2021-07

AUTOMATIC IDIOM RECOGNITION MODEL FOR AMHARIC LANGUAGE

ANDUAMLAK, ABEBE FENTA

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTING

DEPARTMENT OF COMPUTER SCIENCE

MSC THESIS

AUTOMATIC IDIOM RECOGNITION MODEL FOR AMHARIC LANGUAGE

ANDUAMLAK ABEBE FENTA

BDU1100016

JULY, 2021

BAHIR DAR, ETHIOPIA

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTING

DEPARTMENT OF COMPUTER SCIENCE

AUTOMATIC IDIOM RECOGNITION MODEL FOR AMHARIC LANGUAGE

Anduamlak Abebe Fenta

A thesis was submitted to the school of Graduate Studies of Bahir Dar Institute of Technology, BDU in partial fulfillment of the requirements for the degree of MASTER in Science degree in the School of Computing.

Advisor Name: Seffi Gebeyehu (Assis. Prof)

Bahir Dar, Ethiopia

July, 2021

DECLARATION

I, the undersigned, declare that the thesis comprises my work. In compliance with internationally accepted practices, I have acknowledged and refereed all materials used in this work. I understand that non-adherence to the principles of academic honesty and integrity, misrepresentation/ fabrication of any idea/data/fact/source would constitute sufficient ground for disciplinary action by the University and can also evoke penal action from the sources which have not been properly cited or acknowledged.

Name of the student Anduamlak Abebe Fenta

Signature

Date of submission: July ,2021

Place: Bahir Dar

BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
FACULTY OF COMPUTING

Approval of thesis for defense result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student **Anduamlak Abebe Fenta** Signature________ Date 19-11-2013

As members of the board of examiners, we examined this thesis entitled **"Automatic Idiom Recognition Model for Amharic Language"** by **Anduamlak Abebe Fenta** We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of science in "**Computer Science**"

Board of Examiners

Name of Advisor	Signature	Date
Seffi Gebeyehu(Assis. Prof)		26/7/2021
Name of External examiner	Signature	Date
Seid Muhie Yimam (Ph.D)		17.07.2021
Name of Internal Examiner	Signature	Date
Abinew Ali Ayele(Assis. Prof)		26/07/2021
Name of Chairperson	Signature	Date
Gebeyehu Belay (Dr. of Eng)		26/07/2021
Name of Chair Holder	Signature	Date
Haileyesus A.		26/07/2021
Name of Faculty Dean	Signature	Date
Belete B.		19/11/2013 E.C

[iv]

Faculty Stamp

To my families, who helped me

ACKNOWLEDGMENT

First and foremost, we thank GOD and his mother, St. Mariam, for keeping me safe. GOD and his mother, St. Mariam, have helped to solve many issues; without them, nothing would be possible. Second, it would not have been possible without the generous support and assistance of many individuals and organizations; we thank all individuals and organizations for their support, but especially Mr. Seffi Gebeyehu (Assis. Prof), for providing invaluable knowledge, guidance, attention, and time to complete this study. Last but not least, I am grateful to our family, particularly my wife Melkam Enyew, for their unwavering support. Finally, I am grateful to my classmates and best friends for their assistance in offering needed resources and direction, as well as research information, attention, and time to complete the research.

ABSTRACT

Idiomatic expressions are a natural part of all languages and a common part of our everyday conversation. It is difficult to understand the meaning of idioms since they cannot be deduced directly from the word which they are created. Natural Language Processing researche has been influenced by the existence of idioms. It has been shown that idiom affects NLP researches such as machine translation, semantic analysis, sentiment analysis, information retrieval, question answering and next word prediction. Other languages like English, Chinese, Japanese, Indian idioms are identified through different methods in different researches, but for the Amharic language, there is no research to identify idioms. Since there is no standard model for identifying Amharic idioms, this study aimed to develop an idiom identification model for the Amharic language using a supervised machine learning approach. One thousand datasets are collected from Amharic idiom books "የአማርኛ ፈሊጦች" and different Amharic documents. Vector representation of expressions using python programming was used to prepare a compatible dataset for the identification model. We contributed that digitalized the hard copy Amharic idiom book to computerized manner and used different concerned bodies as a look up table to do their own NLP task. This model helps NLP researchers to decide the phrases are idiomatic or literal. The developed model achieved a 97.5% accuracy result in the testing dataset when we employed the KNN algorithm.

Keywords*: idiom recognition, የአማርኛ ፈሊጦች, Amharic idioms*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST of ABBREVIATIONS

2D	Two Dimension
BDU	Bahirdar University
BNC	British National Corpus
BoW	Bag of Words
CPU	Central Processing Unit
GB	Giga Byte
GHz	Giga Hertz
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
MWEs	Multiword Expressions
NLP	Natural Language Processing
NVC	Noun Verb Construction
PCA	Principal Component Analysis
POST	Part Of Speech Tag
RAM	Random Access Memory
SMT	Statistical Machine Translation
SVM	Support Vector Machine
TB	Tera Byte
TF-IDF	Term frequency and Inverse Document Frequency
VNC	Verb Noun Construction
WSD	Word sense disambiguation

CHAPTER ONE

1. INTRODUCTION

1.1. Background

Idiomatic expression is one of the ways of expression, which is made from figurative words. Idiomatic expressions are collecions of words that have a common meaning that is unrelated to the individual word's meanings. Idioms cannot be interpreted from the word which it is formed from directly (Akililu and Worku 1992; Lowri et al. 2015)

Idiomatic expressions are important natural parts of all languages and prominent parts of our daily speech. Idioms are considered as one of the hardest and most interesting parts of Amharic vocabulary. But, they are considered as one of the most peculiar parts of the language; on the other hand, they are difficult because of their unpredictable meanings by people's which are unfamiliar to the nature or meaning of idiomatic expressions (Caillies 2007; Mantyla 2004; Salton 2017). In addressing idioms and idiomatic expressions, it is notable that as idioms are part of the culture, people may not understand the meaning of an idiom because its meaning cannot be determined by knowing the meaning of the words that form it, and all people are not familiar with idioms. The meaning of an idiom is not simply the joint meaning of the individual words. For example, the expression ሰማዩ ተደፋበት (the sky lies on him) has an idiomatic meaning (he confused) that has nothing to do with the meaning of ሰማዩ or ተደፋበት.

Identifying the idioms from literals requires studying both the language part and the mechanisms that used for the automation of NLP expressions. Idioms are one of the main components of a language, developing an algorithm and model for identifying idioms is of great importance to enhance NLP related researches. When it comes to the importance of idioms, without them, languages become boring, because words are the framwork of a language, while idioms are its essence (Ahmadi 2017). As a result, the incorrect algorithm and model result in incorrect recognition, which affects the language's essence.

Since idioms are expressions with special features, recognition is an important and interesting area of study.

Idioms in other languages, such as English, Chinese, Japanese, and Indian, are identified using a variety of approaches in various studies, however there is no such study for the Amharic language. This study used a supervised machine learning approach to construct an idiom identification model for the Amharic language since there is no standard model for identifying Amharic idioms.

1.2. Problem Statement

Most of the Ethiopian languages including Amharic idioms are not collected and well organized yet. Almost all of the Amharic idioms and their definitions are stored manually (on papers), which is difficult to obtain and use easily in digital form. In many Amharic fictions, there are too many idiomatic expressions used by the authors. For example (Alemayehu 1996), one of the famous Amharic fictions called *fikireskemeqabir (ፍቅር እስከ መቃብር)* contains idiomatic expressions. The readers always obtain so many idiomatic expressions in fiction, but they understood the expression contextually because of a lack of opportunities to recognize idioms from the text with collected and organized resources in a digital manner.

The task of learning and evaluating Amharic idiom is left to the Amharic language expert by default. That is why there are still insufficient, well-organized, and computerized Amharic resources available. The situation necessitates making an effort to collect and organize Amharic idioms to make them available.

Other NLP studies, such as bilingual idiom translation, word sense disambiguation, and contextual text similarity, are impacted by the nature of idioms. Another feature of idioms that makes them difficult for the NLP system to process is that idiomatic expressions have both idiomatic and literal (non-idiomatic) usages.

One of the most important NLP applications that are negatively affected by idioms is Statistical Machine Translation (SMT) systems like google translate tools (Zong and Hong 2018). Phrase-based SMT systems extend the basic SMT word-by-word approach

(Koehn et al. 2003). These systems thus limit themselves to a direct translation of expressions without any syntactic or semantic context. Hence, standard phrase-based SMT systems do not model idioms explicitly (Bouamor et al. 2011). Unfortunately modeling idioms to improve SMT is not well studied (Vilar.D.et.al. 2009). It is commonly accepted in the Machine Translation field that the performance of SMT systems degrades when the input sentence contains an idiom as they often try to translate the idiom as "literal text".

The other NLP research affected by idiom is sentiment analysis. Sentiment analysis is the most common text classification tool that analyses an incoming text and tells whether the underlying sentiment is positive, negative, or neutral (Farhadloo and Rolland 2016; Williams et al. 2015) Most sentiment analysis works by looking at words in isolation, giving positive points for positive words, negative points for negative words, and then summing up these points. The sentiment analysis technology classifies a given text as negative if there is a negative word in the text. In Amharic, most of the time the negative word is formed from the prefix አል, አይ, አት... + Root word + postfix ም, ችም. Example አልበላም (አል + በላ+ ም) to mean 'he has not eaten አትጠጣም (አት + ጠጣ + ም) to mean she did not drink አልሄደችም (አል + ሄደ+ ችም) to mean she has not gone. Example

1. ስራ አልሄደም (he has not gone to work). Neutral
2. ምሳ በልተዋል (they have eaten Lunch). Neutral
3. ልብ አቁስል አይደለችም (she is not noisy). Posetive

From the above sentences, the first and the second sentence have no idioms. Therefore, the sentiment analyzer is successful in classifying the sentences as negative, neutral, and positive simply by observing the word-formation from the sentence. The presence of the idiom (ልብ አቁስል) makes the sentiment analyzer classification to be false. If the sentiment analysis fails to know the nature of the phrase "ልብ አቁስል", it fails to classify the given sentence in a wrong way. This problem requires a detailed study of the nature of Amharic idioms.

Semantic analysis is the other NLP research affected by idioms. The semantic analysis of natural language content starts by reading all of the words in the content to capture the real meaning of any text (Singh and Hanumanthappa 2016). Semantic technology processes the logical structure of sentences to identify the most relevant elements in the text and understand the topic discussed. It also understands the relationships between different concepts in the text. Consider the following paragraphs:

> ተስፋ የተጣለበት **መድኃኒት እረከሰ ::** **የመድኃኒቱን መርከስ** ዘግይቶም ቢሆን የተረዳዉ ህመምተኛ ሞቱን ይጠባበቅ ጀመር።

The medication that has been promised is low-cost. The patient, who had been exposed to the medicine for a long time, started to anticipate his death.

As soon as the machine detects the words እረከሰ, it understands as the whole paragraph as the cheapness of the medicine. But, the phrase መድኃኒት እረከሰ and የመድኃኒቱን መርከስ is used in the paragraph is to mean that the medicine is unable to cure the patient.

Our long-term research goal is to investigate how to automate idiomatic expression identification for the Amharic language.

This research is going to answer the following questions throughout and at the end of the research.

1. How to identify idioms from literals for the Amharic language?
2. Why to represent the dataset to before identification model?
3. What supervised machine learning algorithm achieves better performance result?

1.3. Objectives

1.3.1. General Objective

The general objective of this research is to develop an idiom identification model for the Amharic language.

1.3.2. Specific Objectives

- To conduct a literature review and related works to understand the approaches of idiom recognition.
- To collect and prepare dataset
- To study the mechanisms to identify idioms from literals.
- To prepare a dataset that contains idioms and literals
- To represent text into a vector representation
- To implement the appropriate algorithm and develop the model
- To test and evaluate the performance of the model.

1.4. Methodology of the Study

Since the Amharic language has limited resources, the primary action must be to learn the language component, which includes data collection and organization. To perform Amharic idiom recognition, data collection methods would be used to collect and organize Amharic idioms first. The supervised machine learning approach would take place after enough data had been gathered and arranged. This would be an experimental study using supervised machine learning.

1.4.1. Data Collection Methodology

1.4.1.1. Books and Documents

Data related to idioms collected from Amharic Idioms (Akililu and Worku 1992) and literal expressions from different Amharic documents. All the collected data would be cleaned, remove stop words, numbers and normalize characters finally prepared as training data. The models would train with these datasets. To test the model, random expressions which are combination idioms from the Amharic Idiom book and literals would be taken from different Amharic documents.

1.4.2. Preprocessing Data

It is better to think of the presence of some kind of error from the collected data. Therefore, all the collected dataset would not be used as it is; instead preprocessing and further cleaning the data would be performed and requires character normalization, spelling correction, cleaning, tokenization and possible morphological structure of idiomatic expressions would be considered.

To prepare a well-organized dataset, we used hard copy documents, so, it acquired to change from hard copy to soft copy to make digital dataset. During typining misspellings would be observed; these misspellings would result from missed out spaces (e.g. ሆደሰፊ instead of ሆደ ሰፊ to say he is patience), replacing letters with visually similar characters (e.g., ቀንቇጠረ for ቀን ቆጠረ).

1.4.3. Analysis and Design

To accomplish the research, the following points would be analyzed and studied well:

Vector representation: - Machine learning or deep learning algorithms are difficult to process and take text as input. It requires encoding to other representations of texts by using different algorithms. Word2Vec algorithm is one of the text encoding algorithms that represent text with vectors (Grzegorczyk 2019; Salton 2017)

Feature extraction: the research is going to be done with the principle of supervised machine learning. To train the machine, extracting representative features from the labeled data would be done.

The other fact that we can observe from the list of idioms is the number of words forming an expression. The number of words forming Amharic idioms are either of one word or two words or more than two words (Akililu and Worku 1992). From this formation, we used combination of two words to identify idiomatic expressions.

Identifier: we used a supervised machine learning algorithm to train and test the model. The model identifies idiomatic expressions based on the feature which was extracted and the vector representation of expressions.

1.4.4. Evaluation Measures

We used Accuracy, recall, precision, F-score to show and measure the performance of the model.

1.4.5. Development Tools

To build the model and test its performance, we used Python and Matlab programming. We used the Anaconda navigator Jupyter Notebook environment for Python programming, along with various libraries such as tensor flow, pandas, and the Sklearn library.

1.5. Scope and Limitation of the Study

A representative model would be designed for idiom identification for the Amharic language. This research would focus on idioms that are formed from the combination of two words which mean phrase level. We represent expressions into vector or numeric representation using the Word2Vec model.

The research was going to be done primarily by collecting the Amharic idioms found in the Amharic idioms book which are combination of two words only. This thesis is not used one word or more than two words combination of idiomatic expressions.

The study is not considered the nature of idioms which are either pure or literal nature of idiomatic expressions. Our thesis used all pure or semi-pure idiomatic expressions of two words combination.

1.6. Significance of the Research

Idiomatic expressions are frequently used in published books, especially Amharic fiction books, to maximize the degree of the message to be conveyed and to attract the readers' attention. To make the concept stated in the fiction clear, more idiom recognition mechanisms for the Amharic language are needed. This necessitates a computerized Amharic idiom word list that is well-organized and compiled.

A reader may not be able to recognize the combined terms that form idioms and determine the existence of Amharic idiom(s) from the expression at the time of reading. As a result, the readers take the text's idiom(s) literally. The meaning intended to be conveyed would be lost if the idiom was translated word for word. Identifying the existence of idioms from the given text is therefore the most critical and primary action in idiomatically interpreting idiomatic expressions.

In short

1. Writers use an idiomatic expression to maximize the degree of the message to be transmitted and to attract the reader's attention. At this time, it must identify the idioms, to interpret and exchange the message correctly
2. Used to reduce time to check whether a text has an idiomatic word or not using the digital dataset as a lookup table
3. Used to enhance NLP application researches

1.7. Beneficiaries of the Research

Different concerned bodies are beneficial after this research is completed. Thus are **Writers**: - In Ethiopia, there are more book writers in various categories such as fiction, educational books, historical books, people's tradition-based books, and so on. As a result, it is preferable to use idiomatic phrases in the book to draw readers and compose a good book. For example "አለሙ ሁልጊዜ **አፉን ይበለዋል፡፡**", under this example the phrase "**አፉን** ይበለዋል" is an idiomatic expression which means "he is more talkative". "**አለሙ ሁልጊዜ** መለፍለፍ ይወዳል::" the first sentence has a strong sound and attracts the reader's attention due to the existence of idiomatic expressions than the second sentence.

Readers:- They used idioms as literal because there is no mechanism to distinguish idioms from the text when they read various books or reading materials that contain idiomatic expressions. As a result, these readers are missing the intended message or sense of the reading content. However, when idioms for the Amharic language are understood, readers can comprehend the context and message of the book they are reading.

Evaluators:- Due to a lack of mechanisms that identify or extract idiomatic expressions for the Amharic language, it is difficult to understand the context of the reading material They evaluate the document using our digitalized dataset as a lookup table.

Language translators:- They can look for idiomatic expressions in the text before attempting to translate the entire text. As a result, to search for the existence of idioms, they read the entire text and manually extract the idioms. When Amharic idioms are extracted from the entire language, they can be used to translate other tokens word for word or as a complete text.

For example: - to translate from Amharic to English through machine translation we consider the following sentence. "አበበ **ሆደ ሰፊ** በመሆኑ ለሀገራችን አስፈላጊ ነዉ።" when we give the whole sentence to the machine as input, it translates as the following “As Abebe's abdomen is wide, he is important to our country". The Amharic idiom “ሆደ ሰፊ” from the sentence tells about የአበበን ታጋሽነት/“Abebe’s patient”. When we used direct word-by-word translation, it gives about Abebe’s abdomen wideness and importance to our country. Without considering idiomatic expressions, using direct machine translation affects the meaning of the text.

The correct meaning of the above sentence is “As Abebe's patient, he is important to our country"

Professionals and language speakers - The fact that Amharic is a mother tongue language does not imply that he or she is an expert in that language. Speakers encounter language in a variety of ways. They also practice idiomatic gestures, which they attempt to use in their lives. Professionals of every language must understand the meaning of the language's idiomatic expressions and use them appropriately; the same is true for Amharic language professionals. As a result, this research work aids them in understanding the essence of the Amharic language by distinguishing idioms from literals.

1.8. Organization of the Thesis

This thesis is divided into five parts. Context details, a statement of the issue, the study's goal, scope and limitations, and benefits and beneficiaries are all covered in the first chapter. The second chapter includes a literature review section that covers an overview of Amharic language, the nature of Amharic idiom, and related works. The third chapter covers research methodology, which includes data planning for the experiment, models, and algorithms used in the research. The experiment findings and performance results are discussed in the fourth chapter. The final chapter wraps up the conclusions and suggests some feature works for further study.

CHAPTER TWO

2. LITERATURE REVIEW

This chapter aims to demonstrate the most important aspects of existing works, such as substantive observations as well as theoretical and methodological contributions relevant to idiom identification. The overview of the Amharic language, its grammatical nature, and the writing system in the Amharic language was discussed in this chapter. A study of similar works of literature reviewed to establish the model for this research as well as to organize the research concept. The literature on idiom recognition and idiom properties has been reviewed specifically for this study.

2.1. Overview of Amharic Language

Amharic language is a Semitic language and written by using the Fidel system adapted from Ge'ez. According to the Ethiopian central statistics agency census reported in 2007, a country of 73.92 million people has used the Amharic language. Even outside Ethiopia, Amharic is the language of millions of emigrant peoples (notably in Egypt, the US, Israel, and Sweden), and is spoken in Eritrea.

Amharic words are categorized under six basic classes, namely, ስም (noun), ተዉላጠ ስም (pronoun), ግስ (verb), ቅፅል (adjective), ተውሳከ ግስ (Adverb) and መስተዋድድ (preposition) based on morphology and position of the word in Amharic sentence (ይማም 1987)
Noun: a word would be categorized as a noun if it can be pluralized by adding the suffix ኦች/ ዎች ("owch") and used as nominating something like person and animal. It is used as a subject in a sentence.

Pronoun, the following are some of the pronouns in Amharic ይህ, ያ, እሱ, እስዋ, እኔ, አንተ, አንቺ…; quantitative specifiers, which includes አንድ, አንዳንድ, and possession specifiers such as የእ ኔ, የአ ን ተ, የእ ሱ.

Verb: any word which can be placed at the end of a sentence and which can accept suffixes as /ሀ /,/ሁ/,/ሸ/, etc. which is used to indicate masculine, feminine, and plurality is classified as a verb**.**

Adjective: is a word that comes before a noun and adds some kind of qualification to the noun. But every word that comes before a noun is not an adjective.

Adverb: a word that qualifies the verb by adding extra ideas from time, place, and situations point of view. The following are adverbs in Amharic ትናንት, ገና, ዛሬ, ቶሎ, ምንኛ, ክፉኛ, እንደገና, ጅልኛ and ግምኛ.

Preposition: a word that doesn't take any kind of suffix and prefix, that can't be used to create other words, and which doesn't have meaning by itself but can represent different adverbial roles when used with nouns. The different propositions include ከ ፣ ለ ፣ ወደ፣ ስለ ፣ እንደ…፣ ወዘተ.

2.1.1. Characteristics of Amharic Writing

Under this section, we focused on the nature of Amharic language based on Fidel's. Amharic languages took the whole Ge'ez alphabet and uses in its writing system and add some other alphabets like ሸ, ጨ, ኘ, ጀ, ቸ…. There is a redundancy of characters in the Amharic language. However, in Amharic, there is no meaning change but, each alphabet in Ge'ez has its meaning even if alphabets are having the same sound. The table below shows an example of character redundancy.

Table 2. 1:- Amharic characters with the same sound

Consonants	Other symbols with the same sound
ሀ (hä)	ሃ ሐ ሓ ኀ ኃ ኻ
ሰ (sä)	ሠ
አ(ä)	ኣ ዐ ዓ
ጸ (tsä)	ፀ

Spelling variants of a phrase would increase the number of terms describing a text unnecessarily, reducing the efficiency and precision of the subtext categorization and idiom recognition classifiers. Word variants (spelling differences) caused by inconsistent use of redundant characters should be normalized during preprocessing. The different types of a character with the same sound are converted to one common form during the pre-processing stage of Amharic expressions in this work.

2.2. Overview of Amharic Idioms

Amharic idiom is a phrase made up of a sequence of either one word or two words or more than two words that cannot be interpreted from the meaning of the individual words or their normal mode of combination. Amharic idioms are created using single word, two words and more than two words.

Example:-

One word:- ሻለተ፣አደረጋት፣ ሰብቷል

Two words:- ልቡ ሸፈተ፣ብልት አወጣ፣አፈን በዳቦ …

More than two words:- ለአፉ ወሰን የለዉም፣ለአቅመ ሄዋን ደረሰች፣ በረሀብ አለንጋ ተገረፈ ...

One of the complex natures of Amharic idiom is having both idiomatic and literal property. Some Amharic idioms like እጅ ሰጠ and ልበ ቢስ are pure Amharic idioms which cannot be interpreted out of their pure idiomatic meaning. The Amharic idiom እጅ ሰጠ cannot be interpreted in another way except its idiomatic meaning 'ተሸነፈ'. Unlike the pure Amharic idioms, there are semi-pure idioms that can be interpreted in two ways; idiomatically and literally.

For example, የግንባር ስጋ has both idiomatic and non-idiomatic property. Idiomatically, 'የግንባር ስጋ' is used to express someone's openness; on the other way, 'የግንባር ስጋ' is used to express the tissue on our face. Identification of idioms is a challenging problem with wide applications because of idioms having complex nature.

As stated in the book (የአማርኛ ፈሊጦች, 1992) Amharic idioms are related to body parts, culture related and people's traditional practices. By considering the newness of our research, we limit this research, to propose recognition of idiomatic expressions that are a combination of two words for the Amharic language using supervised machine learning.

2.3. Related Works

Different researches are conducted research related to idioms in different languages and their effects were analyzed on language translation (Salton 2017; Williams et al. 2015) Different scholars do idiom identification using different methods like using meaning (Verma and Vuppuluri 2015), VNC part of tag sequence, sentential distribution (Salton et al. 2016), word embedding (Peng and Feldman 2016a).

Most work on the phrase classification stream imposes syntactic restrictions. Verb/Noun restriction is imposed in (Diab and Bhutada 2009; Fazly et al. 2009), and Preposition-Noun-Verb restriction is imposed in (Fritzinger et al. 2010). The latest studies were used word embedding models by vector representation of phrases through different methods, like term frequency of phrases and Word2Vec approach for identification of idioms (Peng et al. 2010; Salton 2017).

The research centered on the meaning of idiom terms, so that properties of individual words in a phrase vary from the properties of the phrase in itself (Verma and Vuppuluri 2015). The researchers used three data sets for the experiment: englishclub.com, the Oxford Dictionary of Idioms, and the Verb Noun Construction (VNC) corpus. The study's success was assessed using a union and intersection methodology. The study developed a model that recognizes idiomatic speech through dictionary-based type as a result of this research. Takes VNC POS tag sequence only and Difficult for Amharic idioms due to the ambiguous of idioms like እጅ ሰጠ, የግንባር ስጋ.

The research was carried out by automatically detecting idiomatic phrases using dictionaries (Muzny and Zettlemoyer 2012). Three lexical features and five graph-based features were included in the analysis. The research focused on identifying English language phrases from web data. The Wiktionary default rule and the Lesk word sense disambiguation algorithm were used to assess the results. If the meaning of a word is unclear, it becomes an idiom, but not all ambiguous words are idioms. The Lesk Word Sense Disambiguation algorithm was used in the research. Their model was limited to dictionary terms and used a pattern of word matching technique, and all ambiguous words were ruled out as idioms.

The experiment was focused on the context of idiomatic expressions and literal expressions (Peng and Feldman 2016a). The researchers proposed two approaches to represent the context of idioms and literals; compute inner product of context word vectors with the vector representing a target expression and compute literal and idiomatic scatter matrices from local contexts in word vector space. The researchers used a dataset of 2984 VNC tokens from BNC as well as a list of VNC tokens that were classified as literal, idioms, or unknown. The study was restricted to the occurrence and frequency of words in context. In this case, idioms were chosen as the word because it appears regularly, but this was not the case all of the time and aimed to predict the idiomatic usage of VNCs.

Based on the distribution of terms, a study was conducted to classify phrases as literal or idiom (Peng and Feldman 2016b). The word distribution for a literal expression differs from the distribution for an idiomatic expression, according to the study assumption. The analysis represented the distribution of words in vector space as a covariance matrix and word vectors obtained from the Word2Vec tool. The researchers used data from the National Science Foundation's Grant No. 1319846 to conduct their research. The research used 12 datasets to assess output throughout 20 runs. The study looked at the frequency of a word's occurrence to determine if it was an idiom or a literal word. To test and train the model, they used a small dataset.

On the idiom dictionary, the research demonstrated lexical knowledge of idioms (Hashimoto et al. 2006). The research discovered two significant obstacles to idiom recognition word ambiguity and idiom transformations. Researchers maintained that transformable idioms required dependency knowledge, while ambiguous idioms required disambiguation knowledge as well as dependency knowledge. A dataset of 100 verbal idioms was used in the analysis. They gathered 300 sample sentences from the Mainichi newspaper of 1995 as an assessment corpus, each containing 100 idioms. The research looked at idiom dictionary lexical skills.

Using linear discriminant analysis, this study was conducted on idiom recognition (Peng et al. 2010). For English language studies, the researchers used the VNC (Fazly et al. 2009) corpus. The 2,550 sentences in a 6,844-dimensional term space were represented term-by-sentence using a bag-of-words model. The dataset used in the analysis contains 2,550 sentences, 2,013 of which are idiomatic and 537 of which are literal. For the sample, 300 literal and 300 idiomatic sentences were chosen at random as preparation, and 100 literals and 100 idioms were chosen at random as testing from the remaining sentences. Thus, the training dataset consists of 600 examples, while the test dataset consists of 200 examples. The research was achieved 80.15% accuracy through the three nearest neighbor (3NN) classifier.

The sent2vec model was used to create distributed representations to encode features that are predictive of idiom token classification (Salton et al. 2016). The study used a collection of sentences from the British National Corpus that included 53 separate Verb Noun Constructions. The sentences were encoded in three different formats: uni-skip, bi-skip, and comb-skip. The research used four expressions on various models and distributed representations to assess results. The research yielded a variety of successful outcomes. The analysis only used VNC expressions to identify idioms from literals, but it also includes another series of parts of speech tags, so this research ignores these idioms even though they are represented in vector space.

In (D.Salton et al. 2017) presented four different models to overcome the limitations of the state-of-the-art model for VNIC type idiom identification. The study proposed a probabilistic approach, Smoothed Probabilities, Interpolated Back-off Probabilities, and Normalized Google Distance. The study used 319 idioms and 319 literals used to train the idiom identification model and 95 idioms and 95 literals for testing the model from British National corpus. The study showed that feeding the fixedness metrics to an SVM also improves the F1-score on the same VNIC type identification task.

2.3.1. Summary of Related Work

There is no research made for idiom recognition for Amharic language before. By considering the negative impact of idioms in many NLP researches, we initiated to deal on idiom recognition model for Amharic language. The negative influences of idioms on the NLP researches have been stated in the foreign languages. What makes idioms in every language is the complex nature it shows. Every idiom in the world has complex behavior. Therefore, if idioms have a negative impact on NLP applications researches in one language, there is no any situation that the Amharic idioms cannot negatively affect Amharic NLP researches.

When we stand to do this research, we reviewed idiom identification approaches using different language to propose the better approach for our research. The researchers in the stated related work used their own methodology, approach and the result of the research. Researchers used a corpus of grammatical part of speech tag sequence of VNC to recognize idioms like; lose face, lose head, make scene… but when we see the tag sequence of Amharic idioms, it followed NV, NN, VN, tag sequence. So, it is difficult to extract one common feature based on part of speech tag sequence like that of English idioms.

When we have seen the work of (Verma and Vuppuluri 2015), difficult due to Amharic idioms are either pure or semi pure expressions, so, in Amharic language it contains the idiomatic meaning and literal meaning of the expressions when the idiom is semi-pure expression. To represent expressions, researchers used term frequency approach of VNC part of speech tag corpus and consider ambiguity of words. Researchers used ambiguity of words approach, rule-based approach and pattern matching approach, but for this all ambiguous words are not idioms and others are static approaches. We proposed machine learning approach idiom identification for Amharic to overcome the negative impact of NLP applications and the gaps that are observed from literatures.

Table 2. 2:- Summary of related works

No	Author	Titles	Method used	Dataset size	Gaps	Result %
1.	(Verma and Vuppuluri 2015)	A New Approach for Idiom Identification Using Meanings and the Web	Meaning of the word idiom, IdiomExtractor	53 VNC tokens	➢ Consider individual words meaning and the phrase itself difference ➢ Takes VNC POS tag sequence only ➢ Difficult for Amharic idioms due to the ambiguous of idioms like እጅ ሰጠ, የግንባር ስጋ	95.04%
2.	(Muzny and Zettlemoyer 2012)	Automatic idiom identification in Wiktionary	Lesk word sense disambiguation, default wikitionery rule	1,300 dictionary definition in Wiktionary	➢ Limited on words which were found on Wiktionary ➢ Used a pattern of word matching technique, All ambiguous words were ruled out as idioms	83.8%
3.	(Peng and Feldman 2016a)	Automatic idiom recognition with word Embedding's	Tf-idf, phrase-idf, phrase-tf-idf, CoVAR, Context	2984 VNC, Word2Vec representations	➢ Consider the distribution of words ➢ Used VNC tag sequence dataset ➢ aimed to predict the idiomatic usage of VNCs	92%

4.	(Peng and Feldman 2016b)	Experiments in idiom recognition	Tf-idf, phrase-idf, phrase-tf-idf, CoVAR, GMM	12 VNC, Word2Vec model	➢ Depend on the frequency of words ➢ They used a small dataset	81%
5.	(Salton et al. 2016)	Idiom Token Classification using Sentential Distributed Semantics	KNN, Linear-SVM-Per-Expression, Grid-SVM-Per-Expression, SGD-SVM-Per-Expression	53VNC Sent2Vec representations	➢ The model tests on a small dataset that are a sequence of VNC	96%
6.	(D.Salton et al. 2017)	Idiom Type Identification with Smoothed Lexical Features and a Maximum Margin Classifier	SVM and probabilistic method	828 idioms and literals	➢ It only considers VNC datasets with probabilistic method ➢ Fixed metrics were improved using SVM on VNC	85%

CHAPTER THREE

3. RESEARCH METHODOLOGY

In this chapter, we presented detailed research methodologies of the study. We presented dataset collection and preparation methodology, proposed model design, and detailed explanation of proposed model activities.

3.1. Data Collection Methodology

As a data set, we used Akililu and Worku's Amharic idioms book (Akililu and Worku 1992) and different Amharic documents. There are over two thousand Amharic idioms in the Amharic idioms book. We gathered idioms and researched the characteristics of Amharic idioms and analysis of idiom properties is needed. Idioms are made up of a combination of single word, two words, or more than two words and are linked to a variety of topics such as body parts, culture, religion, nature, and behavior, among others.

Most Amharic idioms' are created by relating to the body part namely; heart-related, eye-related, stomach-related, ear-related, neck-related, head-related, blood-related, hand-related, bone-related, leg-related, leap-related, intestine-related and peoples tradition or culture-related idioms (Akililu and Worku 1992).

Examples: ሀሞቱ ፈሰሰ, ሀረግ መዘዘ, ሀረግ ጣለ, ሀሳበ ቢስ, ሀሳብ ገባዉ, ሁብተ ስጋ, ሁብተ ሰባራ, ሁብተ ስንኩል, ሁብትሽ በሁብቴ, ሁብተ ቢስ , ሁብተ ነፍስ, ሁብቷ ቀና, ሁለት ምላስ, ሁሉ አማረሽ, ሁሉ አገርሽ, ህሊና ቢስ, ህቅ አለ, ህግ ተላለፈ, ህግ አፈረሰ, ህገ ወጥ, ህግ ገባ, ሰዉ ሆነ, ልቡ ተሰነጠቀ, ነፍስ ሆነ, ዋስ ሁነኝ, አስብቶ አራጅ, የሆነዉ ሆኖ, ቁመተ ስጋ, ሆደ መጋዝ, ሆደ ሰፊ, ሆደ ቡቡ, ሆደ ባሻ, ሆደ ገር, ሆዱ ሻከረ,ሆዱን ቆረጠዉ, ሆዱ ባሰበት, በጥፊ ለጠፈበት, ለፍቶ መና, በዱላ አለፋዉ, ዐይነ ልም, ላስ አደረገዉ, መሬት ላሰ, መሬት አስላሰ, ማርም አልስ, እሳት የላሰዉ, የሚላስ የሚቀመስ, አዕምሮዉ ላሸቋል, ላባ ቀረሽ, ላክ አደረገ, ሲበሉ የላኩት, የሰይጣን መልዕክተኛ, የበላይና የበታች, ላጤ መላጤ, ሊቀ ላጤ, ወንደ ላጤ, የብረት ልጥ, በደረቁ ላጨ, እሳት የላፈዉ, ገንዘቤን ላፈኝ, ዉሸቱን ላፈዉ, ሌላ ነዉ,ሌባ ሚዛን, ሌባ ዉጋት, …others are listed at appendix B.

We may understand the meaning and characteristics of Amharic idioms, even if they are formed about people's lifestyles, body parts, or other topics. This study would look for idioms at the phrase level, which means idioms made up of two words. Due to the nature of the expression, identifying idioms is a difficult and ambiguous activity. This thesis work is new and it would show the possibility of our local language automation by using different methods.

To build the model of the study data is collected from Amharic idiom books and different Amharic documents. A corpus is prepared that contains one thousand expressions that are idiomatic and literal. The collected dataset has an equal number of literals and idioms. All the collected datasets are preprocessed and represented into vectors or numeric values in N-dimensional spaces.

3.1.1. Training and Testing Dataset

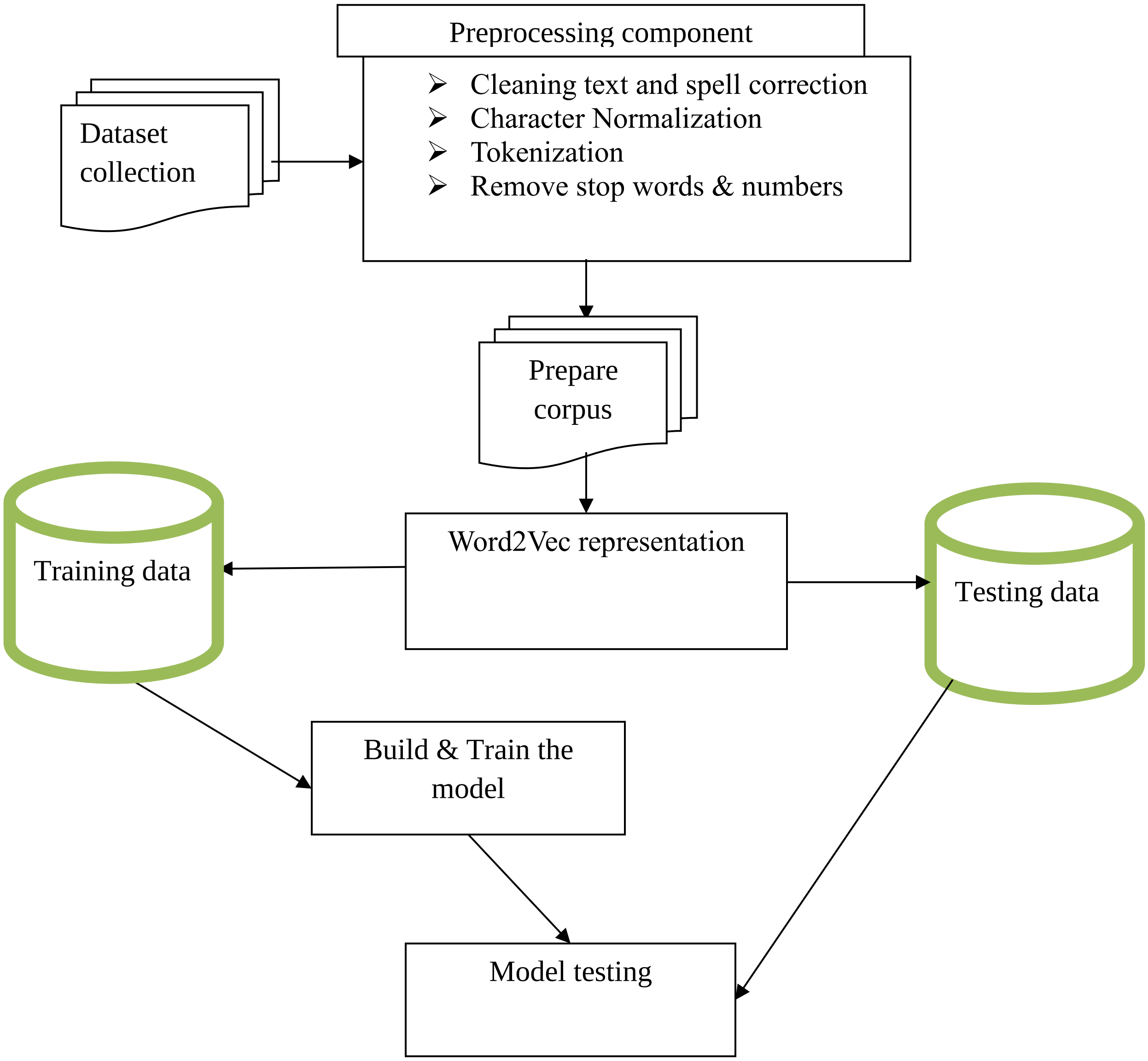
This study used a supervised machine learning approach, which is a method of manually labeling data and assigning the idiom or literal class to expressions. As training data, the collected Amharic idioms and literals are cleaned and normalized with possible morphological structures.

The data set is divided according to the 80/20 rule (Philemon and Mulugeta 2014), with 80% of the dataset going to the training set and 20% to the test set. Eight hundred expressions were used to train the model and two hundred expressions were used to test the model. All the training and testing datasets are expressions that contain either idiomatic or literal classes from Amharic idiom books and different Amharic documents.

3.2. Proposed Model Architecture

The proposed model has the following basic components to identify idioms. The figure 3.1 shows the components with basic activities.

Figure 3. 1:- Proposed model



3.2.1. Data Pre-processing

Since noisy data may slow the learning process and reduce the system's efficiency, preprocessing removes irrelevant data from the dataset, reducing computational time and improving classifier performance (Rase 2020). It is preferable to consider the possibility of some kind of error in the data collected. As a result, the entire dataset will not be used as is; rather, preprocessing and further cleaning of the data will be done, which will necessitate character normalization, tokenization, and the removal of stop words and numbers.

Cleaning and Spell Correction

The collected dataset would be converted from hard copy to soft copy to prepare a well-organized corpus. Misspellings would be observed. Misspellings would result from missed out spaces (e.g. ሆደሰፊ instead of ሆደ ሰፊ to say he is patience), replacing letters with visually similar characters (e.g., ቀን ቖጠረ for ቀን ቆጠረ). During typing such errors are occurred, so, it needs to correct such spellings and also spaces.

Normalization

Character Normalization means normalizing letters that had the same sound as one common letter. Token normalization is the process of canonical tokens (Zhu et al. 2007). In this study, this kind of character-level normalization is done.

Table 3. 1:- character representation

Number	Un-normalized characters	Normalized
1	ሐ/ሓ/ኀ/ሃ/ሀ/ኃ	ሀ
2	ዐ/ዓ/ኣ/አ	አ
3	ጸ/ፀ	ፀ
4	ሠ/ሰ	ሰ

As shown in Table 3.1 above, different characters on the un-normalized characters are normalized to one common representation which is appeared on the normalized character. The table below shows examples of the different word spellings caused by the redundant characters.

Table 3. 2:- Words having spelling variations

Words in English	Words in Amharic	Spelling variations of the word
Stomach	ሆድ	ሖድ ኆድ
Eye	አይን	ዓይን
Sky	ሰማይ	ሠማይ
World	አለም	ዐለም ዓለም ኣለም
Sun	ፀሀይ	ጸሀይ

Tokenization: a recognition of sentence boundaries to segment texts into tokens (Getinet 2015). For Amharic language white spaces and punctuation marks (full stop (::), an exclamation mark (!), question mark (?), and colon (፣)) is used as a word-level separator. For our research work we used the white space to classify the phrase into word level.

Remove Stop Words and Numbers

words that occur too frequently and little semantic in the text (Miretie and Khedkar 2018)**.** Amharic stop words have a negative influence on idiom identification due to their frequency of occurrences. For example, Amharic words like "ግን "(but), "ነው" (is), "ነበር" (was), are considered as stop words. See list of stop words and numbers at appendix A.

In Amharic numbers can be represented by either the Arabic number system or symbol of the Ethiopian number system or alphanumeric representation. For the idiom identification model, they harm the model when they are not properly identified. For example 2 ምላስ፣ 2 ልብ. The following table shows number representations in Arabic, Amharic, and alphanumeric.

Table 3. 3:- Number representation

Arabic	**Ethiopic**	**Alphanumeric**	**Arabic**	**Ethiopic**	**Alphanumeric**
1	፩	አንድ	20	፳	ሀያ
2	፪	ሁለት	30	፴	ሰላሳ
3	፫	ሶስት	40	፵	አርባ
4	፬	አራት	50	፶	ሀምሳ
5	፭	አምስት	60	፷	ስድሳ
6	፮	ስድስት	70	፸	ሰባ
7	፯	ሰባት	80	፹	ሰማንያ
8	፰	ስምንት	90	፺	ዘጠና
9	፱	ዘጠኝ	100	፻	መቶ
10	፲	አስር	1000	፼	ሺህ

For our model, we used an alphanumeric representation of numbers, because our model works on texts.

Morphology of Amharic Language

Amharic is a morphologically rich language with many variations. When the root word combinations of phrases are idioms in the Amharic language, the morphology of the term is also an idiom, and the same is true for literals. When preparing the dataset, we took morphology into account. The phrase's morphology may have an idiomatic or literal interpretation. Both morphological structures in Amharic texts are taken into account.

Table 3. 4:- morphological richness of Amharic idiom

Amharic idioms	Possible morphemes	Status
	እጁ አጠረዉ	Idiom
	የእጁ ማጠር	Idiom
እጁ አጠረ	እጇ አጠረ	Idiom
	እጇ አጠረባት	Idiom
	እጁ አጠረበት	Idiom
	እጁ አጥሮበታል	Idiom
	እጇ አጥሯል	Idiom

3.2.2. Vector Representation

Text as input is difficult to process using machine learning or deep learning algorithms (Grzegorczyk 2019; Salton 2017). It necessitates the use of various algorithms to encode other text representations. One of the text encoding algorithms that use vectors to represent text is the Word2Vec algorithm. Word2Vec is a set of neural network models to represent words in vector space. In vector space vectors with a low cosine gap, identical terms are clustered together, whereas dissimilar words are spread out (Grzegorczyk 2019; Salton 2017). Vector representation of expressions on two dimensional spaces is shown in appendix C. Different vector representation methods were developed for NLP applications.

A. One-hot Representation

In the vector space notation, it is a sparse vector: that is a vector with one at only a single position and zeroes at other remaining positions like [0 0 0 0 0 0 0 0 0 1 0 0 0 0]. Its drawback is space complexity and its failure to represent the similarity between two words (Sharma et al. 2017). no semantic information is getting expressed with this representation system

Example:- corpus= ['ሀብተ ነፍስ, ሀብቷ ቀና, ሁለት ምላስ, ሁሉ አማረሽ, ሁሉ አገርሽ ,ህሊና ቢስ ']

First, generate unique words from the sentence. The length of vectors is the number of unique words in the dataset is 11 and assigns 1 for the expression which is found on the vocabulary like the following

ሀብተ= [1,0,0,0,0,0,0,0,0,0,0]

ነፍስ= [0,1,0,0,0,0,0,0,0,0,0]

ሀብቷ= [0,0,1,0,0,0,0,0,0,0,0] the same is true for others.

B. Bag of Words (BoW) Representation

Bag of Words (BOW) is an algorithm that counts how many times a word appears in a document and quantize each extracted key point into one of the visual words (Xu et al. 2013; Zhang et al. 2010). It creates a vocabulary with unique words and then creates vectors, with the length of the vectors indicating the length of the vocabulary. It is simple to comprehend and put into practice. The sparsity of representations can be affected if the data is too huge and contains numerous unique terms. Sparsity adds to the complexity of both space and time. It makes it more difficult for models to retrieve small amounts of data from a huge representational space.

Example:- corpus= ['ሀብተ ነፍስ ሀብቷ ቀና ሁለት ምላስ ሁሉ አማረሽ ሁሉ አገርሽ ህሊና ቢስ ']

First, generate unique words from the sentence.

Corpus = [”ሀብተ” ፡1,” ነፍስ”:1,” ሀብቷ”:1, “ቀና”:1, ”ሁለት”:1,”ምላስ”:1,”ሁሉ”:2,”አማረሽ”:1, “አገርሽ”:1,” ህሊና”:1,” ቢስ”:1]

C. TF-IDF Representation

It measures the importance of a word in a document by looking at how often it appears in the document (Xu et al. 2013). The frequency of a word is an indication of its importance. If a term appears frequently, it is likely to be significant. IDF (Inverse Document

Frequency) is used to calculate the weight of rare words across all documents. The words that occur rarely in the corpus have a high IDF score (Xu et al. 2013)

$$TF = \frac{\text{number of words that occur in a document}}{\text{Total number of the word in a document}} \quad \text{---------------------equation (3.1)}$$

Example:- Doc1= ['ሀሊናቢስ ሀብተነፍስ ሀግተላለፈ ሀግአፈረሰ ሀገወጥ ሁለትምላስ']

Doc2=['ሀብተነፍስ ሁለትምላስ ሁሉአገርሽ ሀብቷቀና ']

Words	Document one	Document two
ሀሊናቢስ	1/6	0/4
ሀብተነፍስ	1/6	¼
ሀግተላለፈ	1/6	0/4
ሀግአፈረሰ	1/6	0/4
ሀገወጥ	1/6	0/4
ሁለትምላስ	1/6	¼
ሁሉአገርሽ	0/6	¼
ሀብቷቀና	0/6	¼

The total number of expressions in document one is six and document two has four expressions. We have ten expressions. So, the TF of expressions here looks like the following.

TF(ሀሊናቢስ) =1/10, TF(ሀብተነፍስ) =2/10, TF(ሀግተላለፈ)=1/10, TF(ሀግአፈረሰ)=1/10, TF(ሁለትምላስ) =2/10… the same is true for others. Frequently occurred words like ሁለትምላስ፣ ሁሉአገርሽ፣ ሀብተነፍስ are significant. Researchers used this term frequency for idiom identification based on the frequency of words and they concluded as frequently occurred expressions are idioms (Peng and Feldman 2016b).

D. Word2Vec Representation

Word2Vec essentially places words in feature space in such a way that their location is determined by their meaning. Words with similar meanings are clustered together, and the distance between two words has the same meaning (Ma and Zhang 2015; Salton 2017). It is a method/model for generating word embedding for improved word representation. It captures a huge number of syntactic and semantic word associations with great precision. It's a two-layered shallow neural network with only one hidden layer between input and output (Mikolov et al. 2013).

Word2Vec model uses a distributional similarity-based approach for representing each word as a vector of N-dimension, where each element in the vector is a real number (Sharma et al. 2017). Word vectors represent words as multidimensional continuous floating-point numbers where semantically similar words are mapped to proximate points in geometric space (Salton et al. 2016). There are two models in this class used by Word2Vec which convert unsupervised representation to supervised form for model training (Grzegorczyk 2019; Salton 2017).

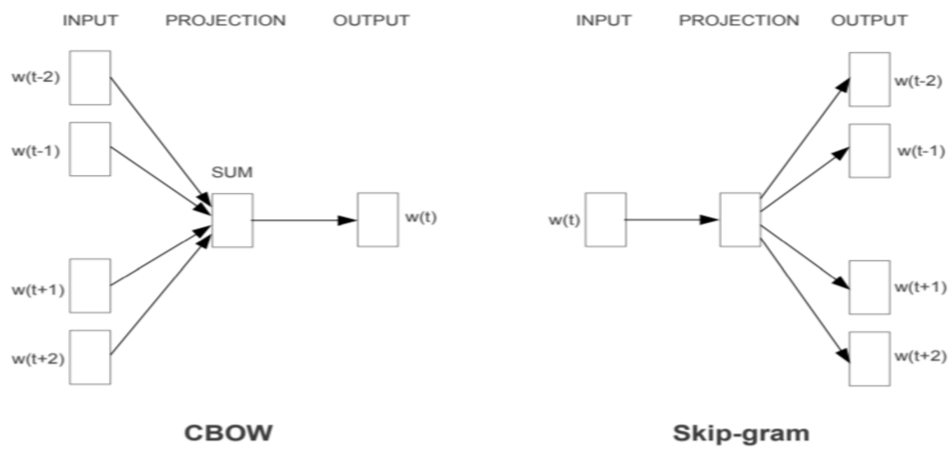
The neural network in CBOW (Continuous Bag of Words) predicts the words that fall in between or predicts the target words from the context, whereas, in Skip-grams, the neural network takes in a word and then tries to predict the surrounding words or predict the target words from the context.

In both models, a predetermined length window is moved along the corpus, and the network is trained using the words inside the window at each step. The learned linear transformation in the hidden layer is used as the word representation once the neural network has been trained. The beauty of representing words as vectors is that they lend themselves to mathematical operators. For morphologically rich languages such as Turkic, Arabic, Chinese, Amharic Word2Vec can treats each word in the corpus like an atomic entity and generates a vector for each word unless we apply morphology analysis before providing a dataset to model (Eshetu et al. Aug-2020).

For this research, we used the Word2Vec method to represent expressions into vectors. Word2Vec produces the probability of words in the output layer. Word2Vec focuses on the idea of a word or term being represented by a vector to represent words in vector space representation. See the represented sample dataset in appendix C.

The following figure shows that the diagrammatic representation of Word2Vec representation structure through CBOW and skip-gram (Mikolov et al. 2013).

Figure 3. 2:- Word2Vec representation



E. Global Vectors (GloVe) Model

Count-based approaches compute word representations using global co-occurrence counts from the corpus. Glove aims to integrate the methodologies of CBOW and skip-gram models, and it has proven to be more accurate and efficient (Sharma et al. 2017). For each word, the models are utilized to construct a vector of a fixed size. As an invariant, each model employs the similarity of two words. They presume that words that appear in comparable circumstances have comparable meanings.

3.2.3. Train and Test Model

To develop the model, we used supervised machine learning methods. We would discuss some supervised machine learning algorithms in detail to implement the better algorithm.

A. K-Nearest Neighbor Classifier

The KNN method is a non-parametric instance-based learning method that stores all available data points and classifies the new data points according to the similarity measure (Bzdok et al. 2018). The idea behind the KNN method is to assign new unclassified examples to the class to which most of their next K belongs. This algorithm is very effective in reducing misclassification errors when the number of samples in the training data set is large.

The KNN is one of the prospective statistical classification algorithms used to classify objects based on the next learning examples in feature space (Thirumuruganathan 2017). It is a lazy learning algorithm in which the KNN function is locally approximated and all calculations are reset to classification. During the learning phase, no model or actual learning is performed, although a learning record is needed, it is only used to fill a sample of the search space with instances whose class is known, this algorithm is also called lazy learning algorithm. This means that training data points are not generalized and that all training data is needed during the test phase. When an instance whose class is unknown, the algorithm computes its nearest K neighbors and the class is assigned by choosing between them. In the KNN algorithm, the training phase is very fast, but the testing phase is expensive in terms of time and memory (Bzdok et al. 2018; Thirumuruganathan 2017)

The KNN algorithm comprises two phases: the training phase and the classification phase. In the learning phase, the learning examples are vectors (each with a class label) in a multidimensional feature space. In this phase, the feature vectors and class tags of the training samples are stored. In the classification phase, K is a user-defined constant, a query or test point (unlabeled vector) is ranked by assigning a label, which is the most recurrent among the K closest training samples this request point. In other words, the KNN method compares the query point or an input feature vector with a reference vector library, and the query point is tagged with the nearest library feature vector class. This way of classifying the query points according to their distance to points in a set of learning data is a simple but effective way of classifying new points (Thirumuruganathan 2017).

K-Nearest Neighbor (KNN) is an automatic learning method in which the classification is performed by determining the nearest neighbors for the determination of the given example class based on the calculation of the minimum distance between the given point and the other points of the distances calculated with; Euclidean, Manhattan, Minkowski, Supremum, and Cosine Similarities (Bzdok et al. 2018) . In the classification, the test pattern is ranked by the largest number of votes of neighbors K, with the sample being assigned to the most commonly used class among its neighbors K-Closer. K is a positive integer determined by a test-and-error method from which the lowest error rate is obtained. In general, the classifier architecture is simple, but as the number of training data increases, the classification time becomes longer.

B. Bayesian Classification

Bayesian classifiers are statistical classifiers that can predict the class membership probabilities of a given tuple. The base for Bayesian classification is the Bayes theorem. Studies comparing classification algorithms have found a simple Bayesian classifier known as the Naive Bayesian Classifier. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence (Sainia et al. 2013).

During pattern classification based on Bayesian classification, there are two kinds of probabilities. The prior probability indicates the probability that the pattern should belong to a class, say Ci, for i=1, 2, 3…N. The posterior probability P(x/Ci), on the other hand, indicates the final probability of belongingness of the pattern x to a class Ci. The posterior probability is computed based on the feature vector of the pattern, class conditional probability density functions P(x/Ci) for each class Ci, and prior probability P(Ci) of each class Ci. Even though, Bayesian classifiers have the minimum error rate in comparison to all other classifiers theoretically. In practice, this is not always the case for inaccuracies in the assumptions made for its use, such as class-conditional independence, and the lack of available probability data.

The training phase and testing phase of Bayesian classifiers operate as follows: using the training samples the method estimates the parameters of a probability distribution. And the prediction of the test sample, the method computes the posterior probability of that sample belonging to each class. Then the test sample is classified according to the largest posterior probability. During training and testing, it is assumed that features are conditionally independent in the given class (Bzdok et al. 2018; Sainia et al. 2013)

C. Support Vector Machine

The support vector machine (SVM) is a machine learning algorithm based on statistical learning theory (Bzdok et al. 2018). A support vector machine builds a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, used for classification (Seetha et al. 2008). Good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (functional margin), generally, the larger the margin lowers the generalization error of the classifier. SVM uses a Non-parametric with binary classifier approach and can handle more input data very efficiently. Performance and accuracy depend upon the hyperplane selection and kernel parameter.

The main advantages of SVM are it gains flexibility in the choice of the form of the threshold, contains a nonlinear transformation, provides a good generalization capability, the problem of overfitting is eliminated, Reduction in computational complexity, Simple to manage decision rule complexity, and Error frequency. Disadvantages of SVM have resulted in transparency is low, training is time-consuming, the Structure of the algorithm is difficult to understand, and the determination of optimal parameters is not easy when there is nonlinearly separable training data (Bzdok et al. 2018).

The calculation complexity and complexity of the decision rule are reduced in SVM. In SVM, training speed depends on the size of the learning data and the separability of the classes. In K-NN the cost of the learning process is zero; no assumptions about the characteristics of the concepts to learn have to be done; complex concepts can be learned by local approximation using simple procedures.

For this study, we used the SVM algorithm due to the above and the following justifications (Bzdok et al. 2018) for our model to identify idioms from literals. SVM is less computationally demanding than KNN and is easier to interpret but can identify only a limited set of patterns. On the other hand, KNN can find very complex patterns but its output is more challenging to interpret.

KNN can create class boundaries that may be less interpretable than those of linear SVM. SVM can achieve good prediction accuracy for new observations despite large numbers of input variables, whereas the classification performance of KNN rapidly deteriorates when searching for patterns using high numbers of input variables because equal attention is given to all variables.

SVM only needs a small subset of training points to define the classification rule, making it often more memory efficient and less computationally demanding when inferring the class of a new observation. In contrast, KNN typically requires higher computation and memory resources because it needs to use all input variables and training samples for each new observation to be classified.

For an SVM, the better hyperplane is the one that has the highest margin between the two groups. The maximum width of the slab parallel to the hyperplane with no interior data points is called the margin. The more distance from the boundary the more chance to be classified but little space means there is high noise for misclassification (Bzdok et al. 2018). Our model uses two classes, so, to classify the new input X as either class one or class two we used the following equation.

$g(x)= W^t X+b=0$ ---equation (3.2)

Where

W=line perpendicular to the hyperplane

b=position of a hyperplane in a feature vector

The above equation draws a straight line to classify the testing dataset into two classes. We used Matlab programing to identify the class of our dataset. The model built based on the following parameters. The simulation result is shown on appendix E.

3.3. Experimental Setup

To develop the model, we used the following parameters

Environment	Version	Remark
Python	Anaconda Navigator 3 python 3.7	
Jupyter note book	For implementation	
Tensor flow library	For model development	
Pandas	For data manipulation and analysis	
Numpy	For array	
Matlab programing	2014a	

3.4. Model Performance Measurement

The model needs performance measurements to analyze the result through Accuracy, precision, F-score, and Recall.

Accuracy:- it is a measure of how closely the experimental results agree with a true or accepted value. It can be calculated as the ratio of correctly predicted values to the total number of prediction

$$\text{Accuracy} = \frac{\text{Tp}}{\text{Total number of predictions}} \quad \text{--------------------------------} \ (3.3)$$

Precision: - is the proportion of instances that are correctly classified which are true positive instances.

$$\text{Precision} = \frac{\text{Tp}}{\text{Tp} + \text{Fp}} \quad \text{---} \ (3.4)$$

Where Tp is the total number of true positives and Fp is the total number of false positives

Recall: - is the proportion of instances that are classified correctly over the total number of instances in the test dataset

$$\text{Recall} = \frac{\text{Tp}}{\text{Total number of instances}} \quad \text{---------------------------------} \ (3.5)$$

Where Tp is the total number of true positives

F-Score: - It is the weighted average of precision and recall

$$\text{F\_score} = \frac{2(\text{Recall} * \text{precision})}{(\text{Recall} + \text{precision})} \quad \text{---} \ (3.6)$$

According to (Peng et al. 2010) we defined the following terms as following due to different reasons.

True Positives (TP):- from the test sentences idiomatic expressions are identified as idiomatic.

True Negatives (TN): from the test sentences literal expressions are identified as literal

False Positives (FP): from the test sentences literal expressions are identified as idiomatic

False Negatives (FN): from the test sentences idiomatic expressions are identified as literals.

CHAPTER FOUR

4. RESULT AND DISCUSSION

Under this chapter, we presented detail experimental processes of the model and we would show the performance that we achieved through this model. In chapter three we have discussed dataset collection, vector representation, and identification mechanisms and in this chapter, we have discussed the implementation.

4.1. Dataset Distribution

To design the model, we used a phrase-level dataset, and then vector representations of our corpus were used to train and evaluate it. According to the analysis (Salton et al. 2016) that is shown below, the distribution of our dataset for training is 80% and for testing 20%.

Table 4. 1:- dataset distribution

	Train (80%)	Test (20%)	Remark
Idioms	400 expressions	100 expressions	
Literals	400 expressions	100 expressions	
Total	800 expressions	200 expressions	

4.2. Word2Vec Representation

We used vector representation of expressions to define Amharic idioms, according to (Grzegorczyk 2019; Salton 2017). Word2Vec representation of texts is used to support word-by-word representation in this case. After representing each word in to vectors, we take combination of two words to prepare our training and testing dataset. Sample prepared datset is placed at appendix D.

We represent expressions in two-dimensional spaces and do not take into account details about their meanings. The two-dimensional spaces are represented by the Word2Vec representation of expressions (Ma and Zhang 2015). To begin, we use Python to implement the *word2int* algorithm, which converts text to real numbers with predefined window size.

Figure 4. 1:- word to integer representations

```
In [117]: word2int = {}
          for i,word in enumerate(words):
              word2int[word] = i

          sentences = []
          for sentence in corpus:
              sentences.append(sentence.split())
          WINDOW_SIZE = 1
          data = []
          for sentence in sentences:
              for idx, word in enumerate(sentence):
                  for neighbor in sentence[max(idx - WINDOW_SIZE, 1) : min(idx + WINDOW_SIZE, len(sentence)) + 1]
                      if neighbor != word:
                          data.append([word, neighbor])
```

```
In [8]: print(word2int)

{'': 0, 'በላዉ': 1, 'ዝግጅት': 2, 'አብርድ': 3, 'ክፉ': 4, 'ምንዝር': 5, 'ህይወት': 6, 'አእምሮ': 7, 'አሟልቶ': 8, 'ክፍት': 9, 'ሙጢ': 10, 'ስማ': 1
1, 'የቃል': 12, 'መማለጃ': 13, 'ቆጡ': 14, 'እቅዳቸዉን': 15, 'በጎደለ': 16, 'ሱሪ': 17, 'ላላ': 18, 'ቦታ': 19, 'አለቃዉን': 20, 'እቅድ': 21, 'ብቻዉ
ን': 22, 'አለቃችን': 23, 'ተመለከቱ': 24, 'ስህተት': 25, 'ገበሬዉ': 26, 'ቀረሽ': 27, 'ሞረደዉ': 28, 'ክረምቱ': 29, 'ከነጻ': 30, 'ድርና': 31, 'ምስክር': 3
2, 'አጠበ': 33, 'ከምታበላሽ': 34, 'ሽታዉ': 35, 'መርቅና': 36, 'ይንሳዉ': 37, 'የሩቅ': 38, 'ዉስጣዊ': 39, 'ፈሊጣዊ': 40, 'ጣና': 41, 'ተማረ': 42, 'ሹ
መት': 43, 'ፈጠራ': 44, 'እንዲያ': 45, 'አያስፈልግም': 46, 'ሰብለ': 47, 'እዝል': 48, 'ዘመናዊ': 49, 'እየነቀነቁ': 50, 'ተያዘ': 51, 'አስተማረ': 52, 'ምታ
ት': 53, 'ያሉት': 54, 'እንጂ': 55, 'ህቅ': 56, 'ለመስከ': 57, 'አጣምሮ': 58, 'ከንፈሯን': 59, 'ወንበዴ': 60, 'ምስጋናዉን': 61, 'አረገፈዉ': 62, 'ወረደ':
63, 'ሰርግና': 64, 'ጠናና': 65, 'ለፍቶ': 66, 'ሻከረ': 67, 'የጌቶች': 68, 'ሀሞቱ': 69, 'አርሶ': 70, 'አካል': 71, 'መጣ': 72, 'ጉዳዩን': 73, 'መዉጣት':
74, 'አንድሮሜዳ': 75, 'መጋረድ': 76, 'ሀብተ': 77, 'ሌባ': 78, 'አማጣኝ': 79, 'ጥኑ': 80, 'የማይታሰብ': 81, 'ሰረሰረ': 82, 'እረስሃለሁ': 83, 'ስነ': 84,
'ምርጊት': 85, 'መለስ': 86, 'አንዲር': 87, 'ዘር': 88, 'ዝናብ': 89, 'ራሱን': 90, 'ለየ': 91, 'ረገበ': 92, 'በዚህ': 93, 'ምቾ': 94, 'አርቶዶክስ': 95, 'ይ
ሸበሩ': 96, 'ተፋችዉ': 97, 'መንጥሮ': 98, 'ይቅለለዉ': 99, 'ከፍሎች': 100, 'ሁለት': 101, 'ፊቱ': 102, 'እንደሱ': 103, 'ጥፉ': 104, 'ነገስት': 105, 'የቆ
ንጆ': 106, 'ቁመተ': 107, 'ይዞ': 108, 'ያለች': 109, 'መሮ': 110, 'ሰአቱ': 111, 'ይግባዉ': 112, 'ታላቁ': 113, 'አስብቶ': 114, 'አነበበዉ': 115, 'ብለ
ዉ': 116, 'የሆነ': 117, 'መስቀለኛ': 118, 'ኢየሱስ': 119, 'ኮሶ': 120, 'አፈር': 121, 'ደካማ': 122, 'እንጠይቃለን': 123, 'አለቀ': 124, 'ታጋሽ': 125,
```

As shown in Figure 4.1 above, the *word2int* algorithm is used to convert total words to real numbers. Each word is represented by a single neighbor expression with a window size of one, which means that each word is assigned an integer value based on one neighbor word.

Example: - Encoding of expressions of window size one

'በላዉ': 1, 'ዝግጅት': 2, 'አብርድ': 3, 'ክፉ': 4, 'ምንዝር': 5, 'ህይወት': 6, 'አእምሮ': 7, 'አሟልቶ': 8, 'ክፍት': 9, 'ሙጢ': 10, 'ስማ': 11, 'የቃል': 12, 'መማለጃ': 13, 'ቆጡ': 14, 'እቅዳቸዉን': 15, ' በጎደለ': 16, 'ሱሪ': 17, 'ላላ': 18, 'ቦታ': 19, 'አለቃዉን': 20, 'እቅድ': 21, 'ብቻዉን': 22, 'አለቃችን ': 23, 'ተመለከቱ': 24, 'ስህተት': 25, 'ገበሬዉ': 26, 'ቀረሽ': 27, 'ሞረደዉ': 28, 'ክረምቱ': 29, 'ከነ

ዳ': 30, 'ድርና': 31, 'ምስክር': 32, 'አጠበ': 33, 'ከምታበላሽ': 34, 'ሸታዉ': 35, 'መርቅና': 36, 'ይንሰዉ': 37, 'የሩቅ': 38, 'ዉስጣዊ': 39, 'ፈሊጣዊ': 40, 'ጣና': 41, 'ተማረ': 42, 'ሹመት': 43, 'ፈጠራ': 44, 'እንዲያ': 45, 'አያስፈልግም': 46, 'ሰብለ': 47, 'እዝል': 48, 'ዘመናዊ': 49, 'እየነቀነቁ': 50, 'ተያዘ': 51, 'አስተማረ': 52, 'ምታት': 53, 'ያሉት': 54, 'እንጂ': 55, 'ህቅ': 56, 'ለመስክ': 57, 'አጣምሮ': 58, 'ከንፈሯን': 59, 'ወንበዴ': 60, 'ምስጋናዉን': 61, 'አረገፈዉ': 62, 'ወረደ': 63, 'ሰርግና': 64, 'ጠናና': 65, 'ለፍቶ': 66, 'ሻከረ': 67, 'የጌቶች': 68, 'ሀሞቱ': 69, 'አርሶ': 70, 'አካል': 71, 'መጣ': 72, 'ጉዳዩን': 73, 'መዉጣት': 74, 'አንድሮሜዳ': 75, 'መጋረድ': 76, 'ሀብተ': 77, 'ሌባ': 78, 'አማጣኝ': 79, 'ጥኑ': 80, 'የማይታሰብ': 81, 'ሰረሰረ': 82, 'እረስሃለሁ': 83, 'ስነ': 84, 'ምርጊት': 85, 'መለስ': 86, 'አንዲር': 87, 'ዘር': 88, 'ዝናብ': 89 ...

However, to represent N-dimensional spaces, we used vector representation. As a result, we converted each real number into a vector. We used a length of words to convert real numbers into vectors of several vocabulary datasets. In the vector space notation, it is a sparse vector: that is a vector with one at only a single position and zeroes at other remaining positions.

When one hot encoding representation has an N-number of input expressions, it is represented in N-dimensional space.

Example: - we take five sample expressions to encode in vector encoding mechanism. It represents through five-dimensional spaces.

Table 4. 2:- one hot encoding representation

Expression	One Hot Encoding
በረዶ	[1, 0, 0, 0, 0]
ራሱን	[0, 1, 0, 0, 0]
መምህር	[0, 0, 1, 0, 0]
ትምህርት	[0, 0, 0, 1, 0]
ብትባል	[0, 0, 0, 0, 1]

On the hidden layer, the algorithm produces vector weight values. The neural network implementation is used to transform input data into output data. The expressions are converted to six-digit floating-point numbers.

Figure 4. 2:- hidden layer vectors generation

```
In [12]: # training operation\n",
         train_op = tf.train.GradientDescentOptimizer(0.05).minimize(loss)

         sess = tf.Session()
         init = tf.global_variables_initializer()
         sess.run(init)
         vectors = sess.run(W1+b1)
         print(vectors)

         [[-1.6749547  -1.9711727 ]
          [-1.9427724  -0.45057726]
          [-2.5033765  -1.5287652 ]
          ...
          [-1.045727   -1.3643637 ]
          [ 1.1842918  -0.37351602]
          [-2.2666798  -1.0364703 ]]
```

As shown in Figure 4.2 above, to encode the input dataset into a numeric value, it produces an N-number of hidden layer weight values. For 2D visualization, we used a window size of one and an embedding dimension of two. The length of words is taken into account in one-hot encoding, and the embedding dimension we used is two. The random normal value of one-hot encoding and embedding dimension is the weight value. We used matrix multiplication of the input vector with the weight value to produce the hidden layer value. The value of the hidden layer is used as input to produce the output value. For optimization of the output vector, we used softmax to do cross-entropy to optimize the model during training. Cross-entropy loss is used when adjusting model weights during training. The objective is almost always to minimize the loss function.

From the above figure 4.2, the lost value is decreased when the iteration is increased means the better model was used to represent expressions by numbers.

Figure 4. 3:- numeric value of the expressions

```
In [31]: #pd.set_option('display.max_rows', None) #used to display all rows
         w2v_df = pd.DataFrame(vectors, columns=['x','y'])
         w2v_df['Expressions'] = words
         w2v_df = w2v_df[['Expressions', 'x','y']]
         print(w2v_df)
```

```
67          ሻከረ  0.299475  2.641170
68        የጌቶች  1.947000  0.302563
69          ሀሞቴ  0.855792  0.965415
70          አርሶ  1.930915 -0.549882
71          አካል  0.769243  0.102756
72           መጣ  2.272143  2.080637
73        ጉዳዩን  1.736586  2.639151
74        መውጣት  1.433603 -0.065078
75     አንድሮሜዳ  2.232865 -0.533422
76        መጋረድ  3.249513  0.572706
77          ሀብተ  2.772535  0.704184
78           ሌባ  1.450898  0.522112
79        አማጣኝ -1.032888  0.825241
80           ጥኑ -0.511198 -0.593538
81     የማይታሰብ  2.178990  0.391314
82        ሰረሰረ  1.688567  0.095116
83     እረስሃለሁ  2.332433  0.654770
84           ስነ  2.559075  2.575535
85        ምርጊት  0.590903 -0.110936
86          መለስ  1.220281  0.269581
```

As shown in Figure 4.3 above, six floats are used to describe expressions numerically in two-dimensional spaces. This expression's numeric value was obtained by multiplying the hidden layer value by the weight values in a matrix. This weight value is the one-hot encoding and embedding dimension's random normal value. The figure shows the numeric representation of expressions to make suitable input for idiom identification using a supervised machine learning algorithm. The representation is generated using the Word2Vec model. To show the representation on two-dimensional space for visualization is using *matplotlib* python library.

4.3. Train and Test the Model

We used the SVM supervised machine learning algorithm to distinguish idioms from literals after converting the dataset into numeric values (Bzdok et al. 2018; Sharma et al. 2017). The algorithm was trained using our labeled dataset and was able to recognize the test dataset's class.

The best hyperplane that distinguishes all data points of one class from those of the other class is found by an SVM.

Figure 4. 4:- parameters used to build the model

Variables - svmmodel

svmmodel

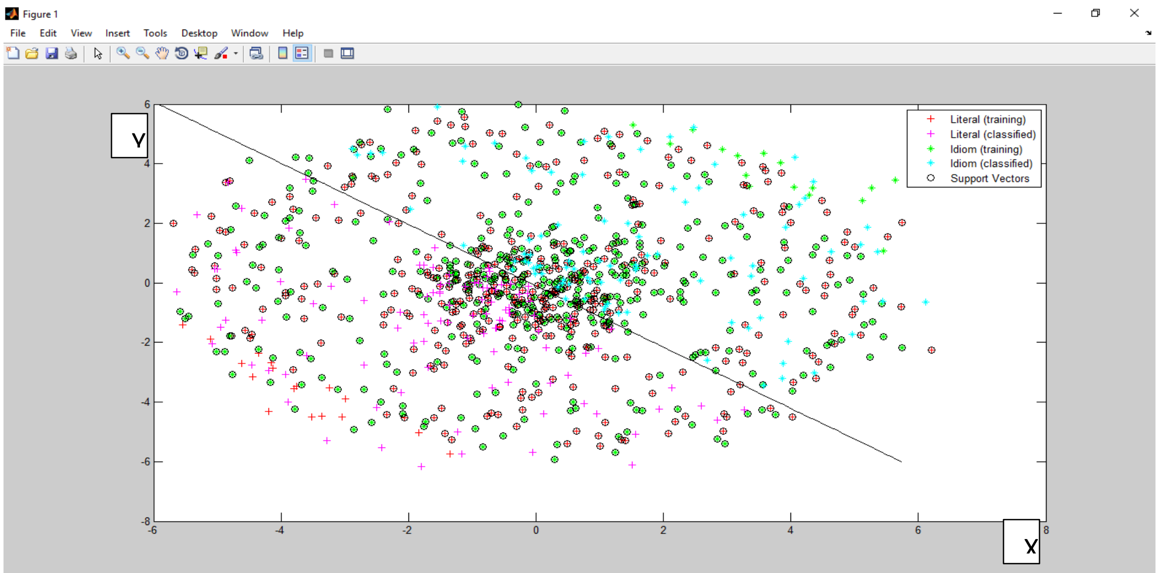
1x1 struct with 9 fields

Field	Value	Min	Max
SupportVectors	*767x2 double*	-2.3920	2.5171
Alpha	*767x1 double*	-1	1
Bias	-0.0117	-0.0117	-0.0117
KernelFunction	@linear_kernel		
KernelFunctionArgs	*0x0 cell*		
GroupNames	*800x1 cell*		
SupportVectorIndi...	*767x1 double*	1	800
ScaleData	*1x1 struct*		
FigureHandles	*1x3 cell*		

As shown in Figure 4.4 above, this model's support vectors had a minimum of -2.3920 and a maximum of 2.5171. The training model used -0.0117 biases to make uniformity dependent on the dataset, and it used a linear kernel function to find the best line to divide the plane into two equal margins. Based on the above parameters, the simulation result for the training and testing dataset is given in Figure 4.5 below.

When we implement the above parameters, our model classified the dataset in to two classes as idiom or literal. The classification is done using a dataset of combination of two words. The model recognized the idiomatic expressions from literals based on support vectors.

Figure 4. 5 :- 2D representation of testing dataset



As shown in Figure 4.5 above, training and testing expressions on two-dimensional spaces are represented. We used SVM supervised machine learning algorithm of linear kernel function to identify the testing dataset. So, the model finds the optimal line based on support vectors (Bzdok et al. 2018). The *svmtrain* function was used to train the model using the training dataset. The model trained based on labeled data because we used SVM supervised machine learning methods (Bzdok et al. 2018). This training model is used to identify the test dataset for the correct class and achieve the objective. We used the *svmclassify* function to identify the class of the testing dataset and the *showplot* function to represent the training and testing dataset on two-dimensional spaces. From Figure 4.5, red crosses represented training literals expression, green crosses represented training idioms expression; orange crosses represented literals which are classified through the SVM model, light green crosses represented idioms which are classified through the SVM model based on support vectors to classify testing dataset to the correct class. Figure 4.5, draws the better hyperplane that has the highest margin between the two groups. Literals are placed below the line and idioms are placed top of the line based on the training support vectors. As to the knowledge of the researchers, we didn't find previous research works on Amharic idiom identification to compare our model to others. So, we focus on the dataset type used, preprocessing, and vectorization to discuss the results of the model. In this study, we have used the one-hot encoding method and

Word2Vec model for the representation of expressions into numeric values.

4.4. Model Performance Evaluation

To assess the model's accuracy, we used a supervised machine learning algorithm. Under the python programming environment, we used accuracy, f1-score, recall, and precision. Finally, using the KNN supervised machine learning algorithm, we were able to achieve a 97.5 percent accuracy rate.

When the KNN supervised machine learning algorithm is used, the overall performance of the model obtained is 97.5 percent accuracy on the given testing dataset. As we have seen in the literature review, identification of idioms using word embedding and idiom token classification using distributed semantics achieves a better result than others (Peng and Feldman 2016a; Salton et al. 2016). The other performance measurement results are listed in the following table.

Table 4. 3:- performance result of model

Class	Precision	Recall	f1-score	Accuracy	Remark
Idiom	95%	100%	98%	97.5%	KNN

When we compared the model's success to that of others discussed in related works,

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

Idiomatic expressions are a natural part of all languages and a common part of our everyday conversation. Idioms are one of the most difficult and fascinating aspects of Amharic vocabulary, as they cannot be deduced directly from the word from which they are created.

Idiomatic expressions are taken from an Amharic idioms book, while literal expressions are taken from various Amharic texts. To delete unrelated and irrelevant symbols from the collected dataset, we used preprocessing. Under preprocessing, cleaning text and spell correction, Character Normalization, Tokenization and join tokens, and Remove stop words & numbers were done. Machine learning algorithms do not process text as input so, they require encoding of texts to another format. For such encoding, we used the Word2Vec model to encode texts into numeric or vector forms. We created an automated idiom recognition model for the Amharic that is used to improve NLP tasks.

Identification of idiomatic expressions is more important for NLP-related applications such as machine translation, sentiment analysis, and semantic analysis, as discussed in the literature. The Word2Vec approach was used in the analysis to represent the expression as vectors. We used a corpus of one thousand idiomatic and literal expressions. A supervised machine learning algorithm was used to identify idiomatic expressions, with 80% of the corpus being used for training and 20% for testing. To assess the model's performance results, we used accuracy, precision, recall, and F-score. We were able to achieve an output accuracy of 97.5 percent. We collected and digitalized the hard copy book of Amharic idiom book, which is important for book readers and writers they used as a lookup table. NLP researchers basically used our model, to recognize the expression is idiom or literal befor doing any task.

FUTURE WORK

The study recommends that researchers and practitioners in the field incorporate WSD and implement deep learning algorithms to enhance the model's accuracy and identifying the nature of Amharic idiom (pure or semi-pure) needs further study in this field. It is preferable to use Amharic spell checkers to correct spelling errors during the preprocessing task. This study aims to find a way to mitigate the negative effects of idiomatic expressions on NLP applications. However, this research only gathered and used five hundred idiomatic expressions that are made up of two terms. As a result, it is preferable to include other Amharic idioms that are single words or a combination of three words, as well as expanding the dataset. We used phrase level idiom identification to develop the model and recognize a combination of two word idiom classes; where as, extract idiomatic expressions from large size corpus is my recommendation for future work on this area.

REFERENCE LIST

Ahmadi, M. (2017). "A contrastive Analysis of Idioms and Idiomatic expressions in Three English and Persian Novels for Translation purpose." *Language Art, Shiraz, Iran* 103-116.

Akililu, A., and Worku, D. (1992). *የአማርኛ ፈሊጦች (Amharic Idioms)*, Addis Abeba, Ethiopia: Kuraz publishing agency.

Alemayehu, A. (1996). *ፍቅር እስከ መቃብር (Fikir eske mekabir)*, Addis Abeba, Ethiopia: Mega Publishing Agent.

Bouamor, H., Max, A., and Vilnat, A. (2011). "Monolingual Alignment by Edit Rate Computation." *Association for Computational Linguistics*, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 395-400.

Bzdok, D., Krzywinski, M., and Altman, N. (2018). "Machine learning: Supervised methods, SVM and kNN." *Nature Publishing Group*, Nature Methods, hal-01657491 1-6.

Caillies, S. (2007). "Processing of Idiomatic Expressions." *Online* 79-108.

D.Salton, G., Ross, R. J., and Kelleher, J. D. (2017). "Idiom Type Identification with Smoothed Lexical Features and a Maximum Margin Classifier"*International Conference Recent Advances in Natural Language Processing* City: Bulgaria.

Diab, M., and Bhutada, P. (2009). "Verb noun construction MWE token supervised classification." *Association for Computational Linguistics*, In Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications., 17–22.

Eshetu, A., Teshome, G., and Abebe, T. (Aug-2020). "Learning Word and Sub-word Vectors for Amharic (Less Resourced Language)." *International Journal of Advanced Engineering Research and Science (IJAERS)*, Vol-7(Issue-8).

Farhadloo, M., and Rolland, E. (2016). "Fundamentals of Sentiment Analysis and Its Applications." *Springer International Publishing Switzerland*, Sentiment Analysis and Ontology Engineering,Studies in Computational Intelligence 639.

Fazly, A., Cook, P., and Stevenson, S. (2009). "Unsupervised Type and Token Identification of Idiomatic Expressions." *Computational Linguistics* 35, 61-103.

Fritzinger, F., Weller, M., and Heid, U. (2010). "A Survey of Idiomatic Preposition-Noun-Verb Triples on Token Level." *European Language Resources Association (ELRA)*, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).

Getinet, Y. (2015). *Unsupervised Part of Speech Tagging for Amharic*, University of Gondar, Ethiopia.

Grzegorczyk, K. (2019). *Vector representations of text data in deep learning*, Doctoral dissertation, AGH University of Science and Technology.

Hashimoto, C., Sato, S., and Utsuro, T. (2006). "Japanese Idiom Recognition: Drawing a Line between Literal and Idiomatic Meanings." *Association for Computational Linguistics*, Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages Sydney,, 353–360.

Koehn, P., Och, J., and Marcu, D. (2003). "Statistical Phrase Based Translation." *Proceedings of HLT-NAACL 2003 Edmonton, May-June 2003*.

Lowri, W., Christian, b., Michael, A., Alun, P., and Irena, S. (2015). "The Role of Idioms in Sentiment Analysis." *Expert Systems with Applications; Journal homepage: www.elsevier.com/locate/eswa*, 7375-7385.

Ma, L., and Zhang, Y. (2015). "Using Word2Vec to process big text data " *Big Data (Big Data), 2015 IEEE International Conference*. City: IEEE: Santa Clara, CA, USA.

Mantyla, K. (2004). "Idioms and Language Users: the effect of the characterstics of idioms on thier recognition and interpretation by native and non-native speakers of English." *Jyväskylän yliopisto*.

Mikolov, T., V.Le, Q., and Sutskever, I. (2013). "Exploiting Similarities among Languages for Machine Translation." *arXiv:1309.4168v1 [cs.CL]*.

Miretie, S. G., and Khedkar, V. (2018). "Automatic Generation of Stopwords in the Amharic Text." *International Journal of Computer Applications (0975-8887)*, 180-No.10.

Muzny, G., and Zettlemoyer, L. (2012). "Automatic idiom identification in Wiktionary ".

Peng, J., and Feldman, A. (2016a). "Automatic Idiom Recognition with Word Embeddings." *Springer Verlag*, Information Management and Big Data - 2nd

Annual International Symposium, SIMBig 2015 and 3rd Annual International Symposium, SIMBig 2016, Revised Selected Papers, 17-29.

Peng, J., and Feldman, A. (2016b). "Experiments in Idiom Recognition." *Computational Linguistics*, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, papers 2752–2761.

Peng, J., Feldman, A., and Street, L. (2010). "Computing Linear Discriminants for Idiomatic Sentence Detection." *Montclair, NJ 07043, USA.*

Philemon, W., and Mulugeta, W. (2014). "A Machine Learning Approach to Multi-Scale Sentiment Analysis of Amharic Online Posts." *HiLCoE Journal of Computer Science and Technology, No. 2*, 2.

Rase, M. O. (2020). "Sentiment Analysis of Afaan Oromoo Facebook Media Using Deep Learning Approach." *New Media and Mass Communication*, 90, 2020(2224-3267 (Paper) ISSN 2224-3275 (Online)).

Sainia, I., Singhb, D., and Khoslaa, A. (2013). "QRS detection using K-Nearest Neighbor algorithm (KNN) and evaluation on standard ECG databases." *Journal of Advanced Research*, vol. 4, no. 4, 331–344,.

Salton, G., Ross, R. J., and Kelleher, J. D. (2016). "Idiom Token Classification using Sentential Distributed Semantics." *Association for Computational Linguistics*, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 194–204.

Salton, G. D. (2017). "Representations of Idioms for Natural Language Processing:Idiom type and token identifification, Language Modelling and Neural Machine Translation." *Association for Computational Linguistics*, Proceedings of The 8th International Joint Conference on Natural Language Processing.

Seetha, M., Deekshatulu, B., and Muralikrishna, I. (2008). "ARTIFICIAL NEURAL NETWORKS AND OTHER METHODS OF IMAGE CLASSIFICATION." *Journal of Theoretical and Applied Information Technology*, 4 No11.

Sharma, Y., Agrawal, G., and Jain, P. (2017). "Vector Representation of Words for Sentiment Analysis Using GloVe." *IEEE*, 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT) Manipal University Jaipur.

Singh, R., and Hanumanthappa, M. (2016). "Techniques of Semantic Analysis for Natural Language Processing: A Detailed Survey." *International Journal of Advanced Research in Computer and Communication Engineering (ICRITCSA)*, 5(2).

Thirumuruganathan, S. (2017). "A detailed introduction to K-nearest neighbor (KNN) algorithm." *wordpress.com weblog*.

Verma, R., and Vuppuluri, V. (2015). "A New Approach for Idiom Identification Using Meanings and the Web." Proceedings of Recent Advances in Natural Language Processing, , 681–687.

Vilar.D.et.al. (2009). "Error Analysis of Statistical Machine Translation Output." *See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/307174612*.

Williams, L., Bannister, C., Ayllon, M., Preece, A., and Spasić, I. (2015). "The role of idioms in sentiment analysis." *Elsevier*, Expert Systems with Applications; Journal homepage: www.elsevier.com/locate/eswa, 7375-7385.

Xu, Z., Chen, M., and Weinberger, K. Q. (2013). "An alternative text representation to TF-IDF and Bag-of-Words." *Springer,* Encyclopedia of Machine Learning,Boston, MA.

Zhang, Y., Jin, R., and Zhou, Z.-H. (2010). "Understanding bag-of-words model: A statistical framework." *Springer*, International Journal of Machine Learning and Cybernetics December 2010.

Zhu, C., Tang, J., Li, H., Tou, H., and Zhao, T. (2007). "A Unified Tagging Approach to Text Normalization." *Association of Computational Linguistics*, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, Association for Computational Linguistics, 688–695.

Zong, Z., and Hong, C. (2018). "On Application of Natural Language Processing in Machine Translation"*2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. City: IEEE: Huhhot, China.

ይማም, ባ. (1987). " የአማርኛ ሰዋሰው." *ት.መ.ማ.ማ.ድ.*

Appendix A: Stop Words & Numbers

ሁሉ

ሁኔታ

ሆነ

ሆኖም

መሆኑ

ማለት

ሲሆን

በጣም

ብቻ

ተባለ

ተገለጸ

ነበሩ

ነበረ

ነው

0,1 ,2,3,4,5,6,7,8,9

Appendix B: Corpus Preparation

ሀሞቱ ፈሰሰ, ሀረግ መዘዘ, ሀረግ ጣለ, ሀሳበ ቢስ, ሀሳብ ገባዉ, ሀብተ ስጋ, ሀብተ ሰባራ, ሀብተ ስንኩል, ሀብትሽ በሀብቴ, ሀብተ ቢስ , ሀብተ ነፍስ, ሀብቷ ቀና, ሁለት ምላስ, ሁሉ አማረሽ, ሁሉ አገርሽ, ህሊና ቢስ, ህቅ አለ, ህግ ተላለፈ, ህግ አፈረሰ, ህገ ወጥ, ህግ ገባ, ሰዉ ሆነ, ልቡ ተሰነጠቀ, ነፍስ ሆነ, ዋስ ሁነኝ, አስብቶ አራጅ, የሆነዉ ሆኖ, ቁመተ ስጋ, ሆደ መጋዝ, ሆደ ሰፊ, ሆደ ቡቡ, ሆደ ባሻ, ሆደ ገር, ሆዱ ሻከረ,ሆዱን ቆረጠዉ, ሆዱ ባሰበት, በጥፊ ለጠፈበት, ለፍቶ መና, በዱላ አለፋዉ, ዐይነ ልም, ላስ አደረገዉ, መሬት ላሰ, መሬት አስላሰ, ማርም አልስ, እሳት የላሰዉ, የሚላስ የሚቀመስ, አዕምሮዉ ላሽቋል, ላባ ቀረሽ, ላክ አደረገ, ሲበሉ የላኩት, የሰይጣን መልዕክተኛ, የበላይና የበታች, ላጤ መላጤ, ሊቀ ላጤ, ወንደ ላጤ, የብረት ልጥ, በደረቁ ላጨ, እሳት የላፈዉ, ገንዘቤን ላፈኝ, ዉሸቱን ላፈዉ, ሌላ ነዉ,ሌባ ሚዛን, ሌባ ዉጋት, ሌባ ዝናብ, ሌባ ጣት, አሰለጥ ሌባ, አይነ ሌባ, ሌባ ሻይ, ልሳነ እሳት, ሁለት ልብ, ልበ ልል, ልበ ሙሉ, ልበ ሙት, ልበ ሞቃት, ልበ ሰፊ, ልበ ቀላል, ልበ ቅን, ልበ ቆራጥ, ልበ ቢስ, ልበ ባህር, ልበ ብር, ልበ ብርቱ, ልበ ተራራ, ልበ ደንዳና, ልበ ድንጋይ , ልበ ደፋር, ልበ ድንጉጥ, ልበ ድፍን, ልበ ገር, ልበ ጎረምሳ, ልበ ጎደሎ, ልበ ጠናና, ልበ ጡል, ልበ ጥልቅ, ልበ ጥሩ, ልበ ጥኑ, ልበ ፈሪ, ልቡ አበጠ, ልቡን አወለቀ, ልቡ ሞቀ, ልቡ ሰባ, ልቡ ላላ, ልቡ ረጋ, ልቡ ራሰ, ልቡ ሸፈተ, ልቡ ቀረ, ልቡ ቀዘቀዘ, ልቡ ቆመ, ልቡ ቆረጠ, ልቡ ቆሰለ, ልቡ ቋመጠ, ልቡ ባከነ, ልቡ ተሸበረ, ልቡ ተቀሰቀሰ, ልቡ ተቃጠለ, ልቡ ታወከ, ልቡ አረፈ, ልቡ አላረፈም, ልቡ ወላወለ, ልቡ ከዳዉ, ልቡ ወደደ, ልቡ ዋለለ, ልቡ ገባ, ልቡ ፈረሰ, ልቡና ይስጥህ, ልቡን መታዉ, ልቡን ማረከዉ, ልቡን ሰለበዉ, ልቡን ሰለበችዉ, ልቡን ሰቀለዉ, ልቡን ሰወረዉ, ልቡን ሰጠዉ, ልቡን ሸነጠዉ, ልቡን ቀረቀረዉ, ልቡን በላዉ, ልቡን ነሳዉ, ልቡን ነካዉ, ልቡን ከዳዉ, ልብ ለልብ, ልቤን ማረከችዉ, ልቤን አልነካኝም, ልብ ከዳ, ልብ ቀረዉ, ልብ ነሳዉ, ልብ አለ, ልብ አሳጣ, ልብ አቁስል, ልብ አብርድ, ልብ አብሽቅ, ልብ አብግን, ልብ አዉልቅ, ልብ አደረገ, ልብ አድርቅ, ልብ አድርግልኝ, ልብ አጠፋ, ልብ አጣ, ልብ አፈረሰ, ልብ ወረደዉ, ልብ የማይገባ, ልብ አዉልቅ, በልብ አዋለ, በእንቅልፍ ልብ, ከልቡ ተናገረ, ከልቡ አገለገለ, ከልቡ ወደደ, ተረከዘ ሎሚ, ምላሽ ጠራ, ሰርግና ምላሽ, ነገር አመላላሽ, መለስ ያለ, ቤት በመለስ, አዕምሮዉ ተመለሰ, የበይ ተመልካች, መላ ጣይ, በመላ ሄደ, መላ መታ, መላ ቅጡ, በመላ ተናገረ, መልከ መልካም, መልከ ጥፉ, የቆንጆ መራራ, ደመ መራራ, ሰኔና ሰኞ, ምርር አለ, ሰነፍ መረቅ, ምርዝ አደረገ, መርዞ ያዘ, ምርጊት ፊት, ሙርጥ አወጣ, መሪዉን ቀጣ, መሬት ይቅለለዉ, መሮ ጥርስ, ምስለ ቢስ, ምስል አፍሳሽ, ተምሳሌተ ብዙ, አይመስሉ መስሎ, አዉራ ምስክር, ዝርዝር ምስክር, ምስጋና ቢስ, ምስጋና ይንሳዉ, ምስጋና ይግባዉ, እርፈ መስቀል, መስከረምሲጠባ, መስከረም ሲብት, መስክ ለመስክ, ምድር መሸበት, መቀስ አፍ, ዘር መተረ, ልቡን መታወ, መሰንቆ መታ, መታ ያለ, መላ መታ, መንገድ መታዉ, ምች መታዉ, ራስ ምታት, ስልክ መታ, በረዶ መታዉ, በር መታ, በገና መታ, ባዝራዎ ተመታች, ቤት መታ, አድማ መታ, አንበጣ መታዉ, እሳት መታዉ, የዉሻ ቁስል, ጠበል አስመታ, መቼ አጣሁት, መቼ ጠፋኝ, የቤት ልጅ, ስሙን መነዘረዉ, አለቃና ምንዝር, ብልት አወጣ, መንጥሮ አጠበ, ጀርባዉን መነጠረዉ, ቆሽቱ አረረ, መና ቀረ, መንታ ልብ, መከራ ስጋ, መከራ ቀመሰ, መከራዉን በላ, የድሃ መከታ, የወላድ መካን, አንገቱን መዘዘ, ዱላ መዘዘ, ምድር አስጋጠ, ምድር ለቀቀ, ምድር መታች, ምድር ቀለጠ, ምድር ተፋችዉ, ምድር ዞረበት, ምድር ዋጠዉ, ምድር ያዘ, ምድር ፊት, ያንበሳ መደብ, ከንፈሯን መገመገዉ, ምግበ ነፍስ, ምግበ ስጋ, ሊጡ መጠጠ, ስጋዉ መጠጠ, ቁስሉ መጠጠ, አይቡ መጠጠ, ጭቃዉ መጠጠ, ፊቱ መጠጠ, ምጥ መጣ, ጠፍር በሊታ, መጥኔዉን ይስጥሽ, አባ ሙላት, እድሜዉ ሞላ, ሱሪ አለበስኩም, ሙሬ አፍ, አፈ ሙዝ, ሌባ ሚዛን, መማለጃ በላ, የሴት መዶለቻ, ማላአፈረሰ, ማርም አልስ, ጉድጓድ ማሰለት, ከማን አንሼ, ማንቆርቆሪያ አፍ, ማድ አረጋጭ, ማአዱን አማገጠ, ብታደራልኝ አማግሁልህ, በጦር ማገረበት, ዘንዶ ማገር, የግንባር ስጋ, ድርና ማግ, ምራቁን ዋጠ, መርቅና ፍትፍት, የሀገር ምሰሶ, የሰዉ ምሰሶ, በጥርስ ሸኘዉ, የማርያም ምሳ, ያይን ምሳ, ምስጢር አወጣ, ምስጢር አየ, ምኑ ቅጡ, ምን ቆርጦት, ምን ተዳየ, ምን አባቱ, ምን አግዶኝ, ምን አለብኝ, ምን ከፋኝ, ምን ይበጀኝ, ምን ገደደኝ, ምን ቆረጠኝ, የእንጨት ምንቸት, በሬ ወለደ, ሙልጭ አለ, ሞልጮኝ ሄደ, ሞላጫ ሌባ, ለአይን ሞላ, ሰአቱ ሞላ, ሙሉ ልጃገረድ, የጉም ሽንት, ሙሉ አካል, በጎደለ ሞላ, ትዳሩ ሞላ, ዉሃ ሙላት, ጉዳዩን ሞላ, ፈቃድ ሞላ, አሟልቶ ሰጠዉ, ረሀብ

ሞረደዉ, በትምህርት ሞረደዉ, ሞቅ አለዉ, ትዳሩ ሞቀ, የሞቀ ቤት, ጨዋታዉ ሞቀ, ልቡ ሞቷል, በቁሙ ሞተ, ሙት አማጣኝ, በሞተ ከዳ, እሳቱ ሞተ, አይኑ ሞቷል, የሞት ሞት, የሞተ እንጀራ, ወገቡን ሞከረ, ሞገስ አገኘ, ሙጢ አፍ, መሬት አሚሸ, አፉን አሚሸ, እጅ ሚሚሻ, ለዛ ሙጥጤ, ሙዋጣጭ ሌባ, ፍቅር አሚጠጠ, ሞት ይርሳኝ, በቅሎ ነች, ርባና ቢስ, ሰዉነቱ ተረታ, አንደበቱረታ, ጉልበቱ ተረታ, መድሃኒቱ ረከሰ, ሴት አስረካሽ, ርካሽ ቦታ, እድሜዉ ተበጠሰ, ረድኤታም ሴት, ረገብ አደረገ, ቂም አረገዘ, አይኑ አረገዘ, ቢረግጥ ይገል, ገበታ ረጋጭ, ሰዉነቴ ረገፈ, እርግፍ ያርገኝ, ጥርሱን አረገፈዉ, ልቤ ረጋ, ባለህበት ርጋ, የረጋ ወተት, ርጥብ ሬሳ, እግረ ርጥብ, በመጠጥ ተራጨ, በእንባ ተራጨ, ባጣ ቆይኝ, ልቡ ራሰ, አርሶ ዘነበ, አንጀቴን አርሰኝ, ራሰ ዘናና, ራሰ ክፍት, ራሱን ሳተ, ለራስ ያሉት, ራሱ ጠና, ራሱን ሰወረ, ራሱን ቀበረ, ራሱን ነቀነቀ, ራሱን አወጣ, ራሱን አዋረደ, ራስ የሌለዉ, ራሱን ጣለ, ሹመት አይደንግጥ, ራስ ስሞሽ, ልባቸዉ ተራርቋል, አሳበ ሩቅ, የሩቅ ዘመድ, የቅርብ ሩቅ, አፅመ ርስት, ሮሮ አይመረዉ, ሲሮጡ ያሰሩት, የህልም ሩጫ, መስቀል ተሰላጢን, ስልጣኑ ተያዘ, ስልጡን አፍ, ስልጡን እጅ, ክፉ ስልጣኔ, ለነገር ተሰለፈ, ሰልፍህን አሳምር, ሰመር ገጠመ, ሰመር ገባ, ምጣዱ ሰማ, ስማ በለዉ, ቀለም ገባዉ, ያልሰማ ጆሮ, ዳኛዉ ሰማ, ሰማንያ ከነዳ, ሰማንያዋ ወረደ, የሰማንያ ሚስት, ሰማይ ተደፋበት, ሰማይ ቆጡ, የሰማይ ቁጣ, የሰማይ ቤት, ሰም ለበሰ, ሰምና ወርቅ, አይኑ ቀላ, ቤት ሰረሰረ, ልቡን ሰረቀዉ, ስርቆሽ በር, አይኑ ሰረቀ, ሰረገላ ቁልፍ, ሰረገላ ዉግ, የነገር ሰረገላ, ቅቤ አንጓች, ሰርቶ አፍራሽ, ስራም አይዝ, ስራተ ቢስ, ብረት ሰሪ, ቅቤ ጠባሽ, የቤት ስራ, ገበታ ሰራ, ግፍ ሰራ, ልቤ ተሰቀለ, መስቀለኛ መንገድ, መስቀለኛ ጥያቄ, የመስቀል ወፍ, ልቡን ሰበረዉ, ህግ ሰበረ, ሁብተ ሰባራ, ሰባራ ስንጥር, ቅስሙ ተሰበረ, ትልከዉ አሽከር, ትጭነዉ ሲጋር, ፊታዉራሪ መሸሻ, የገብሬ አንዷር, እብድ ይለኛል, አበበ ሄደ, በሚኖርበት ማህበር, መሬት ነካ, በዛብህ ተሰናበተ, ይወዳት ነበር, ቤት መጣ, ለነዚህ ልጄን, በፊት የተቃጠለዉ, ትንሽ ጊዜ, የምታምር አንዷር, ከአንዷሮቹ ሁሉ, የተለየየ ሰዎችን, መግዛት ከፈጣሪ, ሀይል ተሰጠዉ, ይመስል ነበር, አልቀራቸዉም ነበር, እኔን ብቻ, የማይለወጥ ልጅ, የማንም ሰዉ, እዉነተኛ ፍቅር, እራት እንዳትቀር, በደረቱ አሰጠግቶ, እጄን ሲነካዉ, በዉሸት ፈገግታ, መቼም ወንድ, አንዳንድ ነገር, መደንገጥ አያስፈልግም, እንደድሮሽ ብታማክሪኝ, መንገድ እመራሻለሁ, ያለሽኝ አንቺ, የማትችለዉ አደጋ, ለመሸፈን ተጣጣረ, የምትወደዉን አታገባም, ፍጥጥ አድርጎ, ያስቀሩባቸዉን ጥቅም, አበጀ በለዉ, የጠላቶቻቸዉ ጥፋት, የበቀል እቅዳቸዉን, አልነካ አለ, ይሸበሩ ጀመር, ይህች ልጅ, መሆኔን እወቂ, ለቅሶ ለመዋጥ, ተነስቶ ሄደ, ሰብለ ታስራ, አባትዋ እልፍኝ, ትጠበቅ ነበረ, ደህና ሁኚ, አይዞሽ አላላትም, በበዛበት ከዚያ, ክፉ ዘመን, የርግቦችን ባህሪ, ይዞ የተፈጠረዉ, በዛብህ የወጣበት, ቤተሰብ ዘር, ማንነት ሳያግደዉ, በባለፀጋ ልጅ, በማይታይ ፍቅር, ፍቅር ጉልበቱ, ከበደ አለቃዉን, ከተናገረዉ በኋላ, ልብ ተሰምቶት, ደስ ብሎት, ለመገናኘት ያብቃን, ያብቃን አለ, ልትበሳጭ አስባ, ልታለቅስ ይገባታል, እንዲያ ሲያስቡ, ደህና ቆይተዉ, አትበሳጭ ማን, እስዋ አታለቅስ, ማን ይበሳጭ, ማን ያልቅስ, በሰፊዉ የለመደች, የነበራት ወይዘሮ, ቀን ቢጥላት, ቀን ቢከፋባት, የጌቶች ልጅ, እመቤት አልነበረችም, ቢርበኝም የእስዋን, ችግር አይቼ, ዝም አለማለቴ, ጥፋተኛዉ እኔ, ልጄን ደጋፊየን, የሚጦረኝን የሚረዳኝን, በስለት ለታቦት, መስጠት የማይሆን, የማይታሰብ ነዉ, በሙብል ጊዜ, በዛብህ ተነሳ, አማካኝቶ ምንም, ሳይበላ ተነሳ, በበዛብህ ጎጆ, ተመልሰዉ ሲፍራሩ, ብዙ ሳይጫወቱ, ምኝታቸዉ ሄዱ, የሰዉ እብድ, ምንም ምላሳቸዉን, ባይጎርሱ እኔን, ብቻ አሞገሰ, ቀንተዉ ነዉ, ለማለት ያህል, ያክል አልቀራቸዉም, እንደምንም ብለዉ, ለጊዜዉ ተቆጡ, ደማቸዉን ለማብረድ, ማብረድ ቻሉ, የሱስንዮስን ልጅ, ለገበሬ ሰዉ, ልጄን ልሰጥ, እነሱ ባያዩ, ባይሰሙ አጥንታቸዉ, አይከሰኝም ይላሉ, በቁጣ ተመለከቱ, ፈልቶ አለቀ, ራሳቸዉን እየነቀነቁ, እርስዎ ምልስጥ, እኔን መቃብር, ሳይጫነኝ አፈር, የልጄ አጥንት, ሰባራ አታገባም, እንዲያዉም አላቻዋ, አግብታ ትቅር, ከምታበላሽ ተርፋለች, የሚመስላት አጥታ, ሳታገባ ቀረች, ብትባል እንጂ, ይጨምረዋል እንጂ, እንኩዋን እንደሱ, የሚሳብ ሀሳብ, ዋጋ ሚሰጡለት, አሳቡን ሚያከብሩለት, ሰዎች ከሆኑ, በጣም ስወዳት, እንግዲያስ ለራት, እንዳትቀር አሉ, በዛብህ ስንብቱን, ምስጋናዉን አጣምሮ, ለመግለፅ ራሱ, መሬት እስኪነካ, ሲወጣ ሁልጊዜ, የማይለወጥ ልጅ, እሱን አየሁ, አሉ ፊታዉራሪ, እንኩዋንስ እንዲህ, እጄን ሲጨብጠዉ, መደንገጥ ነበረበት, የምትፈልጊዉ ነገር, እንዲፈፀምልሽ መንገድ, የሚወዳት ሰብለ, በፍቅር ታዉራ, ልታየዉ የማትችለዉን, አደጋ ስላየ, የተሰማዉን የልብ, ለወዳጅዋ ያላት, ፍቅር ይሉኝታዋን, የሚያሸንፍ ቢሆን, እንኩዋወ ላጀችዋ, ስለማይፈቅዱላት አዝና, አልቅሳ አንጀትዋ , ተቆርጦ የምትወደዉን, ሳታገባ ትቀራለች,

አለና ቦዘዝ, ብሎ በሩ, ላይ ተክሎ, ያየዉ ጀመር, በዚህ ሁሉ, ባጠፌታ እንዲከፍሉዋቸዉ, አሸከራቸዉ በሞተበት, ያድማዉ መሪዎች, ሁሉ እንዲታሰሩላቸዉ, ይፈልጉ ነበር, የሰራሽ ልብ, ሲያቅዱ ጥርሳቸዉን, ነክሰዉ የጠላቶቻቸዉን, ጥፋት በአይነ, ህሊናቸዉ ሲያዩ, ልባቸዉ ጠጥሮ, እንባ አላፈልቅ, ስለዚህ የሚያደርጉት, ጠፍቶባቸዉ ይሸበሩ, ይሸበሩ ጀመር, መላሸዋ እኔ, መሆኔን እወቁ, የመጣበትን ለቅሶ, ለቅሶ ለመዋጥ, እየታገለ ተነስቶ, ተነስቶ ሄደ, መሆኑን እወድለታለሁ, ወዶ አይደለም, አበበ ከወንድሙ, ባለዉ ግንኙነት, ለማብረድ ቻሉ, አለቃችን ሆደ, ሰፊ በመሆኑ, አስፈላጊ ነዉ, የሆነ ሰዉ, የት ይገኛል, ሳይጫነኝ ልጄ, የአማረኛ ፈሊጦች, ፍቅር መቃብር, ማንኛዉም ቋንቋ, የሰመረ ሀሳብ, ፈሊጣዊ ንግግሮች, በብዙ ቋንቋዎች, እየተሰበሰበ ተመዝግቦ, አማረኛ ቋንቋ, በአጠቃላይ ይዘቱ, የሰዋሰዉ ህግጋት, መልካም ምኞት, ፈተና መቀመጥ, በመፅሀፉ ዉስጥ, በአሁኑ ጊዜ, አዳዲስ ፈሊጦች, ደንታ የሌለዉ, ሀሳብ ያዘዉ, ሀላፊ ደስታ, ወንጀል ሰራ, አደብ ገዛ, ጥሩ አይደለም, ክፉ ሰዉ, በነገር ተነካ, ዉስጣዊ ስሜቴን, ፈፅሞ አለመግባባት, መናገር አቃተዉ, መጠን የሌለዉ, በፍጥነት አነበበዉ, ቀማኛ ወንበዴ, ትንሽ ሰከረ, አመሉ ይለዋወጣል, መጠኑ ጨመረ, የሆነዉ ይሆናል, የሰዉን ችሎታ, አልታዘዝ አለ, በጥፊ ተመታ, ትርፍ የለሽ, የመጀመሪያው ወሰነ, ትምህርት መምህር, እንድትመዲቡልን መጠየቅ, አባሪ አድርገን, የላክን በመሆኑ, በትህትና እንጠይቃለን, እንዳጋመስነው እንጨርሰዋለን, ቴክሎጅ ፋካሊቲ, ኢንፎርሜሽን ቴክኖሎጅ, ኮምፒዉተር ሳይንስ, የትምህርት ዘመን, በማታው መርሃ, ትምህርት ክፍሎች, ታሳውቁን ዘንድ, አልሙናይ ዳይሬክቶሬት, ጣእም ለወጠ, ፍፁም ድሃ, የእሳት ነበልባል, ሰነፍ ደካማ, ሚስጢር አጋለጠ, ዝምተኛ ሰዉ, በዝርዝር ይዘቱ, ለሚያገለግል ህዋስ, ሰዉየዉ ቻይ, ታጋሽ መሆኑ, ቁልፍ ቦታ, ሳይሆን የምንገነዘበዉ, የዋለዉ ለማያ, ቃላት ለዋሉበት, ስራዉ ወረታ, ተመሳሳይ ካልተገኘለት, የሚሰሩበት ቃላት, በማይሰራበት ትርጉም, ሊኖረዉ አይችልም, የሚነገረዉ ሀሳብ, በቀጥታ መንገድ, መገለፅ ካለበት, የሚተላለፈዉን መልእክት, አጠናክሮ በማቅረብ, የሚጥስ ይሆናል, ትኩረት ማድረግ, በፈሊጥ መልኩ, መዛግብተ ቃላት, አፍ መፍቻ, ወንድም እህት, ሀገር ኢትዮጵያ, ጣና ሀይቅ, ሰብለ ወንጌል, ማህደረ ማርያም, ጥናት ማድረግ, አምስት አመት, ድንግል ማርያም, በሶ በላ, መራር እዉነት, እንቅልፍ እድሜ, የአበሻሌ ሊቶች, ብቻዉን መቀመጥ, እንቅልፍ ነቃ, መዉረድ መዉጣት, መቃብር ወረደች, አንድ ኢትዮጵያ, ጉረኛ አይደለም, ብይ መጫወቻ, ግጥም ፃፈ, ትንሽ ስህተት, ልብስ የምታጥብልኝ, ምግብ የምትሰራልኝ, ስማቸዉን ሲረሱባቸዉ, የሚቀጥሯት ሰዎች, አንድሮሜዳ ሞተች, መንግስት የሚያደርገዉ, ታሪክ እንስራ, ረጅም ተረከዝ, ዘፈን አልሰራችም, ቀዝቃዛ ድምፅ, ዘበኛ ነበር, ጓደኛ አለችህ, ገንዘብ ሰጠ, ሸሚዜን ለብሸ, አኩርፈህ ነዉ, የመኪናዉን ሞተር, አሜን አሜን, ማየት መጋረድ, ምርጫ ምርጫ, ከእለታት መሀል, በፊታቸዉ በኋላቸዉ, የሶቅራጥስ ሞት, መፅሀፍ መደርደሪያ, ገናዥ ነበሩ, የመንደሩ ሰዎች, አባቶቻችን ሰልጥነዋል, ዜና መጣ, የሀገራችን ችግር, እንቅልፍ ወሰደዉ, መተኛት አለበት, መልአከ ሞት, እንዴት እረስሃለሁ, አንተ አትስቅበትም, ለተማሪዉ ደወለ, የሰማይ የምድር, ለማንሳት ፈቃድ, አስከሬናቸዉ ተማርኮ, አይኑን ጨፈነ, ኀረምሳ ማዉጋት, ልብስህን አዉልቅ, መቅረብ መራቅ, ለስራዉ ማነቃቂያ, ተመሳሳይ ሀሳብ, መርጨዉ አይደለም, ጥቂት ዝምታ, ባደሩ ማግስት, መታወቂያዉ ዉስጥ, ቆርፋዳ ሻንጣ, በመፃፉ ተፀፀተ, ጠዋት መነሳት, ብርቱካን ልቁረጥልህ, ክሊኒክ ስትመጣ, እድሜሽን ልትነግሪኝ, ሳታደርገዉ ቀረች, በለበሰ አይናቸዉ, መልክ መጣልሽ, በማሳየት ፈንታ, ምግብ በልተሸል, ለማወቅ ጣረች, ጓደኝነት ጀመሩ, የመታደስ ሀሳብ, የሚከተለዉን ነገራት, አለቻት ነርሲቱ, ስንት ወር, በወንድ እቅፍ, ማስወረድ መግደል, ስልክ ደወለች, አምላኬ ማረኝ, የኢትዮጵያ ታሪክ, ቅድመ ታሪክ, ድህረ ታሪክ, የታሪካችን ሰበዞች, የኢትዮጵያ ምንጮች, ክብረ ነገስት, የቃል ትዉፊቶች, የኢትዮጵያ ቋንቋዎች, አፍሮ እስያ, የገድል ፀሀፊዎች, የክርስትና ምንጭ, ፀሀፍተ ትእዛዛት, ዳግማዊ ቴዎድሮስ, ምኒልክ ኢትዮጵያ, ስርአተ መንግስት, ዘመናዊ አፃፃፍ, ደማቅ ታሪካችን, ፊደላት አህዞች, ግብረ ገብነት, እስልምና እምነት, ታላቁ ጥቁር, እምየ ምኒልክ, ዘመናዊ ስልጣኔዎች, ፍትሀዊነት የጎደለዉ, ዘመነ ፅልመት, ምእራፍ ሶስት, ማደጎ ይመስል, ጅምላ ጭፍጨፋ, መንግስታዊ ሽብርተኝነት, ማህበራዊ ዲሞክራሲ, አሻንጉሊት ተቋማት, የትምህርት ስርአት, እዉነትን ፍለጋ, የአማራ መነሻ, የአሰብ ወደብ, ሰንደቅ አላማ, ዋቢ መፅሀፍት, ግእዝ አማረኛ, መጥምቀ መለኮት, መንፈስ ቅዱስ, አረጋዊ ዮሴፍ, ቅዱስ ገብርኤል, ያልተወኝ እግዚአብሄር, ኢየሱስ ክርስቶስ, ዮሀንስን በመምህርነቱ, አምላክን ወለደች, ቅድስት ኤልሳቤጥ, ታቦቱን ሰረቁ, መንፈሳዊ ዝግጅት, ቅዳሴ ማህሌት, ተወላጆች ማህበር, ከመፅሀፍት

አለም, ማህበረ ቅዱሳን, ህብር ሬስቶራንት,ማንበብ ለህይወት, ሮኬት ሳይንስ, መካነ አእምሮ, የኢትዮጵያ ስፔስ, ስልጠና ፍላጎት, መሰረታዊ ግእዝ, እዝል ዜማ, ኮምፒዉተር ጥገና, ኢንተርኔት አገልግሎት, ስልጠና መስጠት, ክህሎት ስልጠና, የኢትዮጵያ ትንሳኤ, ግቢ ጉባኤያት, ሰማይ ስሚ, ምድር አድምጪ, ወንዝ ማዶ, ቡና ኢትዮጵያ, ፋሲል ከተማ, ስራ ፈጠራ, ሀገሬን ላሳያችሁ, ማጠቃለያ ፈተና, ትምህርት ሄደ, ስነ ዉበት, ተፈጥሮ መዉደድ, ቤተ ክርስቲያን, ወደ ኢትዮጵያ, ሀገሬ ኢትዮጵያ, ዳታቤዝ ትምህርት, ትምህርት ቤት, እዉነት ሀሰት, ምርጫ ብቻ, እናት ፓርቲ, ሀገራዊ እቅድ, ስነ መለኮት, ቅዱስ ሚካኤል, ገብረ መንፈስ, አቡነ ሄኖክ, ሀዊረ ህይወት, ግቢ ጉባኤ, ደብረ ታቦር, መዘጋጃ ቤት, መልካም እድል, ትንሳኤ በአል, አቢይ ጾም, ዘወረደ ቅድስት, ምኩራብ መጻጉ, ሆሳእና ትንሳኤ, መመረቂያ ጽሁፍ, ኦርቶዶክስ ተዋህዶ, ቅድስት ጉባኤ, ፓለቲካዊ ተሳትፎ, አሁን ላይ, ገበሬዉ አረሰ, መምህሩ አስተማረ, ቤት ሰራ, በስራ መበልጸግ, ቁም እስር, ፎቶ ኮፒ, አማካሪ መምህር, ደሞዝ ደረሰ, ሙብራት ጠፋ, ዉሃ አለች, ሰራተኛ ሆነ, ወረዳ አዘዘ, ኮምፒዩተር ሰለጠነ, መጽሀፍ አዘጋጀ, ትምህርት ተማረ, ሽማግሌ ላከ, ሰዉ ወደደ, ሰርግ ደገሰ, ዝክር አደረገ, ጉዞ ሄደ, ብቅል አዉራጅ, ብድር መለሰ, ብድር ከፋይ, ብድር መላሽ, ምላሱ ተባ, ታቦት ተከለ, አይኑን ተከለ, ልብ አትክን, ኮሶ ተጣባ, ክረምቱ ተጫነ, ዳኝነቱን አነሳ, እንባ አድርቅ, ሆዱ ጠቡቷል, ሆዱ አይበልጠዉም, ሆዱ ጎሽ, ሆድ ሆዴን, ሆድ ለሆድ, ሆድ ሰጠ, ሆድና ጀርባ, ሆድ እቃ, መለመኛ አጣ, የልመና እህል, ልሳኑ ተዘጋ, ዋንጫ ልቅለቃ, ልቅም ያለች, አንደበቱ የተለቀመ, ልቃቂት ለቀቀ, ልቅ ወጣች, ሚስቱን ለቀቀ, አፉን ለቀቀ, ዉሸቱን ለቀቀዉ, ከብቱን ለቀቀ, በነገር ለበለበዉ, ንባቡን ለበለበዉ, ጠጁን ለበለበዉ, ሀፍረት ለበሰ, ልብሰ ተማሪ, ልብስ ገፋፊ, ፀጋ ልበስ, የለበጣ አነጋገር, ለብ አለ, ለከት የሌለዉ, በሽታ ለከፈዉ, ዉሻ ለከፈዉ, ጋኔን ለከፈዉ, የሰዉን ልክ, ከልኩ አያልፍም, ቁልፍ ሰዉ, ነገር ለኮሰ, በዱላ ለወሰዉ, ሀይማኖቱን ለወጠ, ልብሱን ለወጠ, ሰዉነቱ ተለወጠ, ሽታዉ ለወጠ, ለዛ ሙጥጤ, ለዛ ቢስ, ሰዉ ለየ, ፊደል ለየ, ለይቶ አየ, ነገር ለደፈብኝ, ለጋ ጎበዝ, ለጋ ቅቤ, ለጋ ደመና, ለጋ ጨረቃ, ልጓም አጥባቂ, ልጓም ጣለ, በሀሰት ለጠፈበት

Appendix C: Vector Representation

```
In [15]: #pd.set_option('display.max_rows', None) #used to display all rows
         w2v_df = pd.DataFrame(vectors, columns=['x','y'])
         w2v_df['Expressions'] = words
         w2v_df = w2v_df[['Expressions', 'x','y']]
         print(w2v_df)
```

```
     Expressions         x         y
0              -1.674955 -1.971173
1            በላው -1.942772 -0.450577
2           ዝግጅት -2.503376 -1.528765
3           አብርድ -1.981622 -3.655081
4             ከፉ -1.288707 -3.114516
5           ምንዝር -2.266248  0.092966
6           ህይወት -1.640370 -0.614642
7           አእምሮ -0.918380 -1.199284
8           አሟልቶ -0.524949 -0.977453
9            ክፍት -0.790613 -2.087327
10            ሙጢ -1.027744 -2.329200
11            ስማ -0.972780 -1.475834
12           የቃል -0.676758 -1.185385

13          መማለጃ  2.756359 -0.103811
14            ቆጡ  0.787338  0.196069
15        እቅዳቸውን -0.403257 -0.711361
16          በጎደለ  0.730740  0.786475
17            ሱሪ  0.949949  0.056913
18            ላላ -1.435253  0.790522
19            ቦታ  0.440194  0.806477
20         አለቃውን  1.626016  1.709901
21           እቅድ -0.403443  1.788621
22          ብቻውን  0.955321  0.652900
23         አለቃችን  1.394935 -0.804733
24         ተመለከቱ  0.241569  1.601391
25          ስህተት  0.066470  1.781154
26          ገበሬው  1.017706  1.547944
27           ቀረሽ  0.353624  0.637156
28          ሞረደው  0.642169  1.252949
29          ከረምቱ -0.332393  0.454080
30           ከነዳ -0.494714  1.390756
```

Appendix D: Represented Dataset Sample

Expresions	Vectors	Class
ሀረግ መዘዘ	[-2.254647 0.544843 -1.825512 -0.822505]	Idiom
ሀረግ መዘዘ	[-2.254647 0.544843 -1.825512 -0.822505]	Idiom
ሀረግ ጣለ	[-2.254647 0.544843 -0.918860 -1.925873]	Idiom
ሀሳበ ቢስ	[-1.965279 -2.863393 -2.007703 -0.747365]	Idiom
ሀሳብ ገባዉ	[-0.686118 -0.370945 0.856144 -1.926958]	Idiom
ሀብተ ስጋ	[-1.314511 0.694000 -0.817603 -1.098272]	Idiom
ሀብተ ሰባራ	[-1.314511 0.694000 -2.051465 -0.532003]	Idiom
ሀብተ ስንኩል	[-1.314511 0.694000 -1.388195 -3.255751]	Idiom
ልብሰ ተማሪ	[0.168428 0.352119 -0.340268 2.140347]	Lieral
ልብስ ገፋፊ	[1.248360 0.780462 0.036338 -1.027669]	Idiom
ግቢ ጉባኤ	[1.132941 0.770536 1.085537 0.353204]	Lieral
ግቢ ጉባኤ	[1.132941 0.770536 1.085537 0.353204]	Lieral
ደብረ ታቦር	[0.235832 0.466849 0.939056 0.552514]	Lieral
መዘጋጃ ቤት	[0.836765 0.999328 -0.584712 0.881482]	Lieral
መልካም እድል	[0.486902 -1.600112 0.839818 0.536086]	Lieral
ትንሳኤ በአል	[1.324415 1.485414 -0.221301 0.981636]	Lieral

Appendix E: Sample Simulation Code