ISSN NO. 2320-5407

Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

**Article DOI:** 10.21474/IJAR01/15816
**DOI URL:** http://dx.doi.org/10.21474/IJAR01/15816

***RESEARCH ARTICLE***

# AUTOMATIC RELATION EXTRACTION BETWEEN ENTITIES FOR AMHARIC TEXT

**S. Nagarajan[1], Melkamu Genet[2] and Yonatan Negesa[3]**

1. Assistatnt Professor, Department of Computer Science, Ambo University, Ethiopia.
2. Lecturer, Department of Computer Science, BuleHora University, Ethiopia.
3. Lecturer, Department of Computer Science, Werabe University, Ethiopia.

## Manuscript Info

***Manuscript History***
Received: 05 October 2022
Final Accepted: 09 November 2022
Published: December 2022

***Key words:-***
Information Extraction, Relation Extraction, Supervised Machine Learning, SVM and CRF

## Abstract

This research work primarily focused on the automatic relation extraction between entities for Amharic text using supervised machine learning approach.The Walta Information Centre online archive resources were used to create the study's own corpus, which consisted of 2000 sentences and a reasonable quantity of 30,466 words or tokens. The proposed solution has four processes; namely preprocessing, Text labeling, feature extraction and feature selection and Recognition. The tokenization and POS are used as preprocessing. After the text is tokenized and giving POS for each of tokens the next step is text labeling system. For text labeling mechanism BIO scheme is used. The tag features are selected for building the model. The Tag feature consists of name entity type and relation type. The name entity type features are represented by Location (LOC), Organization (ORG) and Person (PER) and the relation type features are identified every word which existed between two entities for instance between location-location relation type or location-organization relation type and all the corresponding entities that are appeared it. Vectorizations are done using DictVectorizer and word2features. Support vector machines and conditional random field machine learning are used to for recognizing the entity relation between Amharic texts. SVM with SGD achieved the weighted precision of 49%, recall 10% and f1-score 13% are scored. SVM with Multinomial Naive Bayes Classifier Algorithm achieve precision of 61%, recall 41% and f1-score 48%.SVM with Passive Aggressive classifier achieved weighted average precision of 55%, recall 19% and f1-score 27%. CRF algorithm achieved precision of 87%, recall 87% and f1-score 86%. The CRF model outperform compared with other SVM algorithms



## Introduction:-

Speaking, writing, and reading texts with their own scripts in their native tongues are a common practice in many nations as a form of instruction and communication. Modern technological science is concentrating on natural language processing in the situation of many types and an infinite quantity of texts being published over the internet. Due to this, information extraction using natural language processing faces a plethora of challenges. Relation extraction from Amharic text is one of the primary issues being studied.The Federal Democratic Republic of

**Corresponding Author:- S. Nagarajan**
Address:- Assistatnt Professor, Department of Computer Science, Ambo University, Ethiopia.

Ethiopia, the ethnicities of the South and the Regional State of Amhara all use Amharic as their official language. Primary and junior secondary schools also use it as their primary instructional medium. At a number of Ethiopian universities, it is also a field of concentration at the diploma, bachelor's degree, and master's degree levels. In addition to this, the language is used in a variety of publications, including books, newspapers, periodicals, educational materials, official documents, and religious texts. Because of the aforementioned factors, research on relation extraction from Amharic text is now being done.

In the globe, Amharic is the second-most spoken Semitic language after Arabic and is an official working language of Ethiopia [1].The writing system used to write Amharic is called Fidel and it was adopted from the Ge'ez language's system. Amharic accepted the entire Geez alphabet, including extraneous characters, and utilizes it as its own writing system. [2].

Natural language processing (NLP) has recently gained much attention for representing and analyzing human language computationally. Based on Message Understanding Conference-7(MUC-7) [3] information extraction activities are decomposed into five sub components such as named entity recognition, relation extraction, co-reference and anaphoric resolution, temporal and event detection and template filling are sub tasks of information extraction.

Once named entities are identified the second important process in an IE system is relation extraction. The extracted entities from step one is used as input to extract relations. A relation can be loosely defined as a relationship between two entities. More formally, a relation is a tuple t = (e1; e2… en) where eiare entities in a pre-defined relation r within a text T. Depending on the number of related entities, a relation can either be binary or complex. A binary relation spans an entity pair that lies within a sentence, whereas a complex relation may span more than two entities. Since the number of related entities in a complex relation is more than two, it is also called an n-aryrelation [4]the relation extraction task identifies various locations, affiliation, and revival and so on between entities from text.

For example, the sentence የኢትዮጵያጠቅላይሚኒስቴርዶክተርአብይአህመድ (Doctor AbyiAhamed prime minister of Ethiopia) carries the semantic relationship ጠቅላይሚኒስቴር(prime minister of) between named entities ዶክተርአብይአህመድ (Doctor AbyiAhamed) person and የኢትዮጵያ (Ethiopia) location.

## Literature Review:-

This section focuses on a review of earlier research on the extraction of relationships from Amharic text. Examination of NER the first step in IE, is necessary since it serves as a requirement for relation extraction.

Recognizing named entities in Amharic using a Conditional Random Field machine learning approach employed in [5]. The system uses a variety of features on their work, including prefixes, suffixes, part-of-speech tags on tokens, word and tag context features, and more. Four distinct situations based on various feature combinations were taken into account during the Experiment. With the exception of the POS tags of the tokens, all features were taken into account in the first and second scenario. All features, save the prefix and suffix, were taken into account in the third and fourth cases. According to their work's experimental findings, they obtained various results for various feature combinations. Precision, Recall, and F-measure in scenario one experiment were 72%, 75%, and 73.47%, respectively. The remaining experiments on scenarios two, three, and four yielded corresponding F-measures of 69.70%, 74.61%, and 70.65%.

The development of the Named Entity Recognition (NER) system for Amharic is discussed in[6]. The author used supervised machine learning called CRF and corpus of 13,538 words in Amharic has been produced with Stanford tagging approach. The greatest F-measure in their work is 80.66%, with a window size of 2 on both the left and right sides, a current word's preceding and next tags, and a prefix and suffix with a length of 4 respectively. They used feature sets that included a window size of two from the left side of a word and the previous and next word of the current token, and their worst performance was 61.97%.

A hybrid technique for Amharic named entity recognition was investigated [7]. The performance of an Amharic NER (ANER) constructed utilizing a hybrid technique and several feature sets to identify and categorize NEs of the type person, location, and organization was examined and reported on by the author. In their work, the performance of the hybrid ANER was examined using two cutting-edge machine learning (ML) algorithms: decision trees and support vector machines (SVM).

Due to the hybrid nature of their research, two rules that rely on their predictions on the presence of trigger words before and after NEs have been integrated into the hybrid ANER. Decision tree (J48) and SVM are used to construct the ML component (libsvm). By employing the NE class predicted from the rule-based component as a feature in the ML component, the hybrid ANER unifies those two components.The experiments led to the development of a high-performing model for the J48 and libsvm algorithms, which they used to achieve average performance of 96.1% highest F-score for decision tree and 85.9% for SVM, respectively, without using the rule-based feature but instead POS feature with nominal flag feature.

System performance was enhanced by using a hybrid technique for Afaan Oromo named entity recognition [8]. It includes rule-based elements including exact matching, parsing, filtering, grammar rules, white lists, gazetteers, and blacklist gazetteers. Conditional random field is the classifier utilized as the foundation for the system architecture. The first researcher also used the same corpus. Recall, Precision, and F1-measure average performances were 81.21%, 84.12%, and 82.52%, respectively.

The authors are studied a survey of relation extraction on several important supervised, semi-supervised and unsupervised RE techniques. And they also cover the paradigms of Open Information Extraction (OIE) and Distant Supervision [9].

The concept of automatic entity relation extraction based on maximum entropy is presented by Suxiang, Juan, Xiaojie, and Lei [10].According to the study, relation extraction is a categorization issue in texts written in Chinese. In order to extract entity relations between named entities from Chinese text, a machine learning approach based on maximum entropy (ME) was applied. The use of thirteen features, including morphological, grammatical, and semantic ones, is made. Two relation types—person affiliation employment and position-of—were used in the work. The system can obtain recall 93%, precision 87%, and f-score of 90% for person affiliation employment relation type, and recall 98%, recall 95%, and f-score of 96% for position-of relation type using just two classification attributes.

A brand-new and simple method for extracting entity relations is proposed by Chen and Peng [11]. Convolutional units and a dilated gate linear mechanism are the foundation of the end-to-end paradigm the author suggests. The study reduced the number of parameters to a low level by introducing dynamic convolutions based on lightweight convolutions to process extended sequences. The NYT and WebNLG datasets are used for the experimental outcomes of the aforementioned methodologies. Accordingly, it is possible to attain accuracy 67%, recall 60%, and f1 63% using the New York Times (NYT) dataset, and precision 68%, recall 50%, and f1 58% using the WebNLG dataset.

Information extraction method from Amharic language text using a knowledge-poor approach is offered by Getasew to address some of the issues with his work [12]. The term "knowledge-poor approach" refers to knowledge engineering methods that do not heavily rely on linguistic and subject-matter expertise.

For things that have a constant pattern, like person names that proceed by title, they have used both rules and gazetteers. On named entity recognition, the system achieves an F-measure of 89.1%, and overall, it achieves an F-measure of 89.8%. Relation extraction, which is a part of information extraction in his work, they are limited in their ability to help Getasew even if they try. Since a large amount of data is now distributed digitally throughout the world, researchers are now using machine learning approaches for modest information extraction, in accordance with their research, which employed a rule-based approach.

It is suggested to automatically create Amharic semantic networks from unstructured text using Amharic WordNet[13].The author employed a word space model to extract concepts that were semantically connected and then used the extracted text patterns to identify relationships between those concepts. The author claims that they attempt to extract semantically related words from unstructured documents using Amharic WordNet, and they investigate relationships such as how Ethiopia and Kenya are conceptually related to one another in African nations, or how Ethiopia and Ethiopian regions like Oromia and Amhara are related. From these works, it may be inferred that the author creates concepts with semantic relationships and identifies their hierarchical relationships. Additionally, the author does not identify the relationships between things.

The work's output is manually checked by linguists, and according to their assessment, the system's average accuracy in extracting the "type-of" and "part-of" link between ideas (cynsets) from free text corpus is 68.5% and 71.7%, respectively. The author's primary goal is to use Amharic WordNet to automatically build Amharic semantic networks from unstructured material. However, the author does not use relational data extraction.

Using a hybrid method, the intriguing work that served as the basis for this study is presented: extracting relationships between Amharic named items [14]. The SVM method is the only one used in the proposed work's machine learning technique, which uses it to create models and perform classification tasks. A total of 764 sentences from the Walta Information Center were used for training purposes. A significant amount of data is required, according to the supervised machine learning algorithm, to extract relationships between items from a text. A set of 163 sentences are used to test the system. The proposed system can perform at an average level of 80% precision, 81% recall, and 82% f-score. So, it is regarded as a foundational paper for this research.

When compared to this study, this linked work uses supervised machine learning as well and requires a lot more data. Two machine learning techniques are used to train this amount of text, which includes 2000 tokenized sentences and 30, 466 words, in order to improve the system's performance and generate models that are appropriate for relation extraction from Amharic text. One contribution of extracting relations from Amharic text is labeling a large volume of data for proper relation extraction and enhancing performance.

Deep-learning-based semantic relation categorization between Amharic entities is suggested [15]. Semantic relationships between entities are categorized in the article. The semantic relationships between Amharic elements are categorized using the BiLSTM network model in a deep learning technique. The author employed weighted attention technique to choose pertinent features and spatial drop out to reject less important feature vectors. Tensor Flow served as the backend while Keras was used to build the model. 4,171 Amharic sentences are the source of the data, which is gathered from Fana Broadcasting Corporation, Walta Information Center, and Ethiopian News Agency.The testing dataset is used to evaluate the proposed semantic relation classification model, and it can obtain 77.34%. The system performs less well when compared to the previous work of extracting relations between Amharic named things using a hybrid method [16]. The reason is because the deep-learning approach requires far more data than what was said previously. The deep learning approach's other flaw is that it requires a vast amount of unstructured data rather than structured data to learn from and make predictions about. The work claims that the data used is insufficient because it is neither structured nor annotated text data. Less unstructured data is trained, which results in lower system performance prediction. However, supervised machine learning relies on structured or labeled data, and it outperforms deep learning techniques in terms of system performance.

## Research Methodology:-

Figure 1 shows the proposed architecture of Relation Extraction from Amharic text-REAT[17].The proposed solution has four processes; namely preprocessing (tokenization, POS tagging, identifying NEs and relations etc.), Text labeling, feature extraction and feature selection and Recognition. Each process has sub components such as in the text preprocessing tokenization and POS Tag are existed. In process two, Annotation is carried out with BIO scheme.feature extraction process is carried out with two different vectorizationmethods are done in third process. Machine learning AREAT model training component are used, after named entity recognition and relation extraction are done in fourth step. Finally the evaluation of the model is done.

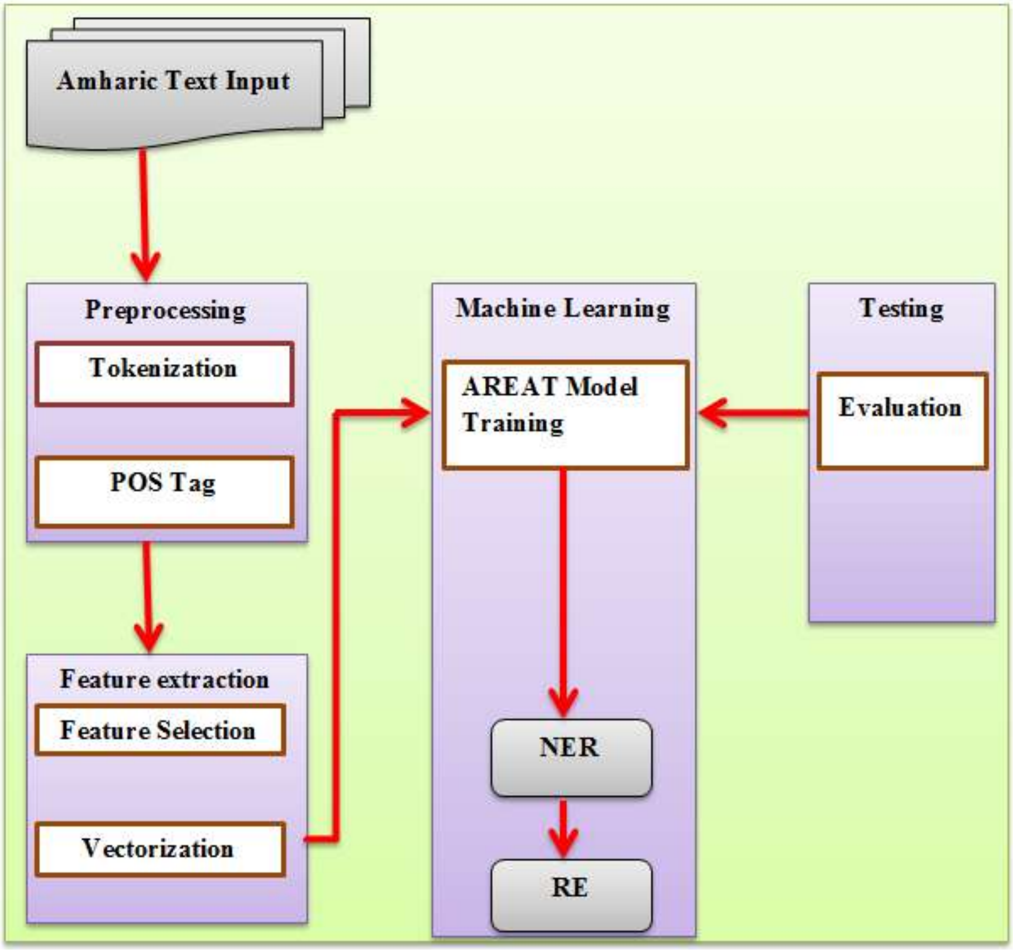

**Figure 1:-** Proposed Architecture.

## Result and Discussion:-

**Corpus Size**

Walta Information Center (WIC) news sources are used to gather the training and testing corpus size. 2000 sentences from the study were broken down into 30,467 tokens, or words, including punctuation.

**Table 1:-** Corpus size.

| Number of Sentences | Total Number of Tokens | Training data | Test data |
|---|---|---|---|
| 2000 | 30,466 | 24,372 | 6,094 |

The data size is displayed in the table above, and this information serves as the experimental case's input. The experimental text that was entered in Amharic is divided up into tokens. It is necessary to break down the text in order to identify the part of speech for each and every word, including punctuation. The Natural Language Toolkit does tokenization (NLTK). Words that have been tokenized are utilized as input for POS tags, as seen in Figure 2.

```
In [4]:  import nltk
         sentence = """ልዩ ዞኑ ያስገነባቸው ፕሮጀክቶች 50ሺ ነዋሪዎችን ተጠቃሚ አደረጉ።
         በአዳማ ልዩ ዞን በተጠናቀቀው የበጀት አመት 6.8ሚሊየን ብር ከተጀመሩት ፕሮጀክቶች መካከል
         tokens = nltk.word_tokenize(sentence)
         tokens

Out[4]: ['ልዩ',
         'ዞኑ',
         'ያስገነባቸው',
         'ፕሮጀክቶች',
         '50ሺ',
         'ነዋሪዎችን',
         'ተጠቃሚ',
         'አደረጉ።',
         'በአዳማ',
         'ልዩ',
         'ዞን',
         'በተጠናቀቀው',
         'የበጀት',
         'አመት',
         '6.8ሚሊየን',
         'ብር',
         'ከተጀመሩት',
         'ፕሮጀክቶች',
         'መካከል',

In [ ]:
```

**Figure 2:-** Tokenization Amharic Text Using NLTK.

POS is the second experimental method that is mentioned in the suggested architecture. POS is now carried out in this experimental approach utilizing the NLTK Python package. This POS tag assigns the proper POS for each token using a list of tokens. In order to know and identify the precise entities in this work, as well as to identify relations between entities during the data annotation process, POS tags are required. Figure 3 demonstrates how to tag Amharic text using POS.

```
Out[2]:  ('ልዩ', 'ADJ'),
         ('ዞኑ', 'N'),
         ('ያስገነባቸው', 'VREL'),
         ('ፕሮጀክቶች', 'N'),
         ('50ሺ', 'NUMCR'),
         ('ነዋሪዎችን', 'N'),
         ('ተጠቃሚ', 'ADJ'),
         ('አደረጉ', 'V',),
         ('።', 'PUNC'),
         ('አሜሪካ', 'NP'),
         ('ኒውዮርክ', 'NP'),
         ('ዋሽንግተን', 'N'),
         ('ትናንት', 'ADV'),
         ('ፕሬዝዳንት', 'NP'),
         ('የደረሱትን', 'VREL'),
         ('አራት', 'NUMCR'),
         ('የአጥፍቶ', 'VREL'),
```

**Figure 3:-** POS Tag for Amharic Text Using NLTK.

The annotation is granted POS tag out. The data is labeled using the "Begin Inside Outside" (BIO) method in order to develop a relation extraction model from Amharic using supervised machine learning. Here, "B" signifies the start of an entity and relation, "I" denotes "inside," which is used for all words other than the first one in the entity and relation, and "O" denotes the lack of entities and the relations that exist between entities.

The encoding process consists of two steps: token or tag generation for a machine learning model and token detection. After the POS tag has been identified, the entities and relations are annotated. If the entities are people, they are represented as B-PER and I-PER, locations as B-LOC and I-LOC, and organizations as B-ORG and I-ORG. If the relations are between people and places, they are represented as B-LOC/PER-RE. O stands in for the remaining words or tokens that are part of those entities. Figure 1 displays the annotation of Amharic text for relation extractions using the BIO approach.

| sentence # | words | POS | Tag |
|---|---|---|---|
| | የግብርናና | NPC | O |
| | የመሰረታዊ | ADJP | O |
| | ልማት | N | O |
| | ፕሮጀክቶች | N | O |
| | ናቸው | AUX | O |
| | የኢትዮጵያ | NPP | B-LOC |
| | ፌደራላዊ | NP | I-LOC |
| | ዴሞክራሲያዊ | NP | I-LOC |
| | ሪፐብሊክ | V | I-LOC |
| | ፕሬዝዳንት | NP | B-LOC/PER-RE |
| | ነጋሶ | N | B-PER |
| | ጊዳዳ | N | I-PER |
| | ገልፀዋል | ADJ | O |
| | ። | PUNC | O |
| sentence:4 | ፕሮጀክቶቹም | NC | O |
| | 25ኪሎ | NUMCR | O |
| | ሜትር | N | O |
| | የውስጥ | NP | O |
| | ለውስጥ | NP | O |
| | የጠጠር | NP | O |
| | መንገድ | N | O |
| | በቻይና | NP | B-LOC |
| | የምትምራዉ | ADJ | B-LOC/LOC-RE |
| | የሆንክ | NP | B-LOC |
| | ኮንግ | NP | I-LOC |
| | ግዛት | NP | I-LOC |

**Figure 4:-** BIO Data Annotation Scheme.

It employed two mechanisms in order to extract more features. The first one is DictVectorizer, a system that can convert tokens into binary digit vectors. The tokens in the examples below have been transformed into vectored features. The vectorization of Amharic text data using this technique is demonstrated in the followingsample.

```
In [6]: X = df.drop('Tag', axis=1)
        v = DictVectorizer(sparse=False)
        X = v.fit_transform(X.to_dict('records'))
        y = df.Tag.values

        classes = np.unique(y)
        classes = classes.tolist()

        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.20, random_state=0)
        X_train.shape, y_train.shape
        print(X_train, y_train)

        [[0. 0. 0. ... 0. 0. 0.]
         [1. 0. 0. ... 0. 0. 0.]
         [1. 0. 0. ... 0. 0. 0.]
         ...
         [0. 0. 0. ... 0. 0. 0.]
         [0. 0. 0. ... 0. 0. 0.]
         [0. 0. 0. ... 0. 0. 0.]] ['B-PER' 'O' 'I-ORG' ... 'B-LOC/LOC-RE' 'B-ORG' 'O']
```

**Figure 1:-** Text Vectorization using DictVectorizer.

You might infer from this that the Tag column represents the chosen features. The class of entity type and relation type can be predicted with great accuracy and suitability.

Word2feature is the mechanism employed in the second feature extraction and selection. By assigning the negative and positive numbers of the bottom word, upper word is title, and is digit features; this word2feature extraction extracts more features. The title, digit features, sample lower, middle, and upper words, as well as the portioned text data, are as follows.

```
In [41]: X = [sent2features(s) for s in sentences]
         y = [sent2labels(s) for s in sentences]
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20, random_state=0)
         print(X_train, y_train)

         ()': False, 'Tag': 'N', 'Tag[:2]': 'N', '-1:word.lower()': 'የሰራዊት', '-1:word.istitle()': F
          '-1:Tag': 'NP', '-1:Tag[:2]': 'NP', '+1:word.lower()': 'ተከፋፍሏል', '+1:word.istitle()': Fa
         1:Tag': 'V', '+1:Tag[:2]': 'V'}, {'bias': 1.0, 'word.lower()': 'ተከፋፍሏል', 'word[-3:]': 'ፍሏ
         r()': False, 'word.istitle()': False, 'word.isdigit()': False, 'Tag': 'V', 'Tag[:2]': 'V',
         rd.istitle()': False, '-1:word.isupper()': False, '-1:Tag': 'N', '-1:Tag[:2]': 'N', '+1:wo
         ()': False, '+1:word.isupper()': False, '+1:Tag': 'PUNC', '+1:Tag[:2]': 'PU'}, {'bias': 1.
         3:]': '።', 'word[-2:]': '።', 'word.isupper()': False, 'word.istitle()': False, 'word.isdig
         2]': 'PU', '-1:word.lower()': 'ተከፋፍሏል', '-1:word.istitle()': False, '-1:word.isupper()':
          'V', 'EOS': True}]]

          [['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-LOC', 'I-LOC', 'B-LOC/ORG-RE', 'I-LOC/ORG-RE
          'O', 'O', 'O', 'O', 'B-LOC', 'I-LOC', 'B-LOC/LOC-RE', 'I-LOC/LOC-RE', 'B-LOC', 'I-LOC'],
          'B-LOC', 'I-LOC', 'B-LOC/LOC-RE', 'I-LOC/LOC-RE', 'B-LOC', 'I-LOC', 'O'], ['B-ORG', 'I-OR
          'O', 'B-PER', 'I-PER', 'B-PER/PER-REL', 'I-PER/PER-REL', 'B-PER', 'I-PER', 'O', 'O', 'O',
          'O', 'O', 'O', 'O', 'B-PER', 'B-PER/PER-REL', 'I-PER/PER-REL'], ['O', 'O', 'B-LOC', 'I-LC
          'B-LOC', 'I-LOC', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-LOC', 'I-LOC', 'B-LOC/L
          'O', 'O', 'O', 'O', 'O', 'B-LOC', 'I-LOC', 'B-LOC/LOC-RE', 'I-LOC/LOC-RE', 'B-LOC', 'I-LC
         L', 'I-ORG/PER-REL', 'B-PER', 'I-PER', 'O', 'O', 'O', 'O', 'O', 'B-ORG', 'I-ORG', 'B-ORG/P
```

**Figure 2:-** Text Vectorization using word2feature.

The SVM and CRF algorithms are used to develop the machine learning model for relation extraction from the Amharic text model in the next experimental step.

**SVM Algorithm Result Analysis.**
The relation extraction from Amharic text model is created using SVM algorithm and the performance of the model is tested byStochastic Gradient Descent (SGD) performance optimizer andclassifier, multinomial Naive Bayes classifier algorithm, and Passive Aggressive classifier algorithm and result is shown in table 2

The system performance is measured and optimized using SVM with SGD. This study uses the SGD optimization technique for text categorization and assesses the effectiveness using training and test data [18]. It can be inferred from the results of Table 2 below that each and every one of the annotated class types is trained, modeled, created, and forecasted. For instance, the above SGD result for the relation type B-LOC/LOC-RE indicates that it can achieve precision of 44%, recall of 13%, and f1-score of 20% from the supporting column, which indicates that 91 location-location relation type classes are predictedThe precise number of predicted from the total number of generated trained entities and relations class types is shown in the support column of the result. This algorithm is being used to enhance the system's performance. However, as you can see, this optimization process yielded results with weighted precision of 49%, recall of 10%, and f1-score of 13%. Classes are Weakley predicted which makes them ineffective for this purpose. That is why it is necessary to use a different algorithm.The results of each and every entity and relation type are shown in Table 2 when utilizing the multinomial NB classifier algorithm, and they are predicted based on the model that was built using the SVM technique that was trained and modeled[19]. The multinomial NB model performs better than the preceding SGD, according to the results. Passive classifier is capable of 61% precision, 41% recall, and 48% f1-score. For this investigation, the performance of the model produced using the SVM algorithm is still unsatisfactory. The result shown using the SVM method model displays weighted average precision of 55%, recall of 19%, and f1-score of 27%.Based on the provided created and trained datasets of entities and relations class of type and their algorithmic performance levels, the aforementioned three classifier algorithms forecast each and every tag set. The produced and trained model is used to calculate the precision, recall, and f1-score for each class, and the weighted average of each classifier is used to calculate the overall measurement.

According to the table below, different classifier algorithms produce different results. All algorithms are used as a weighted average for comparing purposes. The equal amount of data is rejected and predicted true by each method. Even when the same amount of data is expected, the performance of the algorithms used to measure each entity and relation class type varies from one another. Each algorithm receives a distinct score as a result of the evolving shifting performance measurement of an algorithm. This comparison shows that the Multinomial Naive Bayes classification algorithm performs better than the other two algorithms. In general, the outcome of this SVM algorithm model is completely unsatisfactory.

**Table 2:-** SVM Algorithm Result.

| Class | SVM with SGD | | | SVM with Mulitimodal | | | SVM with Passive | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1-Score | Precision | Recall | F1-Score | Precision | Recall | F1-Score |
| B-LOC | 0.52 | 0.08 | 0.13 | 0.56 | 0.35 | 0.43 | 0.45 | 0.26 | 0.33 |
| B-LOC/LOC-RE | 044 | 0.13 | 020 | 0.54 | 0.41 | 0.46 | 0.44 | 0.19 | 0.26 |
| B-LOC/ORG-RE | 0.51 | 0.21 | 0.30 | 0.51 | 0.44 | 0.47 | 0.45 | 0.24 | 0.32 |
| B-LOC/PER-RE | 0.40 | 0.09 | 0.15 | 0.57 | 0.39 | 0.46 | 0.57 | 0.09 | 0.16 |
| B-ORG | 0.60 | 0.05 | 0.09 | 0.71 | 0.36 | 0.48 | 0.58 | 0.13 | 0.21 |
| B-ORG/ORG-RE | 0.00 | 0.00 | 0.00 | 0.57 | 0.35 | 0.44 | 0.58 | 0.17 | 0.26 |
| B-ORG/PER-RE | 0.20 | 0.11 | 0.14 | 0.60 | 0.38 | 0.46 | 0.67 | 0.22 | 0.33 |
| B-PER | 0.69 | 0.05 | 0.09 | 0.63 | 0.42 | 0.51 | 0.67 | 0.15 | 0.24 |
| B-PER/ORG-RE | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| B-PER/PER-RE | 0.67 | 0.03 | 0.05 | 0.61 | 0.40 | 0.48 | 0.55 | 0.23 | 0.32 |
| I-LOC | 0.40 | 0.07 | 0.12 | 0.61 | 0.48 | 0.54 | 0.51 | 0.12 | 0.20 |
| I-LOC/LOC-RE | 0.73 | 0.09 | 0.16 | 0.56 | 0.31 | 0.40 | 0.71 | 0.14 | 0.23 |
| I-LOC/ORG-RE | 0.35 | 0.07 | 0.12 | 0.56 | 0.49 | 0.52 | 0.48 | 0.12 | 0.19 |
| I-LOC/PER-RE | 1.00 | 0.02 | 0.04 | 0.85 | 0.46 | 0.59 | 0.59 | 0.21 | 0.31 |
| I-ORG | 0.19 | 0.35 | 0.24 | 0.57 | 0.42 | 0.48 | 0.54 | 0.21 | 0.30 |
| I-ORG/ORG-RE | 0.00 | 0.00 | 0.00 | 0.36 | 0.45 | 0.40 | 0.43 | 0.27 | 0.33 |
| I-ORG/PER-RE | 0.43 | 0.04 | 0.07 | 0.67 | 0.55 | 0.60 | 0.55 | 0.32 | 0.40 |
| I-PER | 0.52 | 0.06 | 0.10 | 0.70 | 0.42 | 0.52 | 0.57 | 0.24 | 0.34 |
| I-PER/ORG-RE | 0.00 | 0.00 | 0.00 | 0.33 | 0.50 | 0.40 | 0.00 | 0.00 | 0.00 |
| I-PER/PER-RE | 0.67 | 0.04 | 0.07 | 0.63 | 0.41 | 0.49 | 1.00 | 0.02 | 0.04 |
| **Micro Average** | **0.30** | **0.10** | **0.15** | **0.60** | **0.41** | **0.48** | **0.53** | **0.19** | **0.28** |
| **Macro Average** | **0.41** | **0.08** | **0.11** | **0.50** | **0.39** | **0.44** | **0.51** | **0.18** | **0.25** |
| **Weighted-Avg** | **0.49** | **0.10** | **0.13** | **0.61** | **0.41** | **0.48** | **0.55** | **0.19** | **0.27** |

**CRF Algorithm Result Analysis**

CRFs can model the sequential data that is useful in many different applications of natural language processing. Relation extraction, which predicts the order in which they are reliant on one another, is one of the well-known uses of CRFs in NLP [20]. Similar to the SVM technique, the CRF predicted the type of relation based on the trained and test datasets with annotations. This approach can perform better than SVM, as shown in the following table, where CRF is the best algorithm for creating a model that is appropriate for relation extraction from Amharic text. This approach employs the flat-classifier algorithm to evaluate the performance of the model and predict relation types that were generated by the CRF algorithm. Take a look at the example performance measurement below on table 3.

The flat classifier algorithm is used to assess the model's performance. The weighted result can therefore attain precision of 87%, recall of 87%, and f1-score of 86%. Similar to the SVM algorithm model, the conditional random (CRF) field generated all classes of entities and relations. The total number of class types created by the CRF model is smaller than the SVM algorithm from the provided annotated and trained model. However, some trained and produced data outperform SVM in terms of prediction performance.

**Table 3:-** CRF Algorithm Result.

| Class | Precision | Recall | F1-score |
|---|---|---|---|
| **B-LOC** | 0.85 | 0.80 | 0.82 |
| **B-LOC/LOC-RE** | 0.80 | 0.85 | 0.82 |
| **B-LOC/ORG-RE** | 0.80 | 0.73 | 0.76 |
| **B-LOC/PER-RE** | 0.90 | 0.82 | 0.86 |
| **B-ORG** | 0.87 | 0.72 | 0.79 |
| **B-ORG/ORG-RE** | 0.89 | 0.67 | 0.76 |
| **B-ORG/PER-RE** | 0.82 | 0.90 | 0.86 |
| **B-PER** | 0.81 | 0.80 | 0.80 |
| **B-PER/ORG-RE** | 0.50 | 1.00 | 0.67 |
| **B-PER/PER-RE** | 0.59 | 0.67 | 0.62 |
| **I-LOC** | 0.81 | 0.82 | 0.82 |
| **I-LOC/LOC-RE** | 0.78 | 0.89 | 0.83 |
| **I-LOC/ORG-RE** | 0.75 | 0.86 | 0.80 |
| **I-LOC/PER-RE** | 0.92 | 0.86 | 0.89 |
| **I-ORG** | 0.87 | 0.72 | 0.79 |
| **I-ORG/ORG-RE** | 1.00 | 0.50 | 0.67 |
| **I-ORG/PER-RE** | 0.83 | 0.77 | 0.80 |
| **I-PER** | 0.80 | 0.71 | 0.75 |
| **I-PER/ORG-RE** | 0.50 | 1.00 | 0.67 |
| **I-PER/PER-RE** | 0.75 | 0.75 | 0.75 |
| Micro Average | **0.87** | **0.87** | **0.87** |
| Macro Average | **0.80** | **0.81** | **0.79** |
| Weighted Average | **0.87** | **0.87** | **0.86** |

Two models created with the help of the SVM and CRF algorithms are used in this investigation. A range of findings can be obtained from the algorithms by studying the model using different ones. Table 4 illustrates that conditional random field can produce findings that are superior to SVM for the overall evaluation performance of this study.

**Table 4:-** Comparison of SVM and CRF.

| Algorithm | | Precision | Recall | F1-score |
|---|---|---|---|---|
| SVM | SGD Weighted Average | 0.49 | 0.10 | 0.13 |
| | MultinomialNB Weighted Average | 0.61 | 0.41 | 0.48 |
| | Passive Aggressive Weighted Average | 0.55 | 0.19 | 0.27 |
| CRF | Flat Classifier Weighted Average | 0.87 | 0.87 | 0.87 |

Table 5 contrasts the suggested (CRF) work with Salomon's work. With a precision of 87%, recall of 87%, and f1-score of 86%, the CRF model can perform better. Selomon, however, was able to achieve precision of 80%, recall of 81%, and f1-score of 83% using SVM. The findings suggest that supervised machine learning is more effective than a hybrid approach at completing tasks. The utilization of a thoroughly annotated dataset can lead to greatly improved performance.

**Table 5:-** Comparisons CRF and SelomonGetachews.

| Comparision | Precision | Recall | F1-score |
|---|---|---|---|
| **Condition Random Field** | 87% | 87% | 86% |
| **SelomonGetachews** | 80% | 81% | 83% |

## Conclusion:-

The goal of this project is to use supervised machine learning to adopt and develop an appropriate model for relation extraction from Amharic text. Predefined person, location, and organization are extracted as entities and predefined relations, such as the relation between person-person, person-organization, person-location, person-location-organization, etc., are to be done extracted as relation in this document. This work used Python software with NLTK and scikit-learn libraries and Amharic news text that was collected from Walta information center. Support vector machine (SVM) and conditional random field (CRF) algorithms were employed in this investigation.These methods were used to develop relation extraction from an Amharic text model, as well as to classify data and evaluate the system's performance. To the best of our knowledge, the performance of the system can be obtained with precision of (p) of 87%, recall of 87%, and f-score of 87% using the most up-to-date CRF algorithm.

## References:-

1. https://en.wikipedia.org/wiki/Amharic
2. https://en.wikipedia.org/wiki/Ge%CA%BDez_script
3. Sun, B. (2010). Named entity recognition Evaluation of Existing Systems. Master in Information Systems,Norwegian University of Science and Technology Department of Computer and Information Science.
4. Gitansh, K. (October 2017). Supervised Algorithms for Complex Relation Extraction. in the School of Computing and Information Systems Melbourne School of Engineering. Melbourne.
5. MOGES, A. (2010, November). NAMED ENTITY RECOGNITION FOR AMHARIC LANGUAGE. 1-66.
6. Besufikad, A. (2013, June). Named Entity Recognition for Amharic. 1-67.
7. MIKIYAS, T. (2014, August). AMHARIC NAMED ENTITY RECOGNITION USING A HYBRID APPROACH. 1-61.
8. Abdi, S. (2015, March). AFAAN OROMO NAMED ENTITY RECOGNITION USING HYBRID APPROACH. 1-96.
9. Pawar, S., Palshikar, G. K., Pushpak, a., & Bhattacharyya. (Dec 2017). Relation Extraction. A Survey.
10. Suxiang, Z., Juan, W., Xiaojie, W., & Lei, L. (2006). Automatic Entity Relation Extraction Based on Maximum Entropy. Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06).
11. Chen, X., & Peng, G. (October 4 2020). Entity–Relation Extraction—A Novel and Lightweight Method Based on a Gate Linear Mechanism. Electronics, 1-15.
12. Bekele, W. (2015, June). Information Extraction from Amharic language Text : Knowledge-poor Approach.
13. Alelgn, T. (2014). Automatic Construction of Amharic Semantic Networks From Unstructured Text Using Amharic WordNet. in proceeding of the seventh global wordnet conference, (pp. 172-117).
14. Selomon, G. (2020, 01 22). Jimma University Open access Institutional Repository. (JIMMA UNIVERSITY JIMMMA INSTITUTE OF TECHNOLOGY) Retrieved 4 29, 2020, from https://repository.ju.edu.et/handle/123456789/5376
15. TESFAHUN, N. M. (2021, April 30). BDU IR. Retrieved may 08, 2021, from http://hdl.handle.net/123456789/11273
16. Getnet, S. (January 2020). EXTRACTING RELATIONS BETWEEN AMHARIC NAMED ENTITIES USING A HYBRID APPROACH. 1-57.
17. Suxiang, Z., Juan, W., Xiaojie, W., & Lei, L. (2006). Automatic Entity Relation Extraction Based on Maximum Entropy. Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06).
18. Scikit-Learn. (2007 - 2020). (scikit-learn developers) Retrieved May 06, 2021, from https://scikit-learn.org/stable/modules/sgd.html
19. Gautam, S. (2020, Sep 20). Great Learning. Retrieved May 05, 2021, from https://www.mygreatlearning.com/blog/multinomial-naive-bayes-explained/
20. Chavan, A. (2019, June 27). AI TIME JOURNAL. Retrieved from https://www.aitimejournal.com: https://www.aitimejournal.com/@akshay.chavan/introduction-to-conditional-random-fields-crfs.