2021-09

INFORMATION EXTRACTION MODEL þÿFROM GE EZ TEXTS

Worke, Wolde Asfaw

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

FACULTY OF COMPUTING

INFORMATION EXTRACTION MODEL FROM GE’EZ TEXTS

Worke Wolde Asfaw

Bahir Dar, Ethiopia

September 2021

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

INFORMATION EXTRACTION MODEL FROM GE’EZ TEXTS

Thesis Submitted to School of research and Graduate Studies of Bahir Dar Institute of Technology, Bahir Dar University in partial fulfillment of the requirement for the degree of Master in Information Technology in the Faculty of Computing

Advisor: Seffi Gebeyehu (Asst. professor)

Bahir Dar, Ethiopia

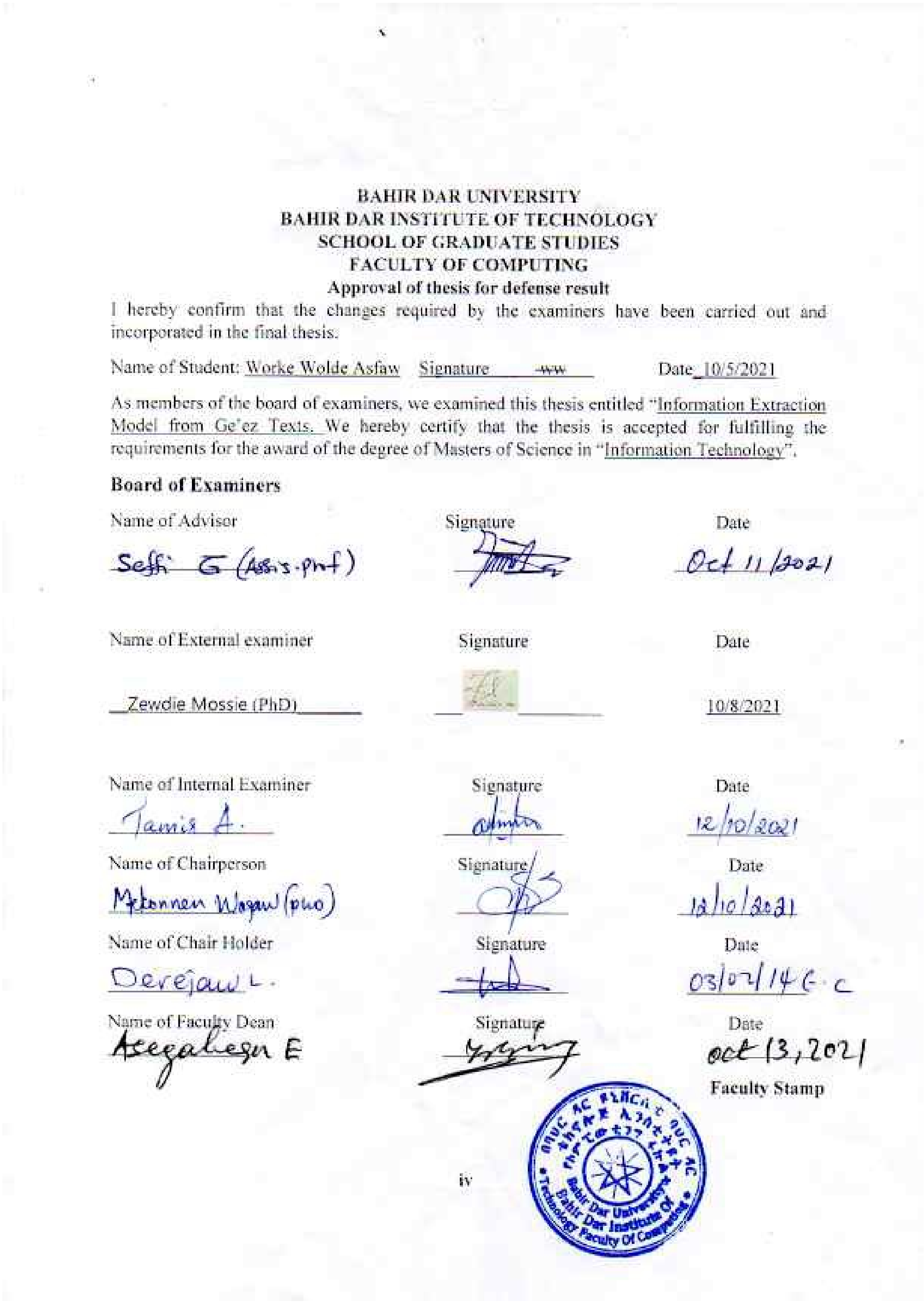
September 2021

DECLARATION

This is to certify that the thesis entitled " **Information Extraction from Ge'ez Text** ", Submitted in partial fulfillment of the requirements for the degree of Master of Science in **Information Technology** under **Faculty of Computing**, Bahir Dar Institute of Technology, is a record of original work carried out by me and has never been submitted to this or any other institution or university to get any other degree or certificates. The assistance and help I received during this investigation have been duly acknowledged.

Worke Wolde ________ 10/5/2021

Name of the Candidate Signature Date



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: Worke Wolde Asfaw Signature ww Date 10/5/2021

As members of the board of examiners, we examined this thesis entitled "Information Extraction Model from Ge'ez Texts. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in "Information Technology".

Board of Examiners

Name of Advisor Signature Date
Seffi G (Assis.prof) Oct 11/2021

Name of External examiner Signature Date
Zewdie Mossie (PhD) 10/8/2021

Name of Internal Examiner Signature Date
Tamir A. 12/10/2021

Name of Chairperson Signature Date
Mekonnen Wagaw (phD) 12/10/2021

Name of Chair Holder Signature Date
Derejaw L. 03/02/14 G.C

Name of Faculty Dean Signature Date
Asegahegn E oct 13,2021

Faculty Stamp

ባህር ዳር ዩኒቨርሲቲ ባህር ዳር ቴክኖሎጂ ኢንስቲትዩት ኮምፒውቲንግ ፋኩልቲ
Bahir Dar University Bahir Dar Institute Of Technology Faculty Of Computing

ACKNOWLEDGMENTS

First of all, I would like to thank God the Almighty and St. Mary for giving me the wisdom, strength, support, and knowledge in exploring things.

Next, I would like to express my gratitude to my advisor Mr. Seffi Gebeyehu (Ass. professor) who has always advised the process of this final thesis work for giving me support, encouragement, and continuous advice, without his support, this final thesis wouldn’t have been completed.

Finally, there is no word to express my family, however, I would express my special thanks to my mother (Aynalem) and father (Woldemikael), my lovely sister (Haderemariyam, Tsgemariyam, Tselotehaymanot & Nigstehlina), and brother (Fsihatsion and Mengistu) for giving me support and encouragement

ABSTRACT

Currently, there is a vast volume of unstructured textual data on the internet that offers diverse useful information for health care, commerce, education, religious, cultural, historical, and other domains. The problem is that, the amount of unstructured data grows, extracting valuable information from unstructured data and the need for tools and techniques for extract (automatic text extraction of text has become a critical task)and explore useful information to address and satisfy the user's needs is a challenging task due to the overloading of information on the internet. Information extraction extracting structured text from unstructured data using Natural Language Processing statistical techniques. In this study, we developed an information extraction model from Ge'ez text for extracting named entity text information. However, the limitation of the study the relation between entity attribution and scanned, image voice, and video text did not considerable. The proposed model has its main component dataset preprocessing, traning or learning and testing phase, and predciton phase. The preprocessing phase performs tokenization of sentence, stop word removal, affix removal or stemming, and paddind sequence, the , traning or learning and testing phase used to train or learn the model, and test the learned model and the prediction phase predict or extract the catagories of texts. In this work, the accuracy of traning, validation, and testing is employed as an evaluation metric for the information extraction model from Ge'ez text. We used a 5270 sentence dataset (63262 tokens) from the Addis Ababa Ethiopian Orthodox Tewahdo Church that was being trained and tested for our research.We used two experimental setup i.e Long short-term memory and bidirectional long short-term memory to demonstrate the experimental evaluation with 80% training and, 20% testing size of dataset split ratio. Finally, the results of an experimental evaluation, evaluates using a long short-term memory accuracy is 98.89% traning, 98.89% validation, and 95.78%, testing and bidirectional long short-term memory 98.59% training, 97.96% validation and 96.21% testing accuracy the proposed model performed. For the design of a full-fledged information extraction system, further researchers incorporating the post of speech tagging for extracting relationships between things, or relation extraction.

Keywords: EOTC, NLP, IE, LSTM, BILSTM.

Table of Contents

LIST OF TABLE

LIST OF FIGURE

LIST OF ALGORITHM

ABBREVIATIONS

AOTIE	Afaan-Oromo Text Information Extraction
BiLSTM	Bidirectional Long Short-Term Memory
EOTC	Ethiopia Orthodox Tewahido Church
IE	Information Extraction
IEGT	Information Extraction for Ge’ez Text
LSTM	Long Short-Term Memory
ML	Machine Learning
MCM	Multi classification model
NLP	Natural Language Processing
SOV	Subject Object Verb
SVO	Subject Verb Object
VSO	Verb Subject Object

CHAPTER ONE: INTRODUCTION

1.1 Background

The amount of textual content available on the internet is rapidly increasing these days (Chiu & Nichols, 2016; Gao et al., 2020; Jang et al., 2020; Yang et al., 2019). This unstructured data found in the form of text, images, video, and audio(Moens, 2006). A significant part of such information like government documents, legal acts, online news, and social media communication is transmitted in unstructured form(Milosevic et al., 2019). To address those problems, the researchers used natural language processing to integrate human language and computer language. Natural language text is now used to store a huge amount of the natural language text or world's knowledge, which may be found on Web pages, news stories, research papers, e-mails, and blogs. The total amount of data on the planet is estimated to be in the thousands of gigabytes. A vast volume of text data on the internet contains diverse useful information for health care, markets, education, and other fields that is not visible in unstructured textual data. (Jayaram, 2017; Panda et al., 2019) to get the accurate and right information from current abundant unstructured text is still now a complex task. NLP (natural language processing), which is one of the artificial intelligence domains that analyzes communication mechanisms through natural language creation and understanding(Limsopatham & Collier, 2016). The work in NLP can be grouped as rationalist (rule-based) and empirical (corpus-based) Numerous general-purpose linguistic capabilities are characteristic of this type of system are POS, parsing, word-sense disambiguation, higher-level (semantic) understanding, dialog systems, natural language interfaces, and queries, (Janevski, 2000)

A major difficulty for early years is the lack of tools to extract and search usable information to address and satisfy the user's demands. To solve the current challenge, A lot of studies have been done in the fields of Information Extraction (IE), Information Retrieval (IR), Question Answering, Text Summarization, and Text Categorization(Valero et al. 2009 The Advanced Research Projects Agency (ARPA) has responded by financing research into the development of information extraction, a new technology (Grishman, 2019; Okurowski, 1993). Information extraction is a type of document processing that extracts and outputs factual information contained inside a document.

Information extraction is the process of extracting structured data from unstructured sources, such as entities, relationships between things, and attributes defining entities(Sarawagi, 2008). This enables significantly more complicated queries on the various unstructured sources than simple keyword searches. When structured and unstructured data coexist, information extraction allows the two forms of data to be combined and queries to be created that cover both. Information Extraction (IE) is a contemporary study topic in the field of natural language processing, and it is the process of extracting useful data from an unstructured document using statistical techniques(Jayaram, 2017). The most important text from a large number of papers is displayed. It is the process of extracting structured data from unstructured sources, such as entities, entity relationships, and entity properties. The amount of information available is growing all the time, resulting in information overload. Due to the widespread availability of electronic documents on the internet, there is currently an information overload. It's difficult to manually search, filter, extract and pick which data should be used for different purposes(Abera, 2018; Hirpassa, 2017; Worku, 2015).

To complete a full-fledged information extraction, several tasks are required, including Named entity recognition (named entity), Co-Reference Resolution, Template Element Construction (TE), Template Relation Construction (TR), Relation extraction, Scenario Template Production (ST), and other sub-tasks such as parsing and tagging. The named entity recognition task was employed in our model to determine the fundamental entity tags, which include persons, places, times, and events. We used the Ethiopian Orthodox Tewahedo Church (EOTC) in Addis Ababa to acquire Drsan, Yehawariyat Sra, and Gedile Aba NOB databases for our research. The entity tags are identified using the same futures.

Ethiopia is one of the East African countries that has its own Fidel (Letter and Numbers) and writing system for identifying the country internationally (Kassa, 2018). The word Ge'ez signifies first in the Ethiopian Orthodox Tewahedo Church's Zema (Gloss)teaching, first in the Alphabet, first in reading style, and first in the Ethiopian Orthodox Tewahedo Church's Zema (Gloss)teaching. It is an ancient South Semitic language that developed in the southern areas of Eritrea and the northern sections of Ethiopia in the Horn of Africa and later became the official language of the Kingdom of Aksum and the Ethiopian imperial court(Kassa, 2018). The Ethiopian Orthodox Tewahedo Church, the Eritrean Orthodox Tewahedo Church, the Ethiopian Catholic Church, the Eritrean Catholic Church, and the Beta Israel Jewish community all use the Ge'ez language as their primary language(Abera, 2018).

Ge’ez bible, Hamer, and religious books are currently available online in EOTC religions(Abebe, 2010; Alemu et al., 2021), because those resources are given in unstructured and semi-structured text formats, users must manually extract or read relevant information from the web, which takes more time and results in lingering activities. The Geez scripting writing system comprises 26 fundamental alphabets (named "Fidel"), each of which has seven forms generated by fusing a consonant for an alphabet, generating 182, and various additional forms are derived from the basic alphabets, such as ቈ ቊ ቍ ቋ ቌ from ቀ, ኰ ኲ ኵ ኳ ኴ from ከ, ዀ ዂ ዅ ዃ ዄ from ኀ and ጐ ጒ ጕ ጓ ጔ from ገ(Kassa, 2018)(Alemu et al., 2021).

The Modern Ge'ez writing system is based on a left-to-right writing system; nevertheless, it was written from right to left before the fourth century(Abebe, 2010). The term Ge'ez refers to an earlier and more ancient language used by the Ethiopian Orthodox Church. It has its qualities and features or is defined by those characteristics. Geez is distinct from other languages. Its letter has an alphabetical order(ሀ-ፐ), structure(ሀ፤ሐ፤ኀ፡ሰ፤ሠ፡አ፤ዐ፡ጸ፤ፀ) (Kassa, 2018). As a result, if the user is extracting Geez text, he or she should be familiar with the linguistic character (Geez writing system). However, some scholars are working on the Geez language before the year is out(Sisay 2016; Tesfasilassie et al., 2017) and Designing A Stemmer for Ge’ez Text using Rule-Based Approach(Abebe, 2010) also some Information Extraction researchers work on different languages by using different approaches and metrics to measure the performance of the system. Although all of the recent approaches and datasets aren’t directly applying to the Ge’ez language(Worku, 2015).

For this reason, we develop an information extraction model from Ge’ez text using a deep learning approach to solve the existing problems. Deep learning is a set of learning methods attempting to model data with complex architectures combining different non-linear transformations. The elementary bricks of deep learning are the neural networks, that are combined to form the deep neural networks.

1.2. Statement of the problem

There is a need for accessing significant sources, extracting, evaluating, and putting them into the proper form, due to the quick increase of utile information such as newspapers, articles, and emotive comments on the Internet. Religious church leaders or authors can submit various texts, videos, audios, and visual objects to people who speak Ge'ez, and readers can extract the data. While the writers categorized the information from the users' text, they manually represented each information category since it is vital to recognize and classify the types of entities for future work. The data that was sent could not be structured. For the past few years, information extraction has been a major focus, and it has been included in several products(Chang et al., 2006).

In Arabic, English, Latin, and other languages, a lot of work has been done. Nonetheless, there is no attempt to extract the categories of information in the Ge'ez writings. As a result, scholars, religious communities, and language communities are unable to easily extract and classify written data. This takes time and requires psychological knowledge. This is owing to a lack of resources for Ge'ez IE and the slow pace at which Ge'ez is progressing. This language has its writing style and grammatically structured language rather than another language like the Amharic language because the Amharic language copies 7 letters from the Ge’ez language. However, to the researcher's knowledge, no research has been conducted or may not be published as Ge'ez IE in general.

As a result, the lack of information extraction motivates us to undertake research on IE for Ge'ez, as well as build and create the Ge'ez (Ge'ez IE) model utilizing a Deep learning technique. In addition to this, starts the process of Ge'ez Information Extraction (Ge'ez IE) will be the focus of researchers, who will also innovate new concepts in this field.

In the Ge’ez language, 9 letters have similar sounds but are not the same and have their mining rather than replacing’” each other. For example, ኃለየ(ዘፈነ/Sings) and ሐለየ(አሰበ/He thought), አመት(ገረድ/The Maid) and ዓመት(ጊዜ/ዘመን/Time). For this reason, we perform the task on NLP the normalizations of this Ge’ez text must be as it was replaceable for the text pre-processing stage. Because the replacement of one character extracts other ideas, rather than users predefined text. As an example, when the user needs to extract a text year(ዓመት/ጊዜ/ዘመን/Time), the replacement (normalization) of a character “ዓ” by “አ” the extracting data is ገረድ(The Maid).

There are a lot of cultural, religious, and historical documents are available on the web page written by Geez language, so the user extracts the required data must knowing the language grammatical (morphological structure) and writing system of Ge’ez language (E.g. the input of the data for searching or extracting the required data like the user extract the date/time must insert the text or string values ፲ወ፪(ዐሥሩ ወክልኤቱ) not write 12) rather than writing some word or data’s (Alemu et al., 2021; Tadesse et al., 2020) This written document or file is unstructured/misstructured the user's get accurate and relevant information from the current vast unstructured text is a challenging task so, this problem shows more time taking, resource expensive, and linguistic knowledge of the language is low(Kassa, 2018). As a result, if the user is extracting Geez text, he or she should be familiar with the linguistic character (Geez writing system)(Hirpassa, 2017).

Language and domain depend on the information extraction system (Singh, 2018). Even if the domain is identical, the IE method established for English/Amharic and in the specific domain may not operate for the Geez language. Each language has its grammar specifications (Subject-Object-Verb agreements) E.g. The home entertainment extraction system is incompatible with traffic accident extraction devices. Knowing the value of data after text extraction, and which set of text is more important for a user, is the most difficult process.

Inappropriate use of the method for extracting information from a measuring system to evaluate the system's performance. As a result of this erroneous use of methods/approaches, the system's performance measures are inadequate. Before the year, there were only a few types of research on information extraction systems in Ethiopian and foreign languages(Abera, 2018; Panda et al., 2019; Ronran et al., 2020; Worku, 2015), because of its grammatical nuances, syntactic structure, lexical structure, writing system, morphological structure, and encoding style, existing approaches created for other languages are difficult to apply directly to the Ge'ez language. As a result, we developed a model for extracting information from Ge'ez text.

To address this problem, the following reasonable research question is drafted.

1. How to develop a model for information extraction model from Geez text?
2. Which deep learning approach(LSTM & BiLSTM) improve the performance of the model?
3. How to test and evaluate the performance of the proposed model?

1.3 Objectives

1.3.1 General Objective

The general objective of this study is to develop an IE model from Ge’ez text.

1.3.2 Specific Objectives

To achieve the general objective of the research, the following specific objectives are set.

- To conduct different literature reviews related to information extraction.
- Collect and identify Ge’ez corpus.
- To prepare a dataset to train and test our model.
- To design an optimal model for Ge’ez text information extraction.
- To compare the performance of two deep learning approaches (LSTM and BiLSTM) in order to improve the model's performance.
- To test and evaluate the performance of the proposed model.

1.4 Research Methodology

A methodology is a set of principles, practices, processes, methods, and techniques that are applied to a body of knowledge to assist the researcher in planning and carrying out research (Ahmadi et al., 2020). Research methodology is a method for using research to solve some problems. It allows researchers to learn which procedures or strategies are appropriate, what they mean, what they indicate, and why they are used. It also gives them the capacity to comprehend the assumptions underlying various techniques and the rules by which they can determine whether or not specific techniques and procedures are appropriate for a given situation. This is an experimental study that aims to construct a model for extracting information from Ge'ez text. The following approaches were examined for the successful completion of this study.

1.4.1 Literature Review

We reviewed different literature, conference paper, and workshops related to information extraction models, information extraction methodology, and performance measurement. (Vacancy, news, big data, and other areas of extraction models) and choose the appropriate approaches that are used for our research work and get the research gap or limitation.

1.4.2 Data collection

To conduct this research work on the proposed model, we have collected all the necessary Ge’ez text not scanned text dataset from Addis Ababa Ethiopian Orthodox Tewahedo Church (EOTC) religious church containing Drsan, YehawariyatSra, and Gedile Aba Nob, Lisane Ge’ez, Likuas Art, and Betremariyam Abebaw Metshaf hads triguame telegragm channel because there is no well-defined Ge’ez text corpus rather than scanned corpus available to extract text. Those datasets were collected manually from that source.

1.4.3 Preprocessing

The preprocessing activities have been done on the collected dataset. The data cleaning is used to remove unwanted handwriting characters, numbers, or punctuation marks from the statement and annotated the dataset set as BIO format in recent sequence to sequence labeling format. All of the text sources are converted its character sets, the file form, and the text structure and also the language-specific problem word, punctuation mark and separate perform on tokenization, character, and number normalization perform on normalization, and the stop word in a text like (articles, prepositions, and conjunctions) removal, stemming, and padding or embedding. After this, split the dataset into train and test data to train the BiLSTM model and test the trained model to predict the targeted class or extract the entity were included in our research.

1.4.4 Tools and Techniques

We used different tools to complete the research work. For development, the Python programming language is utilized to access crucial libraries and modules. To develop the information extraction model, we used Jupyter notebook as a code editor; a powerful scientific environment for Python, Keras for preprocessing the word embedding data extraction and training the model on top of TensorFlow as a backend (a library that includes Keras as a submodule), Pandas is an open-source toolkit for the Python programming language that provides high-performance, easy-to-use data structures, and data analysis tools(McKinney & Team, 2015). It's used to read our dataset files and run various operations on them. We applied long short-term memory(LSTM) and bidirectional long short-term memory(BiLSTM) deep learning techniques in this research work. Deep learning is a set of strategies for learning in deep neural networks that is extremely strong. It's a type of machine learning that employs a series of algorithms to model high-level abstractions in data by employing many processing layers with complicated structures (Zhang et al., 2020).

1.4.5 Evaluation Technique

The term "evaluation" refers to the process of determining how well the information extraction model from Ge'ez text performs. The functionality of the information extraction model has been tested through experiments. Testing datasets were inserted into the constructed model to analyze the model rationally. To measure the proposed system, annotated Ge'ez text or sentences which contain person, place, date, and event tags were used to test the learned model. The performance of the model has been evaluated in terms of training, validation, and testing accuracy or loss of categorical entropy accuracy.

1.5 Scope and Limitation of the study

The scope of this research is to use a deep learning approach to construct an information extraction model from Ge'ez text. The four main entity classes, such as person name, location name, date or time, and event, have been considered. Different deep learning techniques, such as long short-term memory and bidirectional long short-term memory, were used in this work. It is a small corpus based on data acquired from EOTC (Drsane, Gedle Aba Nob, and ACT or ye Hawariyat Sra) text files rather than scanned text to extract the Name entity extraction.
The study's limitations did not include the voice, image, scanned text, and video text data. In addition, the extraction of relationships between entities was not considered.

1.6 Significance of the study

The proposed model for the Ge’ez text has the following benefit for the Ge’ez language community as well as the research community.

For the Ge'ez language community: It is important to grasp the language's morphological properties to extract them and The religious church community extracts major issues in a short amount of time.

For the research community: Researchers working on higher-level NLP (spelling checker, wordnet, POS, NER) to build and improve the Ge'ez language's information extraction modules. To learn about and develop information extraction strategies for the Ge'ez language. To focus the researcher's attention on learning and increasing knowledge of the ancient Ge'ez language to advance other related IR research areas such as question answering, text summarization, machine translation, and others. To conduct additional research using different deep learning techniques and to establish a system

1.7 Organization of the Thesis

This thesis is organized as follows. Chapter 1 presents an introduction, statement of problems, research question, objective, methodology, scope and limitation, and significance of the research. Chapter 2 presents a literature review and different related works on IE. Chapter 3 presents the methodology and implementation. Chapter 4 presents experimental design and results. Finally, in Chapter 5 the conclusion and future works are presented.

CHAPTER TWO: LITERATURE REVIEW

2.1 Overview of Literature Review

In this chapter, we presented the general principles or theoretical concepts of information extraction and the Ge'ez language. To understand our research works, we conducted some of the literature reviews which included an introduction to Ge’ez language, Geez alphabet, Geez numerals, letters, punctuation marks, phonetics, and grammar structure. Furthermore, it discusses some of the most widely used approaches for IE tasks, natural language processing, information extraction subtasks, model performance evaluation metrics or measurement, and finally, it discusses some related work for information extraction systems. For this reason, we applied it throughout the thesis work.

2.2 Ge’ez language (ልሳነ ግዕዝ)

Affixes (prefixes, infixes, and suffixes) are linked to the roots or stems in Ge'ez, just like in any other language. Grammatical information is conveyed through affixes (prefixes, infixes, and suffixes) linked to the roots or stems. Ge'ez is a traditional Ethiopian language with an original name that belongs to the Semitic language family(ABEBE, 2020). In Western speech, the language is referred to as either "Old Ethiopic/Classical Ethiopic" or just "Ethiopic" the names Ethiopic and Ge'ez are used interchangeably in various publications and papers.

However, instead of using the term Ethiopic, the researchers in this study chose the term Ge'ez. Let us return to the beginning (Ge'ez history), as documented evidence indicates that, before the 5th century B.C. old Ge'ez writing was from right to left, similar to Arabic, Syriac, and Hebrew, and the lettsers were simply the Ge'ez (first-order), with no 2nd, 3rd, 4th, 5th, or 6th. Ge’ez means first in order or first letter when writing a text. For example, with Alfawu (አ(a) ገአዘ or ግእዝ. Modified Ge'ez language letters in the 4th century AD (Yacob, 2000) 6 orders from the second to the seventh ((ካዕብ፣ ሣልስ፣ ራብዕ፣ ኃምስ፣ ሳድስ ፣ ሳብዕ) orders; for example, the letter ‘' (hä) was just the first order before modification, but after modification, it became ሀ፣ ሁ፣ ሂ፣ ሃ፣ ሄ፣ ህ፣ ሆ (hä, hu, hi, ha, he, h, ho) and those modified letters are still used in current Ge'ez.

on the other hand, believe that Ge'ez is a dead language(Weninger, 2010) however, a dead language, as it is still studied and used as a classical language by church scholars in Ethiopia and Eritrea(Weninger, 2010). Because it is currently given in a few primary and secondary schools, universities, and is a topic of research in higher educational institutions beyond traditional schools, the investigators of this study agreed with Weninger's idea. The term Lisane Geez is named as the tongue of the free Ge’ez or Ethiopic Script language the only endangered African Script remained the spoken language until the end of the Aksum Empire in the ninth century(Alemu et al., 2021). At the current time, Ge’ez language is given as one sub-subject matter for higher education level.

2.2.1 Ge’ez Alphabet (የግእዝ ቋንቋ ፊደላት)

The writing system of Ge’ez language uses abugida (አቡጊዳ) called in Ethiopian Semitic languages called ፊደል Fidel ("alphabet", "letter", or "character") refers to ‘ፈደለ’ comes from Ge’ez verb means that Writing (ጽሕፈት), the father of words(የቃል አባት), the grandfather of phrases and sentences (የሐረግና የዓረፍተነገር አያት), the ancestors of literature(የሥነ-ጽሑፍ ቅድመ አያት) and it has been the working language of government, the military, and of the Ethiopian Orthodox Tewahedo Church throughout medieval and modern times(Abebe, 2010; Alemu et al., 2021; Kassa, 2018).

Ge’ez language used a consonant-based written system. At the time of Aba sälama vowels were created for it, for you find both types (written in consonant and vowel form(Abebe, 2010; Alemu et al., 2021). In the Ge’ez language, there are four (4) letters that have similar sounds. Even though they are having similar sounds, the letters are different in shape orthographically. Those letters are ሀ፣ሐ፣ኀ(ha), ሰ፣ ሠ(sa), አ፣ዐ(a), ጸ፣ፀ (tse). It has 26 first-order letters and contains 7 letters gives 182 letters.

In general, the Ge'ez writing system contains 202 symbols (i.e. 182 CV syllables = 26*7 + CVW labiovelars(ቈ ቍ ቍ ቋ ቌ from ቀ, ኰ ኲ ኵ ኳ ኴ from ከ, ኈ ኊ ኍ ኋ ኌ from ኀ and ጐ

ጐ ጐ ጓ ጔ from ገ.)=4*5), 20 numerals, and eight punctuation marks(Yacob, 2000) eliminating mathematical operations and St. Yared's rhyme song. The sequence of Ge'ez letters was ‘abegede' () vertical line or ‘abudida' (አቡጊዳ) horizontal line until the 4th century AD, and all letters were referred to as alphabet/ ‘አሌፋት’ combined. However, the present order is ‘heleheme' () when it was altered, and letters are referred to as hohyat/ ‘ሆህያት' at one time.

Table 2.1: Ge’ez Letter

First-order (early)			**Second-order(modern)**		
አ-ግእዝ	የ-ግእዝ	ቀ-ግእዝ	ሀ-ግእዝ	ነ-ግእዝ	ጠ-ግእዝ
በ=”	ከ	ረ	ለ	ኀ	ጸ
ገ=”	ለ	ሰ	ሐ	አ	ፀ
ደ=”	መ	ተ	መ	ከ	ፈ
ሀ	ነ	ጰ	ሠ	ወ	ፐ
ወ	ሠ	ፐ	ረ	ዐ	
ዘ	ዐ		ሰ	ዘ	
ሐ	ፈ		ቀ	የ	
ኀ	ጸ		በ	ደ	
ጠ	ፀ		ተ	ገ	

2.2.2 Chart of Ge’ez letters

Ge’ez language has its character to write or read some useful information. In this case there are 26 letters(Ge’ez/ግዕዝ) with corresponding 7 order of character.

Table 2.2: char of Ge’ez

ግዕዝ (Ge’ez)	ካዕብ (Kaib)	ሣልስ (Salis)	ራዕብ (Rabi)	ኃምስ (Hamis)	ሳድስ (Sadis)	ሳብዕ (Sabi)
ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ

2.2.3 Ge’ez Numeral

Geez also has its numerals for designating numbers. These numbers are used in the Ethiopian yearly calendar. The following tables describe the numeric of Ge’ez numeral it identifies the numbers. Example**:** ዐሠርቱ ወክልዔቱ ሐዋርያት ሰበኩ ወንጌለ in this sentence ዐሠርቱ ወክልዔቱ (፲ወ፪/ 12) indicates how a money person is a participant.

Table 2.3:Ge’ez Numeral

አኃዝ	ተባዕታይ/ አንስታይ(male&femal)	በአረብ	አማረኛ
፩	አሐዱ(አሐቲ)	1	አንድ
፪	ክልዔቱ(ክልዔቲ)	2	ሁለት
፫	ሠለስቱ(ሠላስ)	3	ሦስት
፬	አርባዕቱ(አርባዕ)	4	አራት
፭	ኀምስቱ(ኀምስ)	5	አምስት
፮	ሰደስቱ(ስሱ)	6	ስድስት
፯	ሰብዓቱ(ሰብዑ)	7	ስባት
፰	ስምንቱ(ሰማኒ/ት)	8	ስምንት
፱	ተስዐቱ(ተስዑ)	9	ዘጠኝ
፲	ዐሠርቱ(ዐሡሩ)	10	ዐሥር
፳	ዕሥራ	20	ሃያ
፴	ሠላሳ	30	ሠላሳ
፵	አርባ	40	አርባ
፶	ሃምሳ	50	አምሳ
፷	ስሳ/ስድሳ	60	ስድሳ
፸	ሰብዓ	70	ሰባ
፹	ሰማኒያ	80	ሰማኒያ
፺	ተስዓ	90	ዘጠና
፻	ምእት	100	መቶ
፼	እልፍ	1000	አንድ ሺህ
፲፻፻	አእላፍ(ዐሠርቱ እልፍ)	10,000	ዐሥር ሺህ
፻፻፻	አእላፋት	100,000	መቶ ሺህ
፲፼፻	ትእልፌት	10,000,000	ዐሥር ሚሊየን
፲፼፻፻	ምእልፈት	1.000.000.000	አንድ ቢሊየን

Example:2.1

5458235=5*1000000+45*10000+82*100+30+5=፭*፻፼+፵፭*፼+፹፪*፻+፴+፭=፭፻፼፵፭፼፹፪፻፴፭።

አማርኛ፡-አምስት ሚልየን አራት መቶ አምሳ ስምንት ሺ ሁለት መቶ ሠላሳ አምስት።

ግእዝ፦ኀምስቱ አእላፋት አርብዓ ወኀምስቱ እልፍ ሰማንያ ወክልኤቱ ምእት ሠላሳ ወኀምስቱ።

English፦Five Million four hundred fifty eight thousand two hundred thirty five.

2.2.4 Ge’ez Punctuation Marks

Ge’ez language has its punctuation mark Punctuation, much of it modern includes ፠ section mark, ፡word_separator, ። full stop (period), ፣ comma, ፥ colon, ፤ semicolon, ፦ preface colon, ፧ question mark, and ፨ paragraph separate.

2.2.5 Phonetics of Ge’ez

(Dobrovolsky & Katamba, 1997)phonetics is the study of speech sounds, including how they are created, perceived, and what physical qualities they have. Articulatory phonetics (deals with how the vocal organs produce speech sound),Acoustic phonetics (studies the physical features of speech, such as frequency, duration, and sound intensity), and Audito ry phonetics (observes how the human auditory system perceives speech) are the three ar eas of phonetics. They're linked because changing the articulatory configuration of the vocal tract causes acoustic changes that affect how a sound is perceived (Jurafsky & Martin, 2007).

In the Ge'ez language, there are 37 phonemes (30 consonantal phonemes + 7 vowel phonemes)(Ado, 2021). There are 30 consonants, which are classified as stops, fricatives, approximants, ejectives, nasals, and trills based on how they are articulated. There are seven vowels in Ge'ez, which are divided into long and short vowels. In Ge'ez, the fundamental vowel ä has the most predominance and plays a significant role in the formation of words as short and long vowels (KEBEDE, 2017). when verbs try to show the subject marker and object marker using the subject marker phones (ከ፣ ኩ፣ ኪ፣ ክሙ፣ ክን፣ ነ፣ ት).

2.2.6 Ge’ez grammar structure

The term refers to the set of structural rules governing the composition of clauses, phrases, and words in any given natural language and phonology, morphology, and syntax, often complemented by phonetics, semantics, and pragmatic. This grammar structure is classified into eight major parts in Ge’ez. (Alemu et al., 2021). These are Nouns, Verbs, Adjectives, Adverbs, Pronouns, Prepositions, Conjunction, and Interjection. The content words (open word classes), such as nouns, verbs, adjectives, and adverbs, are words that carry lexical or content meaning. Structure words (closed word classes), such as prepositions, pronouns, conjunctions, and determiners, are words that show grammatical relationships within sentences.

Speakers are endlessly creating new Ge’ez and Amharic open words, especially nouns and verbs. Therefore, the major word or form classes are called open word classes because new words enter the language constantly(Abebe, 2010; Kassa, 2018).

ስም (Nouns): A noun is a name used to identify the name of people, places, things, or others. ስም ብሂል ጽዋኤ ነገር ብሂል ውእቱ። ስመ ሰብእ(person),ሰመ እንስሳ(name of animal)፣ ሰመ ዕፀዋት ወአዝርእት(Plant)፣ ሰመ ነገራት(name of things). Making nouns plural for Ge’ez language to change a singular noun to Plural.

in Ge’ez to plural number by prefixing አ” at the beginning, suffixing “ን ፣ ፣ይ ፣ ያን፣ ል ፣ ት ፣ ም ፣ ው. Example. ነጠላ/ዋሕድ/singular **ሀገር** ቅድስት ትሴብሕ ፈጣሪሃ& ብዙኅ/ Plural አህጉር ቅዱሳት ይጼልያ.

Pronouns (መራሕያን (ተውላጠ አስማት)): In a sentence, a pronoun is a word that replaces a noun. In Ge’ez language, there are different kinds of pronouns. Those are personal pronoun, possessive pronoun, objective pronoun, demonstrative pronoun, reflective pronoun.

Table 2.4: Pronouns

	ግእዝ	አማረኛ	English		ግእዝ	አማረኛ	English
፩	አነ	እኔ	I	፮	አንትን	እናንተ	You
፪	ንሕነ	እኛ	We	፯	ውእቱ	እሱ	He
፫	አንተ	አንተ	You	፰	ይእቲ	እሷ	She
፬	አንቲ	አንቺ		፱	ውእቶሙ	እነሱ	They
፭	አንትሙ	እናንተ		፲	ውእቶን	እነሱ	They

Demonstrative pronouns: it is a type of pronoun that is used to show something. It is used to replace a noun that has been mentioned previously in conversation or in a written work. Instead of repeating the noun in multiple sentences, speakers or writers will sometimes use a demonstrative pronoun to refer to it.

Table 2.5:Demonstrative Pronoun

አቅሩብ አስተማሪ(near)			ዘርሑቅ አስተአማሪ(far)		
ግዕዝ	አማረኛ	English	ግዕዝ	አማረኛ	English
ዝ ዝንቱ	ይህ ይኸው	This	ዝኩ ዝስኩ ውእቱ	ያ ያው ያውና	That
ዛ ዛቲ	ይቺ ይቸው	This	እታክቲ አንትኩ ይእቲ	ያቺ ያቸው ያቸውና	that
እሉ እሎንቱ	እኒህ እኒሁ	These	እሙንቱ እልከቱ ውአቶሙ	እነዚያ እነዚያው እነዚያውና	Those
እላ እላንቱ እሎን	እኒህ እኒሁ	(ለሴት) These	እማንቱ እልክቶን እልኮን	እነዚያ እኒያው	Those

Example: - ዝንቱ ውእቱ እኍየ (that is my brother)

Objective pronoun: The Ge'ez language has its objective pronouns, just like the English language. In a sentence, objective pronouns are pronouns that receive the action. Me, you, him, her, us, them, and who are the pronouns in question.

Table 2.6፡Objective pronoun

ግዕዝ	አማረኛ	English
ኪያየ/ኪየነ	እኔን/እኛን	Me/Us
ኪያከ/ኪያኪ	አንተን/አንቾን	You(singular)
ኪያክሙ/ኪያክን	እናንተን/እናንተን(ሴት)	You(plular)
ኪያሁ/ኪያሃ	እሱን/እሷን	Him/Her
ኪያሆሙ/ኪያሆን	እነሱን/እነሱን(ሴት)	Them

ድርብ ተውላጠ አስማት(Double **Pronoun)፡** አሐደ ነገር(one action) ከመተፈጸመ(perform) በአሐዱ ሰብእ(by one person) ብሂል (called) double pronounውእቱ ።

Example: በአማርኛ፡-እኛ ራሳችን እንሞታለን

በግእዝ፡-ለሊነ ንመውት

English:We ourselves will die.

Table 2.7፡ Double Pronoun

ግእዝ	አማረኛ	English
ለልየ ለሊነ	እኔ ራሴ እኛ እራሳችን	My self Our self
ለሊከ ለሊኪ ለሊክሙ ለሊከን	አንት ራስህ አንቺ ራስሽ እናንተ ራሳችሁ(ወ) እናነተ ራሳችሁ(ሴ)	Your self Your selves
ለሊሁ ለሊሃ	እሱ ራሱ እሷ ራሷ	Himself Herself
ለሊሆሙ ለሊሆን	እነሱ ራሳቸው	Them selves

Verb: It indicates or shows the action word that tells the listener or reader what is happening in the sentence and has more to do with mental processes and perceptions. In

Ge’ez language, four different verbs describe an action. Those are Root verb, Active and passive Verb to be/have: it is a verb that has/ has.

Table 2.8:verb

ግእዝ	አሉታዊ ግእዝ	አማረኛ	አሉታዊ አማረኛ	English	Negative marker
	አልቦ		የለም		Haven’t
ውእቱ-----በ/በቱ ይእቲ------ባኣ/ባቲ	አልቦቱ አልባቲ	አለ/አለው አላት	የለውም የላትም	Is/was-----Has	He/she hasn’t
ውእቶሙ-------በቶሙ እሙንቱ-------በን ውእቶን -----በሙ እማንቱ-----በቶን	አልቦሙ አልቦን	አላቸዉ ነበራቸው አላቸው ነበራቸው	የላቸውም የላቸውም	Are/were---Have	They haven’t
አንተ--------ብከ	አልብከ	አለህ/ነበረህ	የለህም	Are/were Have	You haven’t
አንቲ------ብኪ	አልብኪ	አለሽ/ነበረሽ	የለሽም	Are/were Have	
አንትሙ--ብክሙ	አልብክሙ	አላችሁ/ነበራችሁ	የላችሁም	Are/were Have	
አንትን-----ብክን	አልብክን	አላችሁ/ነበራችሁ	የላችሁም	Are/were Have	
አነ-------ብየ	አልብየ	አለኝ/ነበረኝ	የለኝም	Are/were Have	I haven’t
ንሕነ-----ብነ	አልብነ	አለን/ነበረን	የለንም	Are/were - have	We haven’t

Example. አልቦ ኄር ዘእንበለ አሐዱ ፈጣሪ:

Auxiliary verbs are verbs that provide functional or grammatical meaning to the clause in which they appear, such as tense, aspect, modality, voice, stress, and so on... Helping verbs, helper verbs, and (verbal) auxiliaries are all terms for auxiliary verbs.

Table 2.9: auxiliary verbs

ግእዝ	አማርኛ	English
1 ውእቱ	ነው፥አለ፥ነበረ	Is, Be, was
2 ይእቲ	ናት፥አለች፥ነበረች	Is, Be, Was
3ውእቶሙ	ናቸው፥አሉ፥ነበሩ	are, be, were
4 ውእቶን	ናቸው፥አሉ፥ነበሩ	are, be, were
5 አንተ	ነህ፥አለህ፥ነበርክ	are, be, were
6 አንቲ	ነሽ፥አለሽ፥ነበርሽ	are, be, were
7 አንትሙ	ናችሁ፥አላችሁ፥ነበራችሁ	are, be, were
8 አንትን	ናችሁ፥አላችሁ፥ነበራችሁ	are,be,were
9 አነ	ነኝ፥ነበርኩ፥አለሁ	am,be,was
10 ንሕነ	ነን፥ነበርን፥አለን	are,be, were

Possessive pronoun is a pronoun that refers to something of a specific kind that belongs to someone. In English, the possessive pronouns' mine' and 'yours' are used. They show that a noun is in possession or ownership when they replace nouns in a sentence.

Table 2.10:Possessive Pronoun

ግዕዝ	አማረኛ	English
ዘአሁ/ዘዚአሁ አንቲአሁ/እሊአሁ	የሱ የሱው	His (3rd person possessive pronoun)
ዚአሃ/ዘዚአሃ አንቲአሃ	የሷ የሷው	Her
ዘአሆሙ/ዘዚአሆሙ እንቲአሆሙ/እሊአሆሙ ተባዕታይ ዚአሆን/ዛዚአሆን/ አንቲአሆን/እሊአሆን አንስታይ (female	የነሱ የነሱው	Their Theirs
ዚአየ/ዘዚአየ እሊአየ/አንቲአየ	የኔ የኔው	My Mine
ዚአነ/ዘዚአነ እሊአነ/እንቲአየ		Our Ours
ዚአከ/ዘዚአከ/እንቲአከ/እሊአከ ዚአኪ/ዘዚአኪ/እንቲአኪ/እሊአኪ ዚአክሙ/ዘዚአክሙ/እንቲአክሙ/እሊአክሙ ዚአክን/ዘዚአክን/እንቲአክን/እሊአክን		Your Yours

Reflective pronoun: There are some Kinds of reflective pronouns that asking a question and reply to an answer for the predefined user. However, the user is either a person or thing. In Ge'ez language, there are four different reflective pronouns those are when, who, where, and what.

Table 2.11: Reflective pronouns

እዝ	አማርኛ	English
መኑ (አይ)	ማን	Who
እለመኑ(ብዙ)	እነማን	
ምንት(አይ)	ምን	What
ምንተ	ምንን	At what
ምንታት	ምኖች	
ማዕዜ	መቼ	When
አይ		
በአይ ዘመን	በየትኛው ዘመን	While
አይ	መቼን	When/ Since
አይቴ	የት	Where
ኃበ አይቴ	ወዴት	To where
እም አይቴ	ከየት	From where
ስፍን	ስንት	How money
እስፍንቱ	በስንት	How much
በእስፍንቱ	ስንትን	How/ In how
እስፍንተ	እንዴት	To where
እፎ	እረግ እረግ ምንኛ	Which
እፎ እፎ	የቱ/ማንኛው	
አይ	የትኛው	In which
አይቴሁ	በየትኛው	
አያት	የትኞቹ/ማናቸው	

Adjective(ቅጽል): An adjective is a word that describes, identifies(ምልክት), or further defines a noun or pronoun. things behavior or characteristics, like shape, size, color, type, property. There are different kinds of adjectives in the Ge'ez language.

Table 2.12: Adjective

ግእዝ	አማረኛ	English
ሕቅ/በበሕቅ	ጥቂት/በየጥቂት	Little
ንስቲት/ሕዳጥ	ትንሺ/ትንሽ	A few/Few
ብዙኀ/ንሕኑሕ/ዝህዙሕ	ብዙ/የብዙ ብዙ	Many/ much
ኵሉ/ኵሎሙ	ሁሉ/ሁላቸው	Whole

Example: ሕዳጥ ሕብስት ተርፈት ለነ። የዋህ ብእሲ ኃለፈ ዮም። from this sentence the word የዋህ and ሕዳጥ are adjectives that describe the character/property/ conduct of a person and size of food.

Adverb: In Ge’ez language there are five kinds of adverbs, are adverb of time (ዘጊዜ ተውሳከ ግስ), adverb of frequency (ካዕበታዊ ተውሳከ ግስ), adverb of place (መካናዊ ተውሳከ ግስ), adverb of manner (ኵታዊ(ሳድስ ቅጽላዊ) ተውሳከ ግስ), adverb of reason (ምክንያታዊ ተውሳከ ግስ), adverb of question (መጠይቃዊ ተውሳከ ግስ)

A. **adverb of time (ዘጊዜ ተውሳከ ግስ)፡-** adding some time modifier adverbs for a sentence. Such as tomorrow(ጌሠም), today(ዮም/ይእዜ), yesterday(ትማልም), before(ትካት), morning(በጽባሕ), at night (ምሴት), after (ድኅረ), after tomorrow(ሳኒታ). Example: - መምህር የሐውር ጌሠም. From this sentence the word ጌሠም indicates time.

B. **Adverb of frequency** (ካዕበታዊ ተውሳከ ግስ)፡- the word in a sentence It appears that the quantifier an action/ event how much time repeat at a time or number of does not fit with the uncountable noun frequency. Consider changing it. Some frequency modifiers in Ge’ez language like three times (ሠለስተ ጊዜ), always(ኩለሄ), sometimes (ኣሐደ ኣሐደ ጊዜ/ ኃልፎ ኃልፎ), daily(በበዕለቱ), every time (በበጊዜሁ). Example: - አነ እደግም ጸሎትየ በበእለቱ. In this sentence, the word በበእለቱ indicates the frequency of time.

Conjunctions: Words, phrases, and clauses (both dependent and independent) are held together by conjunctions. There are three types of conjunctions: coordinating, subordinating, and correlative. Each serves a different purpose, but they all work to connect words.

Table 2.13: Conjunctions

ግእዝ	አማርኛ	English
ከመ አምሳለ በዘ	እንደ/እንድ እንዳ እንዲ	As As
በእንተ ህየንተ በይነ አመ	ስለ/ዘ	About
ሶበ ጊዜ	በ	For/ to
አምጣነ እስመ አኮኑ እመ	ከ/ቢ/ባ	
ድኅረ ቅድመ ዘእንበለ እስከ በእንተዝ ወ አው ዓዲ ባሕቱ ዳዕሙ አላ	ኋላ/በኋላ በፊት ----- እስኪ/ድረስ ሲ/ሳ/ስ/እየ እና ወይም ገና ነገር ግን እንጅ	After Before And Or But

A preposition is a word used before a noun, noun phrase, or pronoun to show direction, time, place, location, spatial relationships, or to introduce an object. In Ge’ez, the following proposition is well known.

Table 2.14:Preposition

ግእዝ	አማረኛ	
ዲበ	ላይ፣ በ-ላይ	On
ላእለ	ከ - ላይ	Above
መልዕልተ		
ታሕተ	ከታች	Under
መትሕተ	በታች	
ውስተ	ውስጥ	Inside
ውሳጤ	በ - ውስጥ(ከ - ውስጥ)	
ማእከለ	በመካከል	In the middle
ኀበ		
መንገለ	ወደ	To
እም	ከ	
እምነ	ከ - ይልቅ	From
ምስለ	ከ - ጋራ	With
በ/ለ	በ - ጋራ	
ዘ		For/to/by
እንተ	የ	
እለ		Possessives

Example: እስመ፡ አንተ፡ እግዚአብሔር፡ ባሕቲትከ፡ ልዑል፡ ዲበ፡ ኵሉ፡ ምድር, “For you are the only Lord, high over all the earth

2.2.7 Ge’ez sentence structure

The syntax is the element of grammar that represents a speaker's comprehension of sentences and their structures. When we talk about syntax, we usually mean the sequence of words and the structure of sentences. Every sentence is made up of words, but not every group of words is a sentence. This is because syntactic structures ensure that language is well-formed and grammatical, but disregarding them results in ill-formed and grammatical speech. So, to construct a meaningful full sentence, any human language has syntax or grammar. However, due to differences in syntax, sentence structure may differ from one language to the other. In contrast to Amharic, the Ge'ez language has its own set of sentence patterns (syntax). In the Amharic language, the structure of the sentence is SOV, however, in Ge’ez language the most frequent sentence structure is VOS. Sometimes the sentence in Ge’ez language is flexible.

It might be SOV, SVO, OVS, VOS, OSV and so, where S= subject, O= object, and V=verb.

Example : ኢትዮጵያን ያስተማረ(O) ዮሐንስ(S) የጌታ ደቀመዝሙር ነው(V። John is the disciple of Christ who preaches Ethiopia። ዮሐንስ(S) ውእቱ(V) ረድአ ክርስቶስ ዘሰበከ ኢትዮጵያ(O)።

2.3 Natural language processing

NLP is the automatic processing of human language communication on machine processing. It is the study of analyzing, understanding, and generating the languages that humans beings use to naturally communicate to interface with computers in both written and spoken contexts using natural human languages instead of computer languages(Abera, 2018). Different scholars tried to develop different information management systems(Worku, 2015). The drawing of summarized and relevant information from huge information can be facilitated and the right information for decision making can be acquired. Among the different solutions to the problems are: Information Retrieval (IR), Information Extraction (IE), Question Answering, Text Summarization, machine translation, and Text Categorization(Hirpassa, 2017). In the development of natural language processing systems different researchers work different systems and models by using Ge’ez language among those work(Abebe, 2010; Alemu et al., 2021; Kassa, 2018).

2.3.1 Text Summarization

The study of (Shabbir Moiyadi et al., 2016) the automatic text summarization is one of the numerous jobs in the research area of NLP. The text summary is the process of extracting the most significant information from a text document to create a condensed version. The purpose of this procedure is to arrive at a concise synthesis of the document's ideas without losing their meaning. As the problem of information overload and data availability on the Internet worsens, the quality of user search query results is deteriorating. When it comes to summarizing text, there are two primary approaches: extractive and abstractive. The extractive strategy removes the most important portions of a document and combines them to create a summary, whereas the abstractive approach extracts and paraphrases the input text to generate a summary of the quality of a human-written summary some researchers work on text summarization (Kantzola, 2020; Lieu, 2015; Risne, 2019).

2.3.2 Question Answering (QA)

Question answering (QA) is an area of artificial intelligence that focuses on natural language processing and information retrieval(Samuel et al., 2018). It involves developing systems that respond to queries presented by humans in natural language. Its programs can create replies by querying a knowledge base (a structured database of knowledge) or an unstructured collection of natural-language documents. Closed domain (answering questions from a certain domain) or open domain (answering questions from all domains) question answering systems exist (relying on general ontologies and widespread knowledge). Different researchers are working on a Question-Answering system (Abrahamsson, 2018; Bete, 2013).

2.3.3 Information Retrieval (IR)

Information Retrieval is a process of recovering stored information from large datasets to satisfy user’s informational needs(Yuan, 2020). It is a branch of science concerned with the representation, storage, organization, and retrieval of information items to meet the needs of users and the process of retrieving relevant information resources for a query from a huge collection of data. The huge information collection, unlike typical database retrieval, is largely semi-structured data, such as text. The data is processed using syntactic methods, leaving the semantic understanding of the data behind (NLP).

Information retrieval systems are a first step toward addressing the problem of information overload. When a user makes a query, the IR system runs it through its database and delivers a group of documents as a result. Some researchers are working on information retrieval to overcome some natural language task challenges.

The work in Yuan (2020) used semantic embedding to improve the performance of Information Retrieval (IR) for Covid-19 related tasks by combining multiple popular semantic embedding models and finding a more effective ranking for retrieving a better IR result through a comparative analysis of these semantic embedding technologies based on the entity-based rather than the present embeddings, which are largely based on words, phrases, or documents of IR function were provide (Johar, 2020).

The approaches and performance measures of information retrieval were machine learning, supervised, unsupervised, hybrid and precision, recall, and F-score. The main difference between information retrieval and information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents, whereas information retrieval (IR) is the task of using queries to locate the material (documents) of an unstructured nature (usually text) that satisfies an information need from large collections (usually stored on computers).

2.3.4 Information Extraction

Information extraction is one of the areas of the NLP field dedicated to the general problem of detecting entities referred to in natural language texts, the relations between them, and the events they participate in (Gao et al., 2020). It is to extract or discover structured information from unstructured or semi-structured text from a document or other source. It has subtasks that are very common and closely related: named entity recognition and relation extraction(Yang et al., 2019). It is extracting structural factual data mostly from unstructured text (web pages, text documents, office documents, presentations, and so on). IE usually uses NLP tools, lexical resources, and semantic constraints for better efficiency (Hirpassa, 2017; Valero et al., 2009; Zhou & Qi, 2016).

As stated in Automatic Content Extraction (ACE) is an evaluation conducted by NIST to measure the tasks of Entity Detection and Tracking (EDT) and Relation Detection and Characterization (RDC). The Entity Detection task requires that selected types of entities mentioned in the source data be detected, their sense disambiguated, and that selected attributes of these entities be extracted and merged into a unified representation for each entity. The goal of RDC is to detect and characterize relations of the targeted types between EDT entities (Worku, 2015).

The Automatic Content Extraction (ACE) Program's goal was to create extraction technology that would allow for automatic processing of source language data(Strassel et al., 2008) (in the form of natural text and as text derived from ASR and OCR). At the time, automatic processing encompassed classification, filtering, and selection based on the source data's language content or the meaning provided by the data. As a result, the ACE program necessitated the creation of technology that could recognize and characterize i.e., automatically. The detection and characterization of Entities, Relations, and Events were seen as the ACE research objectives. To complement the ACE Program, LDC created annotation rules, corpora, and other language resources. Some of these tools were created in collaboration with the TIDES Program to aid in the evaluation of TIDES Extraction.

The basic annotation task, Entity Detection and Tracking (EDT) laid the groundwork for the other activities. Person, Organization, Location, Facility, Weapon, Vehicle, and Geo-Political Entity were identified as later ACE tasks (GPEs)(Doddington et al., 2004). Each type was broken down further into subcategories (for instance, Organization subtypes include Government, Commercial, Educational, Non-profit, Other). All mentions of each entity in a document, whether named, nominal, or pronominal, were marked by annotators. The annotator labeled the top of each mention after determining the maximum length of the string that describes the entity. The detection of relationships between items was the focus of Relation Detection and Characterization (RDC). General entity classes in ACE are described as follows.

Table 2.15:ACE

Type	Subtypes
FAC (Facility)	Airport, Building-Grounds, Path, Plant, Subarea-Facility
GPE (Geo-Political Entity4)	Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
LOC (Location)	Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
ORG (Organization)	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non_Governmental, Religious, Sports
PER (Person)	Group, Indeterminate, Individual

Adapted from(Doddington et al., 2004)

2.4 Subtasks of Information Extraction

Applying information extraction directly to a free text is a challenging task. To improve this challenge splitting the tasks such as Named Entities Recognition (NER), Relation Extraction, Event Extraction, Co-Reference Resolution, Template Element Construction (TE), Template Relation Construction (TR), and Scenario Template Production (ST).

2.4.1 Named entity recognition

In general, named-entity recognition (NER) (also known as entity identification or entity extraction) is a subtask of information extraction that aims to find and classify identified entities stated in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, and so on. When extracting information from unstructured text became a significant difficulty in 1996, the term "Named Entity" was coined at the 6th Message Understanding Conference (MUC)(Shaalan, 2014). In the linguistic domain, named entity recognition entails automatically scanning unstructured text for “entities,” which are then normalized and classified, such as person names, organizations (such as companies, government organizations, and committees), locations (such as cities, countries, and rivers), and the date and time expressions (Mansouri et al., 2008).

The term Named Entity was coined in 1996, at the 6th MUC conference, to refer to “unique identifiers of entities”(Marrero et al., 2013). In simpler words, a Named Entity is a real-world object that identifies Entity types. A Named Entity Recognition system was developed by several academics(Zirikly & Diab, 2015) or named entity Extraction (Esteban et al., 2018) developed Deep Bidirectional Recurrent Neural Networks as End-To-End Models for Smoking Status Extraction from Clinical Notes in Spanish.

The research of (Gao et al., 2020) developing a Medical Named Entity Extraction from Chinese Resident Admit Notes Using Character and Word Attention-Enhanced Neural Network by developing bidirectional long short-term memory (BILSTM) and conditional random field (CRF) to extract entities.

2.4.2 Relation Extraction (RE)

This is a subtask of information extraction it extracts the relationships between entities. Entities and relations are used to correctly annotate the data by analyzing the semantic and contextual properties of data(Adnan & Akbar, 2019).

2.4.3 Co-Reference Resolution

It is one sub-tasks of information extraction systems in the natural language processing field. It is a process of finding multiple references to the same thing about a single entity is expressed in different sentences using pronouns(Abera, 2018).

2.4.4 Template Element Construction (TE

It is one subtask of the information extraction system which associates descriptive information for the NER results. The different recognized name entities in association with their co-reference are the input of template element construction(Gao et al., 2020; Hirpassa, 2017).

2.4.5 Template Relation Construction (TR)

In the information extraction task extraction of relations among entities is the main feature. After the template element construction, the template relation task is applying to identify the possible relationship between template elements(Worku, 2015).

2.4.6 Scenario Template Production (ST)

This sub-task is creating a scenario by combining the template element construction and template relation construction.

2.5 Techniques and Algorithms to Solve Information Extraction Problem

2.5.1 Machine Learning based Methods

Information extraction is a domain-specific and task-related problem. To solve this problem the dataset is sequentially labeled and knowing the domain to extract. Currently, researchers use some technical methods for information extraction like, Rule-based and dictionary-based, machine learning-based methods, the neural network method, and the mixed model. In machine learning methods the researchers include HMM, SVM, and CRF. This method needs a large annotated corpus that stores various features.

The researcher uses Machine Learning approaches for Extracting Information from Natural Disaster News Reports (Valero et al., 2009) for extracting information about five different types of natural disasters: hurricanes, earthquakes, forest fires, inundations, and droughts. Experimental results on a collection of Spanish news show the effectiveness of the proposed system for detecting relevant documents about natural disasters get an F-measure of 98% and extracting relevant facts to be inserted into a given database reaching an F-measure of 76%.

2.5.2 The Deep Learning method

The term deep learning was first oriented in the 1980s and has two main reasons at the recent time become popularly useful, it requires large amounts of labeled data and requires substantial computing power. It has three layers: input layer, hidden layer, and output layer. Deep learning is a branch of machine learning which aims to model high-level abstractions in data. This is done using model architectures that have complex structures or those composed of multiple nonlinear transformations (Elfaik & Nfaoui, 2021).

In deep learning when we train a model it needs a large amount of labeled dataset and it contains multiple layers within a neural network architecture that learns features directly from text, image, audio, and video data source without the need for manual feature extraction to achieve the state-of-the-art accuracy. Deep learning is used in computing and modeling multiple levels of abstraction data through multiple processing layers, and output people expected results. It is widely used in computer vision, artificial intelligence, genomics, and biomedical science, and other domains and brought a breakthrough in image, video, audio, and speech processing of deep learning methods for information extraction. The researchers use some of the recent deep learning approaches such as LSTM, BiLSTM, ANN, CNN, and RNN to solve the information extraction tasks.

In (Chiu & Nichols, 2016) a convolutional neural network as a character-level features extractor and combined with the Bi-LSTM network model. The model used a convolution and max layer to extract the feature from pretraining vectors. They used the additional features capitalization feature, lexicons, and a four-dimensional vector representing four-character types. From the study of (Chalapathy et al., 2016) the extraction task aimed at identifying and classifying concepts into predefined categories i.e. treatments, tests, and problems. State-of-the-art concept extraction approaches heavily rely on hand-crafted features and domain-specific resources which are hard to collect and define. For this reason, the researcher used a recurrent neural network i.e., Bidirectional LSTM with CRF decoding initialized with general-purpose, off-the-shelf word embeddings.

The need to extract meaningful information from big data, classify it into different categories and predict end-user behavior. used a Convolutional neural network (CNN) model that uses convolutional layers and maximum pooling or max-overtime pooling layers to extract higher-level features, whereas LSTM models can capture long-term dependencies between word sequences this is better for text classification. (Gao et al., 2020; Jang et al., 2020) proposed hybrid attention Bi-LSTM+CNN model produces more accurate classification results and higher recall and F1 scores. The work of (Nguyen, 2018) developed a deep learning method to increase the effectiveness problems of information extraction that are event extraction and entity linking.

2.5.3 Convolutional Neural Networks

This is one of the recent deep learning approaches to solve information extractions tasks, it can learn more complex tasks with a need for a larger dataset. It is the ability to process/interpreting temporal data that come in sequences and filters within convolutional layers to transform data i.e., sentences whereas failing. In these methods, the model uses the Relu activation function and uses local response normalization. In the work in (Jang et al., 2020) the accuracy of the training set decreases, and the error rate increases because the model is too complicated, Relu optimization becomes difficult in this case the model cannot achieve good learning results. When increases the training speed of the model improves the training effect of the model.

2.5.4 Recurrent Neural Network

The recurrent neural network is a subtype of a deep neural network that processes/ interpret sequential data. It taking input and reusing the activations of previous nodes or later nodes in the sequence to generate the output (Ambalina, |March 09, 2020). RNN algorithm can learn from ‘past’ data and ‘future data’ in the sequence, and the networks are interconnected and interacting with neurons. In this type of network, the direction of propagation of the information is bidirectional it keeps the sequence of data and makes the connection between an input of long sequences i.e. internal memory(Jauregi Unanue et al., 2017).

It is very useful in information extraction, information retrieval, automatic translation, automatic speech recognition, and automatic pattern recognition. Recurrent neural networks (RNNs) are a type of network that uses recurrent connections to build memory. The inputs in feedforward networks are unrelated to one another. In an RNN, however, all inputs are connected. This enables the network to demonstrate robust temporal behavior for a while, making it ideal for sequential classification tasks such as named entity recognition.

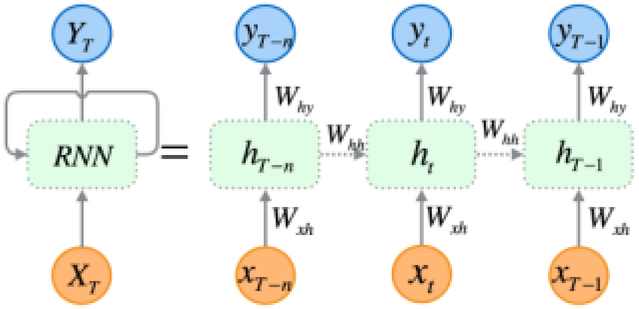


Figure 2.1: RNN adapted from((Cui et al., 2018))

2.5.5 Long Short Term Memory Units

LSTM was firstly proposed by Hoch Reiter and Schmid Huber (1997) to overcome the gradient vanishing problem of RNN(Zhou & Qi, 2016). The main idea is to introduce an adaptive gating mechanism, which decides the degree to keep the previous state and memorize the extracted features of the current data input. (Hochreiter & Schmidhuber, 1997; Sherstinsky, 2020). It can able to store more contextual information. (Jeyakumar, 2018)This layer has a chain-like structure of repeating units.

Also, each unit is composed of a cell, an input gate, an output gate and a forget gate working together. It is well-suited to classify, process, and predict time series with time lags of unknown size and duration between important events (Jeyakumar,2018). The LSTM model is regarded as one of the most prevalent variations of the RNN (Recurrent Neural Network) method. LSTM is designed to handle vanishing gradient issues using the gate mechanism.

Generally, LSTM units help to bypass some errors backpropagate through time in general RNN.

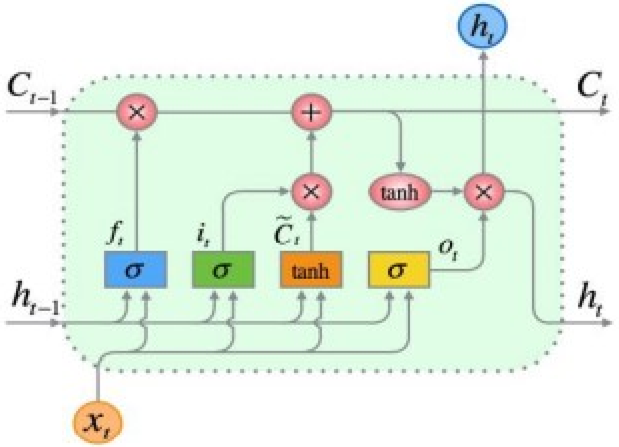


Figure 2.2: LSTM Unit

A neural network's notes are triggered by the notes it receives. LSTM gates, on the other hand, pass or reject input based on its weight. The weights for each of these signals are then grated separately. The weights that control hidden states and input are then modified by the RNN's learning process. Finally, through the steps of guessing, backpropagating error, and gradient descent weight modulation, these cells learn when to allow information to enter, exit, or be deleted(Aziz Sharfuddin et al., 2018).

2.5.6 Bidirectional LSTMs

Bidirectional LSTM is a type of Recurrent Neural Networks (RNNs) that is straightforward (Jeyakumar, 2018) Which is composed of forwarding LSTM and backward LSTM. It is a sequence of processing consists of two LSTMs the forward direction(input) first layer and the backward direction reversed copy output (Chalapathy et al., 2016; Jeyakumar, 2018). It increases the amount of information available to the network, improving the content available to the algorithm (e.g., knowing what words immediately follow and precede a word in a sentence).

It makes it possible to include more context by using the data from the past and the future. It will help to find better feature representations from the input. The stacking of Bi-

Directional LSTM units in each layer helps to capture the non-linear feature representations. BiLSTM based networks are proven to be effective in sequence labeling problems for they have access to both past and future contexts also minimize overfitting problems (Elfaik & Nfaoui, 2021).

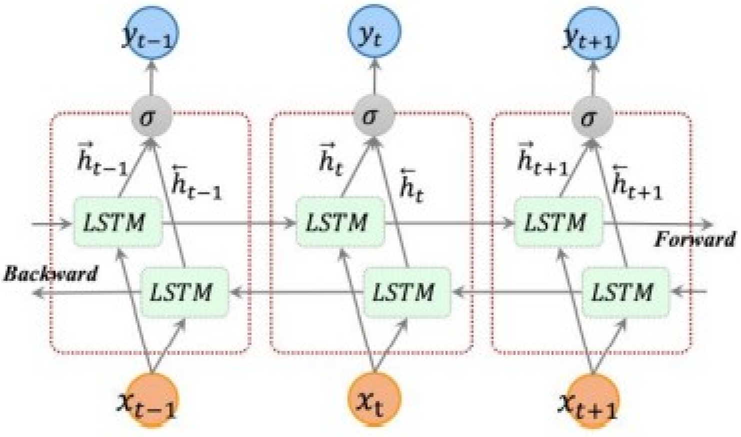


Figure 2.3: BiLSTM (unit adapted from (Aziz Sharfuddin et al., 2018))

BiLSTMs have proven to be very useful when the context of the input is required. Information moves from backward to forward in a unidirectional LSTM. The Bi-directional LSTM uses two hidden states to flow information not only backward to forward but also forward to backward. As a result, Bi-LSTMs have a greater understanding of the context. BiLSTMs were employed to increase the amount of input data that could be utilized by the network.

2.6 Evaluation matrices of Information Extraction

To evaluate the performance of the information extraction system is necessary to use the best-accepted performance measures of accuracy result such measurements are precision, recall, and F-measure. Precision and recall have been and continue to be very useful measures of performance for extraction (Dalianis & Dalianis, 2018; Kumar, 2017; Lee, 1998).

Precision The percentage of correct positive predictions returned by the system is known as precision. The ratio between the number of IE accurately detected by the system True Positives (TP) and the total number of IE returned by the system is used to calculate it. Divide TP by the sum of TP and false positives to get the accuracy (FP).

- **Precision** $P = \frac{\text{True Positive}}{\text{True Positive + False Positives}}$

For example:

Recall: The percentage of affirmative cases identified by the system is referred to as recall. It is calculated as the ratio of the number of NEs correctly identified by the system (TP) to the number of NEs expected to be recognized by the system. A recall is calculated by dividing the number of (TP) by the sum of (TP) and false negatives (FN).

- **Recall:** $R = \frac{\text{True Positive}}{\text{True positive + False Negatives}}$

F-measure: Depending on the weight function:-2*(Recall* Precision/recall + precision), the F-measure is the weighted average of both precision and recalls (Hirpassa, 2017; Jayaram, 2017). The F-measure is defined as the common weighted harmonic mean of Precision and Recall.

- **F-measure**: $F = \frac{\text{2Precision} * \text{Recall}}{\text{Precision + Recall}}$

Generally in the proposed system used accuracy performance metrics.

2.7 Related work for information extraction system

2.7.1 Information extraction system from Amharic text

In the study of (Worku, 2015) using a knowledge-poor approach, he examined how to design and create an autonomous information extraction system for Amharic text. The extraction of named entities, numbers of orthographic coreference was addressed in his research. However, the extraction of relationships between entities and the extraction of co-reference relationships or nominal co-referencing are his research gap or limitations. When the system extracts text from Amharic News text 24760 tokens for train and tests his proposed system by prepared the NER and evaluating the system, it shows promising performance with 90.5% Recall, 89.2% Precision, and 89.8% F1-measure and recommends that nominal co-referencing be used in the future to improve extraction performance

The study of (Hirpassa, 2017) was developing an information extraction system for the Amharic vacancy announcement text. The system was developed by using Python and a rule-based technique was applied to address the problem of automatically deciding the correct candidate texts based on their surrounding context words. He only extracted the selected candidate text. However, the relation extraction or the relation between attributies are not included in his research work. Finally, he obtained results from his experimental work 79.56% precision and 66.6% recall shows from 34 Amharic vacancy announcement texts.

The research presented in (Hirpassa, 2017) used a machine learning approach to design the system and he collects 220 texts from different news and 1225 different related news for training and used MM classifiers and get the performance of the system 67.1%,69.1%, and 72% of result scored. His was used only names and numbers attributes to extracted not include the relation between attributes is his limitation or research gap. The recommended that improve the performance of the system use more data set for training NER, POS, Morphological analyzer, and wordnets.

2.7.2 Information Extraction from Afaan- Oromo News Texts

Using a machine learning method, (Abera, 2018) attempted to construct a model for information extraction from Afaan-Oromo news texts. In his work, only the Names and numbers are considered as candidate texts because names and numbers are the most important facts for extraction. But extracting the relationship between the entities requires different NLP tools which are not publicly available for Afaan- Oromo. Recall, Precision, and F-measure are used as evaluation metrics for AOTIE. Being trained and tested on the dataset of size 3169 tokens, AOTIE performed 79.5% Precision, 80.5% Recall, and 80% F-measure. He also recommends the relation between attributes are in further study will be vital for the development of full-fledged information extraction. And apply the rule-based approach and compare it to the machine learning approach.

2.7.3 Information Extraction from the foreign language

In their study (Panda et al., 2019)used the rule base approaches for their models tested on several user's queries, and in each of the tests conducted and the model successfully obtained the answer to the query and classify all types of WH-type questions accurately. Based on the findings, the performance of the system was performed with 75% precision & 89% of recall. The researchers recommend that improving the model using machine learning techniques to increase accuracy. The authors work which type of Question ask the user and the database system extract reliable answer for the question depending on the user's query. Lastly, the authors recommended using the machine learning approach to train more data. The researcher's gap is only extracted from the WH- type question.

(Jauregi Unanue et al., 2017) aimed to evaluate the current information extraction features by developing good features by inherently heavily time-consuming and the authors use Bidirectional LSTM and the Bidirectional LSTM-CRF to solve the past problem in feature engineering. They obtained the best results with the Bidirectional LSTM-CRF model. This work only solves the gradient problem of information feature extraction did not consider the similarity of words in a text.

Studied by (Gao et al., 2020) developed a Medical entity extraction of RANs For Chinese electronic medical records. When the authors find the previous gap of the research in each medical entity contains not only word information but also rich character information. Although the authors state that an Effective combination of words and characters is very important for medical entity extraction. For this reason, the authors proposed and develop a medical entity recognition model based on a character and word attention-enhanced (CWAE) neural network for Chinese RANs using character-enhanced word embedding (CWE) model and Convolutional Neural Network (CNN) model the study result shows that 94.44% in the F1-score. The research extract of the medical entity does not consider the relationship between medical entities, which is a gap in the research.

Chapter Summary

Depending on the above recent IE research work, the authors have done on different local languages and other languages, rule-based techniques, machine learning approaches, and deep learning techniques. From this review, identify the gap and generate the best-fit approach, corpus size for data set, features set and algorithms, and the result of the system performance for our proposed model. In-text processes the limited data set or corpus size the authors prefer the rule-based approach is good to gain a good performance for different NLP based systems. For this reason, to fill the researcher's gap described in the summary of the related works table 2.16. The researcher proposed a text extraction model using a machine learning approach from an optimized data set and used a deep learning technique, however, the gap is that the problem identification, the approaches, and the state_ art_of_accuracy performance.

Table 2.16:summarize the related work

Author	Title	Problem	Method/tool	Limitation or Gap	Finding
(Worku, 2015)	information extraction system for Amharic text	Information overload. Language and domain-specific issues.	knowledge-poor approach. Incorporating to co-reference resolution and relation extraction.	considered orthographic coreference, not nominal coreference.	90.5% Recall, 89.2% Precision and 89.8% F1-measure.

(Abera, 2018)	Information extraction model from Afaan-Oromo news texts	The IE system developed for English and in the specific domain may not work for Afan-Oromo even if its domain is similar	Machine learning approach.	Extracting relationship between entities	AOTIE performed 79.5% Precision, 80.5% Recall, and 80%F-measure
(Hirpassa, 2017)	IE system for Amharic vacancy announcement text	Unavailability of tools for extracting & exploiting the valuable information from Amharic text.	using Python & a rule-based technique	dataset he used only from one newspaper. Rule-based does not correctly identify candidate text in AVA texts	34 from AVA dataset 79.56% precision and 66.6% recall .
(Panda et al., 2019)	information extraction system for unstructured web data	A huge amount of web data is available online and the majority of which are in the form of unstructured documents & difficult to obtain the exact information from a list of documents quickly as and when required unless the whole document is read	a rule-based with python NLTK tools.	The model excluding auxiliary verb type questions as well as the questions with conjuncts.	performed with 75% precision & 89% of recall

(Jauregi Unanue et al., 2017)	Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition	developing good features is inherently heavily time-consuming	Bidirectional LSTM and the Bidirectional LSTM-CRF	Status name entity extraction	obtained the best results with the Bidirectional LSTM-CRF model
(Gao et al., 2020)	Medical Named Entity Extraction from Chinese Resident Admit Notes Using Character and Word Attention-Enhanced Neural Network	Unappropriated tools to extract entity	character-enhanced word embedding (CWE) model and Convolutional Neural Network (CNN) model.	Character embedding	94.44% in the F1-score
(Chiu & Nichols, 2016)	Named Entity Recognition with Bidirectional LSTM-CNNs	traditionally required large amounts of knowledge in the form of feature engineering and lexicons to achieve high performance.	using a hybrid bidirectional LSTM and CNN architecture	Only knowledge features extraction	F1 score of 91.62 on CoNLL-2003 and 86.28 on OntoNotes

CHAPTER THREE: METHODS AND IMPLEMENTATION

3.1 Introduction

In this chapter, we discussed the architectural design and implementation of the information extraction model. We gathered data from several sources and created a Ge’ez dataset as a sentence for the information extraction model to train the model. The data were gleaned from religious documents, Ge’ez textbooks, religious, telegram channel, and Ge’ez thesis. The proposed architecture methodology includes the preprocessing stage, the data splitting stage, and the prediction stage. The preprocessing algorithms such as Tokenization (divide sentence, remove punctuation mark and numerals), stop word removal, stemming, and padding are some of the preprocessing procedures that have been used. After preprocessing, split the dataset into training and testing phases i.e. classification stage. The training dataset is used to train the created model, which includes the LSTM and BLSTM (bidirectional long short-term memory) stages of sentence encoding. Finally, the trained model predicts or extracts the given text after the trained and testing phases have been completed.

3.2 Data Collection and Preparation

Many natural language comprehension tasks, such as information retrieval, information extraction, question answering, and text summarization are proposed to be supported by the named entity recognition paradigm. Unlike other languages like English and the Amharic language, the Ge’ez language is an under-resourced language(Alemu et al.,s 2021; Kassa, 2018). This resourced language is a complex morphological structure. The challenging task of extracting information from the Ge’ez document has no standard annotated data from the document to extract. To prepare a Ge’ez corpus we gathered data from the EOTC Addis Ababa main office. Such data are Gedle, Yehawariyat sra, Holy Bible, Mezimur, wudase mariyam and Drsan. The IE task does not consider subjective datasets; rather, it is concerned with domain-specific datasets and the significance of the data after extraction. All of the data collection are listed in the table below.

Table 3.1: Dataset Collection

No	Categories
1	Gedle
2	ACT(Yehawariyat sra)
3	Wudase Mariyam
4	Drsan

This dataset was used to do pre-training padding and vector representation. On these data, we used tokenize, stop word, stemming, and padding sequence procedures. The tasks listed above are completed to prepare the Ge'ez dataset.

Gedle, Yehawariyat sra, and Drsan, new testament of the bible contain a wealth of vital information, such as a person's history and the dates of events, the location of the person or event can be used to extract information and use it in a later stage or by another application. Because of this, IE is a domain-specific language, the Gedle, Yehawariyat sra, and Drsan texts were used as training and testing datasets in our research work. The sentences have contains at least two named entities checked manually. The name entity means, person, place, date, and event. We annotated the data set as the structure of the BIO format, we used the letters "B" for the beginning, "I" for the inside of the token, and "O" for the sentence's last (out of entity-tags) token.

The BIO or BILOU schemes are commonly used in NER or entity extraction research. We focused our investigation on the most fundamental and significant elements, such as the person's name, location, time, and event. Take the following sentence as an example: “ደማቴዎስ ሊቀ ጳጳስ ዘለእስክንድርያ ዘይትነበብ በዕለተ በዓሉ ለዐቢይ ወክቡር ወቅዱስ ሊቀ መላእክት ሚካኤል አመ ዐሡሩ ወሰኑዩ ለወርኀ ኅዳር”፡፡ ደማቴዎስ as “B-person፤ እስክንድርያ “B-place”, ዐሡሩ ወሰኑዩ ለወርኀ ኅዳር “B-Time”, and the rest will be outside entities labeled with "O”. Annotation of our dataset contains 9 classes or tags.

Table 3.2: dataset annotation

BIO format	Assign the tags
B (stands for the beginning of the sentence)	B-date (describe date of event(e.g ዐሡሩ ወሰኑዩ ለወርኀ ኅዳር)) B-per(describe a person(e.g ደማቴዎስ)) B-place(place of the person or where an event occurs(e.g እስክንድርያ)) B-eve(an action performed by a person)
I (stands for the inside of the sentence).	I-date I-per I-place I-eve
O (for any attribute found in the document which is not considered as facts).	O ‘assign all words found outside of the tag

Generally, The number of tags are prepared as B-per, I-per, B-place, I-place, B-date, I-date, B-eve, I-eve, and O are placed.

3.3 The Architecture of the Proposed Model

The system architecture of the proposed Model describes that the phase of the overall steps used in the information extraction model from Ge’ez text. The proposed model architecture shows in figure 3.1 below. It has the preprocessing phase (the preprocessing up to model built) and the learning phase(training and testing phase) and the prediction phase predict the information extraction entity i.e. place, person, date, and event).

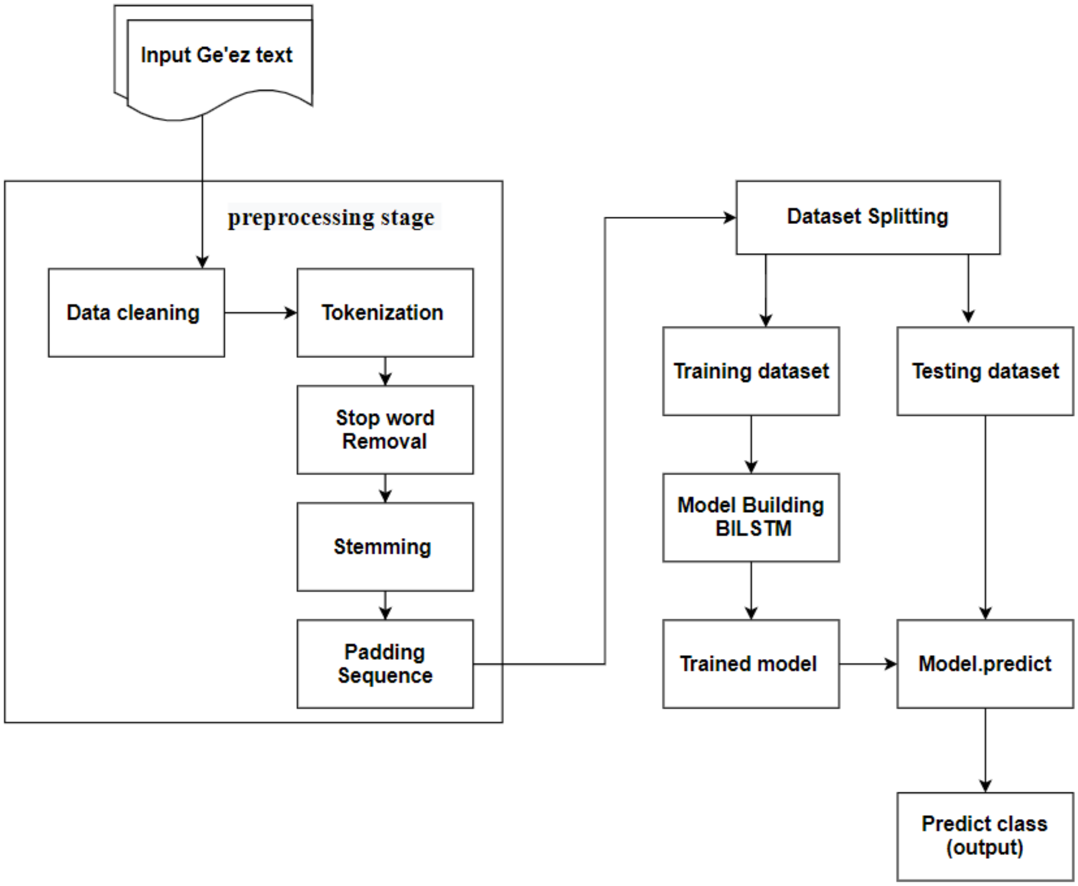


Figure 3.1:Architecture of Information Extraction Model From Ge'ez Text

3.4. Preprocessing of Ge’ez dataset

The dataset collected from different sources containing irrelevant and redundant data, so we need some improvement algorithms to eliminate those unwanted data. Such irrelevant data are URL, HTML, citation, and copyright, etc. This needs the data properly clean and prepare for other preprocessing stages. The data cleaning phase removes unnecessary or noise data like (name of the author, email address, etc.). To develop an information extraction task the dataset was first structured or free from irregularities unless the model or classifier does not understand to learn. At the preprocessing stage, we annotate the dataset using an online annotation tool(tagtog annotation) and assign the labeled or the tag type, and modify the named entity. After this go to the preprocessing stage i.e., tokenization, stop word removal, stemming, and padding. Preprocessing of dataset Perform several activities that we used in our study work not just using all NLP tools.

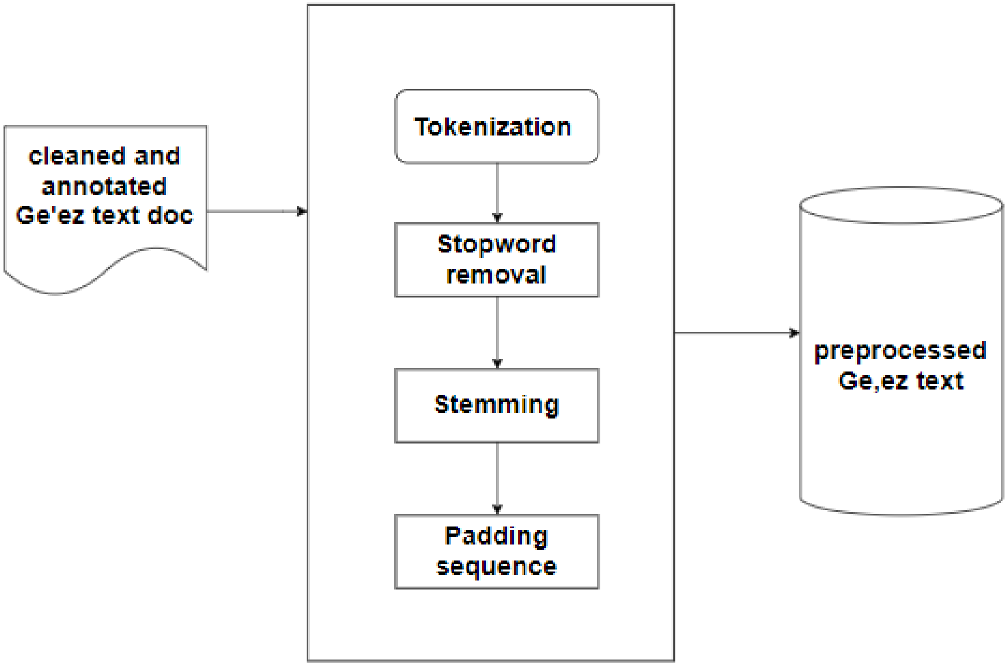


Figure 3.2:Preprocessing of Ge’ez dataset

3.4.1 Tokenization

Naturally, text must be split into linguistic units such as words and sentences before any meaningful text processing can begin. Because the incoming sentence is plain text, we must figure out where sentences begin and end, as well as where words, whitespace, and punctuation marks are located. Because punctuation marks occur at sentence borders in most written languages. It is the first preprocessing stage of natural language processing in text processing. It is the process of breaking down a given text into expected tokens. A text that has been broken down into paragraphs, sentences, or words. However, one of the most difficult tasks in tokenization is figuring out how to split two words into one mining. During this stage token all CSV file datasets both the sentence :# and Entity is split. The entity is describing a word as person, place, date, or event in our dataset has three (3) columns (Word, Tag, and sentence:#). The data from the dataset contain the person name, place name, date, and event in a sentence. The sentence from the dataset contains all the label dataset. Tokenize the text either split or give index for each word. At this stage remove all the punctuation marks and white_space.

Tokenization Algorithm

```
   Accept  Ge'ez input texts
 For  SingleSentence in Sentences
   Split SingleSentence into Words
   For  each Wi in Words
        Add Wi to TokenizedSentence
End For
  End For
```

Algorithm 3.1:Tokenization Algorithm

Ge'ez language has its punctuation marks such as Question mark ፧, paragraph separate ፨, section mark※, word separateor፡, ። full stop ፣ comma ፤ colon ፥ semicolon ፦ preface colon.

Algorithm for punctuation

```
 Open the document
Access punctuation from document
punc= "list of punctuation"
load document
For punct in document:
   if punct in punc:
      remove pun
   End For
End if
```

Algorithm 3.2:Algorithm for punctuation

3.4.2 Stop Word Removal

In a natural language preprocessing algorithm any stop words filter out before starting any startup operations. It's a group of words that appear frequently in a sentence but aren't important or have no bearing on text extraction. Stop words include conjunctions, articles, prepositions, pronouns, and punctuation marks. Common stop words in English, such as ‘of,' ‘a,' and ‘the,' are not used to extract text. Like other languages, the Ge'ez language lacks well-defined stop terms. In this study, we have prepared stop word lists manually from our corpus. Those words such as እስመ, ታሕተ, ማእከለ, መንገለ, አኮኑ, ውእቱ, ውእቶሙ, ውእቶን, ድኅረ, ቅድመ, ዘእንበለ, እስከ, በእንተዝ, አው, ዓዲ, ባሕቱ, ዳዕሙ, አላ, ሶበ and shows in Appendix 1. The NLTK and Keras preprocessors don't recognize Ge'ez stop words. We'll use our corpus to find well-known Ge'ez stop terms and construct a list of them. As a result, during the preparation phase of natural language data processing, those words are filtered out.

```
            Stopword  algorithm
Open the Ge’ez dataset (W)
Access the stop word list
 Stop word list= [‘’]
For stopword found in W
     Remove stop word
     If stop word list found in w
                 Remove stop word list
End for
              End
          End
```

Algorithm 3.3: Stopword algorithm

3.4.3 Stemming

It is an affix removal method that involves removing a small number of prefixes and/or suffixes to get to the root of a word. It is the process of removing prefixes and suffixes from the beginning and ending of words. ‘Stemming is a pain in the neck. In Geez Language, there are several stems. However, in our work only use simplest or common affix i.e., suffixes and prefixes are አ, ን, ዊ/ይ, ያ, ት, ሙ, ው. For example: **አ**ማሰነ ንጉሠ ሀገረ ወአቅተሎ**ሙ** **ለ**ንቡራ**ን** ወአገበረ **በ**ምድር **ዘ**ሞዐ.

Stemming algorithm

```
Open the Ge'ez dataset (W)
Access prefix, suffix
For word found in W
    If word start with prefix
        Remove prefix
    Else
        Word = root word
    If word end with suffix
        Remove suffix
    Else
        Word = root word
End for
    End if
    End if
```

Algorithm 3.4:Stemming algorithm

3.4.4 Padding Sequencing

The last preprocessing task is sequence padding. The need for the padding sequencing task is to equalize the dataset length. In our dataset, the sentences are not all the same length. The maximum length sequence was 88. If the length of the text sequence is less than the maximum length of the sequence, add ‘0' to each sequence until it reaches the maximum length. The Keras library provides a Pad sequence for sentences sequences. Padding can be done before or after the installation. Tag sequences with a maximum length of 88 sequences are being padded means that the entities in the KB are represented in vector space and that these representations are used to predict the missing facts. Because the computer system understands the BIT format data, all dataset sets are changed to vector, i.e., representation of words by numbers, to partition the dataset for training and testing to train the model and predict the test data.

Following the padding sequence of all tags, the word is represented as a vector. The default Keras embedding uses a padding sequence as a word embedding for vector representation of words. For bidirectional long short-term memory networks, the Keras library provides an embedding layer that represents the output of the sequence padding words in a vector to reduce the size input to low dimensional space. The embedding layer in the Keras library takes three arguments: input dimension, output dimension, and input length. The greatest number of sequences in our dataset is approaching 88 but not larger than this, we chose 88 as the input length. The total number of distinct tag sequences is the vocabulary size.

```
    Padding algorithms
Accept Ge’ez text
Count number of token from each sentence
  Find the max and min token num
  Pad the tokens to define the size  n
  Pad the tokens to define the size n
         If token <num
           Add num – token zero pad to t
        Else
          remove token – num tokens from token
         End If
```

Algorithm 3.5:Padding algorithms

3.5 Training and Testing Splitting

It is a model that is used to train and test the functionality of our proposed model. In this paper, we classify the data using assign in the training and testing 0.8 and 0.2 of dataset i.e. the model trained by large dataset for learning, after learn and train the model then test how much present(%) the model learns from the given dataset (Rácz et al., 2021). Therefore, we decided on the model the dataset was properly learned or not. In this case, the model performs both training and predicting(classifying) the entity in a given format. When we predict un training dataset or testing dataset the model predicts correctly.

3.6 Model Architecture

Today, there are some recent and appropriate approaches are available for information extraction. In the field of text processing area information extraction one subtask of NLP extracting information like named entity recognition or named entity extraction, relation extraction, and Event extraction. At the time human beings used the recent and popular machine learning approaches i.e, deep learning. deep learning has wise approaches useful for natural language processing in text processing. Those approaches are ANN, CNN, RNN (LSTM, BiLSTM), BERT, Transformer, etc. are used related to the task. For all deep learning approaches, the only difference is the structures of the neural networks and use cases.

BiLSTM layer: We used the Bidirectional LSTM for the proposed model based on the advancement of using deep learning for NLP tasks. It is a modified variant of the long short-term memory recurrent neural network. This layer maintains the data's sequential order. It enables the detection of linkages between prior inputs and outputs. In both forward and backward directions, the BiLSTM recurrent neural network verifies the sequence of vectors. Because the fault could be at the beginning or end of the sentence, this form of algorithm is preferable for sequence verification.

The bidirectional LSTM model is the combination of two LSTM models that are used to help capture context information from the past as well as the present. The data is processed in two directions using two separate hidden layers. The hidden vectors of both layers are then combined to create the output of the current time step. The Bidirectional LSTM is effectively dealing with sequence information and solves the gradient disappearance problem. This is used to find or obtain large changes in accuracy when compared to machine learning or traditional methods, as the study also supports (Li, 2018).

The study suited the Bi-directional LSTMs model containing input layer, BILSTM layer, dense, dropout, and a softmax activation function for extracted entities from text such as names of persons, places, dates, events in both directions. Deep learning models have made significant progress in natural language processing tasks such as semantic analysis. Using word embeddings, a neural network model can be trained to extract the semantic relationship between two words. After representing the word by vector representation the next path is the BILSTM model. In our study, we used a padding sequence as embedding. The study uses a Keras padding embedding input for a BiLSTM (Bidirectional LSTM) model to form distributed vector representations for the input sentences separately.

Finally, the BILSTM network learns the sequence feature vector in both directions, it learns the sequence of input data to predict the sentence's exact tag. As a result, the embedding layer's output is the input for BILSTM.

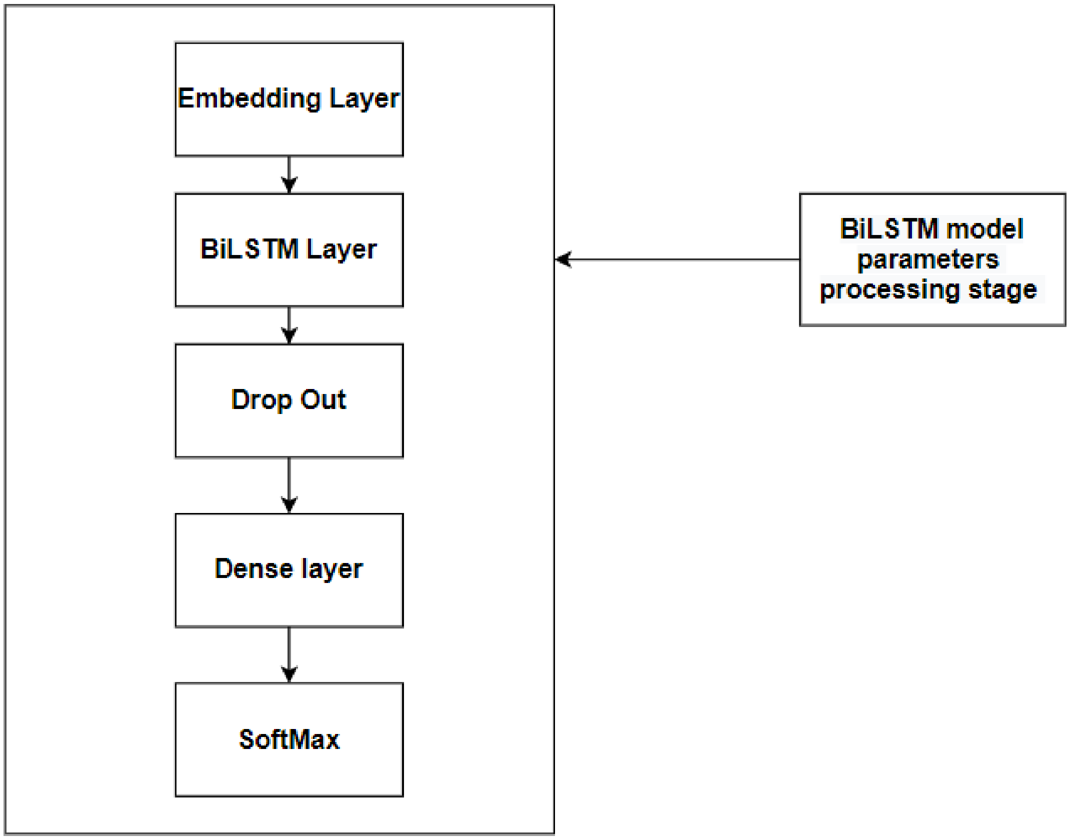


Figure 3.3:Model Architecture

Dropout

It's a regularization method for deep neural networks of all kinds that reduces overfitting and improves generalization error. When the model trains the dataset properly, the test model predicts untrain dataset this says that good generalization has occurred, the training and testing (validation) graph must be at the head or neck of the graphs. It is both computationally inexpensive and extremely effective. For the model, we used a dropout of 0.1 and 0.5 if the generalization gap is poor or good. It is used during the training phase to reduce the overfitting and underfitting of a model. The default dropout layer is employed in our proposed model because the gap between training accuracy and validation accuracy is large and the gap between training loss and validation loss is also large during training. This shows that the graph's generalization gap is inadequate or poor.

Dense

For proposed model used the dense at softmax activation model.add(Dense(9, activation='softmax')) and model.add. By adding bias and various activation functions, the dense layer transforms the input data into output data. As a result, the dense layer receives the output of the BILSTM layer. The dense layer takes 32-dimensional input from the BiLSTM layer and produces 9 dimensions. We have nine(9) different tags i.e. B-per, I-per, B-place, I-place, B-date, I-date, B-eve, I-eve, and O. As a result, the dense layer's job is to reduce the dimensionality of a bidirectional long short-term memory recurrent neural network to a precise number of tags.

SoftMax Activation

The activation function was used as the output vector in our experiment to induce/indicate the probability of distribution. The activation layer is a type of layer. Each convolution layer employs the Rectified Linear Unit (ReLU) function, which allows each negative output to be replaced with a 0 and thus reduces the network's non-linearity. Finally, the SoftMax classifier assigns a probability distribution to each number of tags.

Optimization

In deep learning algorithms, optimization is used to update model parameters (weights and bias values) across iterations. On our model, the **Adam** optimizer is at work. It comes from adaptive moment estimation, which was used to train the model. Adam is an SGD extension method that optimizes networks using an adaptive learning rate that converges quickly and outperforms SGD.

Loss

It is used to compute the quantity that a model should seek to minimize during training. We used a categorical cross-entropy loss function in multi-class classification tasks. To use SoftMax, the model must choose one based on the recommended activation function.

3.7 Prediction phase

This study aims to develop an information extraction model capable of identifying the specified tags to extract text. The prediction phase is used to test this competence. We predicted the entity tag type for the new entity tags in this phase using the learned model. When we add new data from the labeled dataset (train-X) and predict the class (test-Y). In this situation, the model does not know the test dataset (20% of the test dataset is unknown from the train and test split, thus predicts this unknown class using the known train dataset) (80%). Because the model has a softmax activation layer, we were able to easily learn the sequence of text in the built model and categorize the train data into predetermined categories.

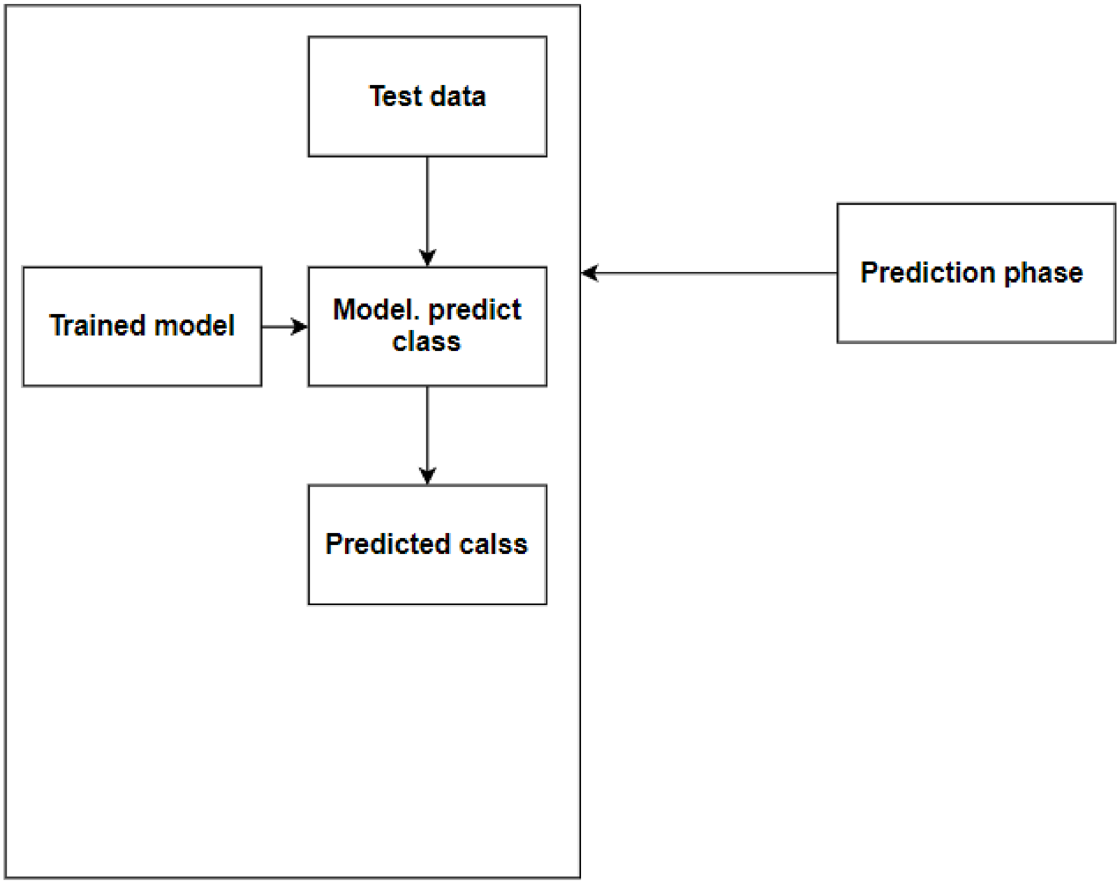


Figure 3.4 Prediction phase

3.8 Evaluation metrics

We calculated the accuracy of the information extraction model from Ge'ez text using the successfully categorized extraction entity tags out of the total training data. For finding the model's loss (training, validation, and testing accuracy), we employed Cross-Entropy Loss accuracy metric rather than confusion matric or precision, recall, and F-score. Within a mulitclass, this loss function employs the Softmax function. It can be used to train a classification model using 'T' tags.

$$\frac{\text{Training}}{\text{Testing/Validation accuracy}} = \frac{\text{Number of correctly categorized}}{\text{Number of correctly categorized } + \text{Number of incorrectly categorized}}$$

3.9 Implementation Tool

The proposed system is designed using different supportive libraries on PyCharm, which is a python editing tool. Among those supportive libraries Kera's, OpenCV, NumPy**,** NLTK, Pandas, and TensorFlow are the most import libraries used. Almost all techniques used for preprocessing of the collected dataset text to remove noise, text correction, text reading, and other relevant tasks are done using functions which are provided by OpenCV. We have used Keras API for developing the neural network architecture and training. Also, Kera's has a graphical representation of the processes of training the neural network. The specifications of the programming tools, supportive libraries, and execution environment are discussed below. we used different tools and programming languages for experimental work from preprocessing tasks to result all over the research work.

Programming language: Python 3.7.11 was used for developing the proposed system because it provides an excellent environment for performing basic text processing and feature extraction(Black, 2017).

Libraries: Some of the relevant and commonly used libraries used during the implementation of the proposed system are listed below.

OpenCV: For preprocessing and other relevant tasks.

Natural Language Processing Toolkit (NLTK) is a library for text processing and built deep learning-based classification algorithm(Seabold & Perktold, 2010).

Kera's: It is a package that is used to learn to preprocess your data, model, evaluate and optimize neural networks. For implanting the BILSTM architecture in a much easier way.

Pandas: Pandas is a valuable data analysis library. It can manipulate and analyze data. Pandas have data structures that are both efficient and simple to use, as well as the ability to easily perform operations on them.

TensorFlow: For normalizing text vectors and as a supporting library for Keras.

NumPy: It is a Python-based general-purpose array processing package. It includes a high-performance multidimensional array object as well as manipulation tools. It is the most important Python package for scientific computing. An N-dimensional array object with a lot of control sophisticated broadcasting roles. For converting the text into integers and storing them as a demission vector.

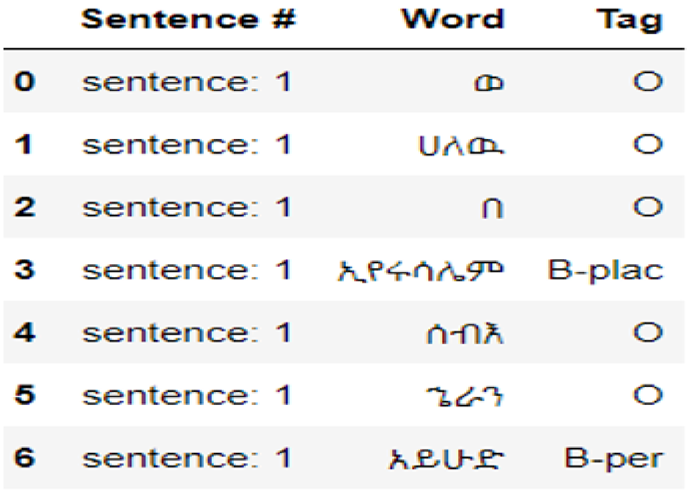
CHAPTER FOUR: EXPERIMENTAL DESIGN AND RESULTS

4.1. Introduction

In this chapter, we discussed the datasets, experimental setup, performance evaluation or evaluation processes, prediction result, and result of the discussion. Around 5270 Ge'ez sentence tags such as a person, place, date, The information extraction model from Ge'ez text was developed using, and event. We used the BiLSTM and LSTM techniques to train the proposed model. We have also used different experimental setups concerning different hyperparameters such as maximum length, dropout, input dimension, output dimension, dense layer, number of the epoch, embedding layer, batch size, activation function, and Adam optimizer. These hyperparameters utilizing both LSTM and BILSTM recurrent neural network models. We have experimented using a single dataset distribution or splitting with the size of the training dataset and the size of the testing datasets being 80% and 20%, respectively.

4.2. Dataset description

The dataset was collected for an experiment and contains a large text file from various sources written in the Ge'ez language, including Ge'ez Drisan, Ge'ez arts, Ge'ez Gedl, and various Ge'ez textbooks. The 63262 tokens in the 5270 sentences were collected, saved, and manually labeled by entity class. We used operations like data cleaning, sentence and word tokenization, stopword removal, affix removal or stemming, and padding sequencing on this dataset. The sample list sentences categorized manually into four entity classes used to create an information extraction model from Ge'ez language are shown in Figure 4.1. The maximum length of words in the Ge'ez dataset is 39.



	Sentence #	Word	Tag
0	sentence: 1	ወ	O
1	sentence: 1	ሀለዉ	O
2	sentence: 1	በ	O
3	sentence: 1	ኢየሩሳሌም	B-plac
4	sentence: 1	ሰብእ	O
5	sentence: 1	ኄራን	O
6	sentence: 1	አይሁድ	B-per

Figure 4.1:Sample of Information Extraction from Ge'ez Text Annotation

4.3 Experiment Setups

For training our model we have used, ThinkPad Lenovo pc with processor: intel(R) Core(TM)i3-6100U CPU @ 2.30GHz 2.30 GHz installed memory (RAM): 4.00GB, and system type 64-bit Operating System, x64-based processor. In the experimental setups, we were performing the overall experimental works from importing the python library to the end of experimental work. Loading of the dataset, assign the dataset for tarin and testing, performance evaluation and testing, test the model and discuss the result. Loading the dataset means importing different packages to load the labeled dataset as save as a "CSV" file because the CSV file is used to easily access quickly by python.

4.3.1 Train test splitting

For this study, the dataset is split into training data and testing data i.e. 80% for Training and 20% used for testing after all the preprocessing stage is completed(Gao et al., 2020). When creating the model train the data and test the data to analyze or know the accuracy performance of the model. When we got less accuracy add more datasets to increase the accuracy to check whether the dataset is accurate or not. The accuracy and its testing accuracy are optimal, the dataset is enough. Splitting the dataset as training and testing for train the model and test the training dataset for further prediction and knowing how much the model train and how much knowledge the training dataset. When the model properly learns the training dataset next test by unknown dataset to predict the true value of the learned dataset. Unless the model does not properly train the dataset, the prediction of the unknown dataset is a false value. At this time the model training accuracy and model loss far apart position i.e., the graph of the two accuracies is training accuracy and validation accuracy displayed at the head and the leg ways this leads to poor Generalization. The following table describes the dataset splitting ratio.

Table 4.1: Train Test Split

Split ratio	Training data	Testing data	validation data	Total data
0.8/0.2	4226	527	527	5270

Table 4.2:Hyperparameter Setup

Hyperparameter	Setup
Max_length	50
Embedding dim	200
Dropout	0.1/0.5
Dense	9
Batch size	32
Epoch	10
Optimize	Adam
Activation	Softmax

4.4 Performance Evaluation

It is critical to evaluate the proposed model's performance to determine whether the method is optimal or not. To evaluate the model, we used the two main experimental setups i.e. LSTM and BiLSTM approaches with the same model parameters such as the number of embedding layers (i.e., the number of input), batch size (i.e., how much dataset to learn a model once a time), epoch (i.e., how many times to iterate the training dataset) learning rate, drop out to minimize overfitting and underfitting(0.1), dense(9), Adam optimizer, and softmax. We used training accuracy, validation accuracy, and testing accuracy to assess the system's performance result. We examined the system using several hyperparameters such as epoch, batch size, learning rate, optimizer, and dropout with an 80/20 data set splitting ratio. Both the long short-term memory(LSTM) and the bidirectional long short-term memory(BiLSTM) networks were evaluated for those parameters. A long-short-term memory network checks the series of tags in a forward direction, whereas bidirectional long short-term memory checks the sequence of tags in both forward and backward directions.

Table 4.3:Split ration vs Hyperparameters setup result

Dataset setups	Setup1(Accuracy In %) for (LSTM and BiLSTM)		
	Training accuracy	Validation accuracy	Testing accuracy
Spit ratio 0.2	96.89/98.59	96.89/97.96	95.78/96.21

4.4.1 Performance of the model with only long short-term memory (LSTM)

In the first experimental process, we have been testing the LSTM methods it only analyzes sequences in the forwarding direction. We also evaluated the performance using an 80% training, 20% testing ratio and trained with 10 epochs. As a result, the model's performance is detailed below. Figure 4.2 depicts the performance of a long short-term memory network with the training set occupying 80% of the dataset and the test set occupying 20% of the dataset. As a result, the model has a training accuracy of 96.89%, a validation accuracy of 96.89 %, and 95.78 testing accuracy respectively.

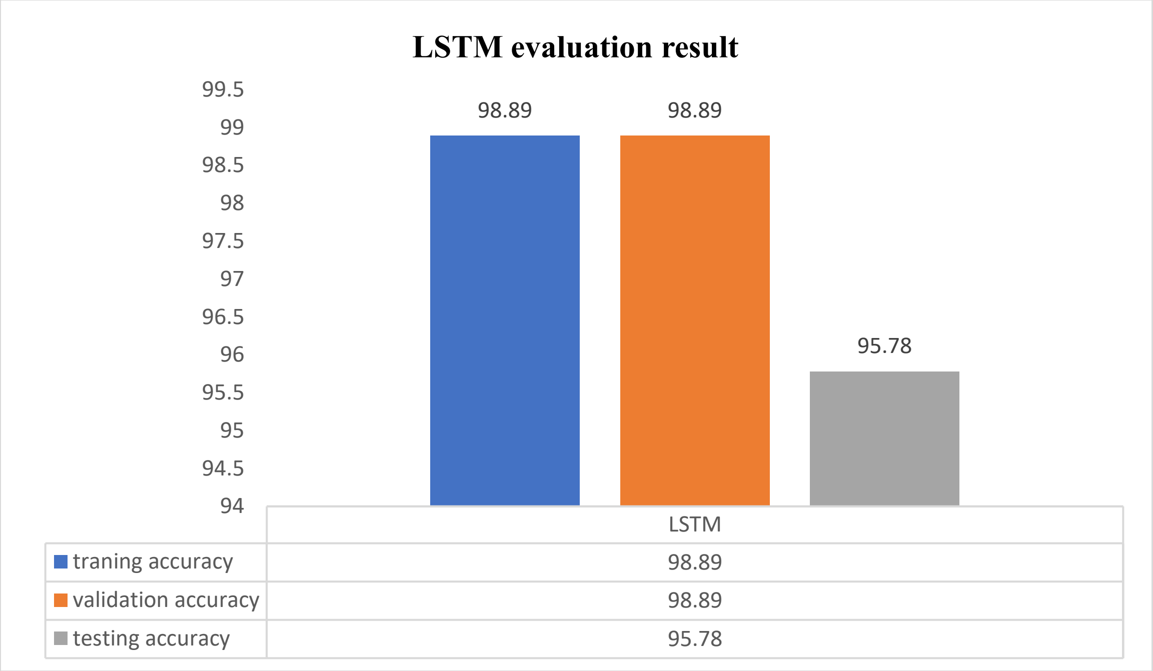


Figure 4.2:Experment1: LSTM evaluation result 80% $ 20% data splitting

In figure 4.2 demonstrates an overfitting problem during the training phase, resulting in the same training and validation accuracy, indicating that the LSTM model is only learning one way at the start or end. As a result, it is unaware that a defect has occurred. In this situation, we'll use the stopped function to begin iteration until the epoch is completed.

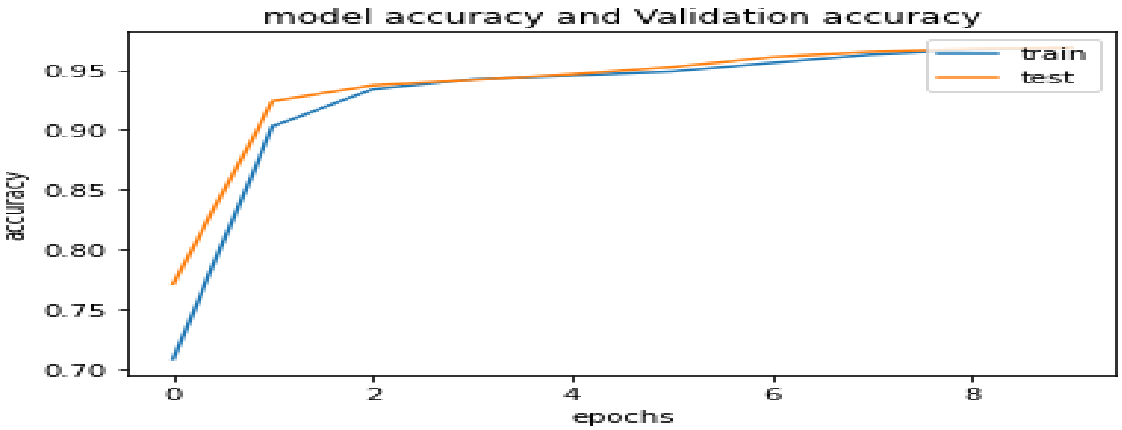


Figure 4.3: LSTM Model Accuracy and Validation Accuracy

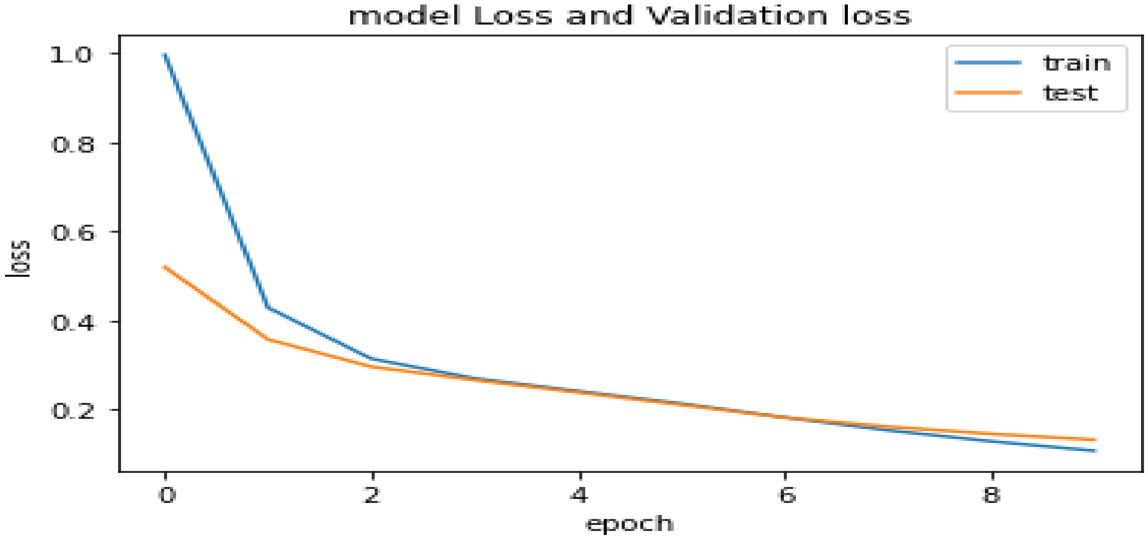


Figure 4.4: LSTM Model Loss and Validation Loss

4.4.2 Performance of the model with only BiLSTM

We have been evaluated the BILSTM methods, we utilized a bidirectional long short-term memory network: we're using 80% of the dataset for training and 20%for validation and testing with some hyperparameters. The training accuracy is around 98.59%, validation accuracy is around 97.96%, and 96.21% testing accuracy respectively.

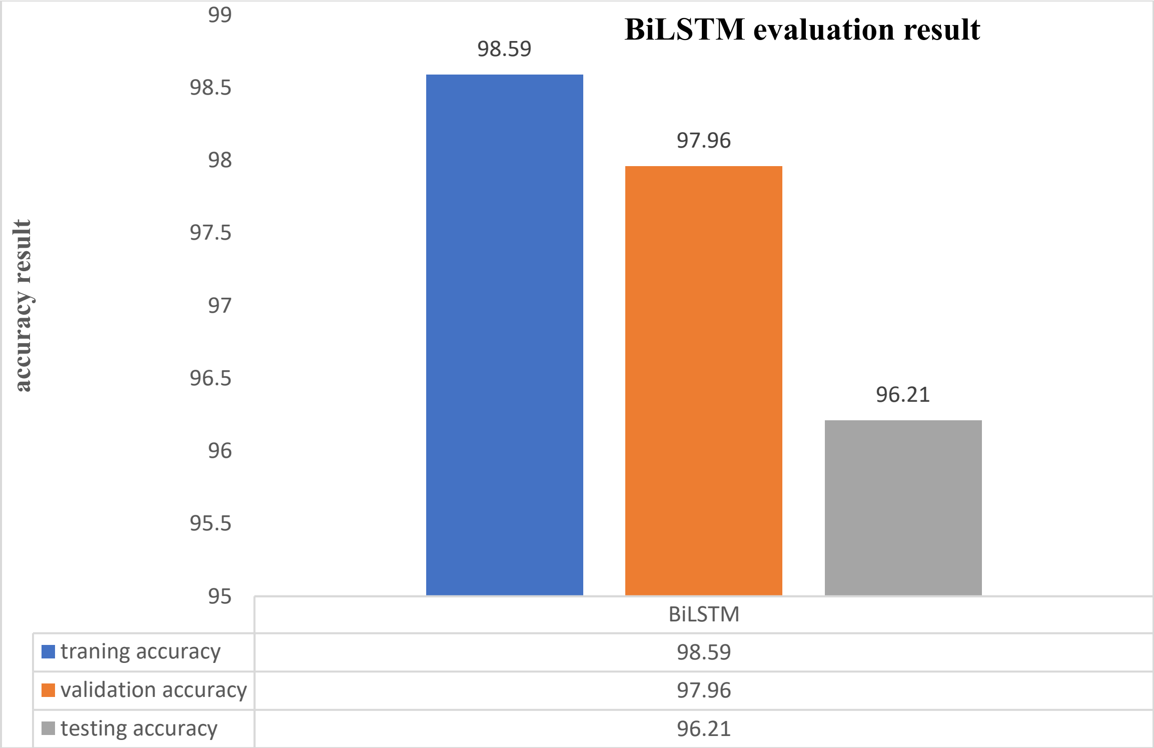


Figure 4.5:BiLSTM evaluation resulyt 80% &20% data splitting

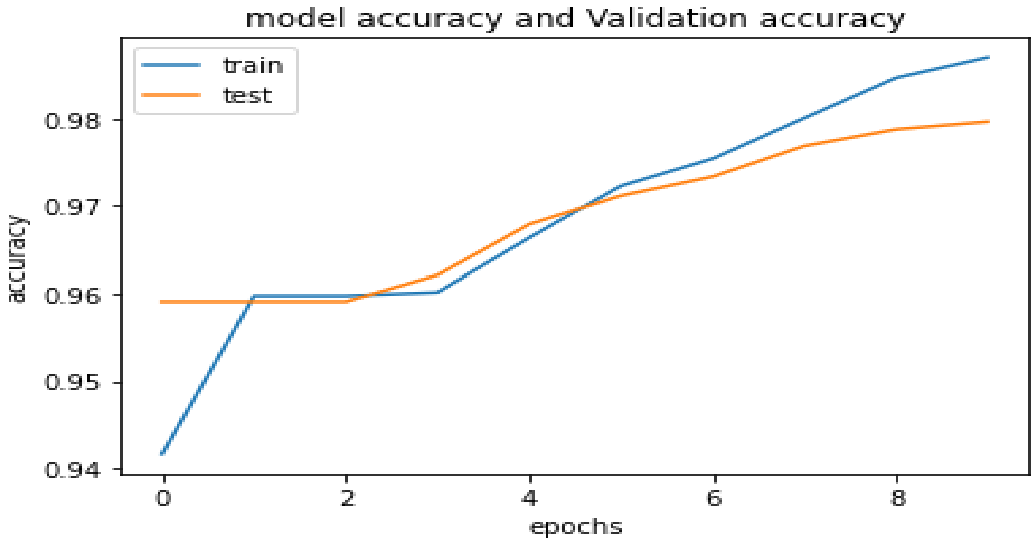


Figure 4.6: BLSTM Model Accuracy and Validation Accuracy

In figure 4.6 indicates a problem with overfitting during the training phase. As a result, the training and validation accuracy graphs are disjointed. We'll use the halted or stopping function at epoch 2 to start iteration until the epoch is finished in this case.

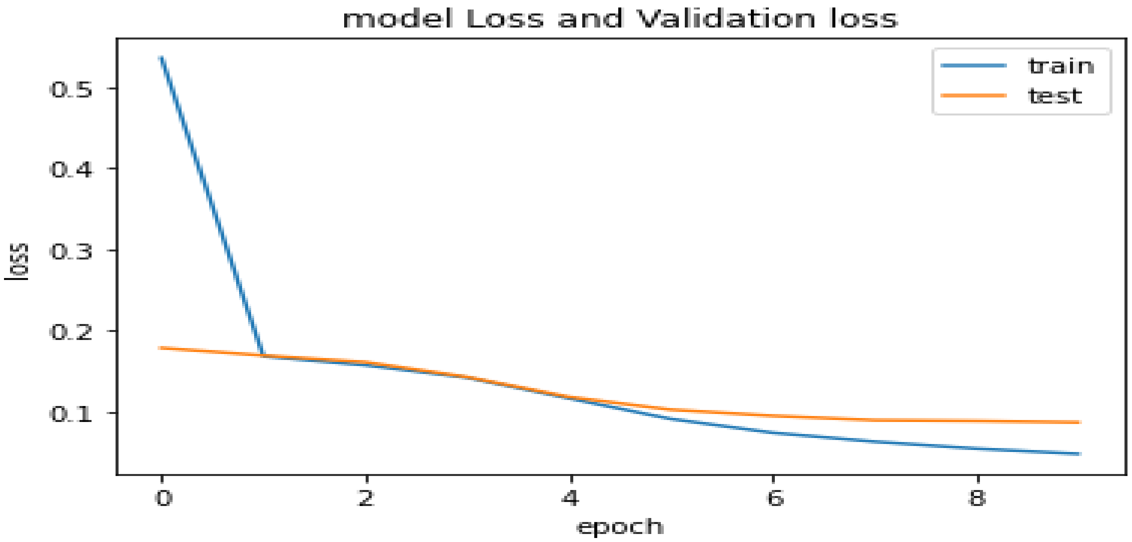
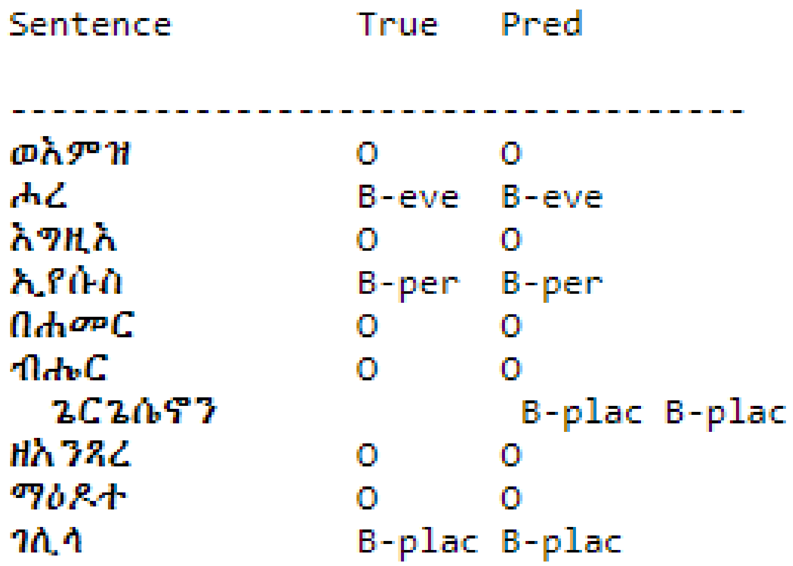


Figure 4.7: BILSTM model loss and validation

Figure 4.7 shows the performance of bidirectional long short-term memory (BILSTM) with 80% of the data set allotted for training and 20% of the data set allocated for testing. So, the training loss and validation loss of the model are 0.51 and 0.86 respectively.

4.5 Prediction Result

When the model has been sufficiently evolved and fits forecast a new row, the prediction of classes in a multi-label classification task must be enabled. The trained model predicts the information extraction texts using new or test sentences. The loaded model will show the class object. The graphs below illustrate a sample result of the suggested model's prediction.



```
Sentence          True   Pred

-------------------------------------
ወእምዝ              O      O
ሖረ                 B-eve  B-eve
እግዚእ              O      O
ኢየሱስ              B-per  B-per
በሐመር              O      O
ብሔር               O      O
  ጌርጌሴኖን                  B-plac B-plac
ዘአንጻረ             O      O
ማዕዶተ              O      O
ገሊላ               B-plac B-plac
```

Figure 4.8: Prediction Result

4.6 Result and Discussion

In this study, we developed an experimental research methodology that allows us to take some hyperparameters as well as dataset distribution. That is to say, every value has been determined through trial and error with various values. We employed several hyperparameter combinations as well as dataset distributions, as mentioned in the experimental setup section (4.3). We used an LSTM and BiLSTM deep learning approach with softmax feature extraction to create an information extraction model from Ge'ez text.

The proposed model evaluated for B-per(person founds at the beginning), B-place(place found at the beginning), B-date(date find at the beginning), B-eve(event find at the beginning) are paced at the beginning of the sentence, I-per, I-place, I-date, I-eve, and O class assign at the inside of the sentence. We evaluated the model by partitioning the data set in an 80% and 20% ratio for the training and testing dataset. We used the LSTM and Bi-LSTM models to train the model. The following hyperparameters are used to train and test the model: epoch, batch size, dropout, and Adam optimizer with default learning rate. As the experiment demonstrates, extended short-term memory and bidirectional memory performance are nearly identical. Bidirectional long short-term memory, on the other hand, is slightly superior to long short-term memory as shown figure 4.9 With an 80/20 splitting ratio, BiLSTM achieved 98.59% of training accuracy and 97.96% of validation accuracy and 96.21 testing accuracy.

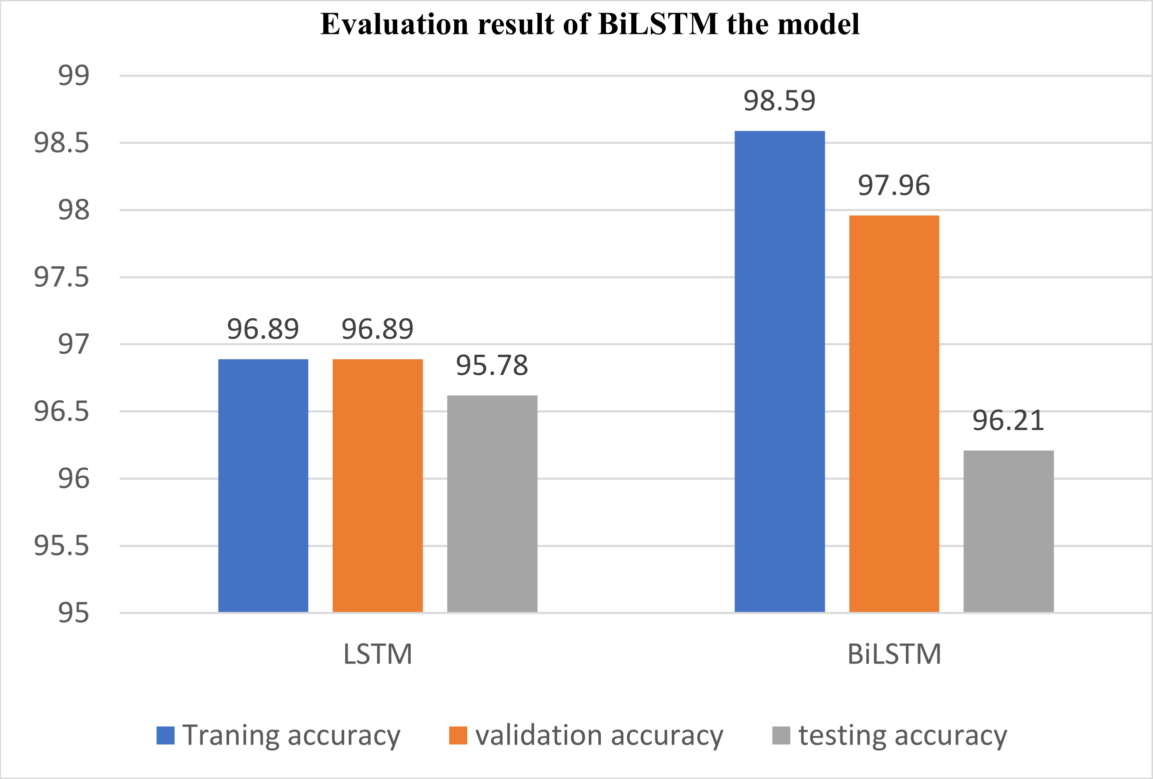


Figure 4.9: Comparision Accuracy Chart

The epoch result graph has a good generalization gap. However, with BILSTM, there is some overfitting that occurred during the training phase. The performance of the LSTM and BILSTM recurrent neural network models differs slightly, as illustrated in figure 4.9. The BILSTM(bi-directional long short-term memory) outperforms LSTM(long short-term memory). Because BILSTM evaluates the sequence of tags in both directions i.e. forward and backward direction. The LSTM (long short-term memory) model only examines sequences in the forward direction.

When we compared the proposed information extraction model from other existing information extraction models, the proposed model outperformed the existing one. To get the experimental result of information extraction from the Ge’ez text model we trained and tested the dataset. The dataset contains 5270 Ge’ez sentences. When trained and tested on the original dataset score 96.89% training, 96.89% validation & 95.78% testing, 98.59% training, 97.96% validation, & 96.21% testing accuracy of LSTM and BiLSTM model respectively. Many researchers have utilized a large dataset and a small dataset to train and test information extraction models designed for various languages and domains. The study of (Abera, 2018) used a machine learning approach that employed a data set of 155 news pieces with a total of 3169 tokens to get an AOTIE score of 79.5 percent precision, 80.5 percent recall, and 80% F-measure. As a result, the size of the data set used for training and testing information extraction from the Ge'ez text model, the results achieved are encouraging.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1 Conclusion

Nowadays, a vast amount of information is available on the internet. However, the availability of this huge amount of information makes it difficult to manually search and acquire the required information for the predefined user. For this problem, Various researchers study information extraction with various languages and domains not include the Ge’ez language. The information extraction system is language and domain-dependent, with the primary goal of extracting important facts from the vast amount of text data available. To achieve the objective of our study or to answer the research question, we developed a BiLSTM (bidirectional long short-term memory) to extract information from Ge’ez text. The scope of our research extracts the named entity text(name of person, place or location, and events that happens) and the number entity(the date or time) from Ge’ez text rather than scanned text (OCR text) or image. The limitation or the gap of this work is doesn't considered the relation between entities, and the image text, audio, and video text for extracted text. In this study, we collected only the Ge’ez text dataset rather than the image, video, audio, or scan dataset. From this dataset, we used a 5270-sentence dataset to train and test our research model. In this study, we annotated the dataset as BIO format (B-per, I-per, B-place, I-place, B-date, I-date, B-eve, I-eve, and O) used to train the model and predict the text. There are some of the primary components of the proposed model such as Preprocessing (data cleaning, tokenization, stopword removal, stemming and padding sequence), train test dataset separation, model building bidirectional long short term memory (BILSTM), and predict class.

To design the proposed model performance evaluation process, the accuracy of the information extraction model derived from the Ge'ez text is tested and the proposed model achieves a BILSTM model 98.59% training, 97.96% validation, & 96.21% testing accuracy. The post of speech tagging is a difficult operation for extracting the relationship between entity attribution for further study work.

5.2 Contribution of the study

We have developed an Information extraction model from Ge’ez text. Some of our contributions to this research work are listed below.

- We've prepared a dataset of Ge'ez sentences that can be utilized for a variety of NLP applications like question answering, text categorization, and textual entailment.
- We prepare Ge’ez named entity from 5270 sentences.
- We apply the recent or deep learning approaches to the information extraction model from Ge’ez text.

5.3 Recommendation

Information Extraction is a subfield of Natural language processing. It is a new research area for the Ge’ez language and other languages. Developing a full-fledged IE requires different NLP processing tools which are post of speech tagging, stemming, wordnet, and grammar checker. In our research work, IE from Ge’ez text is a new research idea it needs some improvements. We recommend the following future works to improve or fill the existing gaps.

- Deep learning models are known to use a large number of datasets to train the model, and we only prepared around 5K sentences. We suggest developing a Ge'ez information extraction system, producing large-scale Ge'ez datasets for above 10k sentences to train the suggested model, and comparing it to our model trained with only roughly 5K phrases.
- We developed an entity extraction task for future work, incorporating the post of speech tagging for extracting relationships between entities, or relation extraction is included for the development of a full-fledged information extraction system.

- We utilized a tautog annotation tool to annotate the named entities from the data, but it's a time-consuming operation. To make it easier, we'll use an automatically named entity recognizer to select the entities' characteristics.
- Incorporating if the Ge’ez spelling checker for solving the problem of manual modification of spelling for annotating entities.
- Finally, we used BiLSTM deep learning approaches, while the researcher can use other deep learning approaches and classification algorithm with a large dataset to increase the model performance accuracy.

REFERENCES

Abebe, B. (2010). Designing a Stemmer for Ge’ez Text Using Rule Based Approach. A Thesis Submitted to School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Science, January, 7–94.

ABEBE, T. (2020). Designing and developing a Speech recognition for Ge’ez Language. Thesis Submitted to School of research and Graduate Studies of Bahir Dar University in partial fulfilment of the requirement for the degree of Master of Science in Computer Science.

Abera Bekele. (2018). Event Modeling from Amharic News Articles Bekele. ADDIS ABABA UNIVERSITY SCHOOL OF GRADUATE STUDIES COLLEGE OF NATURAL SCIENCES DEPARTMENT OF COMPUTER SCIENCE Event Modeling from Amharic News Articles, 151(2), 10–17.

Abera, S. (2018). INFORMATION EXTRACTION MODEL FROM AFAN OROMO NEWS TEXTS Thesis Submitted to School of research and Graduate Studies of Bahir Dar Institute of Technology, Bahir Dar University in partial fulfilment of the requirement for the degree of Master of Science in Computer Science, May, 1-57.

Abrahamsson, F. (2018). Designing a Question Answering System in the Domain of Swedish Technical Consulting Using Deep Learning. Degree Project Computer Science and Engineering.

Adnan, K., & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. In Journal of Big Data (Vol. 6, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-019-0254-8

Ado, D. (2021). Revisiting the status of labialised consonants in contemporary Amharic. Oslo Studies in Language, 11(2), 47–58. https://doi.org/10.5617/osla.8487

Ahmadi, M., Minaei, M., Ebrahimi, O., & Nikseresht, M. (2020). Evaluation of WEPP and EPM for improved predictions of soil erosion in mountainous watersheds: A case study of Kangir River basin, Iran. Modeling Earth Systems and Environment, 6, 2303–2315.

Alemu, A. A., Fante, K. A., & Info, A. (2021). Corpus-Based Word Sense Disambiguation for Ge'ez Language. 8(1).

Aziz Sharfuddin, A., Nafis Tihami, M., & Saiful Islam, M. (2018). A Deep Recurrent Neural Network with BiLSTM model for Sentiment Classification. 2018 International Conference on Bangla Speech and Language Processing, ICBSLP 2018, October. https://doi.org/10.1109/ICBSLP.2018.8554396

Bete, E. B. (2013). SCHOOL OF GRADUATE STUDIES Amharic Question Answering for list questions : SCHOOL OF GRADUATE STUDIES Amharic Question Answering for list questions :

Black, E. (2017). A practical introduction to python using meteorological examples.

Chalapathy, R., Borzeshi, E. Z., & Piccardi, M. (2016). Bidirectional LSTM-CRF for Clinical Concept Extraction. 7–12. http://arxiv.org/abs/1611.08373

Chang, C. H., Kayed, M., Girgis, M. R., & Shaalan, K. F. (2006). A survey of Web information extraction systems. IEEE Transactions on Knowledge and Data Engineering, 18(10), 1411–1428. https://doi.org/10.1109/TKDE.2006.152

Chiu, J. P. C., & Nichols, E. (2016). Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 4(March), 357–370. https://doi.org/10.1162/tacl_a_00104

Cui, Z., Ke, R., Pu, Z., & Wang, Y. (2018). Deep Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction. January. http://arxiv.org/abs/1801.02143

Dalianis, H., & Dalianis, H. (2018). Evaluation Metrics and Evaluation. Clinical Text Mining, 1967, 45–53. https://doi.org/10.1007/978-3-319-78503-5_6

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S. M., & Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. Lrec, 2(1), 837–840.

Elfaik, H., & Nfaoui, E. H. (2021). Deep Bidirectional LSTM Network Learning-Based

Sentiment Analysis for Arabic Text. 395–412.

Esteban, S., Tablado, M. R., Peper, F. E., Terrasa, S. A., & Kopitowski, K. S. (2018). Deep Bidirectional Recurrent Neural Networks as End-To-End Models for Smoking Status Extraction from Clinical Notes in Spanish . 0–8.

Gao, Y., Wang, Y., Wang, P., & Gu, L. (2020). Medical named entity extraction from chinese resident admit notes using character and word attention-enhanced neural network. International Journal of Environmental Research and Public Health, 17(5). https://doi.org/10.3390/ijerph17051614.

Gebregiorgis, Y. S. (2016). Geez letter behaviors in the Amharic language By Yonatan Sisay Gebregiorgis 2016.

Grishman, R. (2019). Twenty-five years of information extraction. Natural Language Engineering, 25(6), 677–692. https://doi.org/10.1017/S1351324919000512

Hirpassa, S. (2017). Information Extraction System for Amharic Text. International Journal of Computer Science Trends and Technology (IJCST), 5(2), 5–15.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Janevski, A. (2000). UniversityIE : Information Extraction From University Web Pages.

Jang, B., Kim, M., Harerimana, G., Kang, S. U., & Kim, J. W. (2020). Bi-LSTM model to increase accuracy in text classification: Combining word2vec CNN and attention mechanism. Applied Sciences (Switzerland), 10(17). https://doi.org/10.3390/app10175841

Jauregi Unanue, I., Zare Borzeshi, E., & Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. Journal of Biomedical Informatics, 76, 102–109. https://doi.org/10.1016/j.jbi.2017.11.007

Jayaram, K. (2017). A Review : Information Extraction Techniques From Research Papers. International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Icimia, 56–59.

Jeyakumar, J. V. (2018). Deep Convolutional Bidirectional LSTM Based Transportation Mode Recognition. 1606–1615.

Johar, A. (2020). Information retrieval system for silte language using BM25 weighting. ArXiv Preprint ArXiv:2012.08907.

Kantzola, E. (2020). Extractive Text Summarization of Greek News Articles Based on.

Kassa, T. (2018). Morpheme-Based Bi-directional Ge'ez -Amharic Machine Translation. Unpublished Master Thesis, Addis Ababa University, Addis Ababa.

KEBEDE, M. (2017). Development of Part of Speech Tagger for Ge ' ez Language Addis Ababa University College of Natural Sciences MULATA KEBEDE ABERA Advisor : YAREGAL ASSABIE (PhD) (Issue October).

Kelemework, W. (2013). Automatic Amharic text news classification: Aneural networks approach. Ethiopian Journal of Science and Technology, 6(2), 127–137.

Kumar, S. (2017). A Survey of Deep Learning Methods for Relation Extraction. ArXiv.

Lee, R. (1998). Automatic information extraction from documents: A tool for intelligence and law enforcement analysts. … Symposium on Artificial Intelligence and Link Analysis, 63–67. http://www.aaai.org/Papers/Symposia/Fall/1998/FS-98-01/FS98-01-011.pdf

Li, S. (2018). Multi-Class Text Classification with Scikit-Learn | by Susan Li | Towards Data Science. https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f

Lieu, nguyen thi. (2015). No Title空間像再生型立体映像の 研究動向. Nhk技研, 151(June), 10–17.

Limsopatham, N., & Collier, N. (2016). Bidirectional LSTM for named entity recognition in ttwitter messages. Proceedings of the 2nd Workshop on Noisy User-Generated Text ({WNUT}), 145–152. https://noisy-text.github.io/2016/pdf/WNUT20.pdf%0Ahttps://www.repository.cam.ac.uk/bitstream/handle/1810/261962/Limsopatham_and_Collier-2016-WNUT2016-VoR.pdf?sequence=1&isAllowed=y

Mansouri, S., Foroumadi, A., Ghaneie, T., & Najar, A. G. (2001). Antibacterial activity of the crude extracts and fractionated constituents of Myrtus communis. Pharmaceutical Biology, 39(5), 399–401. https://doi.org/10.1076/phbi.39.5.399.5889

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named Entity Recognition: Fallacies, challenges and opportunities. Computer Standards and Interfaces, 35(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004

McKinney, W., & Team, P. D. (2015). Pandas-Powerful python data analysis toolkit. Pandas—Powerful Python Data Anal Toolkit, 1625.

Milosevic, N., Gregson, C., Hernandez, R., & Nenadic, G. (2019). A framework for information extraction from tables in biomedical literature. International Journal on Document Analysis and Recognition, 22(1), 55–78. https://doi.org/10.1007/s10032-019-00317-0

Moens, M. F. (2006). Information extraction: Algorithms and prospects in a retrieval context. Information Extraction: Algorithms and Prospects in a Retrieval Context, 21, 1–246. https://doi.org/10.1007/978-1-4020-4993-4

Nguyen, T. H. (2018). Deep Learning for Information Extraction. ProQuest Dissertations and Theses.

Okurowski, M. E. (1993). Information extraction overview. 117. https://doi.org/10.3115/1119149.1119164.

Panda, S. P., Behera, V., Pradhan, A., & Mohanty, A. (2019). A Rule-based Information Extraction System. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(9), 2278–3075. https://doi.org/10.35940/ijitee.I8156.078919

Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. Molecules (Basel, Switzerland), 26(4), 1–16. https://doi.org/10.3390/molecules26041111

Risne, V. (2019). Text summarization using transfer learning Extractive and abstractive summarization using.

Ronran, C., Lee, S., & Jang, H. J. (2020). Delayed combination of feature embedding in bidirectional lstm crf for ner. Applied Sciences (Switzerland), 10(21), 1–22. https://doi.org/10.3390/app10217557

Samuel, K., Savas, O., & Manikonda, V. (2018). A framework for relationship extraction from unstructured text via link grammar parsing. Next-Generation Analyst VI, 10653, 106530K.

Sarawagi, S. (2008). Information extraction. Now Publishers Inc.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference, Scipy, 92–96. https://doi.org/10.25080/majora-92bf1922-011

Shaalan, K. (2014). A Survey of Arabic Named Entity Recognition and Classification. Computational Linguistics, 40(2), 469–510. https://doi.org/10.1162/COLI_a_00178

Shabbir Moiyadi, H., Desai, H., Pawar, D., Agrawal, G., MPatil, N., & Gandhi, R. (2016). NLP Based Text Summarization Using Semantic Analysis. International Journal of Advanced Engineering, Management and Science (IJAEMS), 2(10), 1812–1818. www.ijaers.com

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena, 404(March), 1–43. https://doi.org/10.1016/j.physd.2019.132306

Singh, S. (2018). Natural language processing for information extraction. ArXiv, 1–24.

Strassel, S. M., Przybocki, M. A., Peterson, K., Song, Z., & Maeda, K. (2008). Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction. LREC.

Tadesse, E., Aga, R. T., & Qaqqabaa, K. (2020). Event extraction from unstructured Amharic text. LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, May, 2103–2109.

Tesfasilassie, L., Yigzaw, P. A., & Andualem, M. (2017). The Focus of Teaching Ge ʿ ez Noun and Pronoun Words in Google Newspaper. 1(3), 1–29.

Valero, T., Gómez, M., & Pineda, V. (2009). Using Machine Learning for Extracting

Information from Natural Disaster News Reports. Computación y Sistemas, 13(1), 33–44. https://doi.org/10.13053/cys-13-1-1220

Weninger, S. (2010). Sounds of Gəʿəz--How to Study the Phonetics and Phonology of an Ancient Language. Aethiopica, 13, 75–88.

Worku, B. (2015). Information Extraction from Amharic language Text : Knowledge-poor Approach. Addis Ababa University, Un Published Msc Thesis, Addis Ababa.

Yacob, D. (2000). Considering Ethiopic Character Classes Introduction : Discretizing a Cloud.

Yang, J., Liu, Y., Qian, M., Guan, C., & Yuan, X. (2019). Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. Applied Sciences (Switzerland), 9(18). https://doi.org/10.3390/app9183658

Yuan, Y. (2020). Improving Information Retrieval by Semantic Embedding.

Zhang, S. X., Liu, J., Jahanshahi, A. A., Nawaser, K., Yousefi, A., Li, J., & Sun, S. (2020). At the height of the storm: Healthcare staff's health conditions and job satisfaction and their associated predictors during the epidemic peak of COVID-19. Brain, Behavior, and Immunity, 87, 144–146.

Zhou, P., & Qi, Z. (2016). Text Classification Improved by Integrating Bidirectional LSTM with Two-dimensional Max Pooling. 2(1), 3485–3495.

Zirikly, A., & Diab, M. (2015). Named Entity Recognition for Arabic Social Media. 176–185. https://doi.org/10.3115/v1/w15-1524.

APPENDICES

APPENDIX 1: List of Ge’ez Stop Words

እስመ	ውእቱ	ከመ	ዘአሁ	ዚአሆን
ታሕተ	ውእቶሙ	አምሳለ	ዘዚአሁ	ዛዚአሆን
እለ	ውእቶን	በዘ	አንቲአሁ	አንቲአሆን
ማእከለ	ድኅረ,	በእንተ	እሊአሁ	እሊአሆን
መንገለ	ቅድመ	ህየንተ	ዚአሃ	ዘዚአሆሙ
አኮኑ	ዘእንበለ	በይነ	ዘዚአሃ	እንቲአሆሙ
አላ	እስከ,	አመ	አንቲአሃ	እሊአሆሙ
ሶበ	በእንተዝ	አልቦቱ	እሊአሃ	አልቦን
ዓዲ	አው	አልቦሙ	ዘአሆሙ	አልባቲ
ዲበ	ባሕቱ	ውስተ	መልዕልተ	መትሕተ
ላእለ	ዳዕሙ	ውሳጤ	ታሕተ	

APPENDIX2: List of Ge’ez Affix

እም	ክን	ዋት	ነ	ዘ
እምነ	ክሙ	ያን	ት	እንተ
ምስለ	ኪ	ን	ው	እለ
በ/ለ	ኩ	አ		