Journal of Positive School Psychology
2022, Vol. 6, No. 5, 5482–5495

Hybrid Model to Word Sense Disambiguation for Hadiyyisa Language Using Supervised Machine Learning Models

Abraham wolde

Abstract

The process of determining the most appropriate interpretation of an ambiguous word means Word Sense Disambiguation. WSD has been numerous key applications compared to other NLP applications such as Text Summarization, Text Categorization, and Question Answering. The basic task of this study has been to design the quality and technical word sense disambiguation for Hadiyyisa text. The data collection and data set preparation from different licensed areas was been induced. From the collected data, pre-processing techniques like TF-IDF were been used to convert text into the vector form. In this study Hybrid model and Supervised ML Models have been introduced. These Supervised ML models such as Naive Bayes, Neural Network, Support Vector Machine, and Hybrid Model were been used. In addition to this, to validate this model, an appropriate method adopted like Monte Carlo Cross-Validation (Shuffle Split) was been verified. The result was been found in two models such as SVM, and Hybrid Model. The accuracy of the SVM and Hybrid model for the selected ambiguous words such as Anga, Diinate, Hagara, Hurbaata, Misha and Seera respectively for both models were been investigated. The average accuracy of the model for SVM and Hybrid Model were been verified with their result of 79% and 82% respectively.

Keywords— Natural Language Processing, Supervised ML, TF-IDF, and Word Sense Disambiguation

Introduction

The process of determining the most appropriate interpretation of an ambiguous word means Word Sense Disambiguation (WSD) (Chaplot & Salakhutdinov, 2018)(Escudero et al., 2000). It is one of the challenging areas in the NLP and establishing the works from begging to back MT research projects. Machine translation (MT) was a research project and still difficult for WSD advances MT and discouraged researchers (Popov, 2018). The ambiguity of words in natural languages is difficult to interpret the senses. Depending on the open-class words also the words have more than one meaning either Polysemy (the presence of distinct but coupled senses corresponding to the similar word) or homonym (the phenomenon of semantically separate ideas expressed with the same words (Popov, 2018)(Li & Suzuki, 2020)(Bhadane et al., 2021).

For developers of Natural Language Processing Systems, word sense ambiguity is an unbreakable bad-behaved (Chaplot & Salakhutdinov, 2018). Words take on different meanings in different situations and it remains an outstanding challenge in NLP. Humans have a broad and sensible understanding of the real world, which provides them with the necessary knowledge to make sense of it(Li & Suzuki, 2020). In most circumstances, computers should make the disambiguation decisions without difficulty. However, computers are the ability to understand the environment and make judgments regarding words (Mukti Desai, 2013)(Bhadane et al., 2021)(Chen, 2019).

Many researchers have conducted on word sense disambiguation in different languages using different types of methodologies. One of the better resources to disambiguate the words is WordNet (TESEMA, 2016) (Hassen, 2015)(Chen, 2019) (Nithyanandan & C, 2019).

To improve WSD approaches have used labeled data or sense annotated data based on supervised machine learning techniques (Han & Shirai, 2021) (Bhadane et al., 2021).

Finally, this study has focused on the Hadiyyisa languages. It is one of the Cushitic languages and low resource language. This study addresses the lexical an ambiguous words in the Hadiyyisa language. For example, in the Hadiyyisa language, **Diinate** is ambiguous word. "Lambaami ulluminse **diinate** mine agisukko." In this sentence, the meaning of Diinate is "**Hoora (animals)**". "Lambaami banki **diinate** fisukko." In this sentence the meanings of **Diinate** is "**biira (money)** "for these purposes how to give the correct sense of meaning depending on the context of words.

2. Materials and Methods

Research methods

Annotated corpus or context-based repository techniques: - is one of the more difficult aspects of WSD (word Net). For that reason, data has be collected from several sources, including the Hadiyyisa Bible, Hadiya Zone Education bureau and Wachemo University teaching materials. Collect a number of statements or examples. It also uses feature vectors to represent linguistic evidence or information. It is not necessary to annotate these feature sets for all sentence circumstances (Rahman & Borah, 2021) (Alemu & Fante, 2021)(ASSEMU, 2011).

2.1. Data Collection and Data Preparation

2.1.2. Sources of Data

To collect the data using two main techniques have used for survey of full information about Data set and ambiguous word linguistic expert persons. Data is collected using two methods: primary data collection and secondary data gathering approaches.

Primary Data collection: - To collect the data by interviewing the necessary information about WSD in Selected domain areas like Wachemo University Hadiyyisa Department, Hadiya zone Education Bureau and Hadiya Zone Culture and Tourism Bureau. In this method, try to interview different NLP researchers how to collect and design effective WSD data set and corpus preparation. In this data collection phase the most people probable use for different communication purposes, which is ambiguous words should been selected from Hadiya Zone Culture and Tourism Bureau.

Secondary Data collection: - To collect the data by using document analysis, Related literature review and appropriate Research methods. Based on the above techniques to collect data from these data sources. **Academic documents and student learning textbooks**: - form this data source to collect academic data from the Hadiyyisa Language and Literature Department modules and primary textbooks. Legal paperwork Offline data sources include academic books (grade 1–12 Hadiyyisa language textbooks and modules of higher education), spiritual offline materials, government office norms and guiding principles. **Online Data Sources: -** To acquire data that was previously stored on an internet domain and was freely available. Researchers got this information from free and open source MT platforms. **Hadiya Zone Cultural Bureau and Tourism Office**: - especially since the majority of people are likely to use it for genuine conversation.

Annotation of Data

Data annotation is the process of classification text data. Obtain quality data while maintaining the nature and behaviors of Hadiyyisa language texts.

Table 3. 1 Data Annotation Result of Word Sense Disambiguate for Hadiyyisa Language

No	Ambiguous word	Sense 1	Sense 2	Total senses
1.	Anga(n)	Haraa'mmima(favour) =103	Qaamafeeta(hand)= 184	287
2.	Misha(n)	Hamaama haqqi	Siixo'o tie'm	212

		misha(Fruit)=58	atooota(Product or result)=154	
3.	Diinate(n)	Mine Hoora(domestic animals)=105	Birra (Money)=63	168
4.	Hurbaata(n)	Ichcha(food)=222	Sire'e(Seed)=135	357
5.	Hagara(n)	(colour) moo'akkam haga'l =70	(type, kind) annanni annanni mu'uta, qocca baxo=132	202
6.	Seera(n)	Wi'lonne hawwone hara'manacha(Help organization)=87	Sono'o, ogora(rule, law)=204	291
				Total sense =1,516

Data Set Preparation

The size of the data collection influences the quality of the disambiguate senses, with the general rule being that more data is better. However, statistics for a low-resource language like Hadiyyisa is extremely difficult to come by (TESEMA, 2016)(Eshetu et al., 2020). In this, research how to prepare the data set for new WSD for Hadiyyisa Language. By following in this process has discussed as follows.

Context

You can use this keep fit to test a new WSD technique or NLP system that requires WSD if you are working on one. There are now only a few keep fit options available for analyzing WSD approaches. These data collections have created to make the evaluation technique easier and faster. There are 1516 sentences in this data collection, each with two polysemy terms.

Content

The dataset is an excel spreadsheet (.xlsx). Serial number (SN), sentence/context column, and polysemy word are the three columns. The first and third columns are particularly handy. The sentences are in the second column (context). There could be more polysemy terms in the statement. As a result, the target word (polysemy term) in the second column is required in the third column. Use the sentence to see if your WSD system can distinguish the target word in the third column from the equivalent sentence in the second column.

Data Analysis

Depending on the study scope, various kinds of ambiguity may arise in the suggested investigations. That includes lexical, phonological (sound) ambiguity, as well as referential and semantic ambiguity (ASSEMU, 2011). We select lexical ambiguities have used in this study. Introduce sense at the sentence (instance) level to provide context meanings for terms in lexical ambiguity.

Data Preprocessing

Preprocessing - Tags removal, tokenization, character normalization, stop words removal, and punctuation mark removal are examples of data preprocessing phases. This preprocessing phase has also in Hadiyyisa language it could be used.

Data Cleaning: Data cleaning removes extraneous information that could impair the model's performance. Special letters, punctuations, and symbols have commonly seen in data acquired from many sources. This has cleaned up by using Data preprocessing techniques to remove any unnecessary features. It is also helpful for improving model accuracy. Pseudo code for cleaning Hadiyyisa text shown below:

Input: Uncleaned Hadiyyisa data set

Out Put: Cleaned Hadiyyisa data set

Step 1: Start or initialization

Step 2: Read uncleaned data set

Step 3: If the data set which contains punctuation marks then

A) Remove the punctuation marks
B) Display cleaned data set without punctuation marks

Step 4: Else if the data set which contains sentence

- Tokenize the data set into words
- Display the tokenized data set

Step 5: Else if the data set which contains stop words

A) Remove stop words from the tokenized data set
B) Display without stop words

Step 6: Display cleaned data set

Step 7: Stop

Tags Removal

If we are scraping data from another website, eliminating HTML tags is a must-do step in the preparation process.

Tokenization

It is the process of segmenting of string of characters into words. It divides the raw text into tokens, which are words or sentences. It is use to deduce the meaning of a text by looking at the order of the words. This is an example of a tokenization process using the NLTK Toolkit (python). As a result, several challenges for word tokenization to break sense examples into tokens have discussed in this thesis. "Ergoge waasa hurbaata ito'o," for example. This statement is tokenized as 'Eregoge,' waasa,'hurbaata,' itoo' in Hadiyyisa Language in the right manner.

Stop words Removal

They are any words in a stop list (or negative dictionary) that have filtered out before or after NLP and are often utilized in natural language writings. However, essential for NLP model training. By removing extraneous terms from the dataset, data preparation increases model performance. We used sentences-based window widths to load sentences from datasets in our research. The words that have typically filtered out before processing natural language. These are the most common words in any language (articles, prepositions, pronouns, conjunctions, and so on) and do not add anything to the text. Because there are fewer tokens involved in the training, removing stop words minimizes the dataset size and hence reduces training time. The algorithm is implemented in the steps shown below (Raulji, 2016).

Step 1: Tokenize the target document text and save individual words in an array.

Step 2: From the stop word list, a single stop word has read.

Step 3: Using a sequential search strategy, the stop word has compared to the target text in the form of an array.

Step 4: If they match, the word in the array has eliminated, and the comparison has repeated until the array length has reached.

Step 5: After entirely removing the stop word, another stop word has read from the stop word list, and the algorithm repeats step 2. The algorithm keeps running until the entire stop words have compared.

Step 6: The final text, free of stop words, has displayed, along with the appropriate data such as stop word removal. Because our datasets have created using MS Excel, removing stop words have not modify the position of the ambiguous word.

Normalization

In NLP, normalization is the process of translating texts into a standard format. Some characters in the same words in the corpus and user input are sometimes rendered in uppercase or lowercase in this study, so we have standardized them to lowercase. Why should we normalize? The form of indexed text and query phrases must be the same. M.K. and MK, for example, should been matched. The goal of normalization in this situation is to make the words in this corpus similar in diverse cases (TESEMA, 2016).

Punctuation Mark Removal

This is a crucial text preparation step because it adds no value to the data. A text standardization method allows terms like ‘some’, ‘some,’ and ‘some’ to be treated in the same way.

Punctuation (or interpunction) is the use of spacing, conventional signs (called punctuation marks), and some typographical elements to promote comprehension and accurate reading of written text, whether read silently or aloud. "It is the practice, action, or system of adding points or other small marks into texts in order to facilitate understanding; division of text into sentences, phrases, and other units by means of such marks," according to another definition.

Feature Extraction Techniques

TF-IDF Vectorizer

TF-IDF stands for term frequency-inverse document frequency. It calls attention to a particular issue that, while not prevalent in our corpus, is crucial. The TF–IDF score increases with the number of times a word appears in the document and decreases with the number of documents in the corpus that contain the word. TF-IDF means a near algorithm that uses the occurrence of the words to define how essential those words are to a given documents. It has separated into two sections:

1. Frequency of Term (TF)
2. Inverse Document Frequency (IDF)

Term Frequency (TF)

Term frequency refers to how many times a term appear in the document. It might been compared to the likelihood of discovering a word within a document. It has written as follows:

$$TF_{(t,d)} = \frac{n_{t,d}}{n_d} \quad (3.1)$$

where $TF_{(t,d)}$ is the number of terms t in the document $n_{t,d}$ is the number of occurrences of term t in the document d. whereas n_d is the sum of the terms in the document d (Faisal, 2018).

Inverse Document Frequency (IDF)

The inverse document frequency determines whether a term is rare or common over the entire corpus of texts. It emphasizes phrases that exist in a small number of texts across the corpus or words with a high IDF score in plain English. Divide the total number of documents in the corpus by the number of documents containing the phrase to get the logarithm of the overall word(Faisal, 2018).

$$IDF_{(t,d)} = \log\left[\frac{(1+n)}{(1+df(t,d))}\right]+1 \quad (3.2)$$

Where,

$IDF_{(t,d)}$ Is the frequency of the terms t in document D?

$df(t,d)$ Is the total number of documents in the dataset?

The value of IDF (and consequently tf–idf) is greater than or equal to zero since the ratio inside the IDF's log function must always be bigger than or equal to one.

The ratio inside the logarithm approaches one when a phrase appears in a high number of documents, and the IDF approaches 0.The product of Term Frequency and Inverse Document Frequency (TF-IDF) is TF-IDF. It has written as follows:

$$TF-IDF_{(t,d)} = TF_{(t,d)} * IDF(t,d)$$

The TF-IDF score of a phrase with a high frequency in a document but a low document frequency in the data set is high. For a word that appears in almost all publications, the IDF value approaches 0, pushing the TF-IDF closer to zero. The TF-IDF value is high when both the IDF and TF values are high, indicating that the phrase is uncommon within the document yet prevalent within it.

Model Selection

The purpose of this study is to construct a WSD model for Hadiyyisa Language using supervised machine learning techniques. Due to the lack of Hadiyyisa WordNet and annotated datasets, this study has focus on six ambiguous words. These words:-**Anga**[1](favour) **Anga**[2](Hand), **Misha**[1](fruit) **Misha**[2](product or result) ,**Hurbaata**[1](food) **Hurbaata**[2](crop),**Hagara**[1](colour) **Hagara**[2](type or kind)**Diinate**[1](domestic animals)**Diinate**[2](money) and **Seera**[1](help organization)**Seera**[2](rule or law). These methodologies have used for the suggested research to customize Supervised Machine Learning models: - Neural Network, Support Vector Machine, Nave Bayes, and Hybrid model. In this research after define the result of experiment to decide two models, which are appropriate to handle Hadiyyisa text.

Support Vector Machine

There are several classification algorithms used in machine learning, however **SVM** is better than most of them since it produces more accurate results. For that case we select our study to perform SVM Classifier has a lower computational complexity than other classifiers, and it should been used even if the number of

both senses of instances are not equal, because it can normalize the data or project into the space of the decision border separating the two classes. In addition to, SVM performs admirably when there is a clear separation between classes**.** To perform SVM into our model by following Flow chart procedures in the Figure 1 Below.

Hybrid Models

This hybrid model includes all deterministic rules, allowing us to train it in the same way as any other machine-learning model. Fortunately, Scikit-learn provide a Base Estimator class that we can inherit to quickly create our own scikit-learn models. Only slight gains in prediction accuracy are seen by hybrid approaches (SVM, NN, NB) when compared to standard understandable methods and to integrate by using **Voting Classifier** method (Miškovic, 2014).

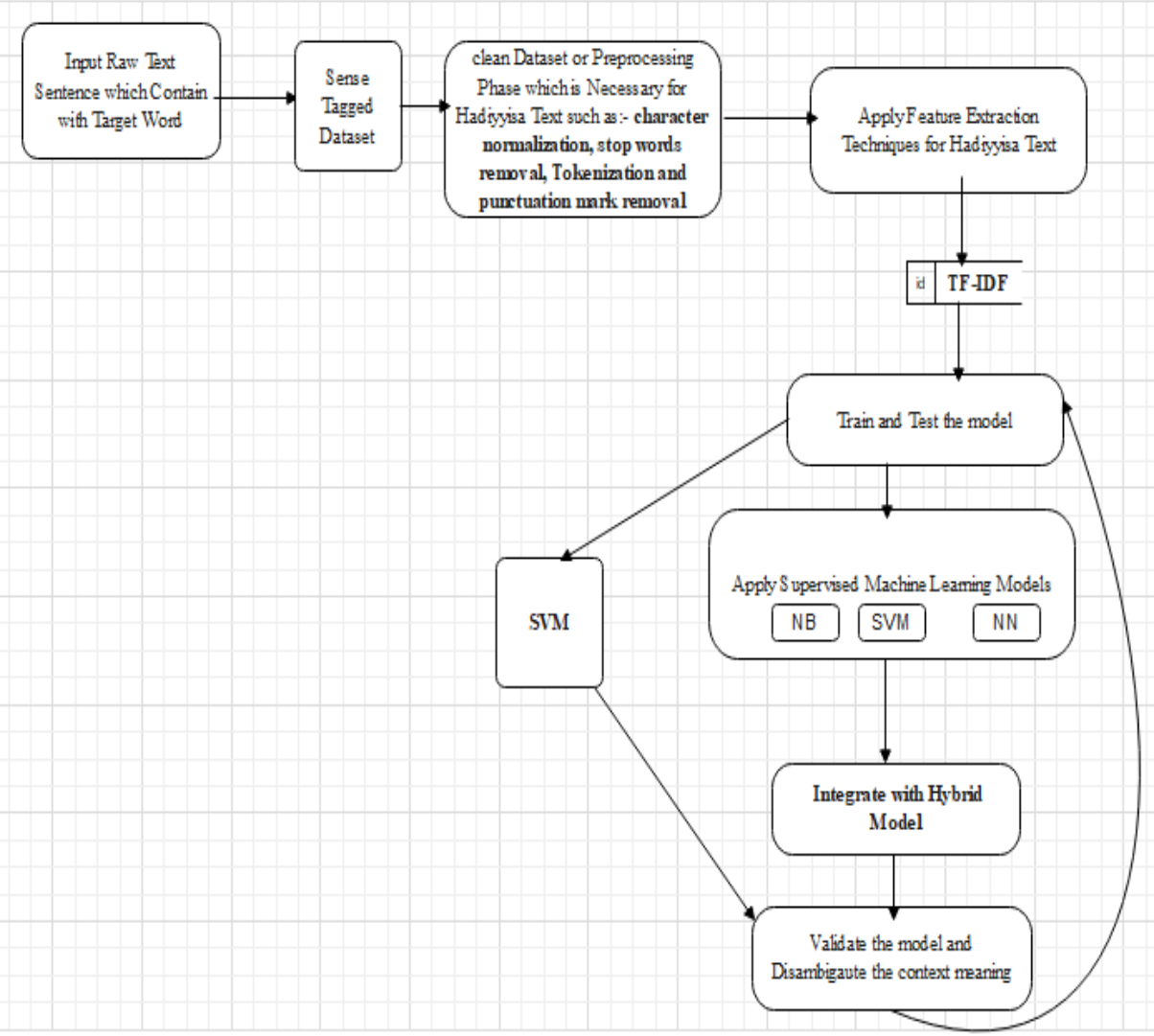


Figure 1 Proposed model for WSD

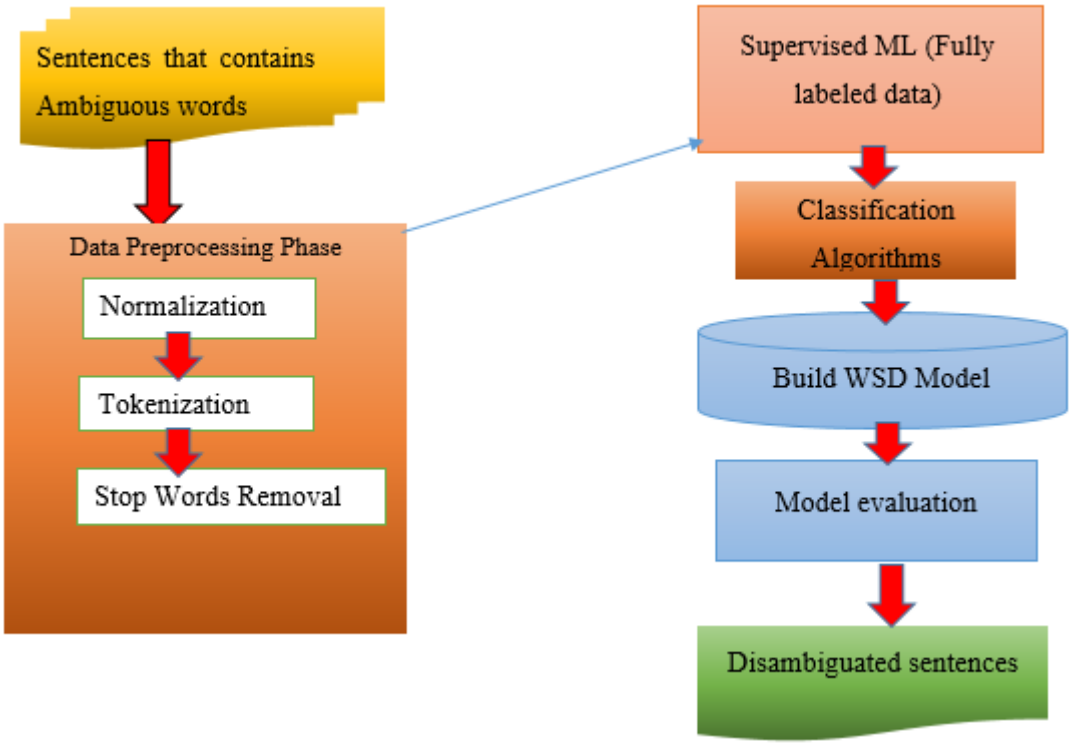


Figure 1 WSD System Architecture for Hadiyyisa Language

Evaluation Metrics

Recall, precision, and F-measure have commonly used to assess the performance of classification algorithms.

Recall is a percentage number that indicates how many right results have found out of a set of results from a processed document have detected (based on the expectations of a certain application).

Recall = the number of CorrectlyDisambiguated Words / the number of Total number of Ambiguous Words

Precision is a percentage metric that indicates how many of a set of outcomes from a processed document are right (based on the expectations of a certain application).

Precision = the number of CorrectlyDisambiguated Words / the number of Disambiguated Words

The **F-Score** is the harmonic mean of a system's Precision and Recall scores. A somewhat high F-score can be the result of an imbalance between Precision and Recall, according to critics of using F-score values to judge the quality of a prediction system.

$$F-Score = 2 \text{ x } \left[(Precision \text{ x } Recall) \right] / (Precision + Recal$$

Accuracy means the performance factors of the number of correctly classified and incorrectly, categorized cases retrieved from the output confusion matrix.

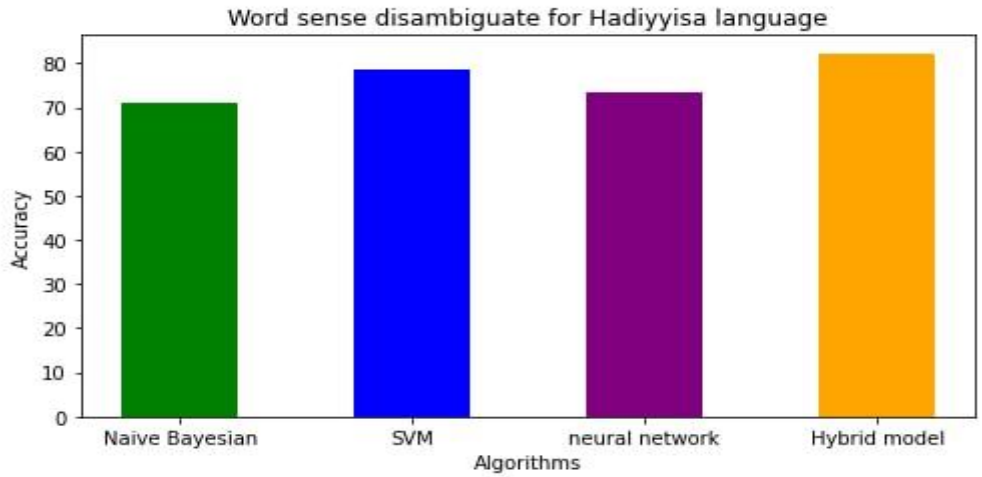
RESULT AND DISCUSSION

This part, discusses about the overall results and discussion of the Hadiyyisa language WSD study. This chapter also covers the fundamental tasks of our research, such as the implementation of WSD to handle ambiguous words, feature extraction mechanism and the evaluation metrics for WSD. Take as a starting point a paper with accuracy and technique, even if language habits have analyzed.

Accuracy of Models

In this section discuss about the four selected models such as SVM, NB, NN and Hybrid Model and based on the accuracy.


```
Naive Bayes accuracy_score: 0.7105263157894737
Neural Network accuracy_score: 0.7335526315789473
Support Vector Machine accuracy_score: 0.7861842105263158
Hybrid Model accuracy_score: 0.8223684210526315
```



Classification Report and Confusion Matrix

In this part, discuss about classification report and confusion matrix for above compare the result of accuracy for selected four classifier models such as SVM, NB, NN and Hybrid Model and based on the accuracy. We select two better models such as SVM and Hybrid models to handles WSD for Hadiyyisa language and describe both models deeply based on evaluation metrics and cross validation methods such as Monte Carlo Cross-Validation (Shuffle Split) is appropriate to validate actual accuracy of the models. Machine learning algorithms use classification reports as one of their performance evaluation indicators. It has used to display our trained classification models' Precision, Recall, F1-Score, and Support. **Precision** means the ratio of true positives to the sum of true and false positives. **Recall** means the ratio of true positives to the sum of true positives and false negatives. The weighted harmonic mean of Precision and Recall is the **F1-Score**. The amount of actual instances of the class in the data set has referred to as **support**. The macro average between the classes is the average of Precision, Recall, and F1-Score. Between the classes, the weighted average of Precision, Recall, and F1-Score is calculated. Monte Carlo Cross-Validation (Shuffle Split) is

an extremely bendable procedure of cross-validation. By using the report accuracy results on the test set training set and for parameter tuning. There in technique, the datasets predispose willy-nilly divided into experiencing and establishment sets.

	precision	recall	f1-score	support
birra	0.80	0.27	0.40	15
hamaama	1.00	0.36	0.53	11
haraa'mmima	0.91	0.45	0.61	22
ichcha	0.73	0.93	0.82	43
mine hoora	0.62	0.95	0.75	21
moo'akkam haga'l annannooma	1.00	0.38	0.55	8
qaamafeeta	0.76	0.95	0.84	37
qocca baxo	0.78	1.00	0.88	18
siixo'o	0.82	1.00	0.90	31
sire'e	0.86	0.48	0.62	25
sono'o	0.84	0.96	0.90	55
wi'lonne hawwone hara'manacha	0.90	0.50	0.64	18
accuracy			0.79	304
macro avg	0.83	0.69	0.70	304
weighted avg	0.81	0.79	0.76	304

```
cross Validation scores:ShuffleSplit[0.84210526 0.79605263 0.78618421 0.80263158 0.78289474 0.80263158
 0.8125     0.79934211 0.79276316 0.79934211]
Average Cross Validation score of ShuffleSplit:0.8016447368421054
Wall time: 500 ms
```

```
Out[9]: <AxesSubplot:xlabel='actual', ylabel='predicted'>
```

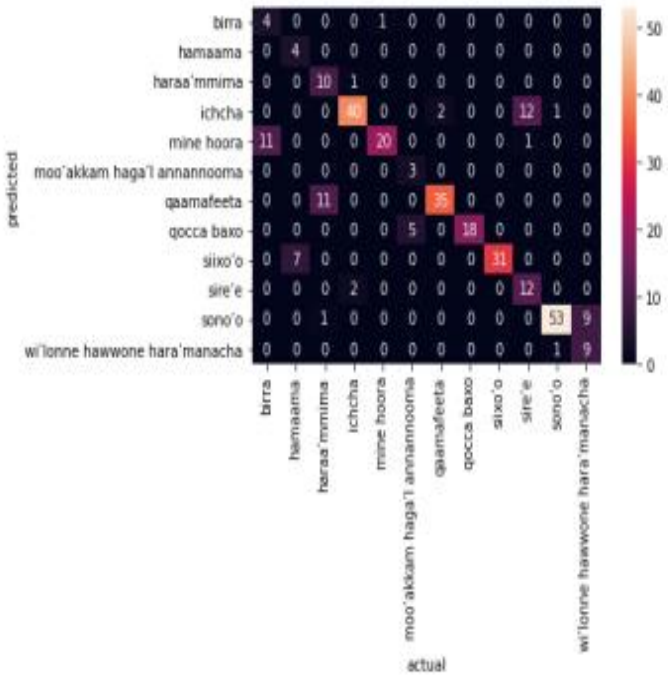


Figure 4. 1 Support Vector Machine Classification report for multi class

The confusion matrix and confusion report for the target word, such as multi class, have shown in the above Figure 4.21. In generally, **65** data points have incorrectly predicted and **239** data points have correctly predicted out of **304** tested data point. By using **Shuffle Split** cross validation test score is almost near to actual accuracy result

	precision	recall	f1-score	support
birra	0.62	0.53	0.57	15
hamaama	0.80	0.36	0.50	11
haraa'mmima	0.86	0.55	0.67	22
ichcha	0.73	0.95	0.83	43
mine hoora	0.74	0.81	0.77	21
moo'akkam haga'l annannooma	0.88	0.88	0.88	8
qaamafeeta	0.80	0.95	0.86	37
qocca baxo	0.94	0.94	0.94	18
siixo'o	0.76	0.90	0.82	31
sire'e	0.87	0.52	0.65	25
sono'o	0.96	0.93	0.94	55
wi'lonne hawwone hara'manacha	0.94	0.94	0.94	18
accuracy			0.82	304
macro avg	0.82	0.77	0.78	304
weighted avg	0.83	0.82	0.81	304

```
cross Validation scores:ShuffleSplit[0.83552632 0.79276316 0.81907895 0.83881579 0.80592105 0.80263158
 0.83881579 0.79605263 0.80921053 0.78947368]
Average Cross Validation score of ShuffleSplit:0.812828947368421
```

Out[24]: <AxesSubplot:xlabel='actual', ylabel='predicted'>

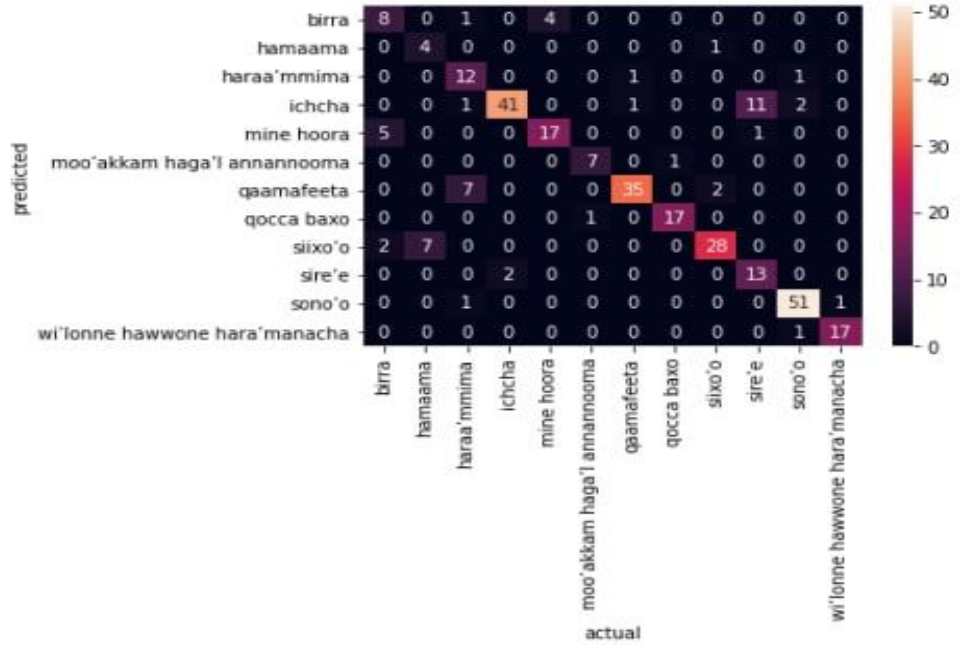


Figure 4. 2 Hybrid Model Classification report for multi class

The confusion matrix and confusion report for the target word, such as multi class, have shown in the above Figure 4.22. In generally, **57** data points have incorrectly predicted and **247** data points have correctly predicted out of **304** tested data point. By using **Shuffle Split** cross validation test score is almost near to actual accuracy result.

4.1.1. Contribution of the Study

The main outcomes or Experimental Outcomes of the study are as follows:

1. Create the required data-preprocessing phase for the Hadiyyisa Language. Stop word removal, lower letter removal, punctuation removal, number removal, and tokenization are some examples.
2. Design WSD Data Set preparation.
3. Customize WSD Model for Hadiyyisa Language.
4. The result for Six ambiguous words to give their contextual sense or meanings

The main contribution of this thesis is to investigate WSD model, for Hadiyyisa language using supervised Machine Learning methods and integration of supervised ML models. Hadiyyisa is a commonly spoken language with a scarcity of natural language corpora. As stated in this paper, our contribution is a newly established and publically accessible benchmark Data set for the Hadiyyisa Lexical Sample WSD challenge. There are 1516 sentences in the data collection, with six target terms (6 nouns) with their two sense only. The results of our WSD tests show

that the TF-IDF method produces the best results. **Comparative study**: First, we have performed a comparison of the most used algorithms in the research literature. Currently the latest paper review about WSD for any language. **Generalization across Data set**: The experiments of chapter 4 empirically show the dependency of supervised Machine Learning on the training Data set.

Summary of the Result and Discussion

As previously stated, Supervised Machine learning was chose for this study. The researcher used sense annotated labeled datasets because of this Classification algorithms were used to complete classification jobs. In this work, four trials were carries out utilizing four different classification methods. This classification algorithms like **Naive Bayes, Support Vector Machine, Neural Network and Hybrid Model**. In addition to use to validating our research study using appropriate methods such as: - Monte Carlo Cross-Validation (Shuffle Split). In addition, these classification algorithms should been applied on the six selected ambiguous words in Hadiyyisa language. These selected ambiguous words such as - **Anga**, **Diinate**, **Hagara**, **Hurbaata**, **Misha** and **Seera**. Moreover, with the total data set contain 1516 sentences for these ambiguous words should been too used. In this study finally to answer the following questions with experimental evaluation:-

Q.1.Which Algorithm better performance for Hadiyyisa language from the supervised machine learning models?

Q.2.What is the model accuracy for Individual Data set and combined data set?

Q.3.Which data preprocessing features procedures should be used to deal with WSD in Hadiyyisa and find the best performing classification algorithm models for Hadiyyisa dataset?

Experiment I: - Which Algorithm better performance for Hadiyyisa language from the supervised machine learning models with their Cross validation evaluation?

Table 4. 1 Summary of Accuracy for each Supervised Machine learning with their Average of cross validation

	NB		**SVM**		NN		**Hybrid model**	
Target words	Accuracy	Monte Carlo Cross-Validation	Accuracy	Monte Carlo Cross-Validation	Accuracy	Monte Carlo Cross-Validation	Accuracy	Average Monte Carlo Cross-Validation
Anga	0.88	0.87	**0.91**	0.85	0.84	0.82	**0.88**	0.85
Diinate	0.76	0.76	**0.79**	0.78	0.76	0.78	**0.82**	0.79
Hagara	0.88	0.88	**0.88**	0.90	0.90	0.92	**0.88**	0.91
Hurbaata	0.82	0.82	**0.85**	0.82	0.85	0.83	**0.85**	0.84
Misha	0.86	0.80	**0.88**	0.87	0.91	0.86	**0.90**	0.80
Seera	0.81	0.81	**0.93**	0.90	0.84	0.86	**0.91**	0.89

To discuss this Experiment depending on above experimentation result using the python programming language and Jupyter Notebook Section 4.2.1 Accuracy of models

to use necessary NLP research developing tools. To do the experimentation for each selected supervised Machine learning methods to compare the better result. We get from experimentation result from three classification Algorithm and Hybrid model for these target words such as:- **Anga**, **Diinate**, **Hagara**, **Hurbaata**, **Misha** and **Seera** is better result from above classification algorithms. To show the best model with accuracy and with their average cross validation score. From above Table 4.2 Four classifier algorithms, SVM and Hybrid model are better result of accuracy to handle for Hadiyyisa language.

Experiment II: - To compare the model accuracy for Individual Data set and combined data set from above selected models. According the implementation of Individual Data set and combined data set result. The individual Data set class result is more appropriate than Multi class result to disambiguate. Because it use binary classifier. From the above Figure 4.23 It shows the overall performance of our study by their accuracy and with cross validation experimentation for both models such as SVM and Hybrid Model. We get the overall performance of WSD for Hadiyyisa language with the feature extraction techniques summarized below Table 4.3 and Table 4.4.

Table 4. 2 Overall Performance Separate Individual Data Set Result

Sno	Model	Target words	Average accuracy
1.	Support Vector Machine	Anga	0.91
		Diinate	0.79
		Hagara	0.88
		Hurbaata	0.85
		Misha	0.88
		Seera	0.91
2.	Hybrid model	Anga	0.90
		Diinate	0.79
		Hagara	0.85
		Hurbaata	0.86
		Misha	0.91
		Seera	0.95

Table 4. 3 Overall performance of Combined Data Set Result

Models	Average accuracy	Average Monte Carlo Cross-Validation(Shuffle Split)
Hybrid model	0.82	0.81
SVM	0.79	0.80

Experiment III: - Which data preprocessing features procedures should be use to deal with WSD in Hadiyyisa and find the best performing classification algorithm models for Hadiyyisa dataset?

To discuss this **Experiment** on the above Section 4.2.1 and 4.2.2 experimentally verify. To review different literature papers that previously done on this research titles. In this study we select TF-IDF is more appropriate to

handle WSD for Hadiyyisa language. Because TF-IDF it has used for to weigh terms for natural language processing tasks depending on the importance of in the sentence scaled by functionality across all sentences in our data sets, which mathematically remove naturally occur words in Hadiyyisa language.

```
Hybrid Model accuracy_score: 0.8223684210526315
Average Cross Validation score of ShuffleSplit of Hybrid Model:0.8167763157894736
Support Vector Machine accuracy_score: 0.7861842105263158
Average Cross Validation score of ShuffleSplit of SVM:0.7983552631578947
```

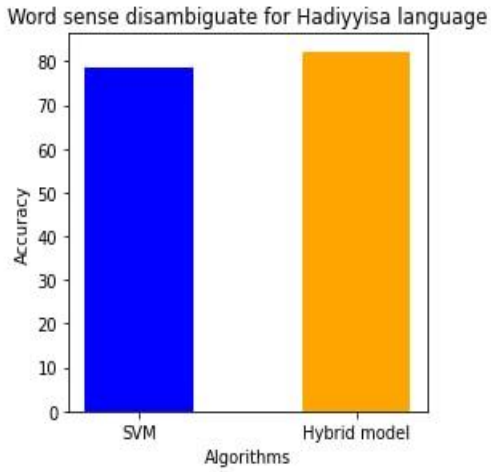


Figure 4. 3 Overall performance Result of model Represented by Bar Graph

Conclusion

There is relatively little study done when looking Afan Oromo, Geez, Tigrinya, Amharic, Wolaita, and other Ethiopian languages. The cause for this is shortage of resources. Because there has been no previous research on Hadiyyisa Language Word Sense Disambiguation in comparison to Oromia and other low-resource Ethiopian languages, this study have investigated on it.

In this study, the WSD of the Hadiyyisa language was supplement by the Supervised Machine Learning Approach and the integration of various supervised ML classification algorithms, resulting in a hybrid model. Machine translation, information extraction, question answering, information retrieval, text classification, text summarization, speech processing, text processing, grammatical analysis, and other NLP applications all require Hadiyyisa WSD. As a result, finding the correct senses (meanings) of polysemy words in context improves NLP applications' efficiency. In this study, the WSD of the Hadiyyisa language was investigate in order to improve the performance of NLP applications undertaken by succeeding researchers who will conduct NLP related studies. The lack of WSD in Hadiyyisa inhibits Natural Language processing systems in future efforts to make computers comprehend Hadiyyisa and construct data sets for computers to learn or understand the proper sense of ambiguous words with correct meanings. As a result, this study attempts to fill a gap in the Hadiyyisa language. The major goal of this research is to provide information on the many resources accessible for the Hadiyyisa language, which would be useful for researchers working on Hadiyyisa WSD. They employ classification methods such as Naive Bayes, Support Vector Machines, Neural Networks, and Hybrid models, which might been applied to six ambiguous Hadiyyisa terms. These ambiguous terms, such as Anga, Diinate, Hagara, Hurbaata, Misha, and Seera, have utilized in conjunction with the whole data set, which contains 1516 sentences for these ambiguous words. Finally, basic reason to select two appropriate classification models such as **SVM** and **Hybrid models** for Hadiyyisa. Frist reason to select depending on the average accuracy to predict results. Second to overcome the disadvantages of other classification classifier such as NN, Naïve

Bayes. The accuracy of SVM are - 0.91, 0.79, 0.88, 0.85, 0.88, and 0.90 and the accuracy of Hybrid model are-0.88, 0.82, 0.88, 0.85, 0.90, and 0.90 for these selected ambiguous words such as: - Anga, Diinate, Hagara, Hurbaata, Misha and Seera respectively. To validate our models such as Monte Carlo Cross-Validation (Shuffle Split).Overall performance of models are 0.82% and 0.79% for Hybrid Model and Support Vector Machine classifier respectively.

As a result, we have the following proposals for WSD of Hadiyyisa texts, which include the development of resources and future research directions:-

- Linguistic resources such as thesaurus, Lexicon, WordNet, and machine-readable dictionaries were employ in WSD research for other languages; however, these resources are not available for Hadiyyisa. Because those resources were unavailable for Hadiyyisa, researchers were unable to perform research on the language's WSD. Other scholars can undertake research for the preparation of these materials with interested institutions or persons by considering their contribution to WSD.
- Extending this research utilizing supervised machine learning for more than six ambiguous terms in this study has advised, as increasing, the data set size has regarded to be more realistic for WSD model.
- Next researcher recommended doing WSD for Hadiyyisa Language, which have not used for this study like it uses Deep Learning approach and Corpus based approach.