Machine Learning Research
2023; 8(2): 17-22
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20230802.11
ISSN: 2637-5672 (Print); ISSN: 2637-5680 (Online)

Vector Representation of Amharic Idioms for Natural Language Processing Applications Using Machine Learning Approach

Anduamlak Abebe Fenta

Department of Computer Science, Gafat Institute of Technology, Debre Tabor University, Debra Tabor, Ethiopia

Email address:
anduabe@dtu.edu.et, anduamlak09@gmail.com

To cite this article:
Anduamlak Abebe Fenta. (2023). Vector Representation of Amharic Idioms for Natural Language Processing Applications Using Machine Learning Approach. *Machine Learning Research*, *8*(2), 17-22. https://doi.org/10.11648/j.mlr.20230802.11

Received: November 21, 2023; **Accepted**: December 14, 2023; **Published**: December 22, 2023

Abstract: Idiomatic phrases are natural components of all languages that cannot be comprehended straight from the word from which they are generated. Vector representations are a key method that bridges the human understanding of language to that of machines and solves many NLP problems. Idiomatic expression representation is necessary for machine learning, deep learning, and natural language processing applications. Machine learning and deep learning techniques have not been used to process text as input for natural language processing applications in previous literature. As such, in order to study natural language processing with machine learning and deep learning methods, vector or numeric representations of idiomatic statements are needed. Therefore, this research aimed at the proposed vector representation of Amharic idioms for NLP applications through vector representation models. Researchers that study natural language processing use this format, and for classification or regression, they employ machine learning and deep learning techniques. Before doing NLP application researches on Amharic idiom, first, it requires vector or numeric representation using suitable methods. We used five hundred idiomatic expressions from Amharic Idioms book as a dataset for this representation, which are comprised of two words. To evaluate performance, we employed the accuracy, precision, recall, and F-score metrics. The dataset produced a result of 95.5% accuracy.

Keywords: Amharic Idiom, Machine Learning, Vector Representation, Word2vector

1. Introduction

Idiomatic phrases are significant, naturally occurring components of all languages and commonplace aspects of everyday speech that are not immediately understandable from the word from which they are derived. Idioms are sophisticated linguistic structures that are imaginatively employed in practically all text categories. Because they are non-compositional, idioms provide challenges for natural language processing (NLP) systems [1, 2].

A phrase known as an Amharic idiom typically consists of two words combined in a way that defies interpretation based on the meaning of the words alone or on how they are typically used together [3]. Amharic idiom hurts NLP applications like machine translation, sentiment analysis, semantic analysis, and next-word prediction. To train idiomatic expressions for machines is difficult due to text-based expressions. A computerized system doesn't process any activity by using text-based inputs. To provide idiomatic expressions for machines requires encoding to other formats using different mechanisms. Researchers implemented different methods to represent words or expressions into vectors or numbers like distribution of words, term frequency, Continuous Bag of Words, Skip-gram, set2vector, one hot encoding method [2, 4-6]. Vector representations of words or phrases are used as input for different NLP applications and machines. So; we did Amharic idiomatic expression vectorization using word2vector and one hot encoding method.

2. Literature Review

Various researchers are researching word representations for different NLP applications for different languages through various models. For word representation in vector space, a neural network is employed to map Arabic vectors to English vectors.

A. M. A. Y. M. H. R. M. R. &. A. A. Mohamed A. Zahran [7] demonstrated that minimizing for cosine error outperforms the standard mean square error minimization for word-to-word similarity using cosine score, indicating that the training procedure's objective function should match the chosen similarity measure. Researchers [8] done Chinese Document Representation and Classification with Word2vec with different models SVM, KNN, and RFB neural network classifier were used. For sentiment analysis, researchers used vector representation of words through the GloVe model [9]. Researchers [10] text classification with semantic features using word2vector and support vector machines and compared with other representation models. for idiom identification also researchers used word embedding, distribution of words, Skip-thought model, sen2vector representation, CBOW [2, 4-6].

3. Contribution of the Research

Considering the vector-based representation of an idiomatic expression is essential for natural language processing applications because depending on the context, many idioms can be recognized literally or idiomatically. We adopt Distributional Hypothesis for the idioms vector representation. We have also stated the usage of KNN based evaluation of the vector-based representation. Collection of idiomatic expression word/phrase lists.

4. Research Methodology

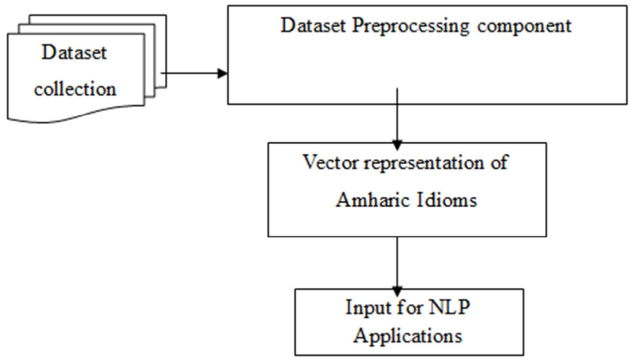

Dataset collection
Dataset Preprocessing component
Vector representation of Amharic Idioms
Input for NLP Applications

Figure 1. *Proposed model.*

4.1. Dataset Collection

We collected five hundred idiomatic expressions from a famous idiomatic book called Amharic idioms book “የአማረኛ ፈሊጦች” (Akililu & Worku, 1992). All the collected idiomatic expressions are a combination of two words like ጆሮ ዳባ, ጆሮ ሰጠኝ, ከአንገት በላይ, አንገት መዘዘ, አንገት ረገጠ, አንገተ ሰባራ, አንገተ ቢስ, አንገተ ወፍራም, አንገተ ደንዳና

4.2. Pre-Processing

After collecting the dataset, we implement preprocessing to remove unnecessary words or symbols. Preprocessing involves cleaning, spell checking, normalization, joining tokens, and removing stop words and numbers.

4.3. Representation of Amharic Idioms

4.3.1. One-Hot Representation

This type of representation is incredibly straightforward and simple to use. Nevertheless, even with such a small sample, many of the flaws would have been obvious. That is, processing and storing such vectors requires a lot of memory. This representation system is unable to convey semantic information since there is zero cosine similarity between distinct words. This representation used many input expressions and dimensional spaces to encode the expression into vectors.

Example: - The vector encoding mechanism, we use six sample expressions. Space with six dimensions is used to represent it.

Table 1. *A single hot encoding representation.*

Expression	One Hot Encoding
ልቡተሰነጠቀ	[1, 0, 0, 0, 0, 0]
ነፍስሆነ	[0, 1, 0, 0, 0, 0]
ሰኔናሰኞ	[0, 0, 1, 0, 0, 0]
አስብቶአራጅ	[0, 0, 0, 1, 0, 0]
ቁመተስጋ	[0, 0, 0, 0, 1, 0]
ሆደመጋዝ	[0, 0, 0, 0, 0, 1]

4.3.2. Word2Vec Representations

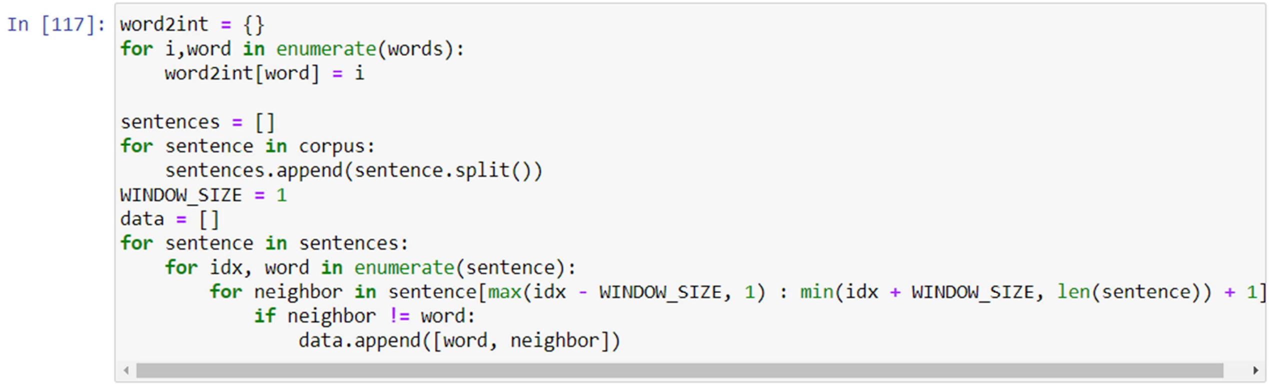
In essence, Word2Vec arranges words in the feature space so that their meaning dictates where they go. This approach creates word embeddings for better word representation. It accurately captures a wide range of syntactic and semantic word relationships. There is just one hidden layer separating the input and output of this two-layered shallow neural network [11]. The concept behind Word2Vec is that a word or term can be represented in a vector space representation by a vector. In order to transform the unsupervised representations into supervised form for model training, word2vec uses two architectures: skip-gram and continuous bag of words (CBOW) [2, 12].

The network is trained using the words inside the window at each step in both models, which involve moving a predefined length window along the corpus. When it comes to context prediction, the Skip-gram model learns from the central word, while the CBOW model learns from the surrounding words to predict the word in the window's center. After the neural network is trained, the word representation is determined by the learned linear transformation in the hidden layer.

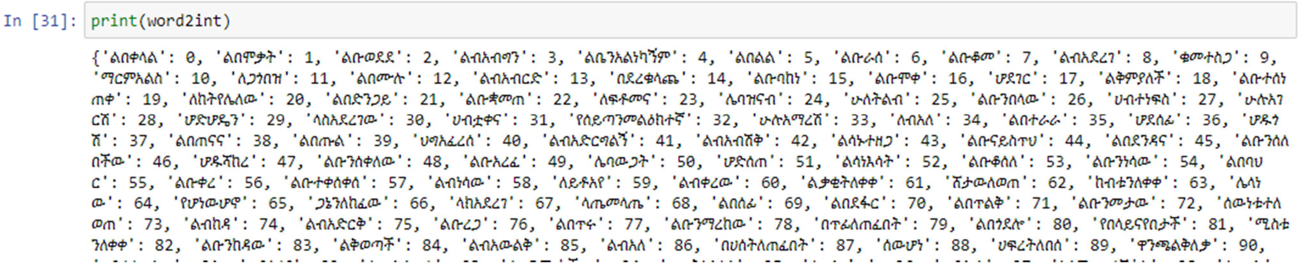
In Skip-grams, the neural network receives in a word and then attempts to forecast the nearby words or the target words from the context. The neural network in Continuous Bag of Words, on the other hand, forecasts the words that lie between or determines the target words from the context. The length of the vocabulary corresponds to the size of the Neural Network's final output vector. Depending on how closely the target words and context words are related, it returns a probability value.

Equipped with mathematical operators, words can be represented as vectors, which is beautiful. For Turkic, Arabic, Chinese, and Amharic, languages with rich morphology [13].

For this analysis, we used the word2vector and one hot encoding method to convert phrases to vectors. To represent numbers into vectors, we used one hot encoding mechanism with a length of word dimensional space. The hidden layer has two weight values which are used for visualization on two-dimensional spaces [11].



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

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



```
In [31]: print(word2int)

{'ልበቀላል': 0, 'ልበሞቃት': 1, 'ልቡወደደ': 2, 'ልብአብግን': 3, 'ልቤንአልነካኝም': 4, 'ልበልል': 5, 'ልቡራሰ': 6, 'ልቡቆመ': 7, 'ልብአደረገ': 8, 'ቁመተስጋ': 9,
'ማርምአልስ': 10, 'ሰጋጎበዝ': 11, 'ልበሙሉ': 12, 'ልብአብርድ': 13, 'በደረቁላጨ': 14, 'ልቡባከነ': 15, 'ልቡሞቀ': 16, 'ሆደገር': 17, 'ልቅምያለች': 18, 'ልቡተሰነ
ጠቀ': 19, 'ለከትየሌለው': 20, 'ልበድንጋይ': 21, 'ልቡቋመጠ': 22, 'ለፍቶመና': 23, 'ሌባዝናብ': 24, 'ሁለትልብ': 25, 'ልቡንበላው': 26, 'ሀብተነፍስ': 27, 'ሁሉአገ
ርሽ': 28, 'ሆድሆዴን': 29, 'ላስአደረገው': 30, 'ሀብቷቀና': 31, 'የሰይጣንመልዕክተኛ': 32, 'ሁሉአማረሽ': 33, 'ለብአለ': 34, 'ልበተራራ': 35, 'ሆደሰፊ': 36, 'ሆዱጎ
ሽ': 37, 'ልበጠናና': 38, 'ልበጡል': 39, 'ሀግአፈረሰ': 40, 'ልብአድርግልኝ': 41, 'ልብአብሽቅ': 42, 'ልሳኑተዘጋ': 43, 'ልቡናይስጥህ': 44, 'ልበደንዳና': 45, 'ልቡንሰለ
በችው': 46, 'ሆዱሻከረ': 47, 'ልቡንሰቀለው': 48, 'ልቡአረፈ': 49, 'ሌባውጋት': 50, 'ሆድሰጠ': 51, 'ልሳነእሳት': 52, 'ልቡቆሰለ': 53, 'ልቡንነሳው': 54, 'ልበባህ
ር': 55, 'ልቡቀረ': 56, 'ልቡተቀሰቀሰ': 57, 'ልብነሳው': 58, 'ለይቶአየ': 59, 'ልብቀረው': 60, 'ልቃቂትለቀቀ': 61, 'ሽታውለወጠ': 62, 'ከብቱንለቀቀ': 63, 'ሌላነ
ው': 64, 'የሆነውሆኖ': 65, 'ጋኔንለከፈው': 66, 'ላከአደረገ': 67, 'ላጤመላጤ': 68, 'ልበሰፊ': 69, 'ልበደፋር': 70, 'ልበጥልቅ': 71, 'ልቡንመታው': 72, 'ሰውነቴተለ
ወጠ': 73, 'ልብከዳ': 74, 'ልብአድርቅ': 75, 'ልቡረጋ': 76, 'ልበጥሩ': 77, 'ልቡንማረከው': 78, 'በጥፊለጠፈበት': 79, 'ልበጎደሎ': 80, 'የበላይናየበታች': 81, 'ሚስቱ
ንለቀቀ': 82, 'ልቡንከዳው': 83, 'ልቅወጣች': 84, 'ልብአውልቅ': 85, 'ልብአለ': 86, 'በሀሰትለጠፈበት': 87, 'ሰውሆነ': 88, 'ሀፍረትለበሰ': 89, 'ቀንጫልቅለቃ': 90,
```

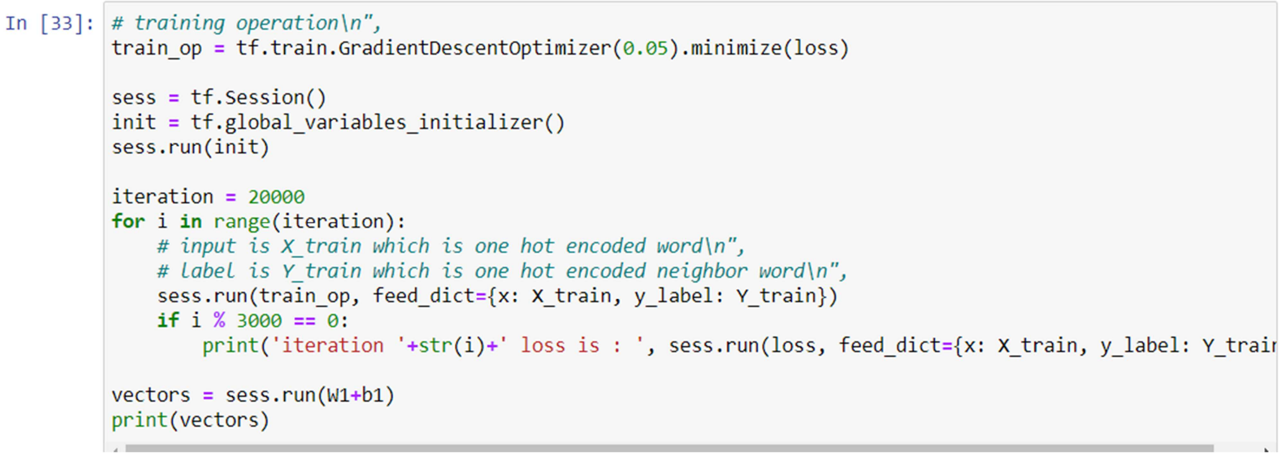
Figure 2. *Representations from words to integers.*

5. Results and Discussion

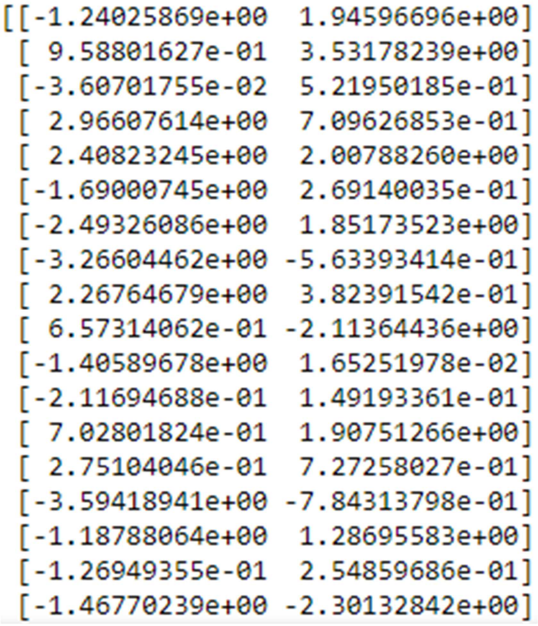
We used one-hot encoding to represent expressions of vectors in N-dimensional space and the word2vector model to generate the numeric value of the expression. The expressions are transformed into numeric values for NLP application researches and machines. For our task, we represent the expression with integer values using the *word2int* algorithm.

As shown in Figure 2, the word2int technique is utilized to transform expressions to integers. Each expression is represented by a single neighbor's expression with a window size of one, which means that one neighbor expression assigns an integer value to each expression.

On the other hand, we employed vector representation to represent N-dimensional spaces, which required us to transform each integer into a vector. We used a one-hot encoding approach to convert integers to vectors. When there are N input expressions, one hot encoding representation is represented in N-dimensional space.



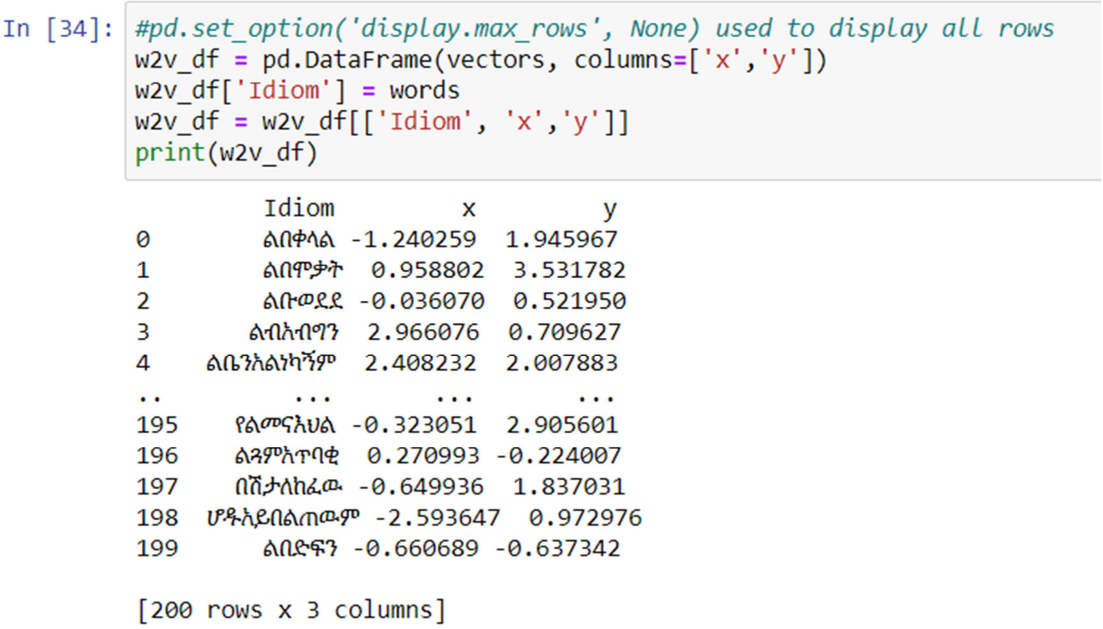
```
In [33]: # training operation\n",
         train_op = tf.train.GradientDescentOptimizer(0.05).minimize(loss)

         sess = tf.Session()
         init = tf.global_variables_initializer()
         sess.run(init)

         iteration = 20000
         for i in range(iteration):
             # input is X_train which is one hot encoded word\n",
             # label is Y_train which is one hot encoded neighbor word\n",
             sess.run(train_op, feed_dict={x: X_train, y_label: Y_train})
             if i % 3000 == 0:
                 print('iteration '+str(i)+' loss is : ', sess.run(loss, feed_dict={x: X_train, y_label: Y_trair

         vectors = sess.run(W1+b1)
         print(vectors)
```



```
[[-1.24025869e+00  1.94596696e+00]
 [ 9.58801627e-01  3.53178239e+00]
 [-3.60701755e-02  5.21950185e-01]
 [ 2.96607614e+00  7.09626853e-01]
 [ 2.40823245e+00  2.00788260e+00]
 [-1.69000745e+00  2.69140035e-01]
 [-2.49326086e+00  1.85173523e+00]
 [-3.26604462e+00 -5.63393414e-01]
 [ 2.26764679e+00  3.82391542e-01]
 [ 6.57314062e-01 -2.11364436e+00]
 [-1.40589678e+00  1.65251978e-02]
 [-2.11694688e-01  1.49193361e-01]
 [ 7.02801824e-01  1.90751266e+00]
 [ 2.75104046e-01  7.27258027e-01]
 [-3.59418941e+00 -7.84313798e-01]
 [-1.18788064e+00  1.28695583e+00]
 [-1.26949355e-01  2.54859686e-01]
 [-1.46770239e+00 -2.30132842e+00]
```

Figure 3. *A hidden layer weight value.*

The aforementioned Figure 3 illustrates how the input dataset is encoded into a numeric value by producing an N-number of hidden layer weight values. For 2D visualization, we utilized a window size of one and an embedding dimension of two. Word length is taken into account in one-hot encoding, and we used an embedding dimension of two. The weight value of the one-hot encoding and embedding dimension is its random normal value. To get the hidden layer value, we multiplied the input vector by the weight value in a matrix. To get the output value, the value of the hidden layer is utilized as an input. We employed SoftMax to update the weight value through cross-entropy optimization of the output vector. The model enters the created weight value into the lookup table and multiplies it by the vector value of the supplied input to obtain the expression's output value.



```
In [34]: #pd.set_option('display.max_rows', None) used to display all rows
         w2v_df = pd.DataFrame(vectors, columns=['x','y'])
         w2v_df['Idiom'] = words
         w2v_df = w2v_df[['Idiom', 'x','y']]
         print(w2v_df)

             Idiom         x         y
0           ልበቀላል -1.240259  1.945967
1           ልበሞቃት  0.958802  3.531782
2           ልቡወደደ -0.036070  0.521950
3          ልብአብግን  2.966076  0.709627
4       ልቤንአልነካኝም  2.408232  2.007883
..           ...       ...       ...
195      የልመናእህል -0.323051  2.905601
196      ልጓምአጥባቂ  0.270993 -0.224007
197      በሽታለከፈዉ -0.649936  1.837031
198   ሆዱአይበልጠዉም -2.593647  0.972976
199         ልበድፍን -0.660689 -0.637342

[200 rows x 3 columns]
```

Figure 4. *The expressions' numerical value.*

This expression's numeric value was obtained by multiplying the hidden layer value by the weight values in a matrix; the weight value is the random normal value of the one-hot encoding and embedding dimension [1, 14]. As can be seen in Figure 4 above, the expressions are numerically described in two-dimensional space using six floats.

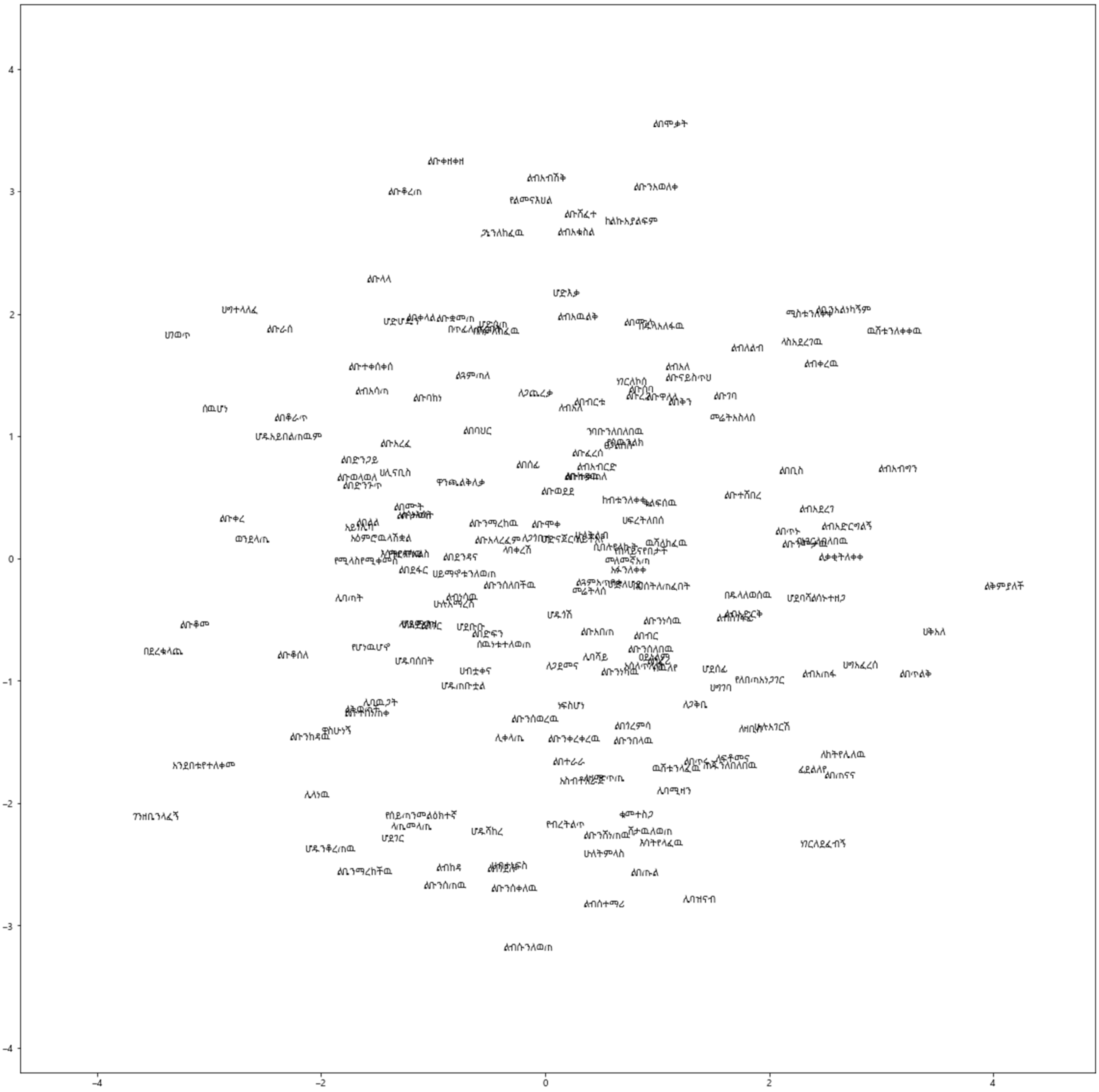


Figure 5. *2D visualization of the expressions.*

Using the KNN supervised machine learning algorithm, the model achieves an accuracy of 95.5% overall on the tested dataset.

6. Conclusions and Recommendations

An Amharic idiom is expressions that are difficult to process on machines and negatively affect natural language processing researchers like machine translation, idiom identification, sentiment analysis, next-word prediction, and information extraction. Vector representations are a key method that bridges the human understanding of language to that of machines and solves many NLP problems. For NLP applications, machine learning algorithms, and deep learning algorithms, representation of idiomatic expressions is crucial. We represent Amharic idiom expressions with numeric values for the NLP applications and machines. We used one hot encoding method and word2vector model to represent the input expression into vectors and represent the expressions into numeric values and show an N-dimensional space. We used a two hundred expressions dataset from the Amharic idiom book. When we used the KNN algorithm, we got a 95.5% accuracy result for the embedding model of Amharic idioms. To mitigate the negative effects of idioms on NLP research, the study suggests that scholars and industry professionals utilize these represented datasets for their research.

ORCID

0009-0001-1853-6984

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix

Sample Dataset

['ህብተነፍስ ህብቷቀና ሁለትምላስ ሁሉአማረሽ ሁሉአገርሽ ህሊናቢስ ህቅአለ ህግተላለፈ ህግአፈረሰ ህገወጥ ህግገባ ሰዉሆነ', 'ልቡተሰነጠቀ ነፍስሆነ ዋስሁነኝ አስብቶአራጅ የሆነዉሆኖ ቁመተስጋ ሆደመጋዝ ሆደሰፊ ሆደቡቡ ሆደባሻ ሆደገር ሆዱሻከረ’, ሆዱንቆረጠዉ ሆዱባሰበት ሆዱጠቡቷል ሆዱአይበልጠዉም ሆዱጎሸ ሆድሆዴን ሆድለሆድ ሆድሰጠ ሆድናጀርባ ሆድእቃ’, መለመኛአጣ የልመናእህል ልሳኑተዘጋ ዋንጫልቅለቃ ልቅምያለች አንደበቱየተለቀመ ልቃቂትለቀቀ ልቅወጣች ሚስቱንለቀቀ', አፉንለቀቀ ዉሸቱንለቀቀዉ ከብቱንለቀቀ በነገርለበለበዉ ንባቡንለበለበዉ ጠጁንለበለበዉ ህፍረትለበሰ ልብሰተማሪ', ልብስገፋፊ ፀጋልበስ የለበጣአነጋገር ለብአለ ለከትየሌለዉ በሽታለከፈዉ ዉሻለከፈዉ ጋኔንለከፈዉ የሰዉንልክ', ከልኩአያልፍም ቁልፍሰዉ ነገርለኮሰ በዱላለወሰዉ ህይማኖቱንለወጠ ልብሱንለወጠ ሰዉነቱተለወጠ ሽታዉለወጠ', ለዛሙጥጤ ለዛቢስ ሰዉለየ ፊደልለየ ለይቶአየ ነገርለደፈብኝ ለጋጎበዝ ለጋቅቤ ለጋደመና ለጋጨረቃ ልጓምአጥባቂ ልጓምጣለ', በሀሰትለጠፈበት በጥፊለጠፈበት ለፍቶመና በዱላአለፋዉ ዐይነልም ላስአደረገዉ መሬትላሰ መሬትአስላሰ ማርምአልስ', እሳትየላሰዉ የሚላስየሚቀመስ አዕምሮዉላሸቋል ላባቀረሽ ላክአደረገ ሲበሉየላኩት የሰይጣንመልዕክተኛ የበላይናየበታች', ላጤመላጤ ሊቀላጤ ወንደላጤ የብረትልጥ በደረቁላጨ እሳትየላፈዉ ገንዘቤንላፈኝ ዉሸቱንላፈዉ ሌላነዉ ሌባሚዛን', ሌባዉጋት ሌባዝናብ ሌባጣት አሰለጥሌባ አይነሌባ ሌባሻይ ልሳነእሳት ሁለትልብ ልበልል ልበሙሉ ልበሙት ልበሞቃት', ልበሰፊ ልበቀላል ልበቅን ልበቆራጥ ልበቢስ ልበባህር ልበብር ልበብርቱ ልበተራራ' ልበደንዳና ልበድንጋይ ልበደፋር', ልበድንጉጥ ልበድፍን ልበገር ልበጎረምሳ ልበጎደሎ ልበጠናና ልበጡል ልበጥልቅ ልበጥሩ ልበጥኑ ልበፈሪ ልቡአበጠ', ልቡንአወለቀ ልቡሞቀ ልቡሰባ ልቡላላ ልቡረጋ ልቡራሰ ልቡሸፈተ ልቡቀረ ልቡቀዘቀዘ ልቡቆመ ልቡቆረጠ ልቡቆሰለ ልቡቋመጠ']

References

[1] A. A. F. &. S. Gebeyehu, "Automatic Idiom Identification Model for Amharic Language," ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22, 8, Article 210 https://doi.org/10.1145/3606864, p. 9, 2023.

[2] G. D. Salton, "Representations of Idioms for Natural Language Processing: Idiom type and token identification, Language Modelling and Neural Machine Translation," Doctotal thesis, DIT, 2017. doi.org/10.21427/D77H8K, 2017.

[3] A. A. &. D. Worku, Amharic Idioms 2nd edition, Addis Abeba, Ethiopia: Kuraz publishing Agency, 1992.

[4] J. P. &. A. Feldman, "Automatic Idiom Recognition with Word Embeddings," in In: Lossio-Ventura, J., Alatrista-Salas, H. (eds) Information Management and Big Data. SIMBig SIMBig 2015 2016. Communications in Computer and Information Science, vol 656. Springer, Cham. https://doi.org/10.1007/978-3-319-55209-5_2, 2017.

[5] J. P. &. A. Feldman, "Experiments in Idiom Recognition," in In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2752–2761. The COLING 2016 Organizing Committee, Osaka, Japan, 2016.

[6] R. R. J. K. Giancarlo Salton, "Idiom Token Classification using Sentential Distributed Semantics," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) doi 10.18653/v1/P16-1019, Berlin, Germeny, 2016.

[7] A. M. A. Y. M. H. R. M. R. &. A. A. Mohamed A. Zahran, "Word Representations in Vector Space and their Applications for Arabic," in International Conference on Intelligent Text Processing and Computational Linguistics: Computational Linguistics and Intelligent Text Processing pp 430–443, 2015.

[8] G. W. &. X. Z. Lei Zhu, "A Study of Chinese Document Representation and Classification with Word2vec," in 2016 9th International Symposium on Computational Intelligence and Design (ISCID) DOI: 10.1109/ISCID.2016.1075, 2016.

[9] G. A. P. J. &. T. K. Yash Sharma, "Vector representation of words for sentiment analysis using GloVe," in 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT) DOI: 10.1109/INTELCCT.2017.8324059, 2017.

[10] Y. Z. Joseph Lilleberg & Yun Zhu, "Support Vector Machines and Word2vec for Text Classification with Semantic Features," in Proc. 20151IEE 14th Internations conference on Cognitive Inlormatics & Cognitive Computing, 2015.

[11] Q. V. L. &. I. S. Tomas Mikolov, "Exploiting Similarities among Languages for Machine Translation," Cornell University arXiv: 1309.4168v1 [cs.CL], 2013.

[12] K. Grzegorczyk, "Vector representations of text data in deep learning," Cornell University arXiv: 1901.01695v1, 2019.

[13] G. T. &. T. A. Abebawu Eshetu, "Learning Word and Sub-word Vectors for Amharic (Less Resourced Language)," International Journal of Advanced Engineering Research and Science, vol. 7, no. 8, 2020.

[14] A. Abebe, "Automatic Idiom identification Model for Amharic language," ir.bdu.edu.et, Bahir Dar, 2021.

Biography

Anduamlak Abebe Fenta: BSC in Computer Science from Arba Minch University, MSC in Computer Science from Bahirdar University. Currently working as lecturer and researcher in Gafat Institute of technology, Department of Computer Science, Debre Tabor University, Ethiopia. He has participated on conference entitled as “IoT Enabled Smart Water Management System” with relatives and he has publication entitled as “Automatic Idiom Identification Model for Amharic Language”. His MSc is focused on the Amharic idiom identification using machine learning. His main research interest is in Natural Language Processing, Machine learning, Cyber Security, and Artificial Intelligence.