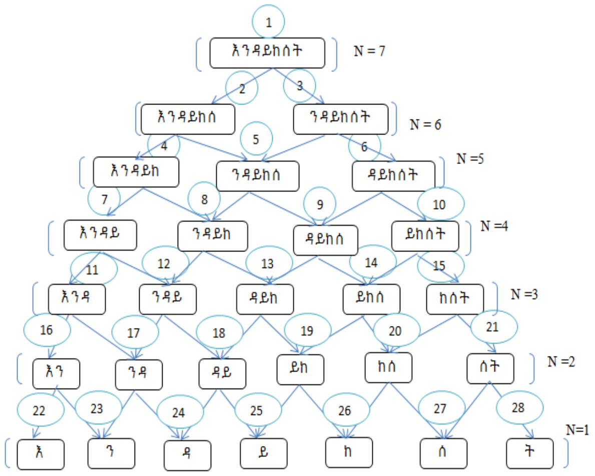


Figure 2: Infinity character ngram representation of word “እንዳይከሰት”

In our study ngram with size of N >=2 is taken to represent a given word, since for N = 1 all Semitic languages are similar pattern, not useful ngram type for our word level language identification task. As explained above, the size of N may vary from word to word, it depends on the string length of a given word, means that it able to extract Ngram N = 2…. |w| as shown in above Figure 2. So, this novel approach used in our study to extract all character ngram types for a given word.

We experimented with a wide range of values to weight the ngrams extracted from the infinity ngram approach, since as showed in above example Figure 2, as N (ngram size) increase, the more capability to express the given word. So, a weighting factor is taken for each character ngrams with N-1. Surprisingly this weighting factory enhances the performance of our infinity ngram approach. We used N-1 as a weighting factor and we name this ngram weight W$_{ngram}$ for the weight of the decision on the language label of a given word. Finally, the probability of each ngram extracted by infinity ngram depending on the ngram size N can be computed as

$$\text{Prob}_{ngram} * (N-1) \qquad (1)$$

Where Prob$_{ngram}$ is an actual ngram probability taken from target language profile and N is ngram size used to extract ngram features from a given word. As mentioned above this weighting value vary depending the size of ngram N, as N becomes large the ngram weight increase and this enhances the weighted probability of a given ngram , means that the ngram has more expressive power on predicting a language category of a given word.

From previous example for Amharic word “እንዳይከሰት”, to compute the ngram probability of a given word all ngram types having ngram size N >= 2 are taken. As explained before, the words “እንዳይከሰ” ,” ንዳይከሰት” with ngram size N = 6 is more expressive the actual word “እንዳይከሰት” than the ngram types “እንዳይ”, “ንዳይከ”,”ዳይከሰ”,”ይከሰት” with ngram size N = 4. So, the language profile having best probability with ngram type with N = 6 is more probable to be language category of given word “እንዳይከሰት” than N = 4 ngram types and to keep this during ngram probability computation of all ngram types is vary depending on size of N.

For N = 6

Prob$_{ngram}$ * (N-1) => Prob$_{ngram}$ * 5

For N =4

Prob$_{ngram}$ * (N-1) => Prob$_{ngram}$ * 3

During our infinity ngram, the language profile as well as words extracted from tested document is represented with all ngrams depending on the word length. So, in order to speed up our searching process of ngram types along each of language profiles we used ngram type length as index. The searching process is the most important and potentially the most time consuming activity in the whole process, since all ngram type extracted from a given word with different N size need to be checked against all ngrams in the language profile. So, in order to enhance our searching process, we used length of ngram types as index. The idea is that ngram type extracted from a given word having N = 3 need to be checked in only with those ngram N = 3 in a language profiles and this makes our searching process faster, since there is no point in search for ngram size N = 3 among other ngram sizes where it will never be found.

Beside this, to speed up the process by organizing the character ngram language profile information by its ngram length as index and this would be a one-time process such that once the language profile is indexed in this way it is only updated as and when necessary.

In order to include those ngrams which are more discriminative for a given word, we exclude ngram size of 1, since it is less discriminative feature for our word level language identification task. Beside this, to achieve language identification at word level using infinity ngram, the relative frequency or probability of each ngram is computed. The ngram probability of each n-gram X_i in a language L_j is computed by a formula in Equation 2.

$$Prob(X_i^j) = \frac{f(X_i^j)}{\sum_{i=1}^{n} f(X_i^j)} \quad (2)$$

Where, $f(X_i^j)$ is the frequency of ngram X_i in the language L_j and $\sum_{i=1}^{n} f(X_i^j)$ is the total sum of ngram occurrence in language L_j.

During language profile construction with infinity ngram the number of times each ngram occurs in the training corpus of each language is computed. It outputs the relative frequency or weight of each unique ngram using a formula in equation 2. This formula also used for both infinity ngram approaches with and without ngram location feature set to compute ngram location relative frequency.

In this investigation, in order to consider the contextual information of current word in a test document during language labelling decision, ngram probability of previous word in each of the target language in a set of domain language is added to the computed ngram probability of current word for a target language and we call this probability as augmented probability. This can be expressed mathematically as following.

$$AgmProb\ (X_i^j) = \frac{Prob(X_{i-1}^j) + Prob(X_i^j)}{2} \quad (3)$$

Where $AgmProb\ (X_i^j)$ is improved probability of a current word X_i for a set of target language Lj,$Prob(X_{i-1}^j)$ is ngram probability of a previous word in language Lj and $Prob(X_i^j)$ refers to the ngram probability of current word X_i in language Lj.

We can observe that, the first word of test document cannot have a previous word and thus no an improvement of ngram probability of a word with contextual information as we explained in equation (3). So, in this case the ngram probability of a word is not improved, rather ngram probability of a word is taken.

F. Classification

This module is concerned to correctly guess the language in which each word of a document is written. To do so, the distance between each word in document with regard to the language models are calculated and the language with minimal distance to the word of a document is chosen as the language of the word.

In this study for classification purpose Bayesian classifier is adopted, which uses the concept of Bayes' theorem [9]. This classifier assigns the most likely classes to an input string based on the highest posteriori probability of given input string. For language identification purpose, a naïve Bayes classifier constructed using ngrams as features. Let T be a set of training samples and each sample be represented by n feature vectors, X = x1, x2… xn, with their class labels. Let there be n classes: L1, L2….Lm. to predict, a sample Xn is selected to belong to class Li, if and only if:

$$P\left({}^{Li}/_{X} \right) > \ P\left({}^{Lj}/_{X} \right); for\ 1\ \leq j\ \leq n; j\ \neq i \quad (4)$$

Where $P({}^{Li}/_{X})$ is the probability of a class Li given a sample X. Bayes' theorem states that:

$$P({}^{Li}/_{X}) = \frac{P(X/Li)P(Li)}{P(X)} \qquad (5)$$

Where $P({}^{Li}/_{X})$ represents the likelihood of a sample X belonging to class Li, and P(X) does not influence model comparison.

The class a priori probability P (Li) represents the count relative frequency in the language profiles, so that P (Li) can be omitted as well. According to the Naive Bayes assumption, statistical independence of features is assumed, and the class Li is selected such that $\prod P(xj \; / \; Li)P(Li)$ is optimized, where $P(xj \; / \; Li)$ is then the likelihood of a specific ngram being observed in a given language profile, and the word being classified consists of j n-grams.

During language classification there may be rare or unseen ngrams which can result in poor probability estimates. In this investigation in order to eliminate this problem additive smoothing technique [10] is adopted because of its simplicity of implementation and suitable for our proposed language identification task.

G. Sentence Level Reformulation

As shown in Figure 1, after the classification module each words of a document is assigned to one of language category in set of domain language and we called this classification result as pre-classified language categories. However, due to very similarity of Ethiopian Semitic languages there is a probability of wrong language category assignment to a given word. To eradicate this problem in our proposed approach we include a module called sentence level reformulation. The sentence level language reformulation is transformed pre-classified language categories into dominance language category if and only if the average occurrence of a particular language is equal or above the defined threshold value per sentence level. The threshold value used to for this decision is selected through the experiment and the average occurrence of each set of language labelled in a given sentence is computed with formula in equation 6

$$avg(L_i^j) = \frac{f(L_i^j)}{\sum_{i=1}^{n} f(L_i^j)} \qquad (6)$$

Where, $avg(L_i^j)$ is the occurrence of language L_i in the sentence $_j$ and $\sum_{i=1}^{n} f(L_i^j)$ is the total sum of language occurrence in the sentence $_{j.}$

This module is concerned to compute language dominance at sentence level from pre-classified language categories result. When a dominance of a particular language satisfies a specified language dominance threshold value then the language categories of each word within a sentence is reformed to dominant language category.

H. Document Level Reformulation

This module is enable the proposed approach effective language identifier in monolingual document setting. To achieve this, after the sentence level reformulation is done, this module is devoted to compute the language reformulation at document level.

Document level reformulation is the process of adding improvements on a language category result reformed by previous sentence level reformulation module through making adjustment at document level as whole. Since, the given test document may be monolingual and this helps to adjust incorrect language labelling of monolingual documents into more than one language category. When a dominance of a particular language occurrence satisfied document level language dominance threshold value then the language category of each word with in text document as whole is reformed to a single dominant language category. The document level threshold value used for document level adjustment is defined based on the experiment result. The average occurrence of each set of language labelled in a given document is computed with formula in equation 7.

$$avg(L_i^j) = \frac{f(L_i^j)}{\sum_{i=1}^{n} f(L_i^j)} \qquad (7)$$

Where, $avg(L_i^j)$ is the occurrence of language L_i in the document $_j$ and $\sum_{i=1}^{n} f(L_i^j)$ is the total sum of language occurrence in the document $_j$.

IV. EXPERIMENT

A. Dataset Collection

For evaluating the effectiveness of proposed language identifier at different level and setting, testing data set is required and to split the total corpus into training and testing data set we used 10-fold cross validation. In this technique dataset is split into 10 mutually exclusive subsets of approximately equal size for each language ash sown in Table 1 and ten iterations were used to conduct the experiments. For each iteration, we isolated one part of the dataset for testing while retaining the remaining nine parts as the training set.

Tests	Amharic # word	Geez # word	Tigrigna # word
Test 1	91,655	55,070	78,726
Test 2	92,870	54,244	77,366
Test 3	91,443	54,421	78,205
Test 4	91,023	52,524	75,226
Test 5	92,786	54,332	75,233
Test 6	91,884	57,063	77,129
Test 7	91,074	51,556	75,430
Test 8	91,470	52,632	72,412
Test 9	92,794	53,300	72,404
Test 10	92,559	55,711	73,033
Average	**91,955.8**	**54,085.3**	**75,516.4**

Table 1: Statistics of test data corpus

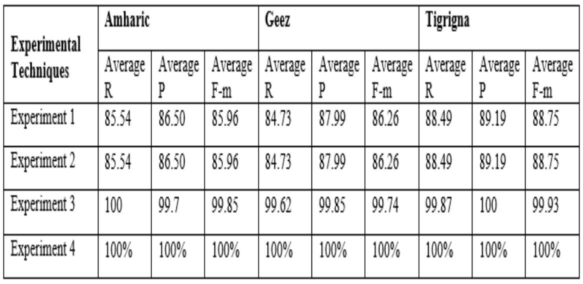
B. Experimental Result and Discussion

In this work, four experiments were conducted to evaluate the effectiveness of the proposed approach with different features: pure infinity character ngram, infinity character ngram with location feature, infinity character ngram with location feature and sentence level reformulation, and infinity character ngram with location feature, sentence and document level reformulation. Beside this, in order to observe the effectiveness comparison of all features of proposed approach we used both monolingual and multilingual document setting at four different test levels (i.e. word, phrase, sentence and document) for experimentation.

The evaluation metrics to measure effectiveness of proposed approach is done by comparing the number of words which are labelled the language category correctly and incorrectly. To achieve this, language labelling for each test document words are done manually and used as reference to cross check with final result of proposed language identifier. Among the different methods of evaluation techniques, in this study we adopt Precision (P), Recall (R) and F-measure (F-m) evaluation parameters.

The main intention of this experiment is in order to compare all variety features used with proposed approach at multilingual document settings. The experimental result illustrates the effect of all experimental techniques (i.e. Experiment 1 for infinity character ngram, Experiment 2 for infinity character ngram with location feature, Experiment 3 for infinity character ngram with location feature and sentence level reformulation, Experiment 4 for infinity character ngram with location feature, sentence and document level reformulation).

Table 2 illustrates to measure effectiveness of proposed approach with all experimental techniques at document level of test input and multilingual document setting.

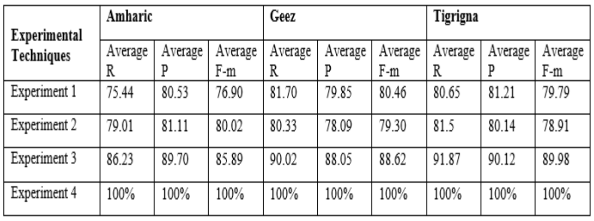


Experimental Techniques	Amharic			Geez			Tigrigna		
	Average R	Average P	Average F-m	Average R	Average P	Average F-m	Average R	Average P	Average F-m
Experiment 1	85.54	86.50	85.96	84.73	87.99	86.26	88.49	89.19	88.75
Experiment 2	85.54	86.50	85.96	84.73	87.99	86.26	88.49	89.19	88.75
Experiment 3	100	99.7	99.85	99.62	99.85	99.74	99.87	100	99.93
Experiment 4	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 2: Experimental result of proposed language identifier with different feature sets at document level with multilingual document setting

From this experimental result, we observe effectiveness of proposed language identifier at document level with combination of all features (i.e. combination of infinity character ngram with location feature, sentence and document level reformulation), that achieve an average F-measure of 100% for all supported Ethiopian Semitic language (i.e. Amharic, Geeze and Tigrigna).

On the other hand, the experiment also done for all experimental techniques at sentence level at multilingual document setting and experimental result is stated in below Table 3

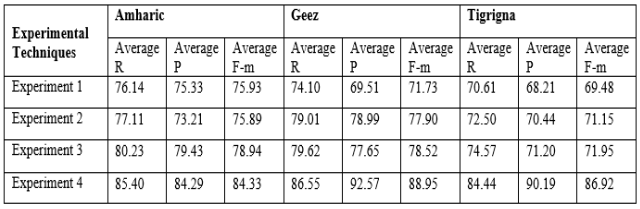


Experimental Techniques	Amharic			Geez			Tigrigna		
	Average R	Average P	Average F-m	Average R	Average P	Average F-m	Average R	Average P	Average F-m
Experiment 1	75.44	80.53	76.90	81.70	79.85	80.46	80.65	81.21	79.79
Experiment 2	79.01	81.11	80.02	80.33	78.09	79.30	81.5	80.14	78.91
Experiment 3	86.23	89.70	85.89	90.02	88.05	88.62	91.87	90.12	89.98
Experiment 4	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 3: Experimental result of proposed language identifier with different feature sets at sentence level with multilingual document setting

Similarly the above experimental result (i.e. Table 3) indicates proposed approach with all features and components achieves an average F-measure of 100% effectiveness of language identification for multilingual document setting at sentence level. This is due to inclusion of sentence level reformulation and other feature in the proposed approach. However, the document level reformulation has not any factor during sentence level multilingual document setting, since test document at multilingual setting and sentence level not reach the document level dominance threshold value.

Moreover, we also conduct an experiment for all experimental techniques at phrase level with multilingual document setting and result is illustrated in below Table 4.

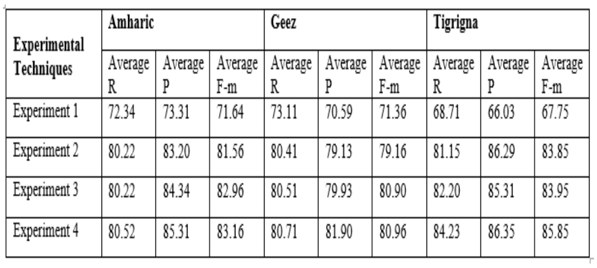


Experimental Techniques	Amharic			Geez			Tigrigna		
	Average R	Average P	Average F-m	Average R	Average P	Average F-m	Average R	Average P	Average F-m
Experiment 1	76.14	75.33	75.93	74.10	69.51	71.73	70.61	68.21	69.48
Experiment 2	77.11	73.21	75.89	79.01	78.99	77.90	72.50	70.44	71.15
Experiment 3	80.23	79.43	78.94	79.62	77.65	78.52	74.57	71.20	71.95
Experiment 4	85.40	84.29	84.33	86.55	92.57	88.95	84.44	90.19	86.92

Table 4: Experimental result of proposed language identifier with different feature sets at phrase level with multilingual document setting

As we have seen from above experimental result in above Table 4, proposed approach with inclusion of all features achieve better language identification result at phrase level multilingual document setting for all supported languages.

Finally we also conduct an experiment for proposed approach at word level multilingual document setting and effectiveness result indicates almost similar for Experiment 2, 3 and 4. This due to lack of re-adjustment of the language category result at word level , since the threshold value is not fulfil the sentence and document level dominance threshold value.



Experimental Techniques	Amharic			Geez			Tigrigna		
	Average R	Average P	Average F-m	Average R	Average P	Average F-m	Average R	Average P	Average F-m
Experiment 1	72.34	73.31	71.64	73.11	70.59	71.36	68.71	66.03	67.75
Experiment 2	80.22	83.20	81.56	80.41	79.13	79.16	81.15	86.29	83.85
Experiment 3	80.22	84.34	82.96	80.51	79.93	80.90	82.20	85.31	83.95
Experiment 4	80.52	85.31	83.16	80.71	81.90	80.96	84.23	86.35	85.85

Table 5: Experimental result of proposed language identifier with different feature sets at word level with multilingual document setting

V. CONCLUSION AND RECOMMENDATION

The digital documents written in different language getting more and more available on the global network and in order to use this content for further processing language identification is required. To solve this, in past decades a number of research works have been conducted in the area of language identification, but there are issues still not solved (i.e. language identification for multilingual documents , for very closely related languages and for very short texts like at words or phrase level).

Hence, in order to eradicate such language identification difficulty this investigation proposed an infinity character ngram approach to identify the language of a text at word level. This feature enables the proposed model to classify the document language category in different levels (i.e. word, phrase, sentence and document). Moreover, it also capable of identify the language of a document in any document setting (i.e. monolingual and multilingual). In order to train the language identifier no need of multilingual dataset rather it needs only any monolingual row text of all supported language. Hence, this makes the proposed approach very flexible to extend and include other language domains.

In this investigation, the corpus of each language is divided into training and testing data set. The training set for infinity character ngram consists of 90% of the corpus and the testing set consists 10% of the corpus. As explained before, we conduct four experiments through combining different features of proposed approach. The experimental result is evaluated based on basic evaluation metrics: precision, recall and F-measure. Ultimately, based on the experimental result the combination of location feature set in infinity character ngram with sentence and document level reformulation achieves better result, which is an average F-measure of 100% for word, phrase, and sentence and document level in monolingual document setting. As well, for multilingual setting also attains an average F-measure of 100% for both sentence and document level, but for phrase level achieves 84.33%, 88.95% and 90.19% For Amharic, Geeze and Tigrigna respectively. Beside this, at word level achieves 83.16%, 80.96% and 85.85% for Amharic, Geeze, and Tigrigna respectively. Finally, based on this finding researchers recommend infinity ngram at character and word level for other classification tasks.

VI. REFERENCES

[1]. Marco Lui, Jey Han Lau and Timothy Baldwin. (2014). “Automatic Detection and Language Identification of Multilingual Documents”, Transactions of the Association for Computational Linguistics, pp. 27–40.

[2]. Hughes, Baden, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay, (2006). “Reconsidering language identification for written language resources“, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Italy, Genoa, 485–488, pp.

[3]. Prager, John M. (1999). Linguini: language identification for multilingual documents. In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences (HICSS-32), Maui, USA.

[4]. Teahan, W. J. (2000). Text classification and segmentation using minimum crossentropy. In Proceedings of the 6th International Conference Recherche d’Information Assistee par Ordinateur (RIAO’00), 943–961, College de France, France.

[5]. Řehûrek R, Kolkus M. (2009) Language Identification on the Web: Extending the Dictionary Method. In Computational Linguistics and Intelligent Text Processing, 10th International Conference, CICLing 2009, Proceedings. Vyd. první. Mexico City, Mexico: Springer-Verlag, 2009. ISBN 978-3-642-00381-3, pp. 357-368.

[6]. Yamaguchi, Hiroshi, and Kumiko Tanaka-Ishii. (2012). Text segmentation by language using minimum description length. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), 969–978, Jeju Island, Korea.

[7]. King, Ben, and Steven Abney. (2013). Labeling the languages of words in mixed language documents using weakly supervised methods. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1110–1119, Atlanta, Georgia.

[8]. Nguyen, Dong, and A. Seza Dogruoz. (2013). Word level language identification in online multilingual communication. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), 857–862,bSeattle, USA.

[9]. Disuke O. Jun I.T. (2009).Text Categorization with All Substring Features.

[10]. S. F. Chen and J. Goodman. (1996) “An empirical study of smoothing techniques for language modeling,” in Proc. ACL, pp. 310–318.

[11]. Zampieri, Binyam Gebrekidan Gebre, and Holland Nijmegen. (2012). Automatic identification of language varieties: The case of Portuguese. In Proceedings of The 11th Conference on Natural Language Processing (KONVENS 2012), 233–237, Vienna, Austria.

[12]. Elfardy, Heba, and Mona Diab. (2013). Sentence level dialect identification in Arabic. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 456–461, Sofia, Bulgaria.

[13]. Diwersy, Sascha, Stefan Evert, and Stella Neumann. (2014). A weakly supervised multivariate approach to the study of language variation. In Aggregating Dialectology, Typology, and Register Analysis. Linguistic Variation in Text and Speech, ed. by Benedikt Szmrecsanyi and Bernhard Wälchli. Berlin: De Gruyter.

[14]. Zampieri, Marcos. (2013). Using bag-of-words to distinguish similar languages: How efficient are they? In Proceedings of the 2013 IEEE 14th International Symposium on Computational Intelligence and Informatics (CINTI), 37–41, Budapest, Hungary.

Cite this article as : Kidst Ergetie Andargie, Tsegay Mullu Kassa, " Language Independent and Multilingual Language Identification using Infinity Ngram Approach", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 4, pp. 217-228, July-August 2019.
Available at
doi : https://doi.org/10.32628/CSEIT195414
Journal URL : http://ijsrcseit.com/CSEIT195414