# Information extraction model from Ge’ez texts

**Seffi Gebeyehu[1], Worke Wolde[2], Zelalem S. Shibeshi[3]**
[1]Department of Software Engineering, Faculty of Computing, Bahirdar Institute of Technology, Bahirdar University, Bahir Dar, Ethiopia
[2]Department of Information Technology, College of Engineering and Technology, Samara University, Afar, Ethiopia
[3]Department of Computer Science, Rhodes University, Grahamstown, South Africa

**Article Info**

***Article history:***

Received May 24, 2022
Revised Dec 20, 2022
Accepted Dec 24, 2022

***Keywords:***

Bi-LSTM
Deep learning
Entity extraction
Ge’ez text
Information extraction

## ABSTRACT

Nowadays, voluminous and unstructured textual data is found on the Internet that could provide varied valuable information for different institutions such as health care, business-related, training, religion, culture, and history, among others. A such alarming growth of unstructured data fosters the need for various methods and techniques to extract valuable information from unstructured data. However, exploring helpful information to satisfy the needs of the stakeholders becomes a problem due to information overload via the internet. This paper, therefore, presents an effective model for extracting named entities from Ge'ez text using deep learning algorithms. A data set with a total of 5,270 sentences were used for training and testing purposes. Two experimental setups, i.e., long short-term memory (LSTM) and bidirectional long short-term memory (Bi-LSTM) were used to make an empirical evaluation with training and a testing split ratio of 80% to 20%, respectively. Experimental results showed that the proposed model could be a practical solution for building information extraction (IE) systems using Bi-LSTM, reaching a training, validation, and testing accuracy as high as 98.59%, 97.96%, and 96.21%, respectively. The performance evaluation results reflect a promising performance of the model compared with resource-rich languages such as English.

***Corresponding Author:***

Seffi Gebeyehu
Department of Software Engineering, Faculty of Computing, Bahirdar Institute of Technology
Bahirdar University
Bahir Dar, Ethiopia
Email: gseffi2010@gmail.com

## 1. INTRODUCTION

The amount of unstructured textual content on the Internet is rapidly increasing [1]–[4]. This unstructured data is found in the text, images, video, and audio. An essential part of such information, including government documents, court cases, online and news reports, and social media conversation, is broadcasted in the unstructured form [5]. Voluminous and unstructured textual data is found on the internet that could provide varied valuable information for different institutions such as health care, business-related, training, religion, culture, and history, among others [6], [7].

Finding detailed information from vast unstructured text data is still a complex task. It's also difficult to manually search, filter, extract and pick the type of data that could be used for different purposes [8]–[10]. Natural language processing (NLP) techniques could be applied to address the abovementioned problems. NLP, one of the artificial intelligence domains, helps computers to communicate with human beings to read text, hear speech, evaluate sentiment and identify the critical part of communication [11]. A significant difficulty in the early years was the lack of tools to extract and search for useful information to address and fulfill the stakeholder’s demands. To solve the current challenge of information extraction from vast textual data, various studies have been done in the fields of information extraction (IE), information retrieval (IR), question answering, text summarization, and text categorization [12].

Information extraction is the automated extraction of specific information from unstructured sources, such as documents, databases, and attributes defining entities [13]. To build a full-fledged information extraction model, several tasks are required, including named entity recognition (NER), 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 [14]. In this paper, the named entity recognition task is employed to determine the fundamental entity tags, which include persons, places, times, and events. The Ethiopian orthodox tewahedo church (EOTC) in Addis Ababa, Ethiopia, is the data source for acquiring Amharic church documents, such as Drsan, Yehawariyat Sra, and Gedile Aba. Ethiopia is an East African country with a writing system called fidel [15].

The Ethiopian orthodox tewahedo church is the first in Zema (Gloss) teaching and in inventing alphabet for writing, with its reading style. Ge’ez bible, Hamer, and other religious books are available online in the EOTC religions repository. However, those resources are presented 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. As a result, if a user is extracting a Ge’ez text, they should be familiar with the linguistic character (Ge’ez writing system). In Arabic [16], English [17], Latin [18], Chinese [19], and other languages, a lot has been done related to information extraction. However, no research has been conducted to explore constructing an information extraction model for Ge’ez free text. The following paragraphs explain related works.

## 2. RELATED WORK

In his study Hirpassa [9] developed an IE model for Amharic text. The model was developed using the general architecture for text engineering (GATE) text processing technique using a knowledge-poor approach to a specific domain of infrastructure. By knowledge-poor approach, we use simple rules and a gazetteer list for entities used in identity identification. The study used 24,760 instances for training and testing the model. The evaluation was done on the name entity recognition component separately, and the system achieved a performance of 89.1 %, which was considered promising.

A recent study introduced a method for information extraction using entity and entity relation for Arabic text collected from Egyptian Arabic newswire [20]. The experiment contained nearly 625,368 entries; the number of sentences was 36,423 and they selected a sample of 3,400 sentences representing the crime news. The study showed that IE depends on annotating the target entities, which supports extracting hidden information from the massive Arabic text that could help decision making.

From recent literature, there has been little research on information extraction systems in Ethiopian and foreign languages [3], [7], [8], [10]. 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 [21]. This is because the IE system has to be trained in the different natures of the language and the domain for which they are developed. As a result, the study is aimed to build IE model from Ge'ez text by applying a deep learning technique.

Valero *et al*. [12] used machine learning approaches for extracting valuable information from natural disaster news magazines. As a branch of machine learning, deep learning is used to model high-level abstractions in data extraction [22]. In deep learning, training a model needs a large number of labeled datasets and contains multiple layers within a neural network architecture. In recent times, researchers used known deep learning approaches such as long short-term memory (LSTM), bidirectional long short-term memory (Bi-LSTM) [23], convolutional neural network (CNN), and recurrent neural networks (RNN) to solve the IE tasksThe known techniques LSTM and bi-directional LSTM were used to set the experiment and choose the optimal, since LSTM is a gated recurrent neural network, and Bi-LSTM is just an extension to that model for future cell processing [24].

Since IE is language-specific [25], the IE model developed for Amharic text cannot be applied to Ge’ezlanguage even though they are in the same domain area. For instance, IE model built for a document in the terrorism domain doesn’t work for housing advertisements. Thus, this study is aimed to develop a suitable IE model for Ge’ez text, and, finally, evaluate the performance and usability of the model using the evaluation metrics.

The rest of the study is organized as shown in: the next section presents the materials and methods used to build the IE model including data preprocessing, tokenization, normalization, stop word removal, stemming, and padding. The discussion of the findings of this research is presented in section three. Section 4 presents the conclusion, recommendations, and limitations of the study and the last section presents the contribution of the study.

## 3. RESEARCH METHODS

The proposed model is concerned with named entity recognition in the context of information extraction subtasks. The module for information extraction from Ge'ez texts is described in this section. A deep network with Bi-LSTM layers underpins our method. The padding sequence is the first Bi-LSTM layer's input. On the other hand, a softmax function is used in the final layer to find the most appropriate labels for each token. The data were gleaned from religious documents, Ge’ez textbooks, religious telegram channels, and Ge’ez thesis. The proposed model architecture (see Figure 1) includes the preprocessing stage, the data splitting stage, and the prediction stage.

Preprocessing algorithms (or tasks) such as tokenization (divide sentences, remove punctuation marks and numerals), stop word removal, stemming, and padding are some of the preprocessing procedures that have been used. After preprocessing, we split the dataset into training and testing phases. The training dataset is used to train the proposed model, which includes the LSTM and Bi-LSTM stages of sentence encoding. Finally, the trained model predicts or extracts the given text after the training and testing phases have been completed. We annotated the data set with a structure of the BIO format, where 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.

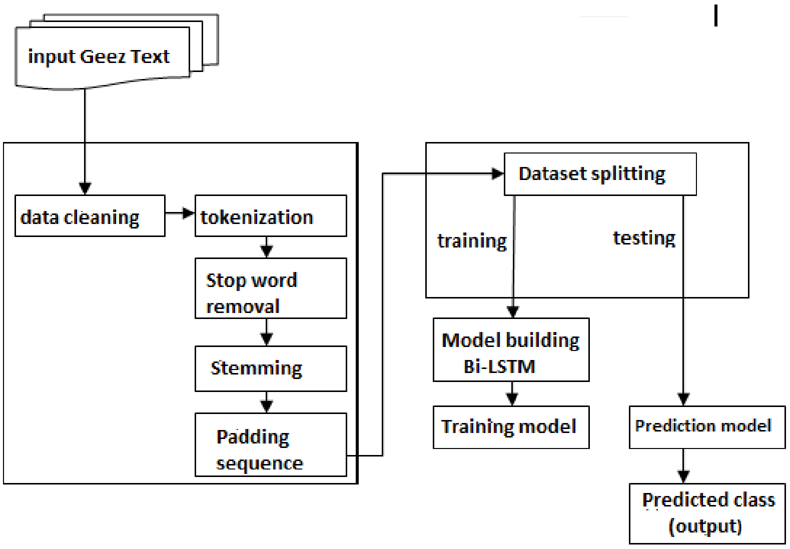

Figure 1. The proposed model architecture

The BIO schemes are commonly used in NER or entity extraction research [26]. This paper is focused on investigating the most fundamental and significant elements of NER, such as the person's name, *“ደማቴዎስ ሊቀጳጳስ ዘለ እስክንድርያ ዘይትነበብ በዕለተ በዓሉ ለዐቢይ ወክቡር ወቅዱስ ሊቀመላእክት ሚካኤል አመ ዐሡሩ ወሰኑዩ ለወርኀ ኅዳር”, ደማቴዎስ* is labelled as “B-person”, *እስክንድርያ* as “B-place”, and *ዐሡሩ ወሰኑዩ ለወርኀ ኅዳር* as “B-Time”, and the rest will be outside entities labeled with "O”. After representing the word using vector representation, the next step is modeling the sentence using the by bidirectional LSTM (Bi-LSTM). In this study, a Keras padding is used as embedding input for a Bi-LSTM model to form distributed vector representations for the input sentences separately. 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 Bi-LSTM recurrent neural network verifies the sequence of vectors. Because of this, this form of algorithm is preferable for sequence verification [27]. The bidirectional LSTM model is the combination of two LSTM models that are used to capture context information from the past as well as the present. Finally, the Bi-LSTM 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 Bi-LSTM (see Figure 2).

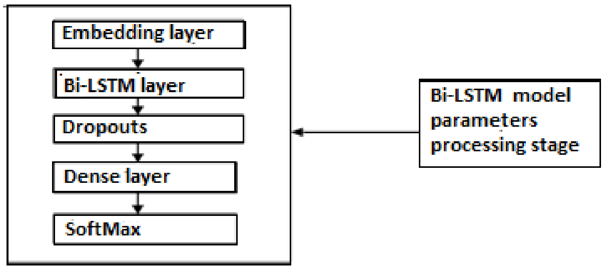

Figure 2. Bi-LSTM model parameters

### 2.1. Data preprocessing

The data preprocessing stage is the data cleaning and preparing the labels dataset for training and testing the model. The data cleaning step removes unnecessary or noisy data like (the name of the author, email address, phone and city, and print date). Using an online annotation tool, the dataset is annotated to assign the class type, and modify the named entity. After preprocessing the raw data, the other steps are: tokenization, stop word removal, stemming, padding, and word2vec (vectorization), which are discussed in next sub-section.

### 2.2. Tokenization

Naturally, a text must split into appropriate linguistic units such as words and sentences before any meaningful text processing can begin. Because the incoming news story 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 NLP in text processing. Tokenization is the splitting of textual data into expected tokens. As a result, the text that was collected for processing has been broken down into paragraphs, sentences, or words using punctuation marks. Ge'ez language has its punctuation marks, where ። is used as full stop, ፣ is used as comma, ፤ is used as semicolon, and ፡ is used as colon [28].

### 2.3. Stopword removal

Stopword removal is another task in preprocessing stage to filter out stopwords before any startup processes in this natural language preprocessing. Conjunctions, articles, prepositions, pronouns, and punctuation marks are examples of stop words because they add no value to the IE system. There is a list of stop words in the Ge'ez language, such as *'እስመ','እለ', 'ታሕተ', 'መትሕተ', 'ውስተ', 'ማእከለ', 'ኀበ', 'መንገለ', 'እም', 'እምነ', 'ምስለ', 'በእንተዝ', 'ባሕቱ', 'ዘእንበለ', 'በእንተዝ'. E.g., 'ወወረደ' 'ፊልጶስ', 'ሀገረ', 'ሰማርያ', 'ወሰበከ', 'ሎሙ', 'በእንተ', 'ክርስቶስ'* [28]. By preparing a stop word list, any word that is in the stop word list is removed from the bag of words identified from the Ge'ez text.

### 2.4. Stemming

The use of stemming is to reduce redundancy. For example, stemming will bring the different forms of the word index, which are indexing, indexable, indexers to index. In its simplest form, stemming can be considered as an affix removal method that involves removing (a small number of–remove) prefixes and/or suffixes that are added to conflate a word from its root (conflation is adding–stemming is removing). It is the process of removing prefixes and suffixes from the beginning and end of words. Stemming is a pain in the neck [28]. In Ge'ez language, there are several stems. However, this study used simplest or common suffixes (such as *ኣን፤ኣት፤ያን፤ያት*) and prefixes (such as *ኣ,ወ) ወብዙሀን ነፍሳተሀጥን እለ ኣውፅኦሙ በእንተ ትንሳኤሁ ቅድስት ወውድስት እንተይእቲ ሰንበተክርስቲያን* [29].

### 2.5. Padding sequencing

Embedding/padding describes a vector space representation for the entities and relations in the knowledge base (KB) and uses these representations to predict the missing facts. Latent feature models do reasoning over the knowledge bases (KBs) via latent features of entities and relations. The intuition behind such models is that the relationship between two entities can be derived from the interactions of their latent features. Recently, the research trend is moving towards using tensors for KB completion tasks. Tensors are multidimensional arrays that represent multi-relational data quite easily. At this stage, all the dataset are changed to vector i.e., representation of word by number to splitting the dataset to training and testing and predict the test data, because the computer system know the bit (or bio) format data.

## 4. RESULTS AND DISCUSSION

In this section, the study findings from each module of the information extraction model are presented. Finally, a discussion of the outcomes of the proposed model is presented. The conventional information extraction evaluation metrics of transcription, validation and testing accuracy are also presented.

### 4.1. Dataset description

The dataset collected for this experiment 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. A total of 63,262 tokens in the 5,270 sentences were collected, saved, and manually labeled by entity class shown in Table 1. We used preprocessing tasks like data cleaning, sentence and word tokenization, stopword removal, affix removal or stemming, and padding sequencing on this dataset. The sample list sentences are categorized manually into four entity classes used to create an information extraction model from Ge'ez language. The study used an 80%/20% training and testing ratio consistent with other research [9]. The hyperparameters are chosen as in Table 1 in such a way that better performance evaluations could be achieved as also supported in other similar research [8].

Table 1. Experimental 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 |

### 4.2. 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 deep learning algorithms, i.e. LSTM and Bi-LSTM with the same model parameters (defined in Table 2).

Table 2. Evaluation result

| Algorithm | Accuracy in % | | |
|---|---|---|---|
| | Training accuracy | Validation accuracy | Testing accuracy |
| LSTM | 96.89 | 96.89/ | 95.78/ |
| Bi-LSTM | 98.59 | 97.96 | 96.21 |

#### 4.2.1. Performance of the model with only long short-term memory

In the first experimental process, we tested the LSTM methods; which only analyze sequences in the forwarding direction. We also evaluated the performance using an 80% training, and 20% testing ratio, and trained with 10 epochs, as shown in Figures 3 and 4. Plotting the scores for the training and testing datasets is a useful indication of whether the model is over-fitting to the training data. The graph in Figure 4 shows how the accuracy of the LSTM approach responds as the number of epochs in the training and testing data isincreases.

The figure shows that LSTM can achieve excellent accuracy with only 8 epochs. The generalization gap of the epoch result graph is good, reflecting no overfitting occurring between the training and testing dataset as shown in Figure 3. The graph of the model accuracy and loss is shown in Figure 4. The graph of the model accuracy and loss demonstrates the non-existence of the overfitting problem during the training phase indicating that the LSTM model is only learning one way at the start or end. Figure 4 also highlights the testing sets exhibit highly similar variations in accuracy and loss score.

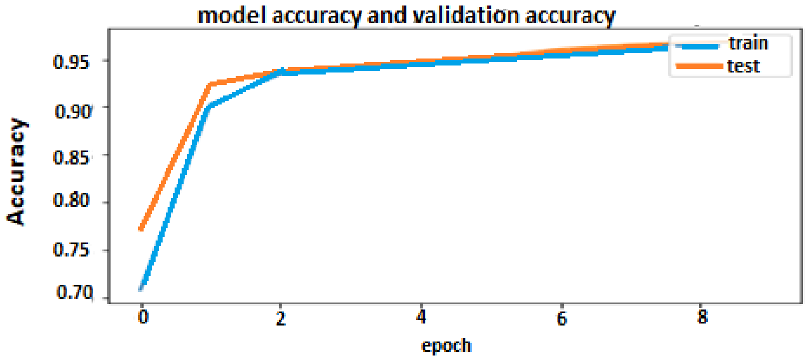

Figure 3. LSTM model accuracy

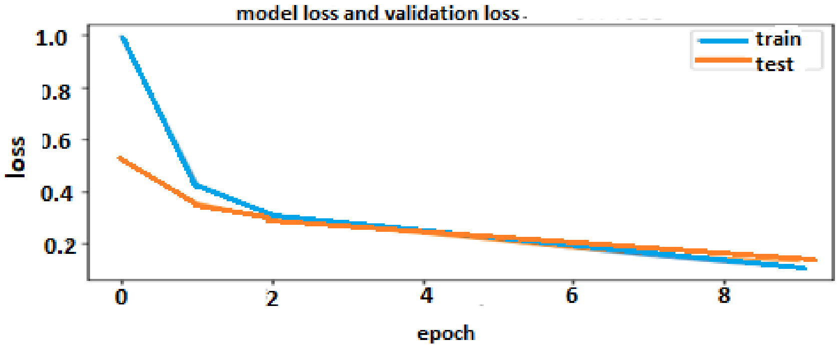

Figure 4. LSTM model loss

### 4.2.2. Performance of the model with only BiLSTM

We have also evaluated the Bi-LSTM methods, which utilize a bidirectional long short-term memory network, using 80% of the dataset for training and 20% for testing with some hyperparameters. The experimental result showed high achievement in training accuracy of 98.59%, validation accuracy of 97.96%, and testing accuracy of 96.21%. From the plot in Figure 5, the accuracy first rises and settles at one epoch.

After training, the test is done and the curve settles reflecting the accuracy increases after the model is trained a little more as the trend for accuracy on both datasets increases for every rise in epochs. The model has not yet overlearned the training dataset, showing comparable skills on both datasets.

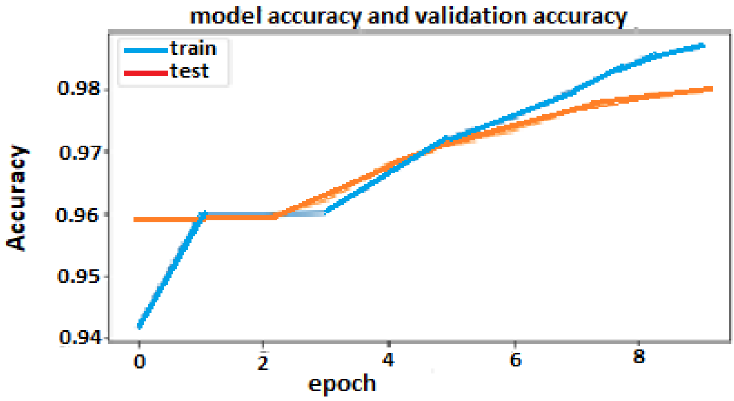

Figure 5. Bi-LSTM model accuracy

From the plot of loss in Figure 6, it can be observed that the training and validation datasets have shown a comparable performance. The loss decreases and settles around one epoch. After training, the test is performed and the curve settles near zero reflecting that the loss decreases after training.

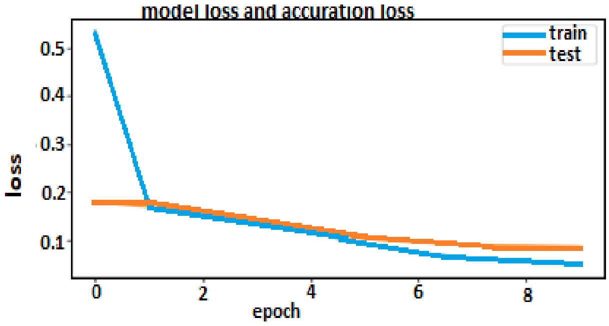

Figure 6. Bi-LSTM model loss

### 4.2.3. 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 shows the class object. Table 3 illustrates a sample result of the suggested model's prediction.

Table 3. Prediction results

| Experimental Result | no. of instances | % |
|---|---|---|
| Correctly Classified Instances | 1007 | 96.21 |
| Incorrectly Classified Instances | 193 | 13.79 |

### 4.3. Discussion of result

In this study, a comparison is done between two models LSTM and BiLSTM. To train both models the same parameters are used. However, the BiLSTM model understands (would be good to use another term) the sentences' semantic features from the beginning and end of the sentence, but the LSTM model only knows the starting features not knowing the end word features.

So, Bi-LSTM is knowing directional features (the similarity of the word within a sentence). Both models have a good generalization gap because the graph lies on the same bands (on the head or the neck), as shown in Figures 5 and 6. The experimental result shows that the model accuracy and model loss performances is 0.968% and 0.13, respectively.

## 5. CONCLUSION

The availability of enormous amounts of textual data over the internet motivates researchers and practitioners to find various ways to automatically index and process text documents. There has also been a growth in the accessibility of text data in local languages such as Amharic and Ge'ez. However, the availability of voluminous information makes it challenging to search and extract valuable information from excessive unstructured data. Therefore, it is required to build an IE model to extract meaningful information and summarize the content. The study aims to build a model using Bi-LSTM to extract information from Ge'ez text. To build the IE model, the study passes through various steps, including preprocessing (data cleaning, tokenization, stop word removal, stemming and padding sequence), training and testing dataset separation, and model building using bidirectional long short-term memory (Bi-LSTM), and predict named entity. The model could help users search and extract specific information (text, sentence, or paragraph) from the available state of voluminous unstructured data. Based on the evaluation result, the proposed Bi-LSTM information extraction model from Ge'ez text achieves 98.59%, 97.96%, & 96.21% accuracy for training, validation, and testing, respectively. The performance evaluation results reflect a promising performance of the model compared with resource-rich languages like English.

## 6. CONTRIBUTION OF THE STUDY

In practice, the study contributes to developing and improving the information extraction modules of the Ge’ez language and, in theory, initiates other researchers to know and develop knowledge about the ancient Ge’ez language for developing other related IR research areas like question answering, text summarization, machine translation, and soon.

## REFERENCES

[1] J. P. C. Chiu and E. Nichols, “Named entity recognition with bidirectional LSTM-CNNs,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 357–370, Nov. 2016, doi: 10.1162/tacl_a_00104.

[2] Y. Gao, Y. Wang, P. Wang, and L. Gu, “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*, vol. 17, no. 5, p. 1614, Mar. 2020, doi: 10.3390/ijerph17051614.

[3] C. Ronran, S. Lee, and H. J. Jang, “Delayed combination of feature embedding in bidirectional lstm crf for ner,” *Applied Sciences (Switzerland)*, vol. 10, no. 21, pp. 1–22, Oct. 2020, doi: 10.3390/app10217557.

[4] J. Yang, Y. Liu, M. Qian, C. Guan, and X. Yuan, “Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding,” *Applied Sciences (Switzerland)*, vol. 9, no. 18, p. 3658, Sep. 2019, doi: 10.3390/app9183658.

[5] N. Milosevic, C. Gregson, R. Hernandez, and G. Nenadic, “A framework for information extraction from tables in biomedical literature,” *International Journal on Document Analysis and Recognition*, vol. 22, no. 1, pp. 55–78, Feb. 2019, doi: 10.1007/s10032-019-00317-0.

[6] K. Jayaram and K. Sangeeta, “A review: Information extraction techniques from research papers,” in *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Feb. 2017, pp. 56–59, doi: 10.1109/ICIMIA.2017.7975532.

[7] S. P. Panda, V. Behera, A. Pradhan, and A. Mohanty, “A rule-based information extraction system,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 9, pp. 1613–1617, Jul. 2019, doi: 10.35940/ijitee.I8156.078919.

[8] S. Abera and T. Tegegne, “Information extraction model for afan oromo news text,” in *Communications in Computer and Information Science*, vol. 1026, 2019, pp. 327–340.

[9] S. Hirpassa, “Information extraction system for amharic text,” *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 5, no. 2, pp. 5–15, 2017.

[10] B. Worku, “Information extraction from amharic language text : Knowledge-poor Approach,” Master's Thesis, Departement Computer Science, Addis Ababa University, 2015.

[11] N. Limsopatham and N. Collier, “Bidirectional LSTM for named entity recognition in ttwitter messages,” in *Proceedings of the 2nd Workshop on Noisy User-generated Text ({WNUT})*, 2016, pp. 145–152.

[12] A. T. Valero, M. Montes y Gómez, and L. V. Pineda, “Using Machine learning for extracting information from natural disaster news reports,” *Computacion y Sistemas*, vol. 13, no. 1, pp. 33–44, 2009, [Online]. Available: http://cys.cic.ipn.mx/ojs/index.php/CyS/article/view/1220.

[13] S. Sarawagi, “Information extraction,” *Foundations and Trends® in Databases*, vol. 1, no. 3, pp. 261–377, 2007, doi: 10.1561/1900000003.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, Mar. 2011.

[15] A. Gebremariam, “Amharic-to-Tigrigna machine translation using hybrid approach,” Master's Thesis, Departement Computer Science, Addis Ababa University, 2017.

[16] H. Elsayed and T. Elghazaly, “Information extraction from Arabic news,” *International Journal of Computer Science Issues (IJCSI)*, vol. 12, no. 1, pp. 114–118, 2015.

[17] J. Zavrel, P. Berck, and W. Lavrijssen, “Information extraction by text classification: Corpus mining for features,” in *Proceedings of the workshop Information Extraction meets Corpus Linguistics*, 2000.

[18] R. Guarasci, “Developing an annotator for Latin texts using Wikipedia,” *Journal of Data Mining and Digital Humanities*, pp. 0–6, 2017, [Online]. Available: https://hal.archives-ouvertes.fr/hal-01279853v2.

[19] Y. Yu, X.-L. Wang, and Y. Guan, “Information extraction for Chinese free text based on pattern match combine with heuristic information,” in *Proceedings. International Conference on Machine Learning and Cybernetics*, 2002, vol. 1, pp. 214–218, doi: 10.1109/ICMLC.2002.1176742.

[20] H. Elfaik and E. H. Nfaoui, “Deep bidirectional LSTM network learning-based sentiment analysis for Arabic text,” *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 395–412, Dec. 2021, doi: 10.1515/jisys-2020-0021.

[21] B. Piper and A. J. van Ginkel, “Reading the script: How the scripts and writing systems of Ethiopian languages relate to letter and word identification,” *Writing Systems Research*, vol. 9, no. 1, pp. 36–59, Jan. 2017, doi: 10.1080/17586801.2016.1220354.

[22] M. Shu, “Deep learning for image classification on very small datasets using transfer learning,” *Creative Components*, pp. 14–21, 2019.

[23] W. Saad, W. A. Shalaby, M. Shokair, F. A. El-Samie, M. Dessouky, and E. Abdellatef, “COVID-19 classification using deep feature concatenation technique,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, pp. 2025–2043, Apr. 2022, doi: 10.1007/s12652-021-02967-7.

[24] U. Hahn and M. Oleynik, “Medical information extraction in the age of deep learning,” *Yearbook of medical informatics*, vol. 29, no. 1, pp. 208–220, Aug. 2020, doi: 10.1055/s-0040-1702001.

[25] J. Ravikumar and R. Kumar, “A framework for named entity recognition of clinical data,” *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 18, no. 2, p. 946, May 2020, doi: 10.11591/ijeecs.v18.i2.pp946-952.

[26] M. Noll, J. Lete, and P. E. Meyer, “Gene entity recognition of full text articles,” in *Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science*, Jun. 2017, vol. Part F1309, pp. 162–167, doi: 10.1145/3121138.3121167.

[27] D. Feng and H. Chen, “A small samples training framework for deep Learning-based automatic information extraction: Case study of construction accident news reports analysis,” *Advanced Engineering Informatics*, vol. 47, p. 101256, Jan. 2021, doi: 10.1016/j.aei.2021.101256.

[28] M. G. Gurmu, “Offline handwritten text recognition of historical Ge’ ez manuscripts using deep learning techniques,” *Jimma University*, 2021.

[29] F. A. Demilew and B. Sekeroglu, “Ancient Geez script recognition using deep learning,” *SN Applied Sciences*, vol. 1, no. 11, p. 1315, Nov. 2019, doi: 10.1007/s42452-019-1340-4.

## BIOGRAPHIES OF AUTHORS

**Seffi Gebeyehu** B.Sc. (CS) and M.Sc. (IT). Currently working as Assistant Professor in the Faculty of Computing, Bahirdar Technology Institute, Bahir Dar University, Ethiopia. He is a Ph.D. student at the Sudan University of Science and Technology. His Ph.D. is focused on the design and development of sharing economy platforms. He has published nine research papers in various international journals and one international conference paper. His main research interest is digital government, data mining, data science, artificial intelligence, and machine learning. He can be contacted at email: gseffi2010@gmail.com.

**Worke Wolde** B.Sc. (IT) and M.Sc. (IT). Currently working as a lecturer at Samara University, Institute of Engineering and Technology's Department of Information Technology in Afar, Ethiopia. She received the American Ethiopian Embassy of the United States of America award and recognition on August 26, 2010, E.C., for representing the University of 9 Ethiopian Regions in the 2010 E.C. Solve IT innovation competition, with the title of E-voting of her final project. Her research interest is in artificial intelligence and big data. She can be contacted at email: workewolde15@gmail.com.

**Zelalem S. Shibeshi** M.Sc. (Information Science) and Ph.D. (CS). Currently working as Senior Lecturer at the Computer Science Department, Rhodes University, South Africa. He has over 10 years of teaching experience in higher education institutes and over 10 years of experience in the IT industry. His research interests include machine learning (NLP), information retrieval (IR), multimedia service development and real-time communication, APIs for TelcoServices, and IoT in agriculture. He can be contacted at email: z.shibeshi@ru.ac.za.