

DSpace Institution

DSpace Repository

<http://dspace.org>

Computer Science

thesis

2021-06

AGRICULTURAL DOMAIN-SPECIFIC JARGON WORDS IDENTIFICATION IN AMHARIC TEXT

MELAKU, LAKE TEGEGNE

<http://ir.bdu.edu.et/handle/123456789/12653>

Downloaded from DSpace Repository, DSpace Institution's institutional repository



BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

FACULTY OF COMPUTING

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

AGRICULTURAL DOMAIN-SPECIFIC JARGON WORDS

IDENTIFICATION IN AMHARIC TEXT

BY:

MELAKU LAKE TELEGNE

JUNE, 2021

BAHIR DAR, ETHIOPIA



BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING

AGRICULTURAL DOMAIN-SPECIFIC JARGON WORDS IDENTIFICATION

IN AMHARIC TEXT

BY

MELAKU LAKE TEGEGNE

**A Thesis Submitted to the School of Research and Graduate Studies of
Bahir Dar Institute of Technology, BDU in Partial Fulfillment for the
Degree of Master of Science in Information Technology in the Faculty of
Computing.**

ADVISOR: TESFA TEGEGNE (PhD)

JUNE, 2021

BAHIR DAR, ETHIOPIA

DECLARATION

This is to certify that the thesis entitled “**AGRICULTURAL DOMAIN-SPECIFIC JARGON WORDS IDENTIFICATION IN AMHARIC TEXT**”, submitted in partial fulfillment of the requirements for the degree of Master of Science in Information Technology under faculty of computing, Bahir Dar Institute of Technology, is a record of original work carried out by me and has never been submitted to this or any other institution to get any other degree or certificates. The assistance and help I received during the course of this investigation have been duly acknowledged.

Name of the candidate **Melaku Lake** Signature: _____ Date: _____.

©2021

MELAKU LAKE TEGEGNE

ALL RIGHTS RESERVED

BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
FACULTY OF COMPUTING

Approval of Thesis for Defense Result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student Melaku Lake Tegegne Signature _____ Date _____.

As members of the board of examiners, we examined this thesis entitled “Agricultural domain-specific jargon words identification in Amharic text” by Melaku Lake. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in “Information Technology”

Board of Examiners:

Board of Examiners:

Tesfa Tegegne (PhD)


Signature

14/11/2013 E.C
Date

Advisor Name

External Examiner:

Michael Melese (PhD)


Signature

July 20, 2021
Date

Name

Signature

Date

Internal Examiner:

Birhanu Hailu (PhD)


Signature

23/07/2021
Date

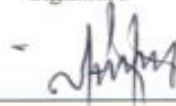
Name

Signature

Date

Chair Person:

Alemu Kumilachew


Signature

16/11/2013 E.C
Date

Name

Signature

Date

Chair Holder:

Derejaw Lake


Signature

16/11/2013 E.C
Date

Name

Signature

Date

Faculty Dean:

Belete Biazen (Ass. Prof.)


Signature

16/11/2013 E.C
Date

Name

Signature

Date



ACKNOWLEDGEMENTS

First and foremost, I would like to thank my **God** and **Ever-Virgin, St. Marry** for your blessing and for giving me the wisdom and directions to accomplish this thesis.

Next, I would like to thank the '**BDU talent**' program of **Bahir Dar University** for providing me a **scholarship** in my field of study.

Next, I would like to thank Dr. **Tesfa Tegegne**, my advisor, for his insightful view, constructive comment, and patience from the preproposal to the end. The routine tasks inline to propose this paper and complete incases are very tedious and challenging; the simplicity of the explored idea is granted by the insightful view of my advisor. His pleasure made things right.

Many thanks have gone to Mr. **Belete Nibret**, a linguistic expert at Colonel Tadesse Muluneh preparatory school in west Gojjam Merawi, for his support during title formulation; and agricultural domain experts in Amhara regional bureau, and south Achefer agricultural office for your unlimited support during data collection and labeling. I would also like to thank you experts and the other selected respondents who filled the prepared questionnaire for the survey we conducted.

Next, my thanks go to **Computing** Postgraduate students for your direct and indirect supports. Next, I would like to thank Mr. **Yosef Bogale** for your moral development and unforgettable encouragement. Next, I would like to thank Computer Science staffs in **Mekdela Amba University** for your encouragement and committed support.

Next, I would like to thank **Dr. Gebeyehu Belay** and **Dr. Mekuanint Agegnehu** for your interesting course delivery and your advice to complete with the intended time.

Finally, I would like to thank my father, **Ato Lake Tegegne** for your strong-minded support and commitment, and my mother, **W/ro Shashitu Ayalew**, my sensor, for your patience and all other my families. The almighty God bless you.

ABSTRACT

Domain-specific jargon words are lists of words used in formal communication of a particular profession between experts of the same field; however, it is difficult to understand by non-experts and society. Experts of an organization use domain-specific Amharic jargon words in scientific and science communication to keep the protocol of the communication within a domain. The domain-specific Amharic jargon words negatively impact people out of the domain to understand the main theme of the disseminated content. We followed a design science research approach to conduct our study and come up with solutions; hence, domain-specific Amharic jargon words are required to convey prominent information to understand the writer's discourse and for further lexical processing. Machine learning classifiers algorithms are employed to develop a model and train the dataset, and predict a text as jargony or non-jargony. We employed three popular machine learning classifiers for text classification with Support Vector Machine, Artificial Neural Network, and Naïve Bayes to develop models with TFIDF feature selection. We labeled the dataset based on the two-way classification. We developed a hybrid system with machine learning and knowledge-based for domain-specific Amharic jargon words identification. We prepared a knowledge source with a list of domain-specific Amharic jargon words and the words meaning. The developed machine learning models with SVM, ANN, and NB show a classification accuracy of 96.2%, 95.2%, and 94.7% respectively. The knowledge-based of the proposed system best performs when a smaller number of input sentences are entered into the knowledge base system. For the input of 20, 40, 60, and 80 test data, an accuracy of 88.2%, 86.7%, 85.4%, and 83.1% is observed. Therefore, we observed the promised result with the hybrid of machine learning and knowledge base for the identification of jargon words in the jargony text.

Keywords: Language, Natural Language Processing, Domain-specific jargon words, Science communication, Knowledge base, Machine Learning

ABBREVIATIONS

ADL: Architecture Descriptions Language

AI: Artificial Intelligence

AJMRD: Amharic Jargon Machine Readable Dictionary

ANN: Artificial Neural Network

BDI: Bilingual Dictionary Induction

BERT: Bi-directional Encoder Representation from Transformers

CCLA: Cross-Context Lexical Analysis

CHV: Consumer Healthcare Vocabulary

DSAJW: Domain-Specific Amharic Jargon Word

DSAJWI: Domain-Specific Amharic Jargon Word Identification

DSR: Design Science Research

DT: Decision Tree

FDRE: Federal Democratic Republic of Ethiopia

FN: False Negative

FP: False Positive

FPR: False Positive Rate

GDP: Gross Domestic Product

GOE: Government of Ethiopia

GPU: Graphics Processing Units

IDF: Inverse Document Frequency

KB: Knowledge Base

KL: Kullback-Leibler Divergence

MIMIC: Medical Information Mart for Intensive Care

ML: Machine Learning

MLP: Multilayer Perceptron

NB: Naïve Bayes

NLP: Natural Language Processing

NN: Neural Network

OSH: Optimal Separating Hyperplane

PAC: Probably Approximately Correct

RF: Random Forest

ROC: Receiver Operating Characteristic

SMS: Short Message Service

SVM: Support Vector Machine

TF: Term Frequency

TFIDF: Term Frequency Inverse Document Frequency

TN: True Negative

TP: True Positive

TPR: True Positive Rate

WSD: Word Sense Disambiguation

WWW: World Wide Web

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT.....	vi
ABBREVIATIONS	vii
TABLE OF CONTENTS.....	ix
LIST OF FIGURES	xiii
LIST OF TABLES	xiv
CHAPTER ONE: INTRODUCTION.....	1
1.1. Background.....	1
1.2. Motivation of the study.....	5
1.2.1. Amharic agricultural jargon words justification.....	5
1.3. Statement of the problem	8
1.4. Objective of the study	10
1.4.1. General objective	10
1.4.2. Specific objective	10
1.5. Methodology	11
1.5.1. Research design	11
1.5.1.1. Problem identification and motivation	11
1.5.1.2. Define objectives of the solution	11
1.5.1.3. Design and development	11
1.5.1.4. Demonstration	12
1.5.1.5. Evaluation.....	12
1.5.1.6. Communication	12
1.6. Scope and limitation of the study.....	13
1.7. Significance of the study.....	14

1.8.	Thesis organization	15
CHAPTER TWO: LITERATURE REVIEW		16
2.1.	Jargon words	16
2.2.	Agriculture	17
2.2.1.	Agriculture in Ethiopia	18
2.3.	Knowledge sources in jargon words identification.....	19
2.3.1.	Lexical Knowledge sources.....	19
2.3.2.	Learned knowledge.....	20
2.4.	Approaches of jargon words identification.....	20
2.4.1.	Knowledge-based approach.....	21
2.4.2.	Machine learning approach	22
2.4.3.	Hybrid approaches	31
2.5.	Amharic Language.....	32
2.5.1.	Amharic writing system.....	32
2.5.2.	Amharic variant characters	33
2.5.3.	Amharic punctuation	33
2.5.4.	Amharic language inflection and derivation	34
2.5.5.	Amharic jargon words	34
2.6.	Related work	36
2.6.1.	Jargon words identification for medicals.....	36
2.6.2.	Jargon words identification for scientists	38
2.6.3.	Jargon words identification for others	39
2.7.	Summary	41
CHAPTER THREE: DESIGN OF THE STUDY		42
3.1.	Design requirements	42

3.1.1.	Machine learning	42
3.1.2.	Knowledge base.....	43
3.2.	Dataset preparation	44
3.2.1.	Labeled corpus for machine learning	46
3.2.2.	Knowledge sources for knowledge base	46
3.3.	System architecture	47
3.4.	Proposed system.....	48
3.5.	Preprocessing	51
3.5.1.	Tokenization	51
3.5.2.	Normalization	51
3.5.3.	Stop word removal.....	52
3.5.4.	Stemming.....	52
3.6.	Machine learning/Model development	54
3.6.1.	Model development	55
3.6.2.	Model training	55
3.6.3.	Model testing	55
3.6.4.	Machine learning classifier algorithm	56
3.7.	Knowledge base	58
3.7.1.	Knowledge Source.....	59
3.7.2.	Jargon word identification	59
3.7.3.	Meaning extraction	60
3.8.	Summary	61
CHAPTER FOUR: EVALUATION AND DISCUSSION OF RESULTS		62
4.1.	Experimental setup.....	62
4.2.	Evaluation of the proposed system	63

4.2.1.	Evaluation metrics	63
4.2.2.	Machine learning evaluation	64
4.2.3.	Knowledge-based evaluation.....	75
4.3.	Discussion.....	78
4.3.1.	Discussion of machine learning evaluation result	78
4.3.2.	Discussion of knowledge base evaluation result	79
4.4.	Summary.....	79
CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS		80
5.1.	Conclusion	80
5.2.	Contributions of the study.....	82
5.3.	Recommendations.....	82
REFERENCES		84
Appendix I:		92
Appendix II:		93
Appendix III:.....		94
Appendix IV:		95
Appendix V:.....		96
Appendix VI:		98

LIST OF FIGURES

Figure 1.1: Agricultural Amharic jargon word justification for domain-experts ...	6
Figure 1.2: Agricultural Amharic jargon word justification for non-domain experts.....	7
Figure 1.3: Word cloud for randomly selected and surveyed DSAJW.....	8
Figure 2.1: System architecture of SVM supervised ML classifier.....	25
Figure 2.2: The architecture of ANN with Multilayer Perceptron (MLP).....	30
Figure 3.1: Sample dataset for domain-specific Amharic jargon words identification	45
Figure 3.2: Flow chart for domain-specific Amharic jargon words identification	48
Figure 3.3: Proposed system for domain-specific Amharic jargon words identification	50
Figure 4.1: F-measure comparison of SVM, ANN, and NB	67
Figure 4.2: Accuracy comparison of SVM, ANN, and NB models	67
Figure 4.3: Confusion matrix of outperformed SVM model	72
Figure 4.4: ROC curve for SVM, ANN, and NB model.....	73
Figure 4.5: Comparison of models with correctly and incorrectly classified test data.....	74
Figure 4.6: Machine learning models comparison using boxplot.....	74
Figure 4.7: Performance result of knowledge base with test 1, test 2, test 3, and test 4.....	77

LIST OF TABLES

Table 3.1: Dataset prepared for machine learning and knowledge base	45
Table 4.1: Performance of jargon identification with SVM classifier	65
Table 4.2: Performance of jargon identification with ANN classifier	65
Table 4.3: Performance of jargon identification with NB classifier.....	65
Table 4.4: Hyperparameters of machine learning models	66
Table 4.5: Comparison of models on TP, FP, FN, and TN.	72

CHAPTER ONE: INTRODUCTION

1.1. Background

Language is an organized combination of sounds with an orthographic structure used for prominent communication between people to share information and to foster political, economic, and social development. A language is a tool of interaction between people to communicate and decide on common topics. Social interactions between communities highly depend on the degree of understanding the theme of the communication. Each community has its language to express ideas, values, and attitudes to members of a particular group of language users (Clayton, 1998).

Members of a particular group of society communicate using a common language upon the issues concerning the objective of communication. A language is a tool of communication to share culture and values; and used to guarantee the continuity of a community's identity that strongly correlates with the social nature of a language (Sirbu, 2015). Considering the speaking and writing aspect of a language is important to takes place a communication.

The field of natural language processing (NLP) that involves the computational processing power, engineering of computational models, and understanding of human language is an emerging area of research in artificial intelligence (AI). The advancement of NLP is used to increase the application area of human language in various domains of service. The computational power of NLP to solve practical problems involves statistics, probability, and machine learning for data-driven computation (Sparck Jones, 1994).

The research in the area of NLP addresses problems in line with language modeling, morphological processing, syntactic processing, and semantic processing (Sparck Jones, 1994). These days, NLP strives to obtain effective communication and accurate knowledge like human beings with increased use of human language in computational language processing (Kevitt et al., 1992). NLP with a range of computational processing power techniques is applicable for the analysis and representation of natural language in different levels of linguistic analysis to obtain human-like language processing.

Communication takes place within some sort of interaction between the communicating people in which people are required to have a common language. Communicants promote effective communication by considering the relationship between communicating people and the language. Knowing the literacy and social context of communicating with people helps individuals to choose the known language for a clear understanding of the communication (Danesi, 1995). In science communication, jargon words are used intentionally or unintentionally. Usage of jargon words in formal communication makes the communication cumbersome as the meaning of the jargon words are unknown for the communicants; therefore, jargon words hamper the interaction (Rakedzon et al., 2017).

Jargon words are defined by the Oxford English dictionary as “words or expressions that are used by a particular profession or group of people, and are difficult for others to understand”. As defined by (Brown et al., 2020), jargon words in a particular language are professional words used by a specialized group of a particular field with less formal alternatives which are broadly accessible by employees of an organization; however, difficult to understand for non-experts and the society.

The authors (Helmreich et al., 2005), defined jargon words as a list of domain terminologies used by experts of an organization that is necessary for the communication of a particular field; however, it needs meaning for users of text out of the field. The authors develop common ways of understanding the belief of experts of an organization that uses jargon words. Experts use the words to explore their ideas besides organizational related tasks and also the readers of the text have a common understanding of the words used in a text.

Scientists that use domain terminologies to explore findings are stressed on how to reach a non-expert reader (Rakedzon et al., 2017). The use of domain-specific jargon words within a domain is a measure of status compensation for employees in an organization, hence low-status employees use more domain terminology. The authors develop a de-jargonizer to provide the meaning at right next for every occurrence of scientist’s jargon words in disseminated content (Rakedzon et al., 2017). In organizational comprehension, it is recommended to use up to 2% of domain-specific jargon words to explore issues in organizational discourse (Schmitt, 2000).

Employees in an organization with low status and technical staff are more likely to use domain-specific jargon words in their communication as compared to high-level managers of an organization to keep the protocol of the communication (Brown et al., 2020). Domain-experts understand the theme of the text that contains domain-specific words in organizational related discourse, nevertheless, the words themselves confuse individuals out of that particular domain (Crémer et al., 2007).

Mapping professional medical jargon words to laymen's terms is necessary to decrease the communication gap between laymen and medical experts in the treatment and consultation process. The authors generate new Consumer Healthcare Vocabulary (CHV) using predefined lexical source or ontology for the medical jargon in the online consultation process to increase the understanding of patients (Ibrahim et al., 2020).

Domain experts, spokesman, journalists, corporate level managers of an organization disseminate organizational related information in Amharic that contains domain-specific Amharic jargon words (DSAJW). For example, in a text, ‘ከብቶች ለግጦሽ ሲጠቀሙበት የነበረውን መስክ ለደረቦ እንዳይጋለጥ በስምምነት መስክደሳ መስራት እንደሚያስፈልግ በውይይት መግባባት ያስፈልጋል’; ‘አዝመራ ከቀልዝ እና ብላፅዋት በሚያገኘው ጥቅም ከፍተኛ ምርት ያስገኛል’; ‘መለያማሬ አርብቶ አደሮች ብዙ ጊዜ የሚጠቀሙበት ስልት ሲሆን በተለይም በጎሌ እንስሳት ባለቤቶች የተለመደ ድርጊት ነው’; ‘እንስሳት ቃቃቃ በሚያሳዩ ጊዜ አስፈላጊውን ክትትል በማድረግ ብዛዘር እንዲሰጣቸው ማድረግ ያስፈልጋል፤ ለቀበሌያችን የተመደበውን ብዛዘር በተለያየ ጊዜ ቃቃቃ ላሳዩ እንስሳት በመስጠት አጠናቀናል’, the society doesn’t have common understanding for the word ‘መስክደሳ (meskdesa)’, ‘ቀልዝ’ (qeliz), ‘ብላፅዋት (bilatsiwat)’, ‘መለያማሬ (meleyamarie)’, ‘ጎሌ (golie)’ ‘ደረቦ (derebo)’, ‘ቃቃቃ (qaqata)’, ‘ብዛዘር (bizazer)’.

Amharic is the dominant language spoken by the Ethiopian people in which characters in a language evolved from Geez's character ‘fidel’. Next to Arabic, Amharic is the second-largest Semitic language spoken in the world. Amharic is a morphologically complex and under-resourced language; few computational linguistic resources were developed (Hudson, 1999). The Amharic language contains 34 base characters with seven orders of consonant-vowel combination (Mindaye & Kassie, 2018).

Agriculture is the dominant source of income for 85% of the Ethiopian people that live in rural areas and also the dominant profession for huge customers in Ethiopia. The agricultural sector plays a vital role for the Ethiopian people's economy that helps with

a variety of guidance to return high yields of domestic food products to sustain the huge Ethiopian people. The number of food-insecure households in Ethiopia is increasing and domestic food production has failed to satisfy the food requirement of the country (Demeke & Ferede, 2014). Though Ethiopia is currently food insecure, the country has a great potential to increase agricultural production and productivity and thereby ensure food security.

The language issue in the agricultural process negatively influences the communication between experts and the people. So that clear and precise communication between agricultural experts and the people in Ethiopia is fundamental to maximize domestic food production to satisfy the food requirement of the country and to achieve food security. (Demeke & Ferede, 2014).

The use of Amharic agricultural jargon words in a text between agricultural domain experts and the people is one of the challenges that hamper the communication and leads to the loss of the target theme. The problem happens when agricultural experts communicate with non-experts, customers, and society to disseminate institutional information and to proceed with agricultural activities using the available resources.

Agricultural societies received the text ‘በእርሻ ውስጥ የበቀሉትን አረሞች ለማጥፋት ደጋግሞ ማረስና ማሳውን አርፋዶር ማድረግ አስፈላጊ ነው’ , from agricultural experts confused with the word ‘አርፋዶር (arifador)’. The word ‘ቀልዝ (qeliz)’ is unknown for the agricultural society in a text ‘በእርሻ ማሳ ላይ ቀልዝ መጨመር ምርት እና ምርታማነት ለመጨመር ያስችላል’. The agricultural society uses the English word ‘compost’ instead of the Amharic word ‘ቀልዝ (qeliz)’. The use of English words instead of Amharic words in Amharic text negatively influences the Amharic language development. So that the meaning of domain-specific jargon words is required for non-expert readers of the text.

Therefore, we define jargon words as a list of words used in formal communication of a particular domain and mostly used between experts of the same field; however, it is difficult to understand for non-experts and society. So that we are required to provide the meaning of domain-specific Amharic jargon words for non-expert readers.

1.2. Motivation of the study

Communication with a particular language requires the combination of words for communicants. The selection of words is the responsibility of writers for prominent communication between communicants. However, the usage of words that cause communication barriers is accustomed.

Domain-specific jargon words are used in organizational related discourse to keep the protocol of communication of a particular domain. Because the existence of domain-specific jargon words hampers the communication between communicants, domain-specific jargon word identification is an emerging area of research in natural language processing (NLP) to provide the necessary information with a text. Though the usage of DSAJW helps for handling simple communication for experts, the way to understand the customer side is a challenge.

So that the study of DSAJW is required to identify the existence of words in Amharic text to minimize communication barriers. The study attempted to alienate communication barriers between experts of a domain and non-expert readers. The study considered the agricultural domain because of the high vulnerability for the occurrence of domain-specific Amharic jargon words.

1.2.1. Amharic agricultural jargon words justification

We conduct a survey on non-experts such as farmers, non-domain experts, and agricultural domain experts at different governmental management positions to know the level of knowing and using agricultural jargon words in communication. We use judgmental sampling and random sampling techniques for selecting samples from a large size population to fill the prepared questionnaire. We prepared a close-ended questionnaire for selected respondents and we collected and analyzed the responses.

We randomly selected 40 Amharic agricultural domain-specific jargon words around 9% of the total from the knowledge source. We prepared different close-ended questionnaire formats for the agricultural society (farmers), non-experts, and agricultural domain experts at different government management positions in agricultural offices. We selected 3 samples from each group of respondents of randomly selected farmers (Lalibela kebele), non-domain experts (Bahir Dar), and agricultural domain experts at the region (Bahir Dar, Amhara), woreda (south Achefer, west Gojjam, Amhara), and kebele (Lalibela, south Achefer, west Gojjam, Amhara) since

the Ethiopian government structure is organized as federal, region, zone, woreda, and kebele.

The following figure depicts the analysis of collected data from the respondents. The rate illustrated in the figure is based on the agreement of all respondents. For example, 92.5% of the randomly selected jargon words are known by all of the regional bureau expert respondents; all regional bureau expert respondents use 70% of the randomly surveyed words for communication. The respondents are requested to fill the questionnaire without consideration of the frequency of using words for communication. So that we observed equal treatment for usual and rare usage of words on the respondents.

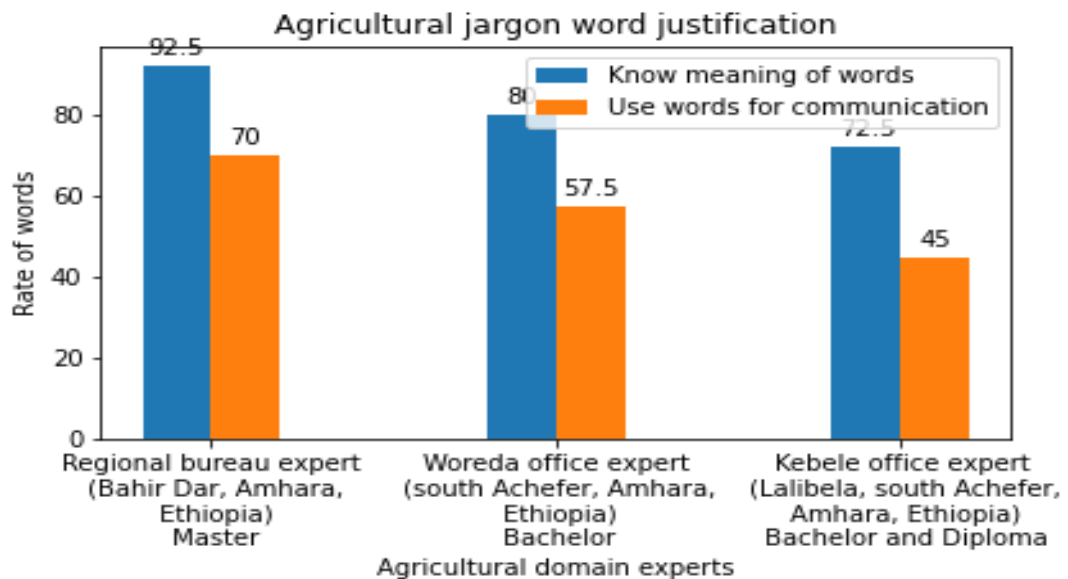


Figure 1.1: Agricultural Amharic jargon word justification for domain-experts

Though, individual differences of experts to know the meaning of words are there, the use of domain-specific agricultural jargon words in organizational reports and other discourses decrease at a decrease rate from regional bureau to kebele office. The above chart depicts the decreasing knowing and using domain-specific Amharic agricultural

jargon words between experts of the agricultural domain in Amhara, Ethiopia.

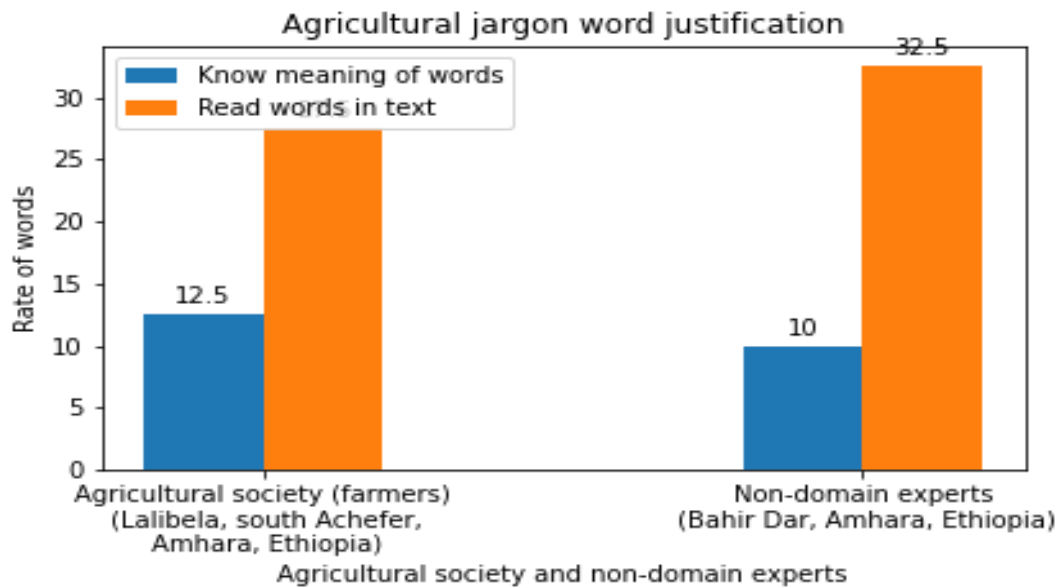


Figure 1.2: Agricultural Amharic jargon word justification for non-domain experts

The above figure 1.2 shows that knowing the meaning of domain-specific Amharic agricultural jargon words is a weighty problem for the agricultural society and non-domain experts. Though knowing the meaning of jargon words is a challenge for agricultural society, the words are available in a text that confuses readers to understand the target theme.

Word cloud

Domain-specific jargon word is a language that is aimed at an intelligent audience to provide clear and precise information. The following figure shows the word cloud of randomly selected and surveyed DSAJW from the knowledge source for the justification.

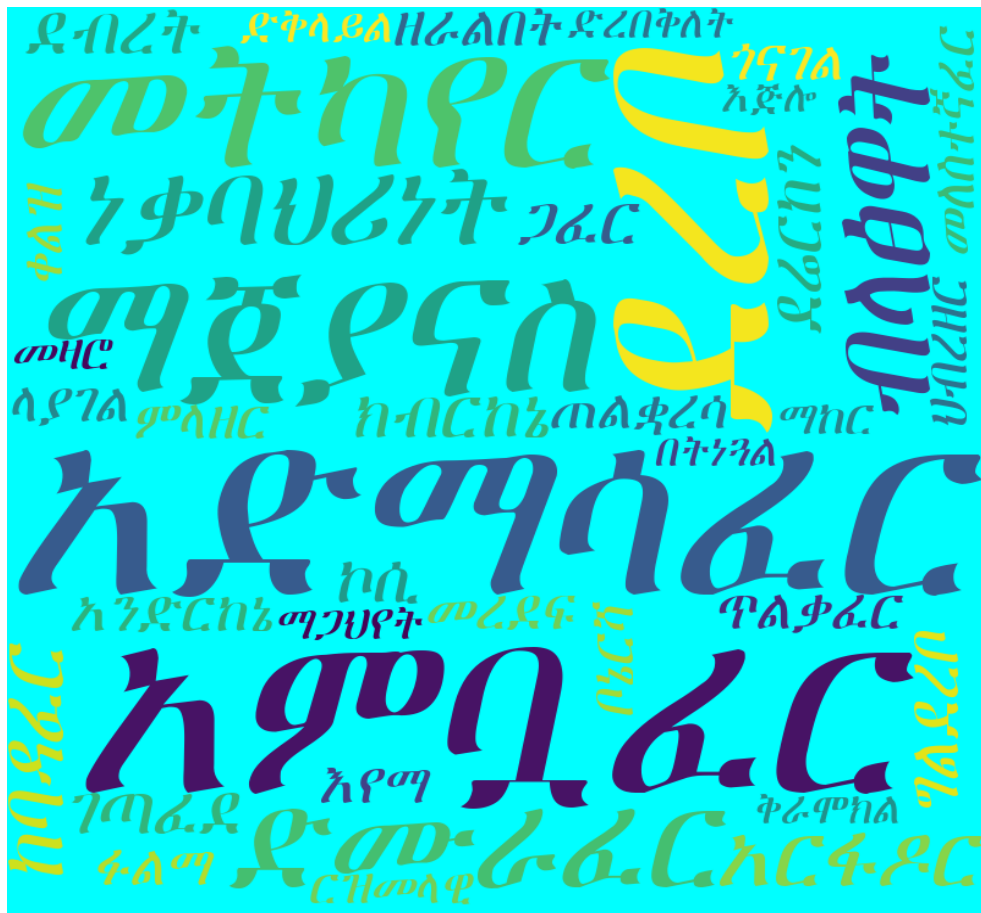


Figure 1.3: Word cloud for randomly selected and surveyed DSAJW.

Therefore, based on the survey we conducted at different levels of domain experts, non-domain experts, and the agricultural society; we found the use of DSAJW in agricultural discourse is a weighty problem for domain experts (because of individual differences), non-domain experts, and the agricultural society.

1.3. Statement of the problem

Communication is the way of delivering and receiving information between communicating parties. Effective and efficient communication is vital to create a positive relationship between the concerned bodies. Written communication helps experts of an institution to disseminate institutional information to society. Scientific and science communications are used for written communication (Burns et al., 2003). In scientific communication, experts of an organization share institutional information to the employees inside their organization; however, science communication refers to experts of a domain share textual information to users outside the domain.

Therefore, scientific communication is written by experts for experts and it is not a challenge to understand the concrete idea of the text. In science communication, experts of a domain in an institution communicate to the people outside the domain, and non-experts using domain-specific jargon words to keep the protocol of communication in the organizations.

The use of domain-specific Amharic jargon words helps experts of an organization to handle simple communication. Communication without the use of DSAJW becomes a challenge for experts, and also avoiding the words is impossible to keep the protocol of communication in a domain. However, non-domain experts and society face a challenge to understand the discourse upon the domain. Understanding the theme of texts that contain domain-specific jargon words in various fields becomes a challenge to society and non-domain experts. These hamper the communication between domain experts and non-domain experts.

Besides, the justification of the survey we conducted on the extent to know and use the domain-specific Amharic jargon words in a text, experts use the textual form of agricultural domain-specific jargon words that are common in the agricultural domain. However, the words are unknown that cause confusion and misperception to non-domain expert and agrarian society readers. The problem is committed when agricultural domain-experts communicate with the customers, non-domain experts, and the agricultural society alone providing service in science communication. So that domain-specific Amharic jargon words for non-domain experts and agricultural society cause clarification problems.

Experts that use domain-specific Amharic jargon words on letters, news, advertisements, reports, and social media like Facebook, Telegram, Twitter are stressed to reach a non-expert reader with a clear theme of the content. The existence of jargon words in a text return wastage of time and money for the reader, business opportunities lost by misunderstanding, untrustworthy readers, potential customers left from the organization, increase the traffic of searching words.

The use of domain-specific agricultural Amharic jargon words such as ‘መላያማሬ (meleyamarie), አንሸልቶ (anishelto), አምቧፈር (ambuafer), አንክክ (ankik), ነገሎነት (negelonet), ቃርጥ (qaremo), ነፀረገይ (netseretsay), ትነተክል (tinetekil)’ are a weighty problem in the agriculture domain. Non-experts, customers and the agricultural society

are confused to understand the targeted content. To the best of our knowledge, there is no prior work for agricultural domain-specific jargon word identification. Besides the aforementioned problem, this work has the following research questions.

1. To what extent domain-specific Amharic jargon word identification system works to identify the existence of domain-specific Amharic jargon words and provide meaning?
2. Which hyperparameters of machine learning classifier influence the performance of agricultural jargon words identification system in Amharic text?
3. Which model outperformed for domain-specific Amharic jargon word identification?

1.4. Objective of the study

1.4.1. General objective

The general objective of this study is to develop domain-specific Amharic jargon words identification systems in a text.

1.4.2. Specific objective

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

- To prepare a dataset with and without Amharic jargon words for training, and knowledge sources for meaning extraction from the target domain.
- To design the proposed system for domain-specific Amharic jargon identification system in Amharic text.
- To discover the meaning of domain-specific Amharic jargon words from the predefined explanatory lexical knowledge source.
- To identify hyperparameters of machine learning classifier that influence the performance of agricultural domain specific jargon words identification.
- To train and evaluate the performance of the domain-specific Amharic jargon words identification system, and recommend future research work.

1.5. Methodology

Research methodology is a technique of solving real world problems by research. It is a way of dealing technical problem solving. Research methodology is the approach of solving explored problems thoroughly. It is the philosophy and scientific approach adopted for conducting the research systematically (Profile & Profile, 2019).

1.5.1. Research design

We followed a Design Science Research (DSR) approach to conduct our study and come up with the solution for domain-specific Amharic jargon word identification (Carstensen & Bernhard, 2019). DSR approach is used to create and evaluate artifacts that are intended to solve real-world problems with its various computational steps. The steps of the DSR approach include problem identification and motivation, defining objectives of the solution, design and development, demonstration, evaluation, and communication (Peppers et al., 2007).

1.5.1.1. Problem identification and motivation

Problem identification is the process to define the specific problem by the researchers to be resolved and also might be recommended from research future work. With the problem identification, the researcher understood the problem in line with the state of the art. Experts of the agricultural domain helped the researcher when exploring the problems and the way how to come up with the intended solution.

1.5.1.2. Define objectives of the solution

The objectives of a solution defined as per the problem of the research work. The explored domain-specific Amharic jargon word problem is the source to define the performed list of objectives with the conducted research work. A neatly defined list of objectives is the boundaries of our research work.

1.5.1.3. Design and development

This stage helped the researcher to design the architecture, develop a model, develop a system, implementation of the algorithm with domain knowledge of theory that guides a researcher to come up with the solution.

1.5.1.4. Demonstration

The designed architecture with the domain knowledge of theory implemented and the way how to use the developed system was demonstrated. The demonstration includes experimentation and other relevant activities for the solved explored problems.

1.5.1.5. Evaluation

Evaluation of this research work done on the collected Amharic agricultural text documents with the help of domain expert curators. Evaluation performed to know how the system we developed works to identify domain-specific Amharic jargon words.

1.5.1.6. Communication

We developed a hybrid system to identify the existence of domain-specific Amharic jargon words in Amharic text. Therefore, we used machine learning models for the classification of the input text, and also, we used a knowledge-based system to provide the meaning of words in line with the agricultural domain. We prepared a lexicon and developed a Machine-Readable Dictionary (MRD) with the help of agricultural domain expert curators. We developed an interactive Amharic Jargon Machine-Readable Dictionary (AJMRD) as an explanatory lexical resource for our work. Binary lexical mapping between jargon words from the input text and the words meaning in the knowledge source is performed to make users full of information.

1.5.2. Literature review:

We reviewed various literatures done on domain specific jargon words identification for more vulnerable domain and resourceful language. Literature review is helpful to have deep understanding on the problem and the way we followed to come up the solution. The domains health sectors, commercial organizations, scientists are more vulnerable for the problem.

1.5.3. Data source

We considered a domain that provides services for huge customers and society from institutions in the Federal Democratic Republic of Ethiopia (FDRE). We collect domain-specific Amharic jargon words from organization's business reports, working guidelines prepared by corporate level managers of agricultural organization, training

manuals for employees and customers, advertisements of products and services, and telegram accounts to prepare the corpus for training with machine learning (ML) and to develop the Amharic Jargon Machine Readable Dictionary (AJMRD) for knowledge source.

1.5.4. Tools and techniques

We used Python programming language which is a cross-platform, interpreted, and the former programming language in the current computing that is applicable for a wide range of NLP applications. Python supports many open-source libraries that are compatible for our experiment and to work with NLP applications. Users can easily understand the code and everyone can become productive in python very quickly.

1.5.5. Evaluation

The evaluation of domain-specific Amharic jargon words identification (DSAJWI) performed using F-measure and accuracy to know how the system we developed works to identify domain-specific Amharic jargon words. The mentioned performance metrics are mostly used for the evaluation of the performance of text classification.

1.6. Scope and limitation of the study

The study focused on the identification of the existence of agricultural domain-specific Amharic jargon words (DSAJW) in the agricultural domain and provide meaning for the words based on the predefined explanatory lexical knowledge resource. The study focused on the agricultural domain; hence agriculture is the main source of income for an estimated 85% of the people who live in rural areas and also, the main source of export products for Ethiopia. Agricultural sectors in Ethiopia are more vulnerable to use domain-specific jargon words for communication. The work focused on the preparation of dataset from documents of agricultural institutions in the Federal Democratic Republic of Ethiopia (FDRE). Amharic texts that contain domain-specific jargon words with the words meaning in the agricultural domain are vital for the researchers to accomplish the objectives.

In this work, we collected a list of agricultural domain-specific words and the words meaning to construct a knowledge source, and also, we collected a dataset with and without jargon words for the two-way classification with the learned knowledge. So

that we limit our study to provide meaning for domain-specific Amharic jargon words (DSAJW) in the agriculture domain of Amharic text. Therefore, identification of Amharic domain-specific jargon words in a text is considered in our study.

The availability of insufficient dataset affects our work to use the state-of-the-art deep learning approach. The static meaning of jargon words with a particular domain motivates us to use the knowledge source with lexical binary mapping of words. So that the semantic behavior of words in a text is not considered. The unavailability of effective and efficient Amharic stemmers affected the performance of the knowledge base system.

1.7. Significance of the study

The theoretical and implementation of Natural Language Processing (NLP) provides many applications. Amharic domain-specific jargon words that existed in Amharic text cannot be removed. Domain-specific Amharic jargon words identification provides much importance for non-experts of the domain and also, used as an input for further natural language processing application such as information retrieval (IR), machine translation (MT), information extraction (IE), question answering (QA), dialogue system, text summarization, word sense disambiguation (WSD) (Kevitt et al., 1992).

Experts of the agricultural domain handle simple communication with non-experts and society. Readers of news, advertisement, business reports, working manuals, periodicals, magazines, social media like Facebook, telegram are full of information besides the theme of the content. Confusion to understand the main theme of domain discourse decreases at a decreasing rate because we provide meaning to the DSAJW. Organizations become profitable with effective communication between domain experts and non-domain experts in the product and service delivery process.

Agricultural organizations become profitable by delivering the required information for customers. Non-domain experts, customers, society, and domain experts (because of individual differences) are the beneficiary of our DSAJWI system towards understanding the targeted content of the organization. The developed DSAJWI system motivates experts to use domain terminology in organizational discourse to handle simple communication with customers. For the Amharic language, the usage of domain terminology is used for the development and usage of Amharic language in a domain.

1.8. Thesis organization

The rest of the thesis is organized as follows. Chapter two presents' precise definition of jargon words, agricultural science, agriculture in Ethiopia, approaches employed for domain-specific jargon words identification, design requirements, Amharic writing system, kinds of Amharic jargon words, and reviews of related work on domain-specific jargon words. Chapter three presents the design and implementation of the system architecture, proposed system, and description of the function of the part and phases of the proposed system. Chapter four discusses the experimentation and discussion of performance results. Chapter five presents the conclusion, contributions and recommendations for future research work.

CHAPTER TWO: LITERATURE REVIEW

In this chapter, we begin with a brief explanation of jargon words. The importance of the agricultural domain, knowledge sources required for domain-specific jargon words identification, design requirements, approaches that have been employed, Amharic writing system, and types of jargon words are discussed. Related literature concerning domain-specific words that have been done on more vulnerable domains is discussed.

2.1. Jargon words

Domain-specific jargon words are defined by different scholars based on their findings of research work. Because jargon words are formal languages for a particular domain such as agriculture, they are well-known for employees of a domain.

Jargon words are well-known for experts working in the same organization, and they are necessary for communication. Experts use domain-specific jargon words to explore their ideas besides organizational related tasks. However, the existence of jargon words in a text needs meaning with clear words for non-expert users (Helmreich et al., 2005).

The use of domain-specific jargon words in a text increases the frustration of the people during reading documents, emails and wasting time and money to understand the required meaning of words in a text. Removing jargon words is impossible because domain-specific jargon words are formal languages of organizations and jargon with new concepts are invented in various domains at different times. Besides, experts are expected to minimize the use of more jargon in organizational discourse to increase the content to be understandable by the targeted group (Willoughby et al., 2020).

Jargon is special words, abbreviations, or expressions that are intended to be used in a particular profession and the words or expressions are not simple to understand to the people outside the profession. Professions have jargons to work with their discipline and to commit simple communication with the employees. Jargon words in a particular language are domain-specific words used by a specialized group of a particular field that is broadly accessible by employees working in the same organization; however, the words are difficult to understand for non-expert users (Brown et al., 2020).

The usage of domain-specific jargon words is used to define and describe unique situations and phenomena of a domain. Domain experts communicate with professional

jargon; however, the words break the communication between professionals and lay people. The usage of professional jargon in a domain-specific discourse is mocked by the people outside the domain. So that the use of domain-specific words impedes the communication to be understood by outsiders. Individuals have to know the meaning of the words to acquire professional identity and group membership (Gallo, 2018).

Managers and employees working in the same organization are required to have solid communication to make sure that an organization is running smoothly to achieve its predetermined objectives. Managers have the responsibility to ensure that less professional jargon is being used in daily communication with employees to avoid misperception and miscommunication. Therefore, organizations come up with success because of clear communication between employees and customers. However, with insufficient communication, employees become demoralized as a consequence of high employee turnover, wastage of organizational resources, and finally, organizations drop behind from the objectives (Patoko & Yazdanifard, 2014).

The use of domain-specific jargon words with different industrial knowledge negatively influences the investment willingness of investors. The problem happens when investors are not well-informed with industrial knowledge as the usage of industrial jargon decreases the understandability of investors on the concerned topics. However, when investors are well-known with the meaning of industrial jargon words, they are motivated and increase investment willingness besides their business (Tan et al., 2019). Different professions have domain-specific jargon words with any language that are used within a domain to handle effective communication between employees for the overall achievement of objectives. (Ong & Liaw, 2013).

Medical terminology hampers the communication between clinicians and patients in the medical diagnosis and treatment process. Unexplained domain-specific medical terms halt the communication in a pediatric surgical consultation for parent decision-making on the consultation. So that people in the online medical treatment process are confused to make decisions on the case consulted (Links et al., 2019).

2.2. Agriculture

Agriculture is the domain of preparing human consumption from the products of mainly animals and plants. Agriculture is an art and science that uses soil for growing crops

and producing livestock and is the dominant source of income for the globe. Agriculture had become the primary source of income for a sizable portion of the global population. When people began growing crops, they began herding and breeding wild animals. Domestication refers to the adaptation of animals and plants for human consumption. Many animals on the planet produce milk, cheese, and butter (Demeke & Ferede, 2014).

Domesticated animals like oxen were eventually used for plowing, pulling, and transportation. People were able to produce an abundance of food thanks to agriculture. If crops failed, the people could eat the extra food or trade for other goods. People were able to work on non-farming tasks because of food surpluses. Agriculture kept previously nomadic people close to the fields, resulting in the formation of permanent villages.

2.2.1. Agriculture in Ethiopia

An estimated 85% of the people are employed in agricultural production. The major agricultural exports are coffee, hides and skins, pulses, oily seeds, beeswax, and more often tea. In domestic production for livelihoods, meat and milk are essential. The socialist Derg agricultural reforms included agricultural reforms, which led to fair land tenure patterns. With monopolistic procurement and sale rights of farm commodities, the state retained full ownership. State marketing boards have been established (Demeke & Ferede, 2014).

Currently, although most marketing councils have been abolished, the government maintains ultimate ownership of land in the agricultural sector. Marketing boards allowed farmers to sell their products to the highest bidder. Ethiopia has a variety of ecological areas and a wealth of agricultural resources. Farming is the backbone of the economy and the government also views the agricultural processing sector as one of the driving forces for future economic growth (Welteji, 2018).

The GOE has implemented a series of interventions to assist the development of the agricultural sector concerning increasing productivity, in collaboration with international partners. These activities led to higher crop yields and higher production of livestock. The sector's main binding constraints are insufficient earnings due to inefficient input and service delivery, unclear land rental rights, small irrigation

investments, marketing, logistical problems, and an absence of agricultural-based financial services(Demeke & Ferede, 2014).

Ethiopia has growth potential and opportunities for potential capital investments in some of its cash crops including coffee, olive seed, and pulse, fruit and vegetables, sweetheart, tea, and spices. To generate foreign exchange, the majority of these crops are exported. In the future, the government plans to work with private sectors in developing the capability to process and generate value-added and higher export prices for certain products, such as fruit and vegetables (Welteji, 2018).

The agriculture economy in Ethiopia accounts for 40% of GDP, 80% of exports, and an estimated 75% of the workforce. However, only 5% of the land is irrigated, and farms are under average crop yields. There are shaky market ties and there is still less use of improved seeds, fertilizers, and pesticides. Despite these barriers, farming-led economic growth in connection with improved livelihoods and nutrition may be a long-term solution to chronic food insecurity in Ethiopia (Welteji, 2018).

2.3. Knowledge sources in jargon words identification

The process of domain-specific jargon word identification required knowledge sources to extract the meaning of words based on the domain of words from the lexical knowledge or learned knowledge. Because domain-specific jargon words have unique meanings in the domain, the word's meaning can be extracted from a lexical resource such as Machine-Readable Dictionary, Thesauri, WordNet, ontology. The learned knowledge obtained from the trained labeled corpus (Antonic, 2008; Gasson, 2003). Though knowledge is also used for jargon word identification, lexical knowledge sources are sources of knowledge for domain-specific jargon word identification (Antonic, 2008; Gong et al., 2017).

2.3.1. Lexical Knowledge sources

Lexical knowledge is a predefined list of domain-specific jargon words in line with the word's meaning in the domain. The conventional meaning of words given by prior experts in the domain is used to create the meaning of words in the lexical knowledge source. We constructed a knowledge source with interactive Machine Readable Dictionary.

MRD organizes the lexical information as a list of jargon words and the words meanings in the targeted domain. Though various researchers addressed different issues in the Amharic language, the agricultural Amharic Jargon Machine Readable Dictionary (AJMRD) is not available to the best of our knowledge. So, we developed an interactive Amharic Jargon Machine Readable Dictionary (AJMRD) for our study.

Therefore, the meaning of domain-specific Amharic jargon words (DSAJW) provided with the developed AJMRD. AJMRD for domain-specific Amharic jargon words identification system used to evaluate the effectiveness and efficiency of the developed system because no more work was done for Amharic jargon words identification with AJMRD.

2.3.2. Learned knowledge

Learned knowledge is used to identify automatically from the context of training corpus using various machine learning techniques (Gasson, 2003; Seyler et al., 2020). Learned knowledge considers the nearest words both on the left and the right of the target word using methods like the fixed-size window.

Therefore, we use the hybrid of the combination of learned knowledge and lexical knowledge sources for domain-specific Amharic jargon word identification system (DSAJWI). We use AJMRD as a lexical knowledge source to identify Amharic jargon words that exist in organizational discourse of Amharic text. With the availability of Amharic Jargon Machine Readable Dictionary, identification of Amharic jargon words possible.

Therefore, a text that contains domain-specific Amharic jargon words can be separated with the machine learning model using the learned knowledge and entered into the knowledge-based system to extract the meaning of words from the knowledge source. Finally, a text that contains jargon words with the meaning of the words is returned to the target user.

2.4. Approaches of jargon words identification

Various approaches of domain-specific jargon word identification are used with resourceful languages for various domains such as medicals, scientists, e-commerce websites. The approaches are focused on the classification of domain-specific concepts,

and retrieving the meaning of words from the predefined explanatory lexical resource. Domain-specific jargon words have one meaning as per the usage of words in a particular domain, hence jargon words convey common and clear understanding for employees in the same domain.

The approaches used for domain-specific jargon word identification are required to return predefined knowledge for the word. The meaning induction of words is highly dependent upon the acquisition of prior knowledge with domain experts based on convention. Based on the acquisition of knowledge, domain-specific jargon words identification approaches classified as knowledge-based with lexical resources (MRD, Thesauri, WordNet), machine learning with the corpus (supervised, unsupervised, and semi-supervised), and hybrid (combination of knowledge-based and machine learning) approaches (Pal & Saha, 2013; Seyler et al., 2020).

2.4.1. Knowledge-based approach

Knowledge-based approaches use predefined lexical sources prepared with the help of domain expert curators using external knowledge sources such as MRD, Thesauri, or lexical databases such as WordNet. Besides, no more training of the dataset is required to identify domain-specific jargon words. So that the meaning of words is available and returned from the knowledge resource.

Knowledge-based approaches are used to return the meaning of words. The lexical resources are scalable and we can add any new induced benign-looking domain-specific jargon words for the society to convey a prominent theme of organizational discourse. The number of newly induced domain-specific jargon words has increased at a very decreased rate over time. The experts of a domain are required to provide the meaning of the word as per the expert's convention in a concerned domain. So that the domain-specific jargon words are organized and stored in a lexical resource (Gong et al., 2017).

The knowledge-based approach provides the meaning of words by mapping domain-specific jargon words with the words' meaning in a specific domain with predefined explanatory knowledge resources. Knowledge-based approaches use lexical resources such as Machine-Readable Dictionary, Thesauri, or WordNet (Gong et al., 2017).

Machine Readable Dictionary: Machine-Readable Dictionaries (MRDs) is an organized collection of lexical knowledge which is useful for Natural Language

Understanding (NLU) of the language. MRD is helpful for further analysis and development of various applications with the natural language. The provision of the meaning of words is sourced from the predefined steady MRDs. MRDs are used for researchers to return meanings from the source (Calsolari, 1984)

Thesauri: thesaurus is used to provide the synonymy relationship between words (Kilgarriff & Yallop, 2000). The words with their meaning can be maintained with the construction of Thesauri. Thesauri is capable of processing large-scale language with a rich network of word associations. The meaning of words can be provided from the thesaurus by looking at the pair of dictionaries meaning without the need of any further corpus training with a large dataset (Ryan, 2014).

WordNet: identification of word can be performed with the predefined list of words and their meaning in the language. WordNet can be developed with the help of domain experts inline to provide the relationship between words (Antonic, 2008; Piasecki et al., 2009).

Ontology: ontology constitutes a knowledge base with a set of instances that is helpful to analyze domain knowledge, share a common understanding of the structure of information among people, reuse domain knowledge, make domain assumptions, share domain knowledge from the operational knowledge (Lamy, 2017).

2.4.2. Machine learning approach

Machine learning is an embodiment of future prediction from experience that focuses on automatic learning methods. So that machine learning is the improvement of algorithms learned from experience without the help of human experts.

Corpus-based word identification is based on statistical and machine learning algorithms and the meaning of the word is predicted from the available set of alternatives inline to the word's context in a text. Meaning induction of the words can be returned from large annotated data using learning and classification phases of the developed model. The learning phase consists of learning the context of words from the corpus training whereas the classification phase consists of the application to return output senses for the words. The four machine learning approaches are namely supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Bakx, 2006).

Supervised machine learning

Supervised machine learning algorithms learn a dataset to recognize words from the learned knowledge. Training labeled corpus can be prepared manually with the help of domain expert curators. After the training of data, a developed system automatically returns the prediction labels of input test data with the learned model based on the training. Though it requires a manually curated labeled corpus for each domain-specific jargon word which is expensive to create, supervised learning methods return performance results with high accuracy. Annotated data with the help of a domain expert curator can be split into training and testing data (Gasson, 2003). Because they have high-dimensionality of the feature-space, selecting the most preferable machine learning technique among support-vector machines (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN) based on the nature of the dataset and the problem is left for the user. Supervised machine learning algorithms are better as compared to unsupervised machine learning algorithms in performance based on the relation of words with probability distribution among the training and testing dataset (Seyler et al., 2020).

Learning and testing are the two main parts of supervised machine learning. Learning attempts to learn a developed model with the training data, and testing focuses on testing a model with unseen test data to appraise the model's accuracy. So that a machine learning model learns from experience to capture best-learned knowledge to make accurate decisions, prediction, and maximize payoff.

A supervised machine learning approach infers an objective function from the labeled training corpus. The approach requires prior information about the environment to infer the output besides the input. The main function of supervised machine learning is classification by the classifier function that fits the characteristics of the trained labeled corpus. So the trained classifier is used to map and classify the new coming text. The prediction of the label for test data in supervised machine learning is performed by the classification with the help of the process of finding models that describe data classes using the learned knowledge (Ikonomakis et al., 2005).

So that the developed model performs prediction of unknown labels based on the training of data. For example, a text that contains words such as ‘ኮሲ (kosi), ደረቦ (derebo), ደሬርክን (derierken), ድምራፊር (dimurafer), ደብረት (debret)’ are jargony, hence

the model first trained as the words are jargony. Some popular machine learning classifiers are discussed as follows.

Support Vector Machine (SVM)

Support vector machine (SVM) used for linear classification of the input data. A linear classifier maps the original input vectors to high-dimensional feature space with a separable training set was the aim of the original SVM approach (Sang et al., 2008). Non-linear SVM classification using the kernel trick was proposed (Boser et al., 1992). SVM is applicable in a wide range of supervised machine learning applications because of its robustness to noise and errors, accurate predictions, and fast evaluation of the target function (Byun & Lee, 2002).

SVM is a popular supervised machine learning technique that takes labeled training corpus for classification with the labels with the hyperplane. The kernel of SVM performs both linear and non-linear classification. SVM develops an optimal hyperplane with the input training data and the decision plane turns test data into labels (Byun & Lee, 2002).

Therefore, we use the SVM supervised ML classification algorithm because it offers the best classification performance on the large labeled training data. SVM provides a pure classification of future test data with more efficiency and the problem of overfitting of data controlled with the powerful kernel such as linear, poly, RBF, sigmoid and the regularization parameter (C). SVM works on the principle of margin calculation for classification purposes by drawing margins of the maximum possible distance between the margin and the nearest data points (support vectors) of classes. Quadratic optimization algorithms can identify which training points from the subsets of x_i (training data) are support vectors.

Therefore, the two-way classification for our data is performed with the SVC supervised learning method that takes a set of labeled training data to generate a classification function (Liebowitz, 2010). The SVC takes the labeled dataset to generate an output for the input with the hyperplane to separate different classes with separate training samples. SVC makes decisions based on the support vectors. So that SVC maps the input test data to classes of two-way classification.

SVM works with the principle of computational learning theory with the help of a decision surface to separate the training data points into two classes based on the support vectors. The support vectors are selected as the only effective elements in the training set based on the principle of structural risk minimization and quadratic optimization algorithm (Ananiadou et al., 2009).

SVM separates labeled training corpus using a hyperplane with maximizing the margin, the distance between the hyperplane to the labeled classes. The following figure illustrates the SVM hyperplane and class label.

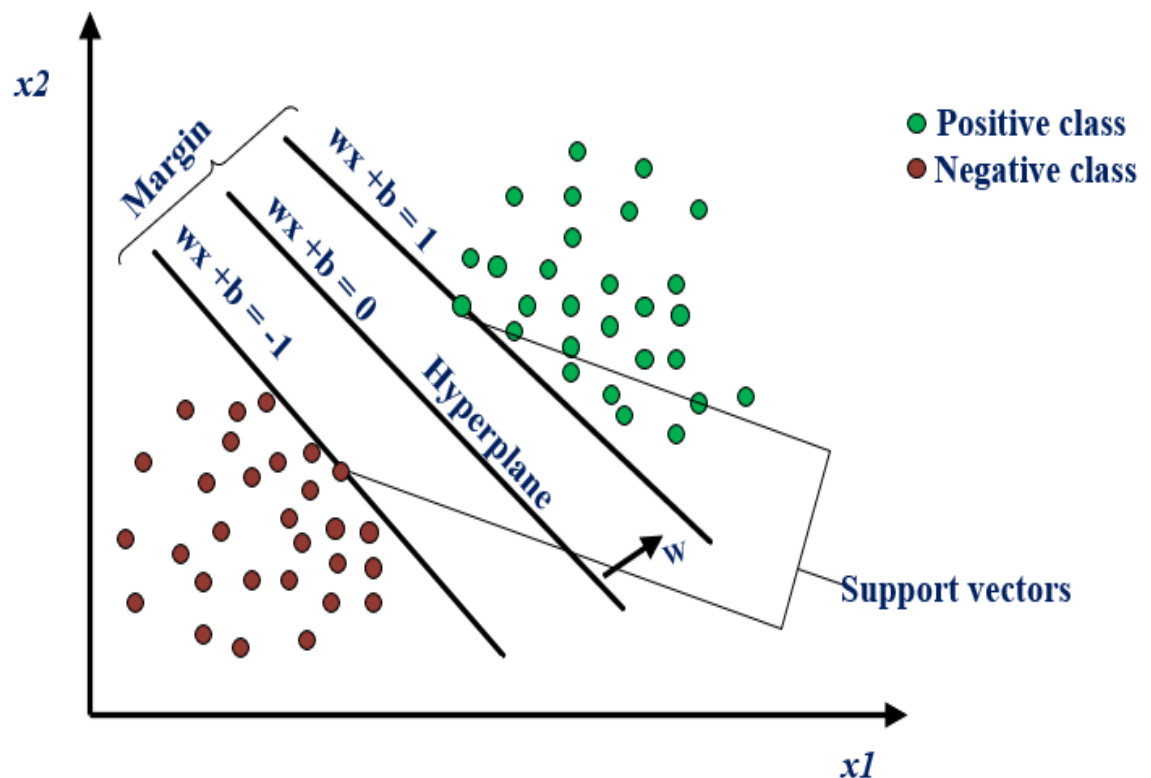


Figure 2.1: System architecture of SVM supervised ML classifier

The small green and red circles in the above figure represent positive and negative training examples respectively, whereas the line represents optimal separating hyperplane (OSH) and the area of the margin for decision surface.

The hyperplane on the above figure is the best possible one as it is the middle element of the widest set of parallel decision surfaces. The training set on the left and right-side lines of the hyperplane indicate support vectors. The final decision of unknown class label for the input test data is made with the trained SVM model based on the OSH.

(Optimum Separation Hyperplane) instead of the whole training set. The decision was performed based on finding out on which side of OSH the pattern of test data is located.

A set of N linearly separable points $S = \{x_i \in \mathbb{R}^n, i = 1, 2, 3, 4, \dots, N\}$, each point x_i belongs to one of the two classes labeled as $y_i \in \{-1, +1\}$. The points in the two-class labels have the same class label in which two sides division can be performed by the separating hyperplane. The separating hyperplane is identified with the pair (w, b) that satisfies:

$$\begin{aligned}
 & w \cdot x + b = 0 \\
 \text{and } & \left\{ \begin{array}{l} w \cdot x_i + b \geq +1 \text{ if } y_i = +1 \\ w \cdot x_i + b \leq -1 \text{ if } y_i = -1 \end{array} \right. \quad \text{for } i = 1, 2, \dots, N
 \end{aligned} \tag{2.1}$$

Where w is a weight vector normal to the line, b is the bias, x_i is the input vector, and y is the label.

The dot product (\cdot) is defined by:

$$w \cdot x = \sum_i^N w_i x_i \quad w, x \text{ are vectors} \tag{2.2}$$

The main goal of SVM learning is to find the optimal separating hyperplane (OSH) that has the maximum margin to both sides of the support vectors. Maximum-margin of the SVM model is helpful according to the intuition and PAC (Probably Approximately Correct) theory. So that the development of maximum margin classifiers is dependent on the position of support vectors. This can be formalized as:

$$\begin{aligned}
 & \text{Minimize margin with } \frac{1}{2} w \cdot w, \text{ and maximize margin } \frac{2}{\|w\|} \\
 \text{Subject to } & \left\{ \begin{array}{l} w \cdot x_i + b \geq +1 \text{ if } y_i = +1 \\ w \cdot x_i + b \leq -1 \text{ if } y_i = -1 \end{array} \right. \quad \text{for } i = 1, 2, \dots, N
 \end{aligned} \tag{2.3}$$

So that SVM is highly competitive and high as compared with other traditional pattern recognition methods in terms of computational efficiency and predictive accuracy (Joachims, 1997; Yang & Liu, 1999). This approach is also applicable in the case in which positives and negatives are not linearly separable. We use SVM for our two-way classification, hence SVM applied successfully in many texts' classification tasks with

the advantages such as robustness in high dimensional space. SVM is robust when there is a sparse sample. The challenge to develop a model with an SVM algorithm is the difficulty to determine the best values of the parameters besides the nature of the dataset.

Naïve Bayes (NB)

Bayesian employs Bayes' Theorem of conditional probability with a Naïve assumption that every pair feature is mutually independent. NB is used to solve classification and regression problems. Naïve Bayes is a strictly supervised machine learning algorithm in which features are learned independently and the final decision is based on the learned independent features. The advantage of making predictions with NB is that a classifier works well with small labeled training data (Berrar, 2018).

The Bayesian networks are the representation of the probability distribution over the set of features used for the learning process. The features such as F_1, F_2, \dots, F_n in a dataset are independent of each other; however, these features are dependent on the class labels like jargony, and non-jargony (Taheri & Mammadov, 2013). The Naïve Bayes classifier is the simplest probabilities classifier. We can observe the best performance of the Naïve Bayes classifier in various real-world applications; hence the algorithm is easy, fast, and well performed in high dimensional data.

The prediction of Naïve Bayes with Bayes rule and available features uses the formula to get the highest posterior probability.

$$P(C|F) = \frac{P(C)P(F|C)}{P(F)} \quad (2.4)$$

The Naïve Bayes (NB) classifier is a supervised learning algorithm with the probabilistic model that uses the joint probabilities of the terms and class labels to estimate the probabilities class labels in a given input test data (Peng et al., 2019).

The speed of computation operations of the NB classifier increased because the parameters for each term learned separately from the term's independence behavior. The three models for NB are Multinomial NB, and Bernoulli NB, Gaussian NB. The models apply Bayes' rule for classification (Mccallum & Nigam, 1997; Peng et al., 2019).

$$P(ci|dj) = \frac{P(ci)P(dj|ci)}{p(dj)} \quad (2.5)$$

where d_j is a test data and c_i is a class.

The posterior probability of each category c_i given the test data d_j , $P(c_i|d_j)$ is calculated, and the category with the highest probability is assigned to d_j . The value of $P(c_i)$ and $P(d_j|c_i)$ have to be estimated from the training set of documents to calculate the posterior probability, $P(c_i|d_j)$. $P(d_j)$ is the same for each category from the computation. The category of prior probability $P(c_i)$ can be estimated as follows.

$$P(c_i) = \frac{\sum_{j=1}^N y(d_j, c_i)}{N} \quad (2.6)$$

Where N is the number of training data and $y(d_j, c_i)$ is defined as follows:

$$y(d_j, c_i) = \begin{cases} 1 & \text{if } d_j \in c_i \\ 0 & \text{otherwise} \end{cases} \quad (2.7)$$

So that the prior probability of category c_i is estimated by the fraction of data in the training set belonging to c_i . $y(d_j, c_i)$.

Multinomial Naïve Bayes

A document d_j in a multinomial model is an order sequence of term events drawn from the term space T . The NB assumption is that the probability of each term event is independent of the term's context, position in the document, and length of the document. So that each document d_j is drawn from a multinomial distribution of terms with several independent trials equal to the length of d_j (Joachims, 1997).

Bernoulli Naïve Bayes

A document is represented by a vector of binary features indicating the terms that occur and that do not occur in the document; hence the document is the event, absence or presence of terms is the attributes of the event. NB works with the principle that the probability of each term being present in a document is independent of the presence of other terms in a document. The absence or presence of each term is dependent only on the category of the document. The probability of a document given its category is simply the product of the probability of the attribute values of all term attributes.

The various NB models are compared, and the multinomial NB model is almost uniformly better than the Bernoulli model and reduces error. We used and evaluated our work with the multinomial NB model (Mccallum & Nigam, 1997).

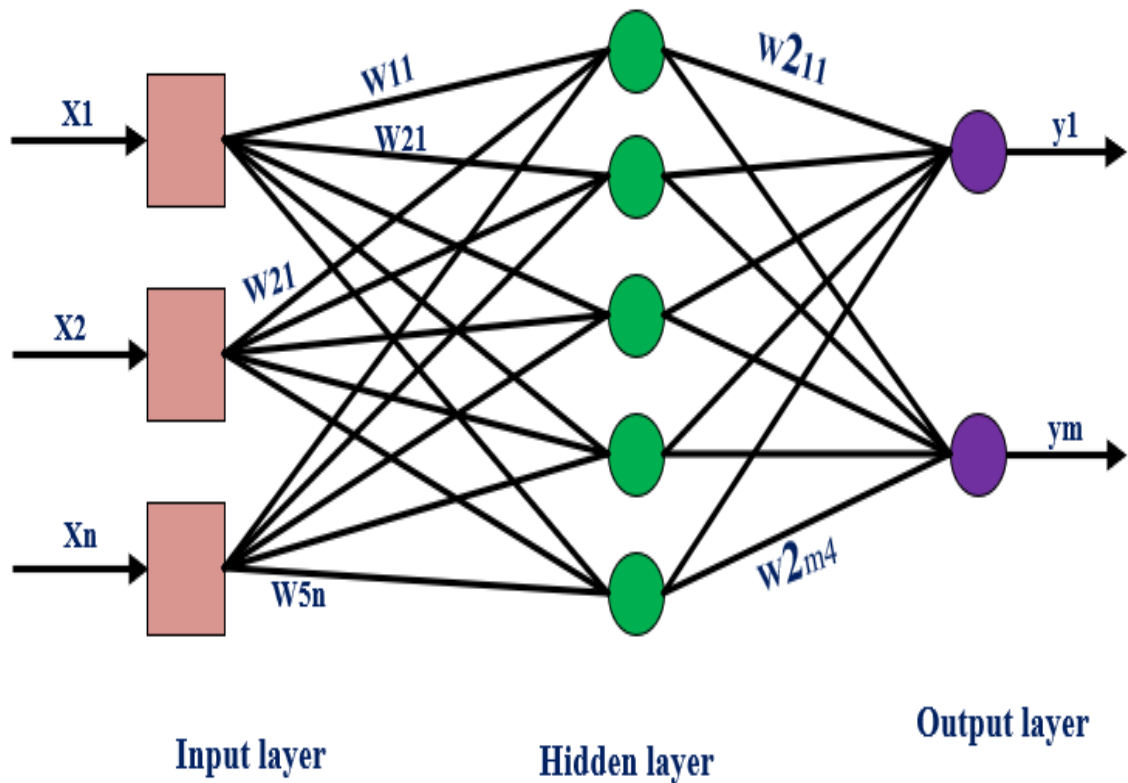
Gaussian Naïve Bayes

The Gaussian Naïve Bayes algorithm is an efficient type of Naïve Bayes algorithm introduced for the representation with the probability distribution over the set of features $f_1, f_2, f_3, \dots, f_n$ which are used for the learning process. The set of features are independent of each other; hence the features are dependent only on the class. So the Naïve Bayes classifier is used for various real-world applications and performs well for high-dimensional data (Taheri & Mammadov, 2013).

Artificial Neural Network (ANN)

Artificial Neural Networks (ANN) are advanced techniques of machine learning applicable in various areas of interest and can be learned irrespective of the type of data. ANN is an essential part of deep learning; hence it is a subset of machine learning. Neural networks are a collection of computational units interlinked by the system of connections. It is applicable in many applications such as pattern classification, and pattern recognition. We use ANN for our purpose of two-way classification of the input text and to return the classified text for further analysis (Cheng, 2015). ANN is the simplest type of neural network based on the Feed-Forward strategy. So that the data flows in MLP are with forwarding direction from the input to the output layer like feed-forward network. The neurons in the MLP are designed to approximate any continuous function and can be non-linearly separable (S. Abirami, 2020).

The following figure shows the architecture of a multilayer perceptron with input, hidden, and output layers.



(Sharkawy, 2020)

Figure 2.2: The architecture of ANN with Multilayer Perceptron (MLP)

The three types of the layers input layer, hidden layer, and output layers are required for the multilayer perceptron with the feed-forward neural network. The input data received at the input layer to be processed; the input data is computed by the arbitrary number of hidden layers of MLP placed between input and output layer; hence prediction, pattern classification, approximation, and recognition are performed by the output layer (S. Abirami, 2020).

Unsupervised machine learning

Unsupervised machine learning techniques are language and domain-independent that use unannotated training data for classification without the necessity of a data curator. Unlike data annotators for supervised machine learning, unsupervised machine learning techniques use clustering algorithms for labeling instances in the training data. The algorithms solely require a dataset with a particular language and any

domain of interest. However, because of no availability of annotated data, the result accuracy of the system is low as compared to supervised machine learning techniques. The approaches follow the principle of that same words have similar neighboring words. So that the meaning of the jargon words can be induced from the input text by clustering word occurrences, and classifying new occurrences into the induced clusters (Weng et al., 2019).

Unsupervised machine learning approaches are applied to the situation in which the prior knowledge is unknown. Unsupervised learning algorithms learn hidden features of the unlabeled input. The main function of unsupervised learning is clustering, dimensionality reduction, blind signal preparation. Unsupervised clustering machine learning algorithms such as K-means, Centroid-based algorithms group similar objects in the same cluster, and different objects in different clusters (J. A. Hartigan and M. A. Wong, 2012).

Semi-supervised corpus-based approach

A semi-supervised approach is an approach for domain-specific jargon word identification that requires a small amount of labeled data and a large amount of unlabeled data for training. The performance result of a semi-supervised approach is between the supervised and unsupervised approach which best performs during scarcity of data (Pal & Saha, 2013).

2.4.3. Hybrid approaches

Hybrid approach a merge of both characteristics of knowledge-based and machine learning approaches. These approaches cannot be categorized as knowledge-based or machine learning because the meaning of the jargon words is from both the knowledge base with knowledge source and the ML with learned knowledge.

The hybrid approach combines lexical information from the knowledge source such as Machine-Readable Dictionary (MRD), Thesauri, and learned knowledge from the trained dataset. So, using hybrid systems is strengthening the developing system and it outperforms the performance result of individual approaches either knowledge-based approach or machine learning approach by overcoming specific limitations faced on individual approaches.

Therefore, we are required to use a combination of supervised machine learning and knowledge-based approach for domain-specific Amharic jargon word identification systems to make our system more robust and flexible. Because this is the first attempt in the area, collecting sufficient data for both knowledge source construction for the KB, and labeled sentences for training is a challenge. So that supervised machine learning techniques are employed to develop our model with labeled training of prepared dataset. The developed machine learning models are used for the predictions of the input test data based on the labeled trained dataset. The integration of a knowledge-based system and machine learning model is employed for the identification of domain-specific Amharic jargon words and to extract the word's meaning for every occurrence of a word in the text. Classification of the whole input text can be performed by the machine learning model; however, identification of a word is performed with the knowledge-based system with the predefined explanatory lexical resources.

2.5. Amharic Language

2.5.1. Amharic writing system

Amharic language that has its unique script is the dominant language in Ethiopia. Amharic is the world's second-largest Semitic language that uses constant-vowel pairs. Amharic is a morphologically complex and under-resourced language so few computational linguistic resources have been developed so far (Gasser, 2011). Amharic with its Ethiopic script called fidel follows a left-to-right writing system. The Ethiopic script 'fidel' is sourced from the Geez language in which the Geez language is the Ethiopian Orthodox Tewahdo church praising language (Mikre-Sellassie, 2000). There are Amharic speakers in the world that include Israel, Canada, USA, Egypt, Eritrea, and Sweden (Hudson, 1999). So that the system developed for the Amharic language is beneficial for the people worldwide.

The volume of electronic Amharic documents is raising hence, Amharic has a wide application in domains of Federal Democratic Republic of Ethiopia (FDRE). The information sharing and transfer between agricultural organizations in Ethiopia is with Amharic language because the language is the national language for the country Ethiopia. So that Amharic is used for communication between agricultural domain

experts, non-domain experts, and agrarian society to proceed with agricultural activity to ensure food security.

2.5.2. Amharic variant characters

Amharic words are the organized combination of phonemes with their orthographic representation in the language. Though many speakers are in the Amharic language, many challenges are there in writing scripts with the language. Though each of the scripts (fidels) has its unique meaning in Ethiopian Orthodox Tewahdo Church and in the language, usually scripts having similar phonemes with different orthographic representations have the same significance in Amharic language processing. So that the use of similar phonemes with different orthographic representations interchangeably to return the same meaning is accustomed. The scripts (*ሀ* (Ha), *ሐ* (Ha), *ኀ* (Ha)), (*አ* (A), *ሐ* (A)), (*ጸ* (Tse), *ፀ* (Tse)), and (*ሰ* (Se), *ሠ* (Se)) have the same phonemes to convey similar meaning in the language in Amharic language processing. For example, the word, ‘sun’ can be written as *ጸሀይ* (tsehay), *ፀሀይ* (tsehay), *ፀሐይ* (tsehay), *ጸሐይ* (tsehay), etc. differently. These challenges can be resolved using normalization to work with NLP systems and applications. Normalization is a challenge for Amharic language processing and got problems for the representation of variant Amharic characters based on the characters meaning hence, Amharic is low-resourced language (Zupon et al., 2021).

2.5.3. Amharic punctuation

The Amharic language does not have uppercase and lowercase representations of letters. The discourse written in the Amharic language includes punctuation marks with different functions that support the written discourse with the relevant meaning to the target user of the information. Some of the punctuation mark used in the Amharic language include *ሁለት ነጥብ* (:) (colon) used for separation of words, *አራት ነጥብ* (::) (full stop) used for separation of the sentence, *ነጠላ ሰረዝ* (፣) (comma) used for separation of Amharic words or phrases with similar concepts, *ድርብ ሰረዝ* (፤) (semicolon) used for separation of Amharic sentences with a similar concept. Some other punctuation marks of Amharic language that are sourced from other languages include *ቃለ አጋኖ* (!) (exclamation mark) used for making attention to the transmitted information, *ጥያቄ ምልክት* (?) (question mark) used for determining a request for the situation and wait for a response for the requested information.

2.5.4. Amharic language inflection and derivation

The various forms of Amharic words with the inflection and derivation of words make Amharic sentences challenging for Amharic text processing. The unavailability of effective and efficient Amharic morphological analyzer and Amharic stemmer increases the challenge of working with Amharic language text processing though, some works are done for Amharic language analyzer (Gasser, 2011). Therefore, domain-specific Amharic jargon word identification without the availability of effective and efficient Amharic spelling checker, morphological analyzer, and stemmer results with less performance.

The Amharic language is a highly inflected language with rich morphology that follows a semi syllabic writing system called Fidel (Eyassu, 2005). The morphological complex behavior of Amharic language is based on consonantal roots with vowel variants describing variants of the root form. Meaningful Amharic words can be generated from Amharic phoneme, morpheme, root, stem, and word. So that the Amharic base characters are phonemes which is the smallest meaningful unit in a word that forms morphemes (Mindaye & Atnafu, 2009).

The Amharic language has 34 base characters called ‘Fidels’ each of the characters occurs in the basic form and six other forms also called orders that follow a regular pattern of vowel usage. The language uses more than 40 other characters that contain special feature usually representing labialization as the basic sounds in the language. The 34 basic characters and their orders give 238 distinct symbols. The base characters are the combination of alphabets from the Ge’ez language and the added alphabetic characters such as ቨ (ve): ኘ (gne): ቸ (che): ዠ (ze): ጪ (che).

2.5.5. Amharic jargon words

Domain-specific Amharic jargon words (DSAJW) are specialized or technical words that have a common understanding and are used frequently by people who are members of a particular profession such as agricultural professionals, health professionals, lawyers. Domain-specific Amharic jargon words are the necessary words for professionals to short-hand much larger concepts to increase the precision of words within the profession to disseminate the targeted information. Though domain-specific Amharic jargon words have hidden meaning for the user out the domain, writer professionals of a particular profession use domain-specific Amharic jargon words to

explore contents and to transmit information to the targeted group. DSAJW is also used to characterize a specific group of people based on their profession; hence most of the time they use these words for everyday business alike the resourceful languages (Helmreich et al., 2005; Pal & Saha, 2013; Rakedzon et al., 2017).

Domain-specific Amharic jargon words are associated with various professions because employees of a particular profession use domain-specific technical terminology in common with common understanding to explore many concepts with a word or a phrase. Understanding jargon words is impossible because we found various types of jargon in various domains and also, jargon is the function of the difference between the communicants. Besides, various Ethiopian institutions use domain-specific Amharic jargon words to cooperate with the employees and business of an organization.

Jargon words are necessary (unavoidable) for good academic writing to explore a certain content in line with the domain (Schmitt, 2000). Besides, the three types of jargon are niche terms of field, acronyms, and erudite vernacular utilized irrespective of necessity (Ong & Liaw, 2015; Oppenheimer, 2006).

Niche terms of field

These words are especially used by experts of a domain to decide with employees and to transmit domain-specific information to the targeted group of receivers. Niche terms are used frequently in a particular domain; however, the terms have hidden meaning to the people out of that particular discipline. The words are the reason for the confusion between people in inter-discipline professionals. For example, the Amharic agricultural jargon words such as ‘ኮሲ (kosi), ደረቦ (derebo), ደሬርክን (derierken), ደምራፈር (dimurafer), ደብረት (debret), ገጣፈደ (getafede), ጋፈር (gafer)’, are frequently used between agricultural experts; however, the words have hidden meaning for non-expert reader and the huge agricultural society. So, our proposed DSAJWI system uses niche terms of the field to identify words for non-expert readers and the huge agricultural society.

Acronyms

Acronyms are short representation words that are common in various domains. People outside of the domain are confused with the acronyms used in a discourse of a particular domain. Because acronyms are the reason for confusion, the target content of the information is not received from the receiver as expected. Organizational discourse initiates experts to use domain-specific acronyms to commit a simple communication

with the target group. Acronyms also happen in the thesis, and dissertation when researchers explore their findings to society. The readers are expected to find the full term of acronyms to return with a prominent understanding of the explored content.

Erudite vernacular utilized irrespective of necessity

These words are harder to understand to the people including some of the employees in the same domain. Jargon words are sourced from extremely erudite people with a particular profession. Words used in the discourse of extremely erudite experts within a particular domain are the reason to hurt the understanding of some employees in the same domain and the people out of the field. The words used in erudite vernacular are extremely experts' words that have hidden meaning to the society including some of the employees of an organization.

Therefore, communicants are required to know that communicating in the same language and working on the same domain doesn't guarantee a common understanding of the content. Though effective communication using clear words with customers is essential for organizations, experts are forced to use domain-specific jargon words to handle simple communication for large concepts. So that the use of jargon words is preferable for experts to simplify concepts.

2.6. Related work

The related work introduces the proposed and implemented works related to domain-specific jargon words identification with resourceful language for various domains. Because domain-specific jargon word identification is an emerging area of research in natural language processing, only few works were done around the globe. Most of the proposed works use a knowledge-based dictionary-based approach, ontology; few works follow the machine learning approach. Most of the authors with the proposed approach in the related work attempt on classifying domain-specific words, provide meaning from lexical resources. Though few works are proposed and implemented for jargon identification, the authors focused on the more likely vulnerable domains like medicals, scientists, and e-commerce websites.

2.6.1. Jargon words identification for medicals

Medical words are challenging to understand by ordinary people (by non-medical people). Biological concepts require induction of meaning to be understandable to non-

experts using predefined ontologies by domain expert annotators. The authors use dictionary-based Variable-step Window Identification Algorithm (VWIA) for biomedical concept classification. Datasets are collected by crawling the URL of the necessary website. After the necessary preprocessing techniques are performed the developed system returns classes of biomedical concepts based on the constructed dictionary with an F-measure of 95%. However, this work is attempted for different classes of biomedical concept classification for further analysis but not meaning identification of concepts. Layman's are required to have meaning of the classified biomedical concept for prominent understanding of the theme (Gong et al., 2017).

The communication between physician and patient requires a clear understanding for efficient diagnosis and treatment. However, the communication of physicians is full of domain-specific professional jargon words that hamper the clear understanding of patients in the treatment and consultancy process. The use of medical jargon diminishes the communication between patients and physicians and increases the social distance between physicians and patients. Professional jargon words used by physicians need to be translated into clear patients' words to improve patient-physician communication. The authors used an embedding alignment method for the word mapping between professional words and patient terms using the data collected from the MIMIC-III database (Roth, 1996). From the two algorithms used for embedding, the Procrustes algorithm with anchors approach outperforms adversarial training for mapping professional jargon words to patient clear words. It achieves an accuracy of 54% using embedding skip-gram algorithm at the word level, and 78% using embedding fastText algorithm at subword level; however, a word with concept level identification is recommended for future work. The authors focused on direct translation at words, and subword level with the data stored in the database. However, classification of a text is required as a text with a domain-specific words or without domain-specific words. So that first classification of a sentence that contain jargon words are required before mapping of a word to the meaning (Weng & Szolovits, 2018).

Physicians and patients require effective communication to come up with the best outcomes of the treatment and consultancy process. Translation of clinical jargon-to-layperson understandable language is essential to improve the communication between physician and patient in the process of treatment, and consultation. This clinical jargon

translation is also used for physicians with the active involvement of patients to increase their decision-making ability concerning the patient's health conditions. The authors use unsupervised learning for unseen datasets using representation learning, bilingual dictionary induction, and statistical machine translation. The embedding space of the words can be learned from unsupervised skip-gram algorithms to preserve the semantic and linguistic properties. The authors use unsupervised bilingual dictionary induction (BDI) to learn a mapping dictionary for the alignment of embedding spaces and return a precision of 82.7% at the subword level (Weng et al., 2019).

Web-based treatment and patient consultation today have increased (Cyr, 2012). In a web-based application, physicians use many medical jargon words for treatment and consultation; this may result in the patient's frustration and confusion. The use of medical words in the digital world using different platforms on the internet is increasing. Because of the confusion and frustration of patients, the authors generate new Consumer Healthcare Vocabulary (CHV) using predefined lexical source or ontology for the medical jargon in the online consultation process to increase the understanding of patients. The authors use word embedding with GloVe Iterative Feedback (GloVeIF) and basic GloVe. The GloVeIF outperforms by 8.7% of the F-measure from the basic GloVe (Ibrahim et al., 2020).

2.6.2. Jargon words identification for scientists

Scientists communicate with scientists of the same department on the progress of the development and innovation of the technology using scientific communication, and also communicate with scientists of other fields to explore their findings on the concerning issue using science communication. The use of professional jargon for scientists makes the target theme of the content hidden from the receiver. Avoiding domain-specific jargon words is a challenge for scientists to convey the required information to the targeted receiver. The authors use over 90 million words from the BBC site for three consecutive years to determine domain-specific jargon words. The developed De-jargonizer with five stages helps scientists to identify the meaning of jargon words to non-experts for science communication. Jargon words were selected by classifying words based on their frequency as high frequency (behavior), low frequency (protein), and jargon (dendritic). So that the De-jargonizer detects the existence of domain-specific jargon words in a text with color code to return the rate of

jargon words in a text, and writers understand the word as a jargon word (Rakedzon et al., 2017).

2.6.3. Jargon words identification for others

Detection of domain-specific words in electronic data in different communication mediums like the internet, mobile services were proposed. The proposed work used semi-supervised learning technique to derive the probability of a suspicious word to be a jargon word by the synset and concept analysis of the text. The current telecommunication system and World Wide Web (WWW) play a vital role in the fast and modern era communication and information sharing via e-mail, chatting, community forums, SMS etc. Communication between people who are far away from us can be handled with a single click or press a single button. However, the facilities have negative influences on the communicants at the time of information sharing with the existence of jargon words. At the time of submission on to the web or any network, the developed algorithm detects the jargon words used in different text. Different countries. The proposed work handles the jargon words with the comparison of word entries in the input text and list of words in the jargon database; hence the database is populated with the jargon words. So that for any word entry in the input text matched with the list of words in the jargon database, the process stops proceeding with the message and the sense of the word derived from the text (Pal & Saha, 2013).

Meetings held by professionals are rich with domain knowledge expressed by domain terminology also called professional jargon terms that positively impact the performance of meeting summarization systems. In this work, the gold-standard annotation for domain terminology from meeting corpus analyzed. The performance of the meeting summarization system with and without the occurrence of domain terms is evaluated. Jargon words or expressions are identified by human annotators. Because meeting summarization enables users to efficiently browse their interests and facilitate information sharing. Domain terminology plays a significant role in determining the salient part of the text on the particular domain. BERT-LARGE (Bidirectional Encoder Representations from Transformer) that contains 24 layers of transformer blocks, 16 attention heads, and 1024 dimensional hidden vectors (Devlin et al., 2019), are used to determine the performance of the summarization system with and without the occurrence of jargon words. Therefore, the occurrence of domain-specific jargon words

improves the performance of meeting summarization systems by 4.3% F-measure as compared to the meeting summarization system performance without jargon words (Koay et al., 2020).

Dark Jargon words are words that appear in a text; however, contain a hidden meaning to the user and it requires clean words that substitute during the understanding of the information. The authors use the word distribution model with Kullback-Leibler Divergence (KL), and cross-context lexical analysis (CCLA) methodology to detect the presence of jargon words in a text and mapped to the word meanings. Binary mapping of dark words to clean words with no hidden meaning is investigated using dark corpus and clean corpus. The word distribution of KL methodology outperforms around 90% of MRR from CCLA for all words and simulated dark words; however, the CCLA performs better for all words of 97.4% and performs worse for simulated dark words. So that KL outperforms the CCLA for the target dark jargon word identification to provide meaning (Seyler et al., 2020).

E-commerce websites such as Amazon are most likely to use fashion jargon words to advertise products available in the database to motivate customers to order products. Data-driven solution with a deep learning approach used to convert high-level fashion concepts into low-level fashion concepts to provide precise information to the customers. 1546 fashion keywords with 5 categories were collected from the corpus to train the deep learning model. After all, prediction of high-level concepts and substitution with low-level concepts was made (Shen et al., 2020).

Therefore, studying domain-specific jargon words in various product and service delivery processes is necessary to benefit non-experts, customers, and society; since non-experts and societies are significantly negatively impacted by the existence of domain-specific jargon words. Identification of domain-specific jargon terms minimizes the communication barrier that will happen during the transmission of information in science communication.

Besides, the previous works inline domain-specific jargon words identification, most of the proposed works use a knowledge-based approach using a dictionary of jargon words in a particular domain for meaning-extraction and classification of concepts. Therefore, for our research work, we followed a hybrid approach with labeled trained dataset for machine learning models and using Amharic Jargon Machine Readable

Dictionary for a knowledge base for classification and extraction of the meaning of domain-specific Amharic jargon words.

To the best of our knowledge, there are no prior works in domain-specific Amharic jargon word identification (DSAJWI) using texts in a particular domain. So that we are motivated to do our research work on domain-specific Amharic jargon word identification systems in the agricultural domain.

2.7. Summary

This chapter discussed the definition of domain-specific jargon words by different scholars based on their findings of research work. The knowledge sources required to develop a jargon word identification system, various approaches followed by scholars in various domains like medicals, scientists are included. The morphologically complex nature of Amharic language script is also discussed. Finally, related works of various domains such as medicals, scientists, e-commerce websites are discussed. Based on the related works, we found the sound problem of domain-specific jargon words on the customer side to understand the main theme of the disseminated content. Some of the authors attempted on domain concept classification using dictionary of jargon words. Some others concerned to provide meaning for dictionary of domain-specific jargon words. However, the works are attempted on either classification of domain-concepts or provide meaning for dictionary of words for the resourceful language. So that our work concerned on the classification of texts with and without jargon words and provide meaning of a jargon words in the classified jargony Amharic text.

CHAPTER THREE: DESIGN OF THE STUDY

In this chapter, the proposed domain-specific Amharic jargon word identification system architecture is discussed. We collected the labeled corpus with and without jargon words from various agricultural documents with the help of domain-expert curators. The meaning of jargon words for knowledge source construction was obtained from domain experts and later reviewed by domain-erudite. The chapter described the way to design the DSAJWI system and the function of each component in the proposed system to recognize and provide the meaning of words are discussed as follows.

3.1. Design requirements

The process to design domain-specific Amharic jargon words identification (DSAJWI) system required to have machine learning models for training of labeled dataset and knowledge base with lexical resources to provide the meaning of words. We used labeled dataset by domain expert curators to train the developed machine learning model, and also, we use Amharic Jargon Machine Readable Dictionary with a collection of jargon words as the lexical resource. The main design requirements in the DSAJWI system are machine learning models and knowledge sources.

3.1.1. Machine learning

Machine learning is used to design any machine to perform capabilities associated with intelligence. We used machine learning models to design the DSAJWI system to acquire learned knowledge from the labeled training data. We employed supervised machine learning algorithms to develop a model for the prediction of unseen test data. So that the model performs the classification process with labeled training data.

The input training data and the corresponding class labels are known before training. Training of supervised machine learning models with known training data is helpful for the algorithm to predict the new unseen data. Supervised machine learning promotes the approach to an advanced level by providing labeled training data essential for the machine to train and predict the new unseen test data. The classification commits besides the label of the training data using the feature vector of unseen data.

3.1.2. Knowledge base

Knowledge sources are the sources of information in which lists of domain-specific Amharic jargon words and the words meaning in agricultural domains are found. Lexical knowledge defined by the domain expert curator and constructed by the researcher is the main knowledge resource of DSAJWI (Gong et al., 2017; Seyler et al., 2020).

The knowledge base component of domain-specific jargon identification consists of phases to identify a jargon word with lexical binary mapping between the word from the input text and the word from the knowledge source. Mapping between the input token and words from the knowledge source can be performed from external knowledge sources like Machine Readable Dictionary.

As discussed, there are different approaches followed for domain-specific jargon word identification in various application domains, we are required to use a hybrid approach with labeled corpus and Amharic Jargon Machine Readable Dictionary (AJMRD).

So that we followed a combination of machine learning and knowledge-based hybrid approach though, no prior DSAJWI works are done using any of the approaches. The combination of a knowledge-based approach using AJMRD and machine learning approaches using learned knowledge are employed. For jargon identification, various works were done for other languages and domains (Gong et al., 2017; Seyler et al., 2020; Shen et al., 2020; Weng et al., 2019).

The process of domain-specific Amharic jargon word identification can be performed after classification has been made by the machine learning model. A text classified as jargony is the input text for the identification phase of the knowledge-based component. The design requirement at this stage uses AJMRD to ensure the existence of jargon words in the input text. So that binary lexical mapping between jargon words in the input text and words from the AJMRD is performed with the approach we are followed. For every occurrence of a jargon word in a text and AJMRD, the meaning of the word is extracted.

3.2. Dataset preparation

Domain-specific Amharic jargon word identification done with a hybrid approach using the labeled trained corpus and the knowledge source. So that we used machine learning techniques to develop a model with labeled trained corpus and also, we used a knowledge base with the constructed knowledge source.

We considered a domain that provides services for huge customers and society from institutions in the Federal Democratic Republic of Ethiopia (FDRE). We collected sentences with domain-specific Amharic jargon words from the agricultural business reports, working guidelines, training manuals, advertisements of products and services to prepare corpus for training and testing with machine learning (ML) and to develop the knowledge source for meaning extraction.

The system is implemented to test and classify the input text with the model and extract the meaning of domain-specific jargon words from the knowledge source. So that the proposed DSAJWI system requires the availability of a labeled training corpus. We used the texts with the existence of jargon words to prepare a labeled training corpus for machine learning, and a list of domain-specific Amharic agricultural jargon words and the meaning to prepare the knowledge source. Although, regardless of availability of a dataset in the target language, we have prepared a dataset to train and test our model.

For the experiment, balanced dataset is organized into two classes for the two-way classification with 80/20 train-test split ratio. The following table shows the dataset prepared for machine learning and the knowledge base.

Table 3.1: Dataset prepared for machine learning and knowledge base

Dataset	No of sentences (Machine learning)			Knowledge base	
	Training	Testing	Total	Testing	AJMRD
Jargony	416	104	520	-	-
Non-jargony	416	104	520	-	-
Total	832	208	1040	-	-
Sentences	-	-	-	80	-
Jargon word	-	-	-	59	358

Sample dataset

		dataset	label
0	በማሳ ላይ ዘር ከመዘራቱ በፊት እንደ ቀልዝ ያሉ ብላፅዋት የሚጨምሩ ነገር...		1
1	ስለዚህ ላቀጣፈር በማሳችን ላይ እንዳይፈጠር በተከታታይ ማረስ እና ከትትል...		1
2	ለቀበሌያችን የተመደበውን ጋፈር ለማክፋፈል ፍላጎት ያላቸው በሙሉ ተመዝገቡ...		1
3	አንድርከኔ መስራት መለስተኛፈር ከማሳ እንዳይንቀሳቀስ ለማድረግ እና የተፈ...		1
4	እርሻ ስናርስ ጠልቋረሳ ድረስ መታረስ ያለበት ሲሆን ይህ ጠልቋረሳ ድረስ ...		1

		dataset	label
1035	የአሁኑ ዘመን ሙሉ በሙሉ የተፈጥሮ ሀብትን አሟጥሞ ለመጠቀም ርኅራኤ በሌላ...		0
1036	በእርግጥም ግብርናው ለዘመናት የአየር ነውጥ እንደ ድርቅ ያሉ አስቸጋሪ ኑ...		0
1037	ዘይቱ ከወጣ በኋላ የተረፈው ዘር ቁጣ በአፍሪካ በተለምዶ ለመኖ ይጠቅማል።		0
1038	ኢትዮጵያ በአፍሪካ ዋናኛ የሽምብራ አምራች ስትሆን፤ ከአንድ ሚሊዮን የሚ...		0
1039	በመሆኑም ኢትዮጵያ ከፕሮጀክቱ በከፍተኛ ደረጃ ተጠቃሚ የመሆን ዕድል አላት።		0

Figure 3.1: Sample dataset for domain-specific Amharic jargon words identification

The experiments are done with SVM, ANN, and NB machine learning classifiers with the TFIDF feature selection technique. This is the first work in the area and collecting

sufficient data for both knowledge source construction for KB, and labeled sentences for training of machine learning is tedious. However, deep learning approach consumes huge amount of data for training and testing. So that we recommend researchers to collect more dataset and improve the performance of the proposed system with the deep learning approach. The performance of the ML classifiers is compared with F-measure and accuracy for the two-way classification. The labeled corpus collected with a maximum of three different jargon words, two different jargon words, two similar jargon words, and with a minimum of one jargon words in a sentence. The maximum length of words in a sentence of the labeled corpus is thirty and also, the minimum length of words in a sentence is five for jargony class. For the non-jargon class the maximum length of words in a sentence is twenty-eight and the minimum length of words in a sentence is five.

3.2.1. Labeled corpus for machine learning

We collected sentences manually from agricultural reports, training manuals, working guidelines, advertisements of product and service delivery processes that contain Amharic agricultural jargon words. Labeled trained corpus was prepared from sentences with and without Amharic agricultural jargon words. We collected 1.04k dataset that comprises jargon and non-jargon words. In this study 80/20 split ratio is used for training and testing sets/phases. The labeled corpus is preprocessed and trained for the model development.

3.2.2. Knowledge sources for knowledge base

The knowledge-based system, the major area of Artificial Intelligence (AI) for our proposed system, captures and uses knowledge from the meaning collection AJMRD with binary lexical mapping.

Amharic agricultural jargon words are collected from different agricultural resources with the help of agricultural domain expert curators. In the Amharic language, there is no available Amharic Jargon Machine Readable Dictionary (AJMRD) for any of the reasons. So that we manually developed AJMRD which is our knowledge source. AJMRD is a predefined explanatory lexical resource having the meaning of domain-specific Amharic agricultural jargon words. Other researchers have developed a dictionary for domain-specific jargon identification system using different languages

(Antonic, 2008; Elkateb & Black, 2005; Gong et al., 2017). So that we developed our AJMRD for our purpose of providing the meaning of Amharic agricultural jargon words.

The steps we followed to develop our AJMRD include the selection of domain-specific Amharic agricultural jargon words with the help of agricultural experts, acquire the meaning of jargon words from domain experts, and store in the AJMRD for further analysis. The prepared AJMRD is reviewed by erudite experts of agriculture. The test sentences are prepared by agricultural experts and also taken from the report generated by experts.

We collected 358 domain-specific Amharic agricultural jargon words from different agricultural sources with the help of agricultural domain expert curators to prepare the Amharic Jargon Machine Readable Dictionary (AJMRD). We randomly prepared a total of 80 test sentences of different lengths for different experiments to test the knowledge base performance.

Because no prior works have been done in line with domain-specific Amharic jargon word identification, no resources for the meaning of Amharic agricultural jargon words were developed so far. Therefore, we prepared a knowledge source for a list of Amharic agricultural jargon words and the meaning of the words.

3.3. System architecture

The system architecture is the conceptual model that describes the structure, behavior, and view of the entire system. The architecture explains various aspects and the data flow of the developed system and it is an Architecture Description Language (ADL) that describes the conceptual parts and their association in the developed system.

The activities in the system architecture defined with principles, concepts, and properties logically related to and also consistent with each other. The architecture consists of features, properties, and characteristics that satisfy the problem or opportunity expressed by a set of system requirements, and life cycle concepts that are implementable with the help of technologies. The following flow chart in domain-specific Amharic jargon word identification starts with a collection of Amharic text documents as input to the developed system. The document is split into training and

testing data to acquire learned knowledge and evaluate the performance of the developed model.

The training of the dataset was performed with the developed machine learning models by employing machine learning classification techniques. Decision on the test data can be committed with the developed model; hence further processes are required for the test data besides the result of the decision.

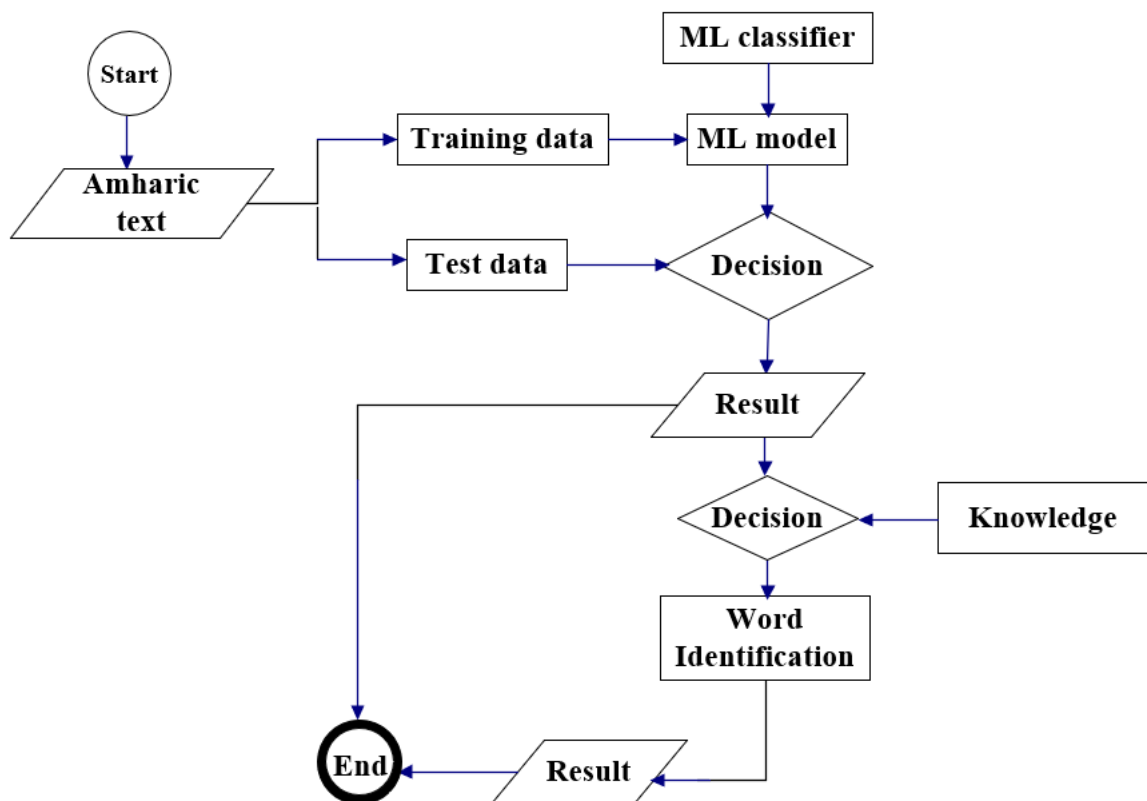


Figure 3.2: Flow chart for domain-specific Amharic jargon words identification

3.4. Proposed system

The main components in the proposed domain-specific Amharic jargon words identification (DSAJWI) system include preprocessing, model development, and knowledge base. The preprocessing performs text operation with tokenization, normalization, stop word removal, and stemming to return preprocessed text suitable for machine learning. The model trains the preprocessed dataset based on the label of the dataset by domain expert curators. The developed machine learning model is trained with the training dataset for further computation capability of the machine.

Classification performed with the testing data and the developed model. The knowledge base consists of a knowledge source, jargon word identification phase, meaning extraction phase, knowledge source update phase to return the final meaningful text. The following figure (3.2) describes the main phases and necessary steps in the domain-specific Amharic jargon identification system.

The proposed system takes a labeled corpus or input sentence as the input to the system followed by preprocessing techniques. Developing a model with a machine learning classifier algorithm and training of labeled dataset to acquire learned knowledge for future testing is performed. At the classification, a text that contains jargon words is the input for the knowledge-based system for extraction of meaning from the lexical resource; however, a text without a jargon word is returned as non-jargon text.

The text classified as jargon can be checked to the developed interactive Amharic Jargon Machine Readable Dictionary (AJMRD). For every occurrence of domain-specific Amharic jargon word from the input text, binary lexical mapping between DSAJW and its meaning from AJMRD performed.

Amharic Jargon Machine Readable Dictionary consists of domain-specific Amharic jargon words and the words meaning in a domain. In the proposed architecture, the knowledge base is responsible to check the existence of words in the classified jargon text and extraction of meaning for every occurrence of the word. The following architecture depicts the proposed hybrid system for domain-specific Amharic jargon word identification systems.

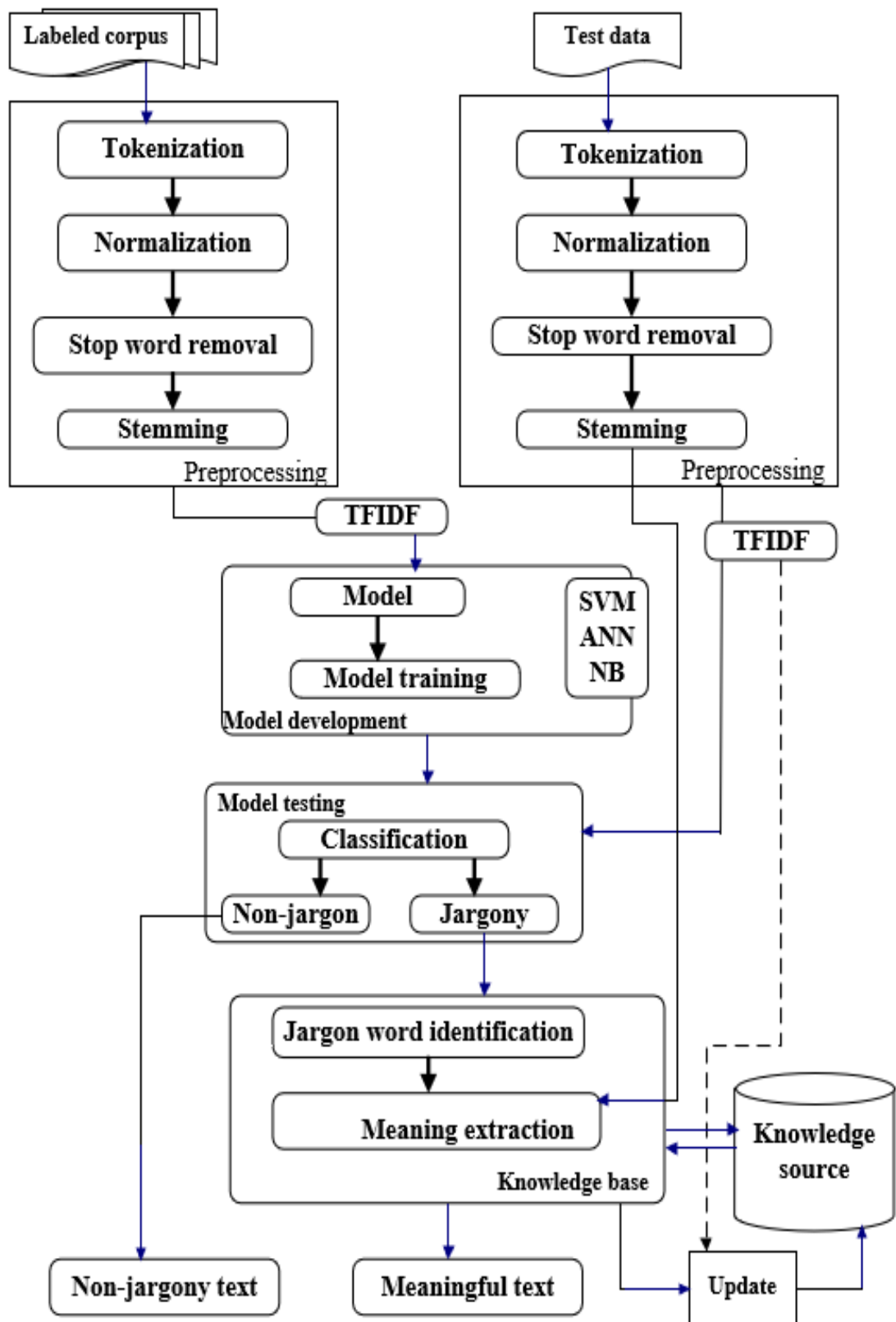


Figure 3.3: Proposed system for domain-specific Amharic jargon words identification

3.5. Preprocessing

Preprocessing makes the input sentence suitable for further analysis with different preprocessing techniques that include tokenization, normalization, stop word removal, and stemming. Therefore, input text can be preprocessed with the preprocessing techniques. The following are the description of the main tasks in the preprocessing for the DSAJWI (Hermawan, 2011). The following section describes how the preprocessing techniques performed for the proposed system to generate content bearing stem words to acquire learned knowledge for the ML model and evaluate the knowledge-based performance.

3.5.1. Tokenization

Tokenization is the first step in the preprocessing technique that can be performed right next to the input labeled corpus to segment the input Amharic text into a list of Amharic tokens. The segmented text (list of tokens) used as an input for the next phase. The process of tokenization splits strings of text into Amharic tokens; hence paragraphs can be tokenized into sentences, and sentences tokenized into the list of words. Tokens are units of text that are sourced from the input text (Hermawan, 2011). White spaces and punctuation marks in the Amharic language that include ‘netela serez (፣) semi colon’, ‘hulet netib (,) comma’, ‘arat netib (#) fullstop’, ‘dirib serez (፡) colon’, ‘tiyakie milikik (?) question mark’, ‘kale agano (!) exclamation mark’ is used for tokenization of Amharic text. The target DSAJW can be generated from the list of tokens.

3.5.2. Normalization

Normalization is the process of making Amharic words having similar pronunciations with different Amharic orthographical structures have similar representation in the preprocessing. One of the issues in the morphologically complex Amharic language is the availability of many letters that have similar pronunciations; however, with different representations. For example, the word ‘ነጻረጻይ (netseretsay):’ can be written as ‘ነፀረጻይ (netseretsay)’, ‘ነጻረጻይ (netseretsay)’, ‘ነፀረጻይ (netseretsay)’. So that with the processes of normalization variant forms with similar meaning Amharic letters called fidels can be handled and they have the same representation. Normalization has been done after the generation of tokens by the tokenizer and before the removal of the stop word (Hermawan, 2011). However, Amharic normalization affects the preprocessing inline

of making similar pronunciation words have similar orthographic representation; hence variant forms of Amharic characters have different meanings in a language. So that further work is required to represent variant forms of Amharic characters as per their meaning.

3.5.3. Stop word removal

Stop words are the most frequent and non-content-bearing words in a text that result in noise in text preprocessing and for the development of applications with the text. The stop words in Amharic text have no discriminatory power for the conveyed information. Stop words include articles, pronouns, prepositions, and conjunction. Amharic language use various forms of morphologically generated stop words that include ‘ማለት (mallet)’, ‘እዚህ (ezih)’, ‘ከላይ (kelay)’, ‘ባለ (bale)’, ‘ያህል (yahil)’, ‘ቢሆንም (bihonm)’, ‘ሌላ (lela)’, ‘ሁሉ (hulu)’, ‘ይህን (yihin)’, ‘እና (ena)’, ‘እስከ (eske)’, ‘ነው (new)’, ‘እንደ (ende)’. Stop words are necessary for sentence construction; however, the words have less importance for the development of NLP application. Stop words are identified manually in the Amharic language to reduce memory usage and recall process. Therefore, Amharic stop words are removed to work with content-bearing words that include DSAJW (El-Khair, 2017).

The various applications of NLP do not require the existence of non-content bearing words (stop words), in text processing to improve the effectiveness and efficiency of the developed application. Because of the morphologically complex nature of Amharic language, removal of frequent words (stop words) has a positive impact on the performance of the DSAJWI system.

3.5.4. Stemming

Stemming is the method of extracting affixes from words to get the stem form. Amharic is a morphologically complex language that requires an effective and efficient morphological analyzer to develop different NLP applications. Stemming is the process of generating morphemes which is the smallest unit of a language that is impossible to divide without losing its actual meaning. Stemming helps to reduce the memory usage of the system developed. On the contrary, many meaningful Amharic words can be generated from a single Amharic morpheme in a language, because of the morphological complex behavior of Amharic language.

Removing affixes of words and generating morphemes from inflectional Amharic words inline to bound morphemes is a challenging task in Amharic morphological analysis. Though it doesn't work well, resources were developed for Amharic, Oromo, and Tigrigna morphological analyzers by Michael Gasser (Gasser, 2011). We used rule-based Amharic stemmer to generate Amharic morphemes in DSAJWI architecture. We collected a list of prefixes and suffixes such as 'ኅ (na), የ (ye), ስለ (sile), በ (be), ም (m), ን (n)' from Amharic language experts, and we removed these affixes to get stem of the Amharic content-bearing word. The list of prefixes and suffixes obtained from experts are removed to work with the stem of content-bearing Amharic words. Stemming for machine learning decreases memory usage and increases performance; however, it results in a decrease in performance for the knowledge-based system because binary lexical mapping is impossible between the over-stemmed words in the input text and the words in the knowledge source.

TFIDF feature selection (TFIDF)

We used a powerful feature engineering technique Term Frequency Inverse Document Frequency (TFIDF) to identify the important and precisely rare words in the text data. The TFIDF feature selection technique with the combination of term frequency (TF), and inverse document frequency (IDF) used for the applications such as classification, information retrieval (IR). For our work, we used the techniques to convert the strings of a text into numbers so that the developed SVM, ANN, and NB machine learning models consume the input data in numerical formats. The TFIDF feature selection technique used for scoring words in machine learning models for the Amharic language processing. TFIDF is the combination of Term Frequency (TF), and Inverse Document Frequency (IDF) (Jing et al., 2002).

Therefore, we selected TFIDF feature selection technique inline to provide convenient data for training and testing of dataset with ML models. So that the ML model is developed and tested with the vectorized data using TFIDF vectorizer. The TFIDF selects features with the conversion of the input string of data into the numerical format.

Machine learning algorithms use the numerical format of strings with the help of text vectorization for analyzing data; hence a document represented with a list of word vectors. The TFIDF score fed to the developed SVM, ANN, and NB models for the jargon word identification.

Term Frequency (TF)

TF is the total count of the unique words available within a document. So that the TF of the term t_i in a document d_j described as the quotient of the number of times a word t_i appears in a document d_j to the total number of words available in a document d_j .

$$TF(t_i|d_j) = \frac{\text{Frequency of } t_i \text{ in a document } d_j}{\text{Total words in a document } d_j} \quad (3.1)$$

Inverse Document Frequency (IDF)

The weight of content-bearing words can be identified by the IDF (Inverse Document Frequency). IDF generates smaller value for frequent words and generates higher value for content-bearing words. So that the generation of content bearing words for the input document can be described by the inverse document frequency (IDF) of words; hence IDF of a term t_i , document frequency df , in the whole document D , described as the logarithmic quotient of a total number of documents (D) to documents containing term t_i (df).

$$IDF(t_i, df, D) = \log \frac{\text{Total number of documents } (D)}{\text{Documents with term } t_i (df)} \quad (3.2)$$

The IDF value of the frequent words in a document collection approach to zero, and the IDF value of content-bearing words in a document collection approach to one. So that highly frequent words in a document with rare occurrence in the document collection return with high value of IDF.

The TFIDF feature selection is a natural language processing technique used to separate distinct words in a document in line with assigned scores by IDF approaches to one. The newly occurring content-bearing words in a document have high IDF value. So that TFIDF in our proposed system is used to update the knowledge source when new jargon words in the text are entered into the system. Though it requires further enhancement, distinct jargon words are ranked at the top and stored in the knowledge source to acquire meaning from the domain expert.

3.6. Machine learning/Model development

The machine learning of our hybrid architecture accepts preprocessed text from the preprocessing to come up with learned knowledge. Preprocessed labeled corpus entered the training phase for model development with machine learning techniques. The final

phase of the machine learning performs classification with the developed model. The following section discusses machine learning phases in our hybrid system architecture.

3.6.1. Model development

Model development can be performed in the first phase of the machine learning. The developed model is used to train the preprocessed data for future prediction on unseen data. The model is applicable for testing unseen data based on the learned knowledge in the training phase. The model is flexible for the learned text and can predict the result of unseen data from the testing data and the user input. The developed model is ready for use of any input to perform classification.

3.6.2. Model training

The training phase of the machine learning can be performed with the developed model and preprocessed labeled corpus. The researchers split the preprocessed labeled corpus to train the model and test the performance of the model using performance metrics. Conversion of the input text into vector form is required to make the corpus understandable by the developed model. So that we used Term Frequency Inverse Document Frequency (TFIDF) to convert the input labeled corpus into its vector form. The training phase learns knowledge from the input labeled corpus for experience and future prediction of unseen data.

3.6.3. Model testing

The model testing of the proposed system was used to test the data. Users are allowed to enter the text and the model predicts based on the learned knowledge. The developed model predicts the label of the input text from learned experience. So that based on the prediction of the model testing, the classification of the input text can be performed as jargony and non-jargony. The dataset was prepared and trained in two-way classification form as jargony and non-jargony. The output of the model testing can be returned as non-jargon text without entering the knowledge-based system. However, the output of the model test might be the input for the knowledge base system for the occurrence of DSAJW in the input text. Texts containing DSAJW require further analysis on the knowledge-based system to return input text with meaningful text of words.

3.6.4. Machine learning classifier algorithm

We used a machine learning classifier algorithm to develop ML model for our two-way text classification of jargon words identification. We developed three models with three algorithms that include Support Vector Machine (SVM), Artificial Neural Network (ANN), and Naïve Bayes (NB). The three machine learning algorithms used for our model development are discussed as follows.

Support Vector Machine algorithm - SVM

SVM is a popular ML algorithm for classification and regression. The best SVM classifier maps the training data by maximizing the margin of the classifier. The more general SVM model can be developed with a large marginal width. The availability of many training patterns maximizes the value of the parameter of the hyperplane to the nearest training patterns from the given class for the SVM classifier. So that we developed a model with SVM classifier for our two-way classification (Abikoye & Omokanye, 2017).

We used the SVM algorithm for our model development; hence SVM bases its theory on the structural risk minimization principle from computational learning to find a hypothesis that guarantees the lowest true error. The SVM model needs both negative and positive training set to seek for optimum separation hyperplane (OSH) that best separates the positive from the negative data in the n-dimensional space. The support vectors are located closest to the hyperplane with the smallest width to the hyperplane as depicted in figure 2.1 (Baharudin et al., 2010).

The two-way classification of our proposed system with the SVM classifier performed using given training data (x_i, y_i) for $i = 1 \dots N$, with $x_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$, learn a classifier $f(x)$ such that:

$$f(x_i) \begin{cases} > 0 & y_i = +1 \text{ (class)} \\ < 0 & y_i = -1 \text{ (class)} \end{cases}$$

For 2D (2-dimensional), the discriminant is a line.

We followed a linear classifier form with $f(x) = w^T x + b$, where w is a weight vector which is normal to the line and b is the bias. Optimal Separating Hyperplane (OSH) is required to separate the data points among multiple hyperplanes. The OSH can be

chosen besides the position of the support vectors, support vectors are a subset of training data points that define the margin.

SVM is helpful to remove irrelevant features and attempt on data points that have high-dimension input feature space. However, SVM has a relatively complex training and categorizing algorithms, its memory consumption, and time usage while training and classification (Baharudin et al., 2010).

The performance of the SVM model decreases when the availability of large training datasets, and more noise on the dataset. As compared to ANN, the SVM model requires more feature engineering, and easier to understand small datasets.

Artificial Neural Network - ANN

The two-way classification with Artificial Neural Network (ANN) with feed-forward neural network strategy is composed of more than one perceptron with an input layer, an arbitrary number of hidden layers, and an output layer to make decisions or predictions for the input test data. A Multilayer perceptron (MLP) is a class of feed-forward artificial neural network (ANN); hence multiple layers of a perceptron are required to develop ANN ML model for the two-way classification of Amharic jargon word identification (S. Abirami, 2020).

We trained the dataset with Multi-layer Perceptron (MLPClassifier) for the two-way classification as jargony and non-jargony. The Multilayer perceptron (MLP) is a supervised learning algorithm that learns from the features $F = f_1, f_2, \dots, f_n$, where n is the number of features. MLP relies on an underlying Neural Network to perform the task of classification. The model developed with MLPClassifier trains iteratively, and the regularization term added to prevent overfitting.

The implementation of MLP is not intended for large-scale applications; hence scikit-learn offers no GPU support; however, building a model with deep learning architecture is helpful for the computation of large-scale applications.

Multinomial Naïve Bayes algorithm - MNB

We developed a model with NB for the two-way classification. NB is a supervised learning and statistical method for classification with a probabilistic classifier based on Bayesian theorem with strong and naïve independence assumptions; hence each term is independent of the other. Scikit-learn implements three Naïve Bayes variants of classifiers based on the same number of different probabilistic distributions with

Multinomial NB, Bernoulli NB, and Gaussian NB. We selected Multinomial NB; hence the classifier uses discrete distribution whenever a feature is represented by the whole number such as frequency of a term in a natural language processing. So that the classifier is useful to model feature vectors using the frequency value in the collection.

The prediction of input test data with the multinomial NB model is designed based on the number of times a term occurs in a document, term frequency (TF), hence a term may be pivotal to decide the label of the input text, and a term is helpful to decide whether a term is useful for the analysis (Mccallum & Nigam, 1997).

The posterior probabilities of the input text t being in category c is given as:

$$P(c|t) = \frac{P(c)P(t|c)}{P(t)} \quad (3.3)$$

$$= \prod_{1 \leq i \leq n} P(w_i|c)P(c)$$

Where $P(w_i|c)$ the conditional probability of the term w_i occurring in a text t of category c . $P(w_i|c)$ describes as a measure of how much w_i contributes that c is the correct category. The list of tokens $w_1, w_2, w_3, \dots, w_n$ in the text t are part of the vocabulary that help for the classification of a text in the expected category, and n is the number of such tokens in the text t . The parameter $P(c)$ is estimated as:

$$p(c) = \frac{\text{Number of texts in } c}{\text{Number of texts}} \quad (3.4)$$

The classification results are not affected because parameter $P(t)$ is independent of category. So that the best class for the text with the NB classifier is the most likely or maximum posterior probability.

Our model is developed with a Multinomial NB algorithm to train and test the dataset. The Multinomial NB model specifies that a document can be represented by the frequencies of a term in the document because a document is represented with the bag of words. In the bag of words approach, individual words in a document constitute its features, and the order of words is ignored. So that the features in a dataset are mutually independent.

3.7. Knowledge base

The knowledge base of the proposed system accepts input from the classification phase of the machine learning when the model predicts a text as jargony. Knowledge-based

include knowledge source, jargon words identification, meaning extraction, and update knowledge source. The meaning of a jargon words is extracted from the knowledge source by the meaning extraction phase when a word is identified as a jargon word with binary-lexical mapping in the jargon word identification phase. The following section describes the function of phases in the knowledge base.

3.7.1. Knowledge Source

Machine Readable Dictionary is a knowledge-based lexical resource used to store words and the words meaning to employ for computational linguistics. Amharic jargon words collected from various agricultural sources and the words meaning obtained from agricultural domain experts stored in the Amharic Jargon Machine Readable Dictionary (AJMRD). The AJMRD helps users to extract the meaning of an exact jargon word with binary lexical mapping, and an over-stemmed jargon word with the help of a close match to the stored words. Because there is no prior AJMRD developed for any of the reasons, we developed interactive AJMRD for our work. The meaning of Amharic agricultural jargon words is sourced from agricultural domain experts. So that the knowledge source is constructed with the list of Amharic agricultural jargon words collected from various agricultural sources and the words meaning obtained with the help of the agricultural domain experts. Agricultural erudite reviewed the constructed knowledge source on the behalf of the meaning of words collected from domain experts.

Update knowledge source: the meaning of collected jargon words in the text are stored in the knowledge source. However, jargon word is invented for different reasons besides the organization's business. The newly invented jargon words by agricultural domain experts and also, the jargon words that are not included in the knowledge source require meaning for users of text. So that the knowledge source becomes updated as new words occur in the input text with the TFIDF value of newly occurred words. Distinct jargon words are ranked at the top value of TFIDF in the collection and stored in the knowledge source to acquire meaning. The domain experts are required to provide meaning for the new collected jargon words in the knowledge source.

3.7.2. Jargon word identification

The Amharic jargon identification phase is the first phase in the knowledge base in the DSAJWI. The input of the jargon identification phase is a list of tokens passed from

the classification phase of the machine learning. So that the existence of a list of tokens in the input text is checked from the constructed knowledge source to extract the meaning of words. Amharic jargon identification phase used to identify a particular jargon word from the input text; hence AJMRD is the main lexical knowledge source for our identification. For example, in a text ‘በእርሻ ማሳ ላይ ቀልዝ መጨመር ምርት እና ምርታማነትን ያሳድጋል’, the DSAJW ‘ቀልዝ (qeliz)’ identified as a jargon word by binary lexical mapping between a word ‘ቀልዝ (qeliz)’ from the input text and a word ‘ቀልዝ (qeliz)’ from Amharic Jargon Machine Readable Dictionary.

3.7.3. Meaning extraction

The meaning extraction phase of DSAJWI extracts the meaning of the identified jargon word in Amharic jargon identification phase of the knowledge base system. Meaning extraction is performed from the knowledge source when a word is identified as a jargon word. So that for the occurrence of DSAJW in a text of domain, the meaning of the word is extracted. The meaning of a jargon word in a domain is unique and was given by prior experts of a domain by convention when a word was invented and available for service. For example, in a sentence ‘በእርሻ ማሳ ላይ ቀልዝ መጨመር ምርት እና ምርታማነትን ያሳድጋል’, a word ‘ቀልዝ (qeliz)’ is identified as a jargon word; hence a word is found in the knowledge source. Therefore, the meaning ‘መሬትን የሚያዳብር ማንኛውም የህያዋን ነገሮች ብስባሽ በተለይም የእንስሳት ፍግ’, extracted from the knowledge source for the word ‘ቀልዝ (qeliz)’. Binary lexical mapping between a jargon word ‘ቀልዝ (qeliz)’ and the meaning ‘መሬትን የሚያዳብር ማንኛውም የህያዋን ነገሮች ብስባሽ በተለይም የእንስሳት ፍግ’ from the knowledge source performed. So that Amharic jargon word meaning extraction is performed after binary lexical mapping between words from the input text and words in the knowledge source. Jargon words have a unique meaning in the knowledge source; hence words have a unique meaning in a domain. Therefore, Amharic text containing DSAJW with prominent meaningful text returned to the user.

Over-stemming: though stemming is a challenge for the meaning extraction from the knowledge source, we handle the problem with entering the over-stemmed word to the proposed system. For example, the over stemmed word ‘ሥተኛፈር (sitegnafer)’, ‘ድረበቅለ (direbekile)’, and ‘ትካየር (tikayer)’ for the exact word ‘መለሥተኛፈር (melesitegnafer)’, ‘ድረበቅለት (direbekilet)’, and ‘መትካየር (metikayer)’ respectively are returned from the stemming phase which in turn no extraction of meaning for the words. So that the

meaning extraction phase of the proposed system extracts the meaning of over-stemmed word with the help of a close match request with the developed system. The user's confirmation is required for the meaning extraction of the most closed word to the over-stemmed word.

3.8. Summary

In this chapter, the data preparation of labeled corpus for the machine learning and also, knowledge source for the Knowledge based that include sample dataset are discussed. We employed machine learning and knowledge based for the design requirements of the proposed system. The proposed DSAJWI system with the function of phases are discussed. Preprocessing, model development, and knowledge-based with TFIDF feature selection technique are the main steps we followed for the proposed system. Support Vector Machine, Artificial Neural Network, Naïve Bayes are the three machine learning classifier techniques we applied to develop a model for the two-way classification. The text classified as jargony with developed machine learning model is the input for the knowledge-based system. The meaning of a jargon word in a jargony text is extracted from the knowledge source to provide meaningful text to the target user of the text.

CHAPTER FOUR: EVALUATION AND DISCUSSION OF RESULTS

In this chapter, experiments are conducted to evaluate the performance of our proposed system with machine learning and knowledge base for domain-specific Amharic jargon words identification, and the results of the experiment are discussed.

Domain-specific Amharic jargon word identification was done with a hybrid approach using a labeled corpus and the knowledge source. So that we used machine learning techniques to develop a model with labeled trained corpus and knowledge sources. The system is developed to classify the input text with the model using Support Vector Machine (SVM), Artificial Neural Network (ANN), and Naïve Bayes (NB) classifier and extract the meaning of Amharic agricultural jargon words from the knowledge source. The meaning of words was collected from experts of the agricultural domain.

We developed an interactive Amharic Jargon Machine Readable Dictionary (AJMRD) from Amharic agricultural words. The manipulation of the knowledge source is performed with the python programming language.

The proposed hybrid DSAJWI system takes an input text, and the necessary preprocessing techniques such as tokenization, normalization, stop words removal and stemming performed by the preprocessing of the system. A text suitable for further analysis is passed to the next machine learning for classification with the developed model. Therefore, the classified jargony text entered into the knowledge-based system. To this end, a text that contains DSAJW with meaningful text can be returned to the user.

4.1. Experimental setup

In our experiment setup, we used python version 3.8 programming language for the implementation because Python is the former programming language in the current computing environment and it supports many open-source libraries. We use anaconda distribution and Jupyter notebook editor for the development of the system. We create our environment in anaconda to install and import the necessary open-source Python libraries for the implementation of the proposed system. We imported various python libraries and machine learning algorithm libraries that are compatible with our experiment. We trained our labeled dataset and also, we loaded our lexical resource

Amharic Jargon Machine Readable Dictionary to the Jupyter notebook for the manipulation of Amharic domain-specific agricultural jargon words.

The following hardware specifications are used to develop and test the domain-specific Amharic jargon words identification system.

- Lenovo with Window 10 Pro 64-bit operating system
- 8 GB RAM
- Hard Disk size 1TB
- Processor Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz 1.99 GHz

4.2.Evaluation of the proposed system

Performance evaluation is required for our developed system to know the effectiveness and efficiency of the system. The evaluation of the developed DSAJWI system can be performed to ensure the classification of input text with the machine learning model and meaning extraction of domain-specific Amharic jargon words from the knowledge base. The performance of the proposed system is evaluated with machine learning to classify the input text as jargony or non-jargony using the learned knowledge from the labeled trained corpus. The system is also evaluated with the knowledge-based system. The evaluation of the knowledge-based is based on the capability of the system to extract the meaning of identified jargon words from the predefined explanatory knowledge source.

4.2.1.Evaluation metrics

We used F-measure and accuracy to evaluate the machine learning and knowledge base of our proposed system though none of the previous works in domain-specific Amharic jargon word identification recommend us to use evaluation parameters (Dalianis & Dalianis, 2018). We used a confusion matrix with F-measure and accuracy that calculated the correctness and completeness of the test set to evaluate the performance of the developed system and finally we conclude the evaluation with F-measure and accuracy.

We used F-measure performance metrics hence most of the paper use this metric to evaluate the text classification (Gong et al., 2017; Ibrahim et al., 2020; Koay et al., 2020; Weng et al., 2019; Weng & Szolovits, 2018). So that F-measure and accuracy is

the performance metric used to evaluate the performance of the machine learning in the proposed system.

F-measure: F-score is defined as the harmonic mean between precision and recall. So that it is the weighted average of both precision and recall.

$$\text{F-measure} = \frac{2PR}{P+R} \quad (4.1)$$

Accuracy: accuracy is the measure of the closeness of the value measured by the developed system to the standard or a known value. So that accuracy is the measure of how close a measured value is to the actual value.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4.2)$$

4.2.2. Machine learning evaluation

The evaluation of our proposed system on the machine learning is committed with the comparison of models developed from the machine learning algorithms. We developed machine learning models to select the most likely model for the classification of the input text. Though numerous machine learning algorithms are introduced and implemented by various researchers, the popular machine learning algorithms are Support Vector Machine (SVM), Artificial Neural Network (ANN), and Naïve Baye (NB). So that we selected SVM, ANN, and NB machine learning algorithms to compare the classification result and select the outperformed model.

The performance result of three supervised ML models is compared for the same labeled input corpus for the agricultural domain. The same algorithm for feature vector representation with TFIDF vectorizer was employed. The algorithms are compared with F-measure and accuracy. The following table shows the performance result of the developed machine learning models.

For our two-way classification of domain-specific Amharic jargon word identification, the following tables (table 4.1, table 4.2, table 4.3) summarizes the performance of the developed machine learning models.

Besides, we observed the following performance for the developed ML models with the labeled trained corpus.

Performance result of the developed system with SVM model

Table 4.1: Performance of jargon identification with SVM classifier

Class (Label)	F-measure	Accuracy
Non-jargon	96	96.2%
Jargon	96	
Average	96	

Performance result of the developed system with ANN model

Table 4.2: Performance of jargon identification with ANN classifier

Class (Label)	F-measure	Accuracy
Non-jargon	95	95.2%
Jargon	95	
Average	95	

Performance result of the developed system with NB model

Table 4.3: Performance of jargon identification with NB classifier

Class (Label)	F-measure	Accuracy
Non-jargon	95	94.7%
Jargon	95	
Average	95	

The following table describes the hyperparameters used in the machine learning model development.

Table 4.4: Hyperparameters of machine learning models

ML models	Hyperparameters	Value of the hyperparameters
SVM	kernel	linear
	C	3.0
	gamma	0.1
ANN	solver	lbfgs
	hidden_layer_sizes	6
	random_state	3
	learning_rate	constant
	momentum	0.9
NB	alpha	0.1
	class_prior	None
	fit_prior	True

The best performance of the models is observed with the hyperparameter value combination of the above hyperparameters for all developed models. The classification result performance of the aforementioned machine learning models with F-measure and accuracy are described in the following figure. Though other computations of the input text can be performed after the model prediction, the developed model shows different classification performances. Therefore, we selected SVM, as it performs better than the ANN and NB for the knowledge base computation of the input text. The following figure depicts the comparison for the performance ML models

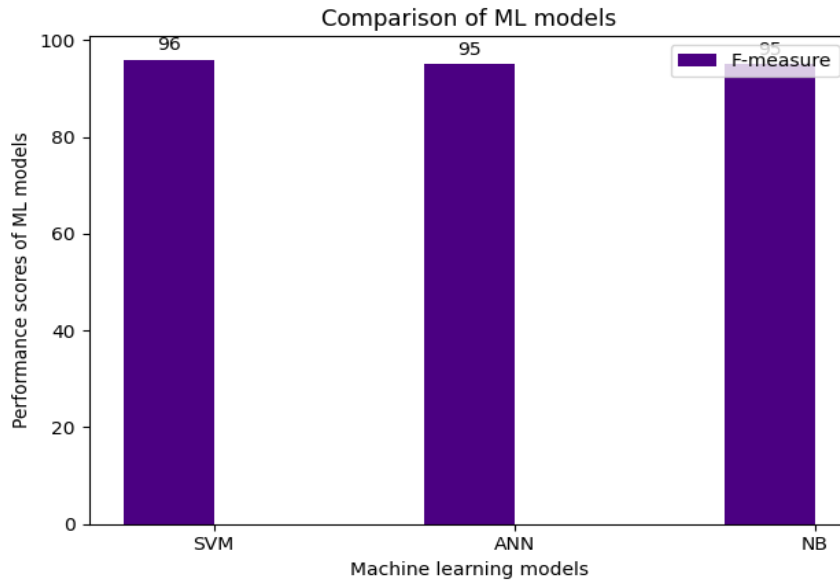


Figure 4.1: F-measure comparison of SVM, ANN, and NB

Figure 4.2 shows that the performance of SVM, ANN, and NB models for the labeled trained data. The accuracy of SVM, ANN, and NB are 96.2%, 95.2%, and 94.7% respectively.

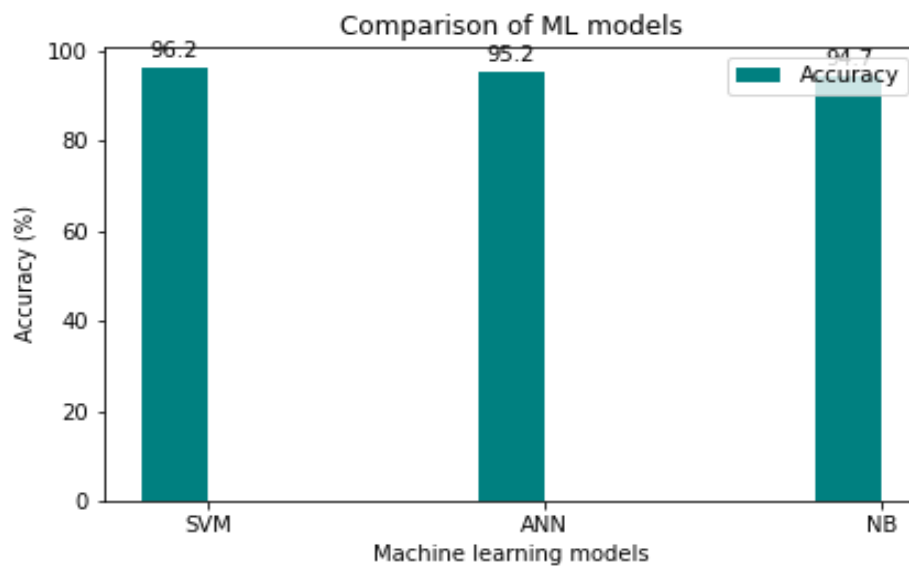


Figure 4.2: Accuracy comparison of SVM, ANN, and NB models

Comparison of performance result: we observed that SVM outperforms the other model with 80/20 train-test split ratio. Because of the performance result of the models, SVM is selected to predict the input test data for the knowledge base. After the prediction is made with the selected SVM model, we evaluate the performance of the

knowledge base with different lengths of sentences. So that we used the SVM model for the prediction of the text as jargony and non-jargony. The text predicted as jargony with the selected SVM model is the input to the next knowledge-based of our proposed system. The performance of the proposed system decreases with the 70/30 train-test split ratio. The number of jargon words and the length of sentences in the training and testing data in 70/30 ratio is alike the preparation of the labeled corpus for the result obtained with 80/20 train-test split ratio.

Outperformed Support Vector Machine model – SVM

SVM classifier is good for binary classification and works well for structured data (Holts et al., 2010). We observed the model developed with SVM gives better results as compared with the models of ANN, and NB for our dataset of binary classification. The risk of over-fitting of SVM is less as compared to other models. The kernel function of the SVM classifier performs well with the strength of the kernel trick. So that the developed SVM model outperformed the ANN model by 1.0%, and the NB model by 1.5%. The model developed by SVM outperformed the ANN model by 1% and the NB by 1.5%.

SVM is applicable in a wide range of supervised machine learning binary text classification because of its robustness to noise and errors, accurate predictions and evaluation for the target function with margin calculation (Byun & Lee, 2002). Support vectors in SVM are helpful for margin calculation hence, support vectors are identified by quadratic optimization algorithm for the better training and testing for the two-way classification. So that support vectors are effective elements in the training with structural risk minimization and quadratic optimization algorithm for the two-way classification (Ananiadou et al., 2009).

The use of increased number of max_features of the TFIDF feature engineering technique in our proposed system result maximum performance of the SVM model as compared to NB and ANN. The fixed number of parameters described in model building is also, the other reason for the better performance of SVM model with as compared to ANN hence, ANN is non-parametric that has infinite parameters to build a model.

A kernel trick function in SVM performs a lot of dot product calculation for the training and testing of binary classification. An interesting and powerful kernel used for pattern

analysis with the dot product of features and later classify the input test jargony and non-jargony text to the expected class with the conversion of the input test data to high dimensional feature space to make the data suitable for classification. So that classification input test data is complex in input space and easy for the feature space with the kernel function. The capability of analysis of the kernel trick increases to the better way for the increase of training labeled corpus.

However, choosing the value of the hyperparameters of SVM is a challenge to develop a good model in line with the nature of the dataset, and a long training time is required for the large dataset. Choosing a good kernel function is a challenge. Classification of the input test data can be performed with many hyperplanes; however, the more general SVM model can be developed with the hyperplane whose distance to the nearest classified data points (support vectors) from each side is maximized. Support vectors are data points closer to the hyperplane that influence the position and orientation of the hyperplane.

Therefore, a maximum margin hyperplane classifier is required with the customized value of weight vector normal to the line (w) and the bias (b). Hinge loss of the SVM classifier helps to maximize the margin. SVM uses hinge loss which is $\max(0, 1 - y_i f(x_i))$ whose approximation is 0-1 loss. The loss function considers the value of $y_i f(x_i)$ either greater than 1, equal to 1, and less than 1. The value greater than 1 describes the data points are outside of the margin, which has no contribution for loss; the value equal to 1 describes the data points are on the margin that has no contribution for loss; however, the value of the function less than 1 describes the data points violets margin constraint that contributes for loss.

The regularization hyperparameter (C) balances the margin maximization and loss; hence the cost function becomes less. So that we used less value of the regularization hyperparameter (C) that helps to develop the maximum-margin SVM model for the best performance. The tradeoff between maximum-margin and the number of mistakes on the training data is observed with the small value of the regularization hyperparameter (C). The small values of C allow constraints to be easily ignored with the maximum-margin classifier; however, the large value of C makes constraints hard to ignore. Therefore, for the given n training dataset $(x_1, y_1), \dots, (x_n, y_n)$, where $y_1, \dots, y_n = (+1, -1)$ that indicate the class of x_i .

The main objective of SVM is to find a hyperplane in N-dimensional space (N-number of features) that distinctly classifies the data points. The dimension of the hyperplane is dependent on the number of features. The hyperplane is a line for two feature inputs and the hyperplane is the two-dimensional plane for three feature inputs. The availability of a limited training dataset is the reason for the best performance of the SVM model for our text classification as compared to ANN with the MLP model.

The SVM model is effective when the more general model is developed with a large marginal distance for clear margin of separation between classes, the number of dimensions greater than the number of samples, and the model is memory efficient. The SVM models put the data points above and below the classifying hyperplane with the need for probabilistic classification of data points. The overfitting challenge of the SVM model solved with the powerful linear kernel and the regularization value of C hyperparameter. We used the linear kernel for our linear text classification problem since the linear kernel outperforms the other kernel with trial and the linear kernel is best for text classification.

High feature engineering techniques for the SVM model are required for its high performance. We observed the best performance of the SVM model as compared to ANN and NB with the increased number of TFIDF features and also, the less performance of the model with decreased number of features.

Low-performed Naïve Bayes model - NB

The SVM model considers the geometric interpretation of the input text classification for the model's best performance, as compared to taking the probabilistic approach with the NB model. The SVM model considers the interaction between terms to understand dense concepts which are impossible with NB; hence NB considers the independence of the terms. The better performance of SVM model as compared to multinomial NB is the treatment of features of data points. NB model treats the features independently; however, SVM looks at the interaction between features to a certain degree. So that for our training data we observed the best performance of SVM for classification. The probabilistic behavior of the NB model is less performed as compared to the geometric behavior of the SVM model.

The presence or absence of one feature does not affect the presence or the absence of the other feature and the classified features are not related to any other feature. The less

performance of the NB model is the assumption of the independent predictor of the test data; however, with the independence predictor of data points, the NB model works well for small training datasets. The NB model assigns the value of zero probability for the data points not observed in the trained dataset. The underperformance of the NB model is the availability of a large training dataset for binary classification; hence the NB model independence assumption returns high performance for multiple class prediction problems as compared to binary class prediction. Multinomial NB (MNB) is better for long documents and large training data and also, better for text classification as compared to other NB classifiers.

Artificial Neural Network model - ANN

The less performance of ANN with the MLP model is the increased number of hyperparameters such as hidden neurons, layers, and iterations. The less performance of ANN with MLP is the limited number of training datasets. The best performance of the SVM model as compared to ANN model is observed; hence SVM is based on the structural risk minimization; however, ANN with MLP classifier is implemented with empirical risk minimization. So that SVM is more efficient and the model obtain optimal separating hyperplane that returns good performance for unseen input test data (Zanaty, 2012).

The SVM model has higher prediction accuracy than multilayer perceptron because SVM has higher runtime as there are computations it performs such as translating n-dimensional space using the linear kernel function and the model finds the perfect hyperplane for classification (Osowski et al., 2004).

The models SVM and ANN have supervised machine learning classifiers and ANN is a parametric classifier that uses hyperparameters tuning during the training phase. However, SVM is a non-parametric classifier that finds the linear vector for linear kernels to separate classes. As we used small dataset for training, we observed the low performance of the parametric ANN model compared to SVM. The ANN is better for multiclass prediction with the probabilities of each class; however, SVM handles this issue with one-versus-all binary classification. So that because our labeled data is for binary classification, the SVM model outperforms the ANN model. The non-parametric SVM model understands the text classification better with the help of the linear kernel as compared to the parametric ANN for our two-way classification.

Confusion matrix: the performance of the developed model is summarized using a confusion matrix with TP, FP, FN, and TN which in turn is used for the evaluation of accuracy and F-measure.

The above machine learning models also returned the number of input test data as True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). The following table describes TP, FP, FN, and TN from the input test data based on the prediction result of the model.

Table 4.5: Comparison of models on TP, FP, FN, and TN.

Performance metrics	Support Vector Machine (SVM)	Artificial Neural Network (NN)	Naïve Bayes (NB)
True Positive	102	102	96
False Positive	4	4	10
False Negative	4	6	1
True Negative	98	96	101
Correctly classified	200	198	197
Incorrectly classified	8	10	11

The following figure depicts the confusion matrix of the outperformed SVM model.

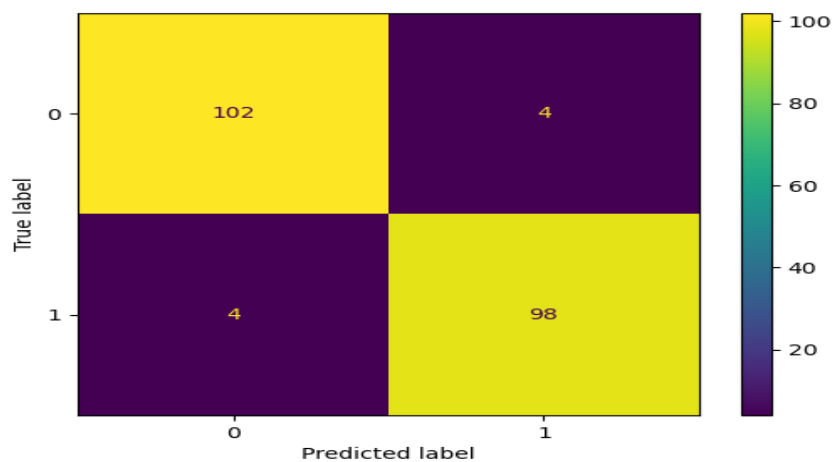


Figure 4.3: Confusion matrix of outperformed SVM model

Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve describes the True Positive Rate (TPR), and False Positive Rate (FPR) of the input test data. The ROC curve shows the trade-off between TPR and FPR of the developed models. The ROC of the model is calculated with the predicted scores. SVM is slightly higher accuracy than ANN and NB; and NB a low performer. We observed the curve of SVM closer to the top-left corner; however, NB is the less performed model because the curve of NB is closer to the 45-degree diagonal of the ROC space. The following figure depicts the comparison of models with ROC using sensitivity (TPR), and specificity (FPR).

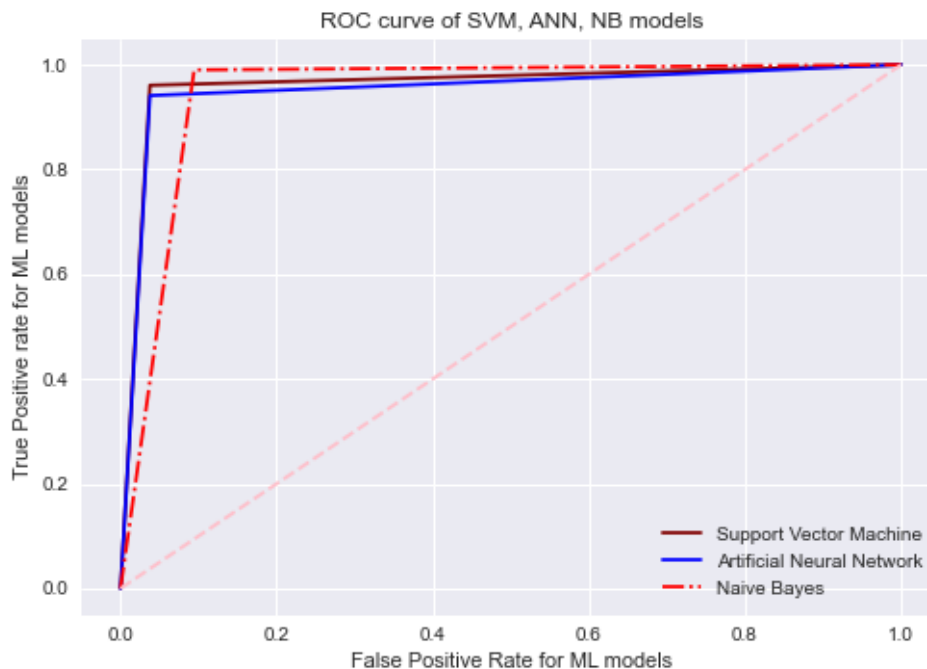


Figure 4.4: ROC curve for SVM, ANN, and NB model

Comparison of models with correctly and incorrectly classified data

The SVM model correctly classified 200 test data out of 208, the ANN model correctly classified 198 out of 208, and the NB model correctly classified 197 out of 208 test data. The following figure 4.5 shows the number of test data correctly and incorrectly classified with the developed model.

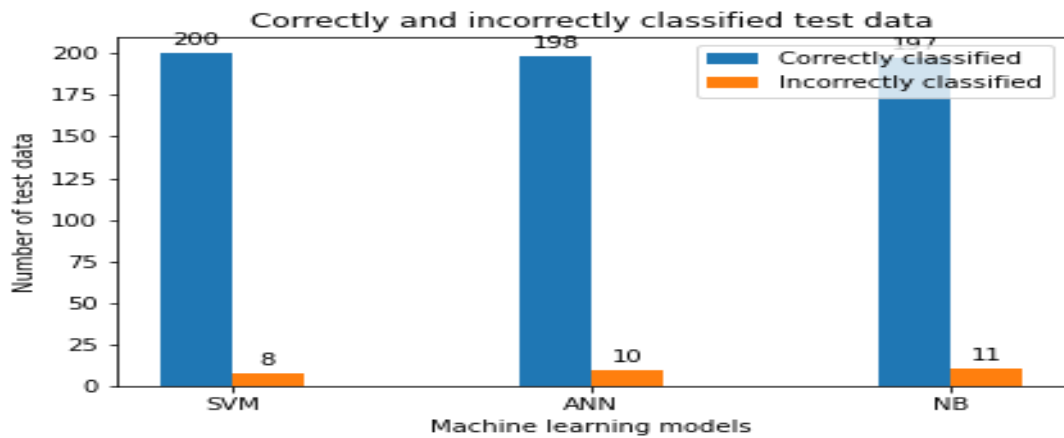


Figure 4.5: Comparison of models with correctly and incorrectly classified test data.

Boxplot of ML models

Boxplot is the chart used to visualize the performance of the developed machine learning models. The boxplot illustrates how a given data is distributed using minimum, maximum, median, first quartiles, and third quartiles of the dataset.

The following figure depicts the boxplot for the performance of SVM, ANN, and NB machine learning models.

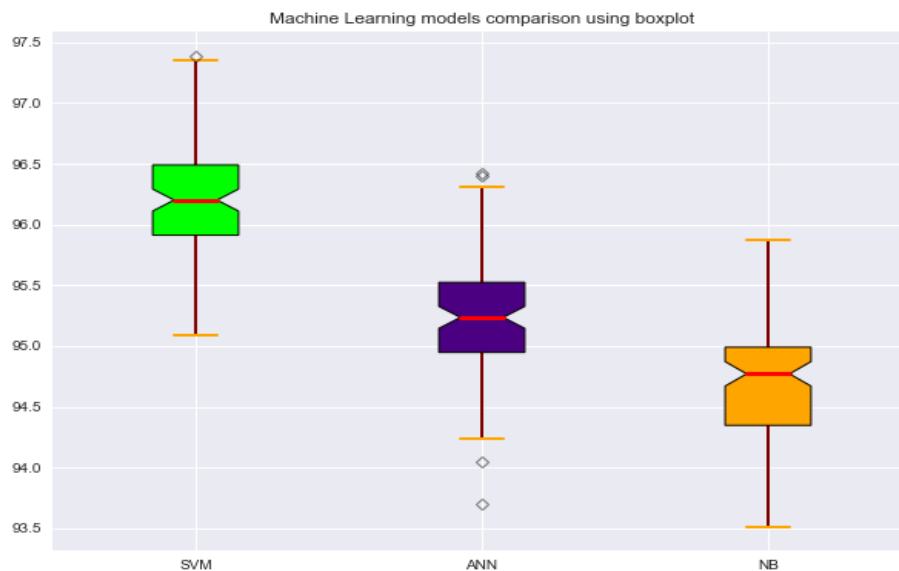


Figure 4.6: Machine learning models comparison using boxplot

4.2.3. Knowledge-based evaluation

The performance evaluation on the knowledge-based measures the capability of the knowledge-based to extract the meaning of jargon words for the input text with the AJMRD. The evaluation of knowledge-based performed after the developed ML model predicts a text as a jargony text. The performance of a knowledge-based is based on accuracy alike to machine learning.

The proposed system considered the agricultural society that read texts with a sentence, and a sentence is classified as jargony and non-jargony with the learned knowledge experience of the ML model. So that a sentence classified as jargony with a list of lexicons entered into the knowledge-based system for meaning extraction. Therefore, the knowledge-based system accepts the sentence with the jargon word and returns the text with the meaning of words.

We randomly used a total of 80 test sentences of different lengths for different experiments to test the knowledge base such as 20, 40, 60, and 80 test sentences. The different lengths of 20, 40, 60, and 80 test sentences used to evaluate the performance on the behalf of the number of input test sentences and the number of jargon words in the input sentences. The knowledge-based extracts the meaning of jargon words returned from the preprocessed with exact-match. However, the meaning extraction of jargon words without an exact match is impossible. We performed different experiments to measure the knowledge base performance. The following experiment describes the number of input test sentences and the number of jargon words in the input test sentences with the performance of the knowledge base.

Test experiment 1

We prepared 20 test sentences for the knowledge-based system to evaluate the capability of the system to extract the meaning of domain-specific Amharic agricultural jargon words from the AJMRD. The input test data includes 17 Amharic agricultural jargon words. The proposed system identifies 15 of the jargon words exactly; however, the meaning words are not extracted. This is because the inflectional behavior of the jargon word and over stemming is committed at the end of the preprocessing of the input test data. So that the performance of the system at this instant is achieved an accuracy of 88.2%.

Test experiment 2

In the second experiment, we prepared 40 sentences with and without the existence of domain-specific Amharic agricultural jargon words. In this experiment, 30 jargon words are included in the input test data, the exact meaning of 26 jargon words are extracted from the knowledge source. Therefore, we observed accuracy of 86.7% for the performance of the knowledge-based part of the developed system.

Test experiment 3

The third experiment considered the 60-test sentences. The prepared test data is entered into the knowledge-based system. Among the 41 jargon words that exist in the test sentences; the meaning of 35 jargon words is extracted exactly from the knowledge source. However, the remaining jargon words are not extracted because of the over-stemmed of the jargon words. So that our proposed system is 85.4% accurate in test experiment 3.

Test experiment 4

In the fourth experiment, we entered 80 test data. A total of 59 jargon words are included in the input test data, and 49 of the jargon words are identified by the developed system. The remaining words are not extracted though the words are available. Therefore, the proposed system performs with an accuracy of 83.1% for test experiment 4. Because binary lexical mapping is the way of extracting meaning from the predefined knowledge source, we faced the challenge of extraction of meaning with over-stemmed jargon words.

Based on the performance result of test experiment 1, test experiment 2, test experiment 3, and test experiment 4 returned from the knowledge-based of our proposed system, the input test data with a small number of test sentences and the small number of domain-specific Amharic agricultural jargon words outperform as compared to other experiments. So that as the number of jargon words in the input test data increases, to some extent, the rate of the performance result of the developed system decreases. The following chart depicts the performance results of test1, test2, test3, and test 4 conducted on the knowledge-based part of the proposed system.

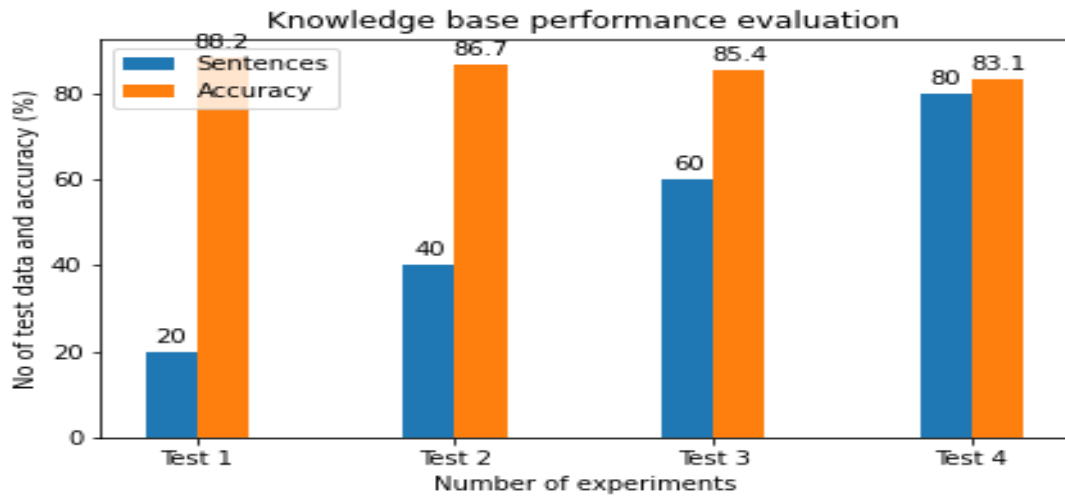


Figure 4.7: Performance result of knowledge base with test 1, test 2, test 3, and test 4

The result of the knowledge base shows as the number of domain-specific Amharic jargon words in the input test data and the number of test data increases, we observed a decrease rate of performance. This is because as the number of jargon words in the test data increases, the use of the inflectional forms of the same stem jargon word becomes increased to explore content. Different morphologically generated of the same jargon words can be used besides the context of the disseminated content. So that over-stemming of a jargon word is committed with the preprocessing of the knowledge base. Therefore, binary lexical mapping between the over-stemmed jargon word in the input text and the jargon word in the knowledge source is impossible. So that it decreases the performance of the knowledge-based system

Therefore, we observed the decreased performance of the knowledge-based system as the number of input test data and the number of domain-specific Amharic jargon words increased. Morphologically generated jargon is used in a different part of a sentence when the number of jargon words in a test data increases. The stemming phase of the preprocessing commits over stemming that results in a problem to extract the meaning of words and causes a decrease in the performance of the system is observed.

Stemming challenge in Knowledge-based system

The stemming phase of the preprocessing for the knowledge-based system commits stemming of a larger part of a word than the required, which in turn leads to stem incorrectly. The challenge of binary lexical mapping between the word in the input text

and the words in the knowledge source is observed which in turn decreases the performance of the knowledge-based system.

We handled the issue of over-stemming of jargon by counting the number of similar characters between the over-stemmed input text and the words in the knowledge source. So that the meaning of the over stemmed word is extracted from the knowledge source with the confirmation of the user based on the input of the over stemmed text.

Therefore, the over stemmed DSAJW from the result of the preprocessing is an input; hence the more related word with the help of close match is extracted from the knowledge source for confirmation with the user. The meaning of words confirmed as a similar word to the over stemmed word is extracted from the knowledge source.

4.3. Discussion

Experimental results described that the hybrid system for DSAJWI is affected by the labeled trained corpus to classify a text as jargony or non-jargony text. The machine learning part of our proposed system minimizes the workload of the knowledge-based system by discarding non-jargon text from entering the knowledge-based system. The developed machine learning model identified the input text as jargony solely entered to the knowledge-based system for further analysis and extraction of meaning.

Therefore, the proposed hybrid DSAJWI system works well for the identification of jargony text and non-jargony text. Jargony text entered to the knowledge-based system returned with the meaning of jargon texts from the knowledge source. We observed the slightly better performance result of the selected SVM machine learning model to predict the input test data as jargony and non-jargony.

4.3.1. Discussion of machine learning evaluation result

We observed machine learning models predict the input text with 96.2%, 95.2%, and 94.7% accuracy using SVM, ANN, and NB respectively. For the prepared labeled trained data, the Support Vector Machine (SVM) model outperformed the other developed models because SVM works well for binary classification (Holts et al., 2010) for our training data. We selected SVM for the model testing phase of our proposed system to classify unseen test data as jargony and non-jargony. So that we test different lengths of text with the SVM model, and the model predicts best for the input test data.

4.3.2. Discussion of knowledge base evaluation result

The knowledge-based of our proposed system best performs when fewer input sentences are entered into the system. As the number of input sentences increased, the inflectional forms of jargon words also increased, and the probability for the occurrence of over-stemming of the words also increased. Besides, the meaning of over-stemmed jargon words with binary lexical mapping cannot be extracted from the knowledge source. We observed accuracy of 88.2%, 86.7%, 85.4%, and 83.1% for the input of 20, 40, 60, and 80 test sentences respectively. For a few sentences entered into the knowledge-based system, the best performance is observed.

Therefore, the proposed hybrid system works well for the identification of jargon and non-jargon text. The machine learning model decreases the workload of the knowledge base by discarding non-jargon text from entering the knowledge base system. Texts classified as jargony text entered into the knowledge-based system. Therefore, for every occurrence of a jargon word in a jargony text, the meaning of the word is extracted from the knowledge source. The reason of the best performance for few sentences entered into the knowledge-based system is because of in a few sentences, small number of jargon words are included and the use of morphologically generated of the same jargon word decrease. In case as the number of input text increase the use of various morphologically generated of the same jargon word increases. So that the performance of the knowledge-based system is better in few sentences of input text.

4.4. Summary

In this chapter, the experimental setup necessary for DSAJWI system is discussed. The experimental results with the developed model using performance metrics are described. The performance result of the developed SVM, ANN, and NB model are discussed and SVM is the outperformed model for our proposed system of two-classification. We used confusion matrix, ROC curve, and boxplot for the description of performance results. We evaluated the performance of the knowledge base and the input with less test data outperforms the other. Therefore, identification of Amharic jargon words is performed with the machine learning for classification and knowledge based for meaning extraction.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATIONS

In this thesis, we developed a DSAJWI system for the Amharic language with a combination of machine learning and knowledge-based. There is no prior work in this area of research in the Amharic language; we selected machine learning models for training our dataset. We used a knowledge source to provide the meaning of words to return clear and precise information to the user of the text. We have studied machine learning techniques to develop models for training the labeled dataset and select better model for the prediction of the test dataset. The selected model used for the prediction of test data and to work with the knowledge-based system.

5.1. Conclusion

This research work attempts to develop a domain-specific Amharic jargon word identification system for the Amharic language to identify a text with a jargon word or prominent text. Institutions in the Federal Democratic Republic of Ethiopia (FDRE) such as agriculture are more vulnerable to the usage of domain-specific Amharic jargon words in organizations' daily business to come up with organizations' success in line to accomplish the seated objectives. Though, the use of domain-specific words in organizational discourse enables communicants to handle simple communication and foster the business development of an organization; however, it hampers the communication when the target receivers are out of the domain. Domain-specific Amharic jargon word identification system involves the steps of preprocessing of labeled training corpus, classification of the text based on the learned knowledge with machine learning, and extraction meaning of jargon word from the predefined explanatory lexical resource. The study focused on identification of jargon words in a text and provide meaning of words for agricultural domains.

We prepared the Amharic Jargon Machine Readable Dictionary (AJMRD) manually for our study with the help of agricultural domain experts for the agricultural domain. We collected Amharic agricultural jargon words from agricultural business reports, working guidelines, training manuals to develop the AJMRD.

The architecture of our proposed system contains preprocessing machine learning and knowledge-based. Our proposed system takes Amharic labeled dataset as the input to the preprocessing. The machine learning model trained with the preprocessed data to

experience learned knowledge for classification. We selected the outperformed model from the developed and trained machine learning models for the knowledge-based system. The text classified as jargon is the input for the knowledge-based system to extract the meaning of the Amharic agricultural jargon words. So that the developed system identifies the existence of domain-specific Amharic jargon words in the input text with the learned experience of the developed model and returns the meaning of jargon from the available knowledge sources in the knowledge base. The knowledge base updated when new jargon words are found that doesn't exist in the knowledge base.

We performed classification of text as jargony and non-jargony using the developed machine learning model and providing meaning of jargon words from the knowledge source in the knowledge-based. Evaluation of the developed machine learning models are performed and we selected the outperformed model. We have developed models with Support Vector Machine (SVM), Artificial Neural Network (ANN), and Naïve Bayes (NB). We evaluated the developed system both with machine learning and knowledge-based. First, we evaluated the machine learning model and we achieved an accuracy of 96.2%, 95.2%, and 94.7% using SVM, ANN, and NB respectively. We selected the outperformed SVM model for the evaluation of the knowledge-based system.

We observed the best performance of the knowledge-based system for the input of the small number of test data. For the input of 20, 40, 60, and 80 test sentences, an accuracy of 88.2%, 86.7%, 85.4%, and 83.1% is observed. So that for a few sentences entered into the knowledge-based system, the best performance of the system is observed. Therefore, we observed the best performance of our proposed system with the knowledge base to extract the meaning of jargon words from the predefined explanatory lexical resource for jargon text with less amount of input test data. So that, based on our experiment with the hybrid of ML and knowledge-based, the proposed system identified the jargony text with the machine learning model, and provided meaning with the knowledge-based. Therefore, with a hybrid approach, we have achieved a promising result for domain-specific Amharic jargon word identification.

5.2. Contributions of the study

In this study, an attempt has been made to identify the existence of domain-specific Amharic jargon words in Amharic text and provide the meaning based on the meaning of words in the domain. This is the first attempt and we employed machine learning and a knowledge-based approach. We confirmed that the problem is weighty with the survey we conducted. We developed the Amharic Jargon Machine Readable Dictionary (AJMRD) for the knowledge source. The study of domain-specific Amharic jargon words has the following main contributions.

- We assured that the problem is weighty for experts and non-experts with the survey conducted on domain-experts, non-domain experts, and the agricultural society.
- We developed a system for the agricultural society for both experts and non-experts to provide the meaning for domain-specific Amharic agricultural jargon words.
- We integrated the machine learning and knowledge-based to decrease the workload of the knowledge base and provide prominent information to the targeted user of the text.
- The agricultural report readers of the text are full of information in line with their target for the text sourced from any agricultural organization and experts.
- We prepared a domain-specific corpus and knowledge source with a list of words and the words meaning for further enhancement of future research.
- We handle the issue of over-stemming for the knowledge base by calculating the number of similar characters between the over-stemmed preprocessed lexicon and the word in the knowledge source with a close-match of words.

5.3. Recommendations

In our current study, we only consider domain-specific Amharic jargon word identification in the agricultural domain with a hybrid of machine learning and knowledge-based. We trained the developed machine learning models with the labeled training and testing dataset prepared with the help of agricultural domain-expert curators for the two-way classification. However, we observed that this work requires further enhancement to improve the performance of the developed system to provide the most robust system to the target user of text. The following are some of the future works.

- ✓ The proposed system created the knowledge base manually to extract the meaning of domain-specific words. This problem can be alleviated with other techniques. We recommend researchers for the automatic generation of the meaning of domain-specific jargon words in various domains.
- ✓ We are targeted on the domain-specific agricultural jargon words because the domain is more vulnerable to use jargon words and most of the life of Ethiopian people is agriculture-dependent. The developed model is applicable for any domains. We recommend other researchers to work on other domains such as health, law, education, technology, construction, etc with the extended features of our proposed system to provide meaningful information to the user and for the improvement of the Amharic language usage in a particular domain.
- ✓ We collected a dataset from various agricultural documents for the machine learning models; we recommend collecting more datasets to work with a deep learning approach to experience learning and improve performance.
- ✓ The model for the proposed system was developed with SVM, ANN, and NB. We recommend researchers select the best hyperparameter values to enhance the performance; hence selecting the best value of hyperparameters makes the developed model outperform.
- ✓ We extract the meaning of words from the meaning collection of the knowledge source with lexical binary mapping between the preprocessed input text and the list of words in the knowledge source. The over-stemmed preprocessed text results in a problem to extract the meaning without an exact match between the over stemmed word and the word in the knowledge source. So further research works are required to handle the issue of stemming for the Amharic language; hence this area of research is advantageous with effective and efficient Amharic stemmer/morphological analyzer.
- ✓ Our research work is focused on the meaning of jargon words. However, multi-word jargon words in various disciplines are available; hence experts of a domain use the words to explore domain-specific themes; in case customers are confused to understand. So, researchers are recommended to address this issue to provide prominent information for the intended users.
- ✓ Our proposed system is the prior work attempted on the identification of DSAJW. However, we recommend researchers work on the domain-concept classification of the domain.

REFERENCES

- Abikoye, O. C., & Omokanye, S. O. (2017). *BINARY TEXT CLASSIFICATION USING AN ENSEMBLE OF NAÏVE BAYES AND BINARY TEXT CLASSIFICATION USING AN ENSEMBLE OF NAÏVE*. January 2018.
- Ananiadou, S., Rea, B., Okazaki, N., Procter, R., Thomas, J., Tsafnat, G., Glasziou, P., Choong, M. K., Dunn, A., Galgani, F., Coiera, E., Joachims, T., O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2009). Text Categorization with SVM: Learning with Many Relevant Features. *Systematic Reviews*, 4(1), 1–15.
- Antonic, S. (2008). *Serbian Wordnet for biomedical sciences*.
- Baharudin, B., Lee, L. H., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1(1), 4–20. <https://doi.org/10.4304/jait.1.1.4-20>
- Bakx, G. E. (2006). Machine Learning Techniques for Word Sense Disambiguation. *Machine Learning*, 1–172. <https://www.lsi.upc.edu/~escudero/wsd/06-tesi.pdf>
- Berrar, D. (2018). Bayes' theorem and naive bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1–3(September), 403–412. <https://doi.org/10.1016/B978-0-12-809633-8.20473-1>
- Boser, B. E., Vapnik, V. N., & Guyon, I. M. (1992). Training Algorithm Margin for Optimal Classifiers. *Perception*, 144–152.
- Brown, Z. C., Anicich, E. M., & Galinsky, A. D. (2020). Compensatory conspicuous communication: Low status increases jargon use. *Organizational Behavior and Human Decision Processes*, 161(July), 274–290. <https://doi.org/10.1016/j.obhdp.2020.07.001>
- Burns, T. W., O'Connor, D. J., & Stocklmayer, S. M. (2003). Science communication: A contemporary definition. *Public Understanding of Science*, 12(2), 183–202. <https://doi.org/10.1177/09636625030122004>
- Byun, H., & Lee, S. W. (2002). Applications of support vector machines for pattern recognition: A survey. *Lecture Notes in Computer Science (Including Subseries*

- Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), 2388, 213–236. https://doi.org/10.1007/3-540-45665-1_17
- Calsolari, N. (1984). *Machine-readable dictionaries, lexical data bases and the lexical system. January 1984*, 460–460. <https://doi.org/10.3115/980431.980586>
- Carstensen, A. K., & Bernhard, J. (2019). Design science research—a powerful tool for improving methods in engineering education research. *European Journal of Engineering Education*, 44(1–2), 85–102. <https://doi.org/10.1080/03043797.2018.1498459>
- Cheng, B. (2015). *Titterington, D. M.: Neural Networks : A Review from a Statistical Perspective . 9*(February 1994). <https://doi.org/10.1214/ss/1177010638>
- Clayton, T. (1998). Explanations for the use of languages of wider communication in education in developing countries. *International Journal of Educational Development*, 18(2), 145–157. [https://doi.org/10.1016/S0738-0593\(98\)00002-9](https://doi.org/10.1016/S0738-0593(98)00002-9)
- Crémer, J., Garicano, L., & Prat, A. (2007). Language and the theory of the firm. *Quarterly Journal of Economics*, 122(1), 373–407. <https://doi.org/10.1162/qjec.122.1.373>
- Cyr, A. (2012). Social Media: Don't Discount the Benefits! *Oncology Times*, 34(8), 1–3. <https://doi.org/10.1097/01.COT.0000414683.49317.3b>
- Dalianis, H., & Dalianis, H. (2018). Evaluation Metrics and Evaluation. *Clinical Text Mining, 1967*, 45–53. https://doi.org/10.1007/978-3-319-78503-5_6
- Danesi, M. (1995). What is Language? *New Vico Studies*, 13, 43–54. <https://doi.org/10.5840/newvico1995132>
- Demeke, M., & Ferede, T. (2014). *Agricultural Development in Ethiopia : Are There alternatives to Food Aid ? AGRICULTURAL DEVELOPMENT IN ETHIOPIA : ARE THERE ALTERNATIVES TO FOOD AID ? BY. October.*
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies - Proceedings of the Conference, 1(Mlm), 4171–4186.

- El-Khair, I. A. (2017). *Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study*. January 2006. <http://arxiv.org/abs/1702.01925>
- Elkateb, S., & Black, W. (2005). *Arabic WordNet and the Challenges of Arabic*. *Tufts 2004*, 15–24.
- Eyassu, S. (2005). *Classifying Amharic News Text Using Self-Organizing Maps*. June, 71–78.
- Gallo, K. (2018). *Understanding professional jargons literature review*. December. <https://doi.org/10.5604/00332860.1234518>
- Gasser, M. (2011). HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. *Conference on Human Language Technology for Development, August*, 94–99.
- Gasson, S. (2003). Human-Centered vs. User-Centered Approaches to Information System Design College of Information Science and Technology. *Journal of Information Technology Theory and Application*, 5(2), 29–46.
- Gong, L., Yang, R., Liu, Q., Dong, Z., Chen, H., & Yang, G. (2017). A Dictionary-Based Approach for Identifying Biomedical Concepts. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(9), 1–12. <https://doi.org/10.1142/S021800141757004X>
- Helmreich, S., Llevadias Jané, J., & Farwell, D. (2005). Identifying jargon in texts. *Identifying Jargon in Texts*, 35(35), 425–432.
- Hermawan, R. (2011). Natural language processing with python. In *Indonesian Journal of Applied Linguistics* (Vol. 1, Issue 1). <https://doi.org/10.17509/ijal.v1i1.106>
- Holts, A., Riquelme, C., & Alfaro, R. (2010). Automated text binary classification using machine learning approach. *Proceedings - International Conference of the Chilean Computer Science Society, SCCC, May 2014*, 212–217. <https://doi.org/10.1109/SCCC.2010.30>

- Hudson, G. (1999). *Linguistic Analysis of the 1994 Ethiopian Census Linked references are available on JSTOR for this article : Linguistic Analysis of the 1994 Ethiopian Census*. 6(3), 89–107.
- Ibrahim, M., Gauch, S., Salman, O., & Alqahatani, M. (2020). Enriching consumer health vocabulary using enhanced glove word embedding. *CEUR Workshop Proceedings*, 2619.
- Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 4(8), 966–974. <https://doi.org/10.11499/sicej11962.38.456>
- J. A. Hartigan and M. A. Wong. (2012). Algorithm AS 136 A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 28(1), 100–108.
- Jing, L. P., Huang, H. K., & Shi, H. B. (2002). Improved feature selection approach TFIDF in text mining. *Proceedings of 2002 International Conference on Machine Learning and Cybernetics*, 2(November), 944–946. <https://doi.org/10.1109/icmlc.2002.1174522>
- Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. *Proceedings of ICML97*.
- Kevitt, P. M., Partridge, D., & Wilks, Y. (1992). Approaches to natural language discourse processing. *Artificial Intelligence Review*, 6(4), 333–364. <https://doi.org/10.1007/BF00123689>
- Kilgarriff, A., & Yallop, C. (2000). *What 's in a thesaurus*. November 2001, 1371–1379.
- Koay, J. J., Roustai, A., Dai, X., Burns, D., Kerrigan, A., & Liu, F. (2020). *How Domain Terminology Affects Meeting Summarization Performance*. <http://arxiv.org/abs/2011.00692>
- Lamy, J. B. (2017). Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine*, 80(August 2017), 11–28.

<https://doi.org/10.1016/j.artmed.2017.07.002>

Liebowitz, J. (2010). Studies in Fuzziness and Soft Computing: Foreword. *Studies in Fuzziness and Soft Computing*, 258.

Links, A. R., Callon, W., Wasserman, C., Walsh, J., Beach, M. C., & Boss, E. F. (2019). Patient Education and Counseling Surgeon use of medical jargon with parents in the outpatient setting. *Patient Education and Counseling*, 102(6), 1111–1118. <https://doi.org/10.1016/j.pec.2019.02.002>

Mccallum, A., & Nigam, K. (1997). *A Comparison of Event Models for Naive Bayes Text Classification*.

Mikre-Sellassie, G. A. (2000). The Early Translation of the Bible into Ethiopic/Geez. *The Bible Translator*, 51(3), 302–316.
<https://doi.org/10.1177/026009350005100302>

Mindaye, T., & Atnafu, S. (2009). Design and implementation of Amharic search engine. *Proceedings - 5th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2009, January*, 318–325.
<https://doi.org/10.1109/SITIS.2009.58>

Mindaye, T., & Kassie, T. (2018). *The Need for Amharic WordNet The Need for Amharic WordNet. January 2008*.

Ong, J., & Liaw, H. (2013). Language Usage of Jargon and Slang in Strategic Studies. *Australian Journal of Basic and Applied Sciences*, 7(4), 661–666.

Ong, J., & Liaw, H. (2015). *Language Usage of Jargon and Slang in Strategic Studies. January 2013*.

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology*, 20(2), 139–156. <https://doi.org/10.1002/acp.1178>

Osowski, S., Siwek, K., & Markiewicz, T. (2004). MLP and SVM networks - A comparative study. *Report - Helsinki University of Technology, Signal Processing Laboratory*, 46(2), 37–40.

- Pal, A. R., & Saha, D. (2013). *DETECTION OF JARGON WORDS IN A TEXT USING SEMI-SUPERVISED*. 95–107.
- Pal, A. R., & Saha, D. (2013). *Detection of Jargon Words in a Text Using Semi-Supervised Learning*. July 2013, 95–107. <https://doi.org/10.5121/csit.2013.3411>
- Patoko, N., & Yazdanifard, R. (2014). The Impact of Using Many Jargon Words, while Communicating with the Organization Employees. *American Journal of Industrial and Business Management*, 04(10), 567–572. <https://doi.org/10.4236/ajibm.2014.410061>
- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Peng, D., Jin, L., Wu, Y., Wang, Z., & Cai, M. (2019). A fast and accurate fully convolutional network for end-to-end handwritten chinese text segmentation and recognition. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*. <https://doi.org/10.1109/ICDAR.2019.00014>
- Piasecki, M., Szpakowicz, S., Wydawnicza, O., Wrocławskiej, P., & Radziszewski, A. (2009). *A Wordnet from the Ground Up*.
- Profile, S. E. E., & Profile, S. E. E. (2019). *Handbook of research methodology*. August 2017.
- Rakedzon, T., Segev, E., Chapnik, N., Yosef, R., & Baram-Tsabari, A. (2017). Automatic jargon identifier for scientists engaging with the public and science communication educators. *PLoS ONE*, 12(8), 1–13. <https://doi.org/10.1371/journal.pone.0181742>
- Roth, R. I. (1996). Hemoglobin enhances the binding of bacterial endotoxin to human endothelial cells. *Thrombosis and Haemostasis*, 76(2), 258–262. <https://doi.org/10.1055/s-0038-1650565>
- Ryan, C. (2014). *M Tif*. January. <https://doi.org/10.3318/DRI.2014.1>
- S. Abirami, P. C. (2020). The Digital Twin Paradigm for Smarter Systems and

- Environments: The Industry Use Cases. In *Advances in Computers*.
- Sang, Y., Zhang, H., & Zuo, L. (2008). Least squares support vector machine classifiers using PCNNs. *2008 IEEE International Conference on Cybernetics and Intelligent Systems, CIS 2008*, 290–295.
<https://doi.org/10.1109/ICCIS.2008.4670890>
- Schmitt, N. (2000). *The Percentage of Words Known in a Text and Reading Comprehension*. <https://doi.org/10.1111/j.1540-4781.2011.01146.x>
- Seyler, D., Liu, W., Wang, X., & Zhai, C. (2020). *Towards Dark Jargon Interpretation in Underground Forums*. 1–8. <http://arxiv.org/abs/2011.03011>
- Sharkawy, A.-N. (2020). Principle of Neural Network and Its Main Types: Review. *Journal of Advances in Applied & Computational Mathematics*, 7(1), 8–19.
<https://doi.org/10.15377/2409-5761.2020.07.2>
- Shen, Y., Kumar, R., Jiang, S., & Wellyanto, M. R. (2020). Can AI decrypt fashion jargon for you? *ArXiv*.
- Sirbu, A. (2015). The significant of language as a tool of communication. *PROQUEST SciTech Journals, XVIII(2)*, 405–406.
<http://www.thefreedictionary.com/dialect>
- Sparck Jones, K. (1994). Natural Language Processing: A Historical Review. *Current Issues in Computational Linguistics: In Honour of Don Walker*, 3, 3–16.
- Taheri, S., & Mammadov, M. (2013). Learning the naive bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, 23(4), 787–795. <https://doi.org/10.2478/amcs-2013-0059>
- Tan, H. T., Wang, E. Y., & Yoo, G. S. (2019). Who likes jargon? The joint effect of jargon type and industry knowledge on investors' judgments. *Journal of Accounting and Economics*, 67(2–3), 416–437.
<https://doi.org/10.1016/j.jacceco.2019.03.001>
- Welteji, D. (2018). A critical review of rural development policy of Ethiopia: Access, utilization and coverage. *Agriculture and Food Security*, 7(1), 1–6.
<https://doi.org/10.1186/s40066-018-0208-y>

- Weng, W. H., Chung, Y. A., & Szolovits, P. (2019). Unsupervised clinical language translation. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 3121–3131.
<https://doi.org/10.1145/3292500.3330710>
- Weng, W. H., & Szolovits, P. (2018). Mapping Unparalleled Clinical Professional and Consumer Languages with Embedding Alignment. *ArXiv*, 1–7.
- Willoughby, S. D., Johnson, K., & Sterman, L. (2020). *Quantifying scientific jargon*.
<https://doi.org/10.1177/0963662520937436>
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, 42–49.
<https://doi.org/10.1145/312624.312647>
- Zanaty, E. A. (2012). Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification. *Egyptian Informatics Journal*, 13(3), 177–183.
<https://doi.org/10.1016/j.eij.2012.08.002>
- Zupon, A., Crew, E., & Ritchie, S. (2021). *Text Normalization for Low-Resource Languages of Africa*. <http://arxiv.org/abs/2103.15845>

Appendix I:

ባሕር ዳር ዩኒቨርሲቲ

ባሕር ዳር ቴክኖሎጂ ኢንስቲትዩት

ኢንፎርሜሽን ቴክኖሎጂ ትምህርት ክፍል

(Questionnaire for domain-experts)

በግብርና የሙያ ቃላት አጠቃቀም ላይ የባለሙያዎችን አስተያየት ለመሰብሰብ የተሰጠ የፅሁፍ መጠይቅ

የሚከተሉትን የፅሁፍ መጠይቆች በጥንቃቄ በማንበብ አስፈላጊውን መልስ ያስቀምጡ፡፡

1. የትኛውን የትምህርት ደረጃ ያሟላሉ?

- የሶስተኛ ዲግሪ (PhD) የመጀመሪያ ዲግሪ (BSc)
 ሁለተኛ ዲግሪ (MSc) የዲፕሎማ ሌላ

2. ከሚከተሉት የግብርና ተቋማት ውስጥ የየትኛው ተቋም ስራተኛ ነዎት?

- ቀበሌ ባለሙያ ክልል ግብርና ቢሮ
 ወረዳ ግብርና ፅ/ቤት መንግስታዊ ያልሆነ ተቋም
 ዞን ግብርና መምሪያ ሌላ

3. የሚከተሉትን የግብርና የሙያ ቃላት ያንብቡ እና መልስዎን ያስቀምጡ፡፡

ተቁ	የሙያ ቃል	የቃሉን ትርጉም ያውቃሉ?		ቃሉን ለስራ ይጠቀማሉ	
		አዎ	አይደለም	አዎ	አይደለም
1	መለስተኛፈር				
2	መትካየር				
3	ማጀያናስ				
4	ማጋህየት				
5	በትነጓል				
6	ብላፅዋት				
7	ነቃባህሪነት				
8	አምባፈር				
9	አርፋዶር				

Appendix II:

ባሕር ዳር ዩኒቨርሲቲ

ባሕር ዳር ቴክኖሎጂ ኢንስቲትዩት

ኢንፎርሜሽን ቴክኖሎጂ ትምህርት ክፍል

(Questionnaire for non-domain experts)

በግብርና የሙያ ቃላት አጠቃቀም ላይ የአርሶ አደሮችን እና ከግብርና ተቋም ውጭ ያሉ ባለሙያዎችን አስተያየት ለመሰብሰብ የተሰጠ የፅሁፍ መጠይቅ

የሚከተሉትን የፅሁፍ መጠይቆች በጥንቃቄ በማንበብ አስፈላጊውን መልስ ያስቀምጡ።

1 ከሚከተሉት ውስጥ የየትኛው ግብርና አገልግሎት ተጠቃሚ ነዎት?

- የገጠር ግብርና የከተማ ግብርና

2 የትኛውን የግብርና መንገድ ይከተላሉ?

- ባህላዊ ግብርና ዘመናዊ ግብርና

3 የሚከተሉትን የግብርና የሙያ ቃላት ያንብቡ እና መልስዎን ያስቀምጡ።

ተቁ	የሙያ ቃል	የቃሉን ትርጉም ያውቃሉ?		ቃሉን በስራ ውስጥ ያገኙታል	
		አዎ	አይደለም	አዎ	አይደለም
1	መለስተኛፈር				
2	መትካየር				
3	ማጀያናስ				
4	ማጋህየት				
5	በትነጓል				
6	ብላፀዋት				
7	ነቃባህሪነት				
8	አምቧፈር				
9	አርፋዶር				
10	አድማሳፈር				
11	ከባዳፈር				

Appendix III:

List of Amharic stopwords

Stop words

ነገር	ሌላ	ሌሎች	ስለ	ቢሆን
አንድ	ብቻ	መሆኑ	ማድረግ	ማንም
ማለት	ማለቱ	የሚገኝ	የሚገኙ	ሲሆን
ማን	ይህን	በተለይ	እያንዳንድ	በሆነ
አንድን	ሲል	እዚህ	እንጂ	በኩል
እና	ከመካከል	ከጋራ	ጋራ	ሲሉ
ና	በውስጥ	በጣም	ወዘተ	ወደ
ወይም	ከዚህ	ከላይ	ከመሀል	ያለ
ሆኑ	በተመለከተ	ሆኖም	ነው	ናቸው
ሁሉንም	ላይ	ተመሳሳይ	ያሉ	የኋላ
የሰሞኑ	አንቀጽ	ሀያ	የሆኑትን	ስድስት
እስከ	ይኸኛው	ሆነ	በሆኑ	ምንም
ብሎ	ከርሱ	እንደሆነ	በኋላ	ጀምሮ
በሆነው	ከሰላሳ	መሆኑን	በአንድ	እንደዚህ
የሚሆኑ	ለዚያው	የሆኑ	በመሆን	ይሁን

Appendix IV:

List of Amharic prefix and suffix

List of prefix and suffix

ከነ	እስከ	መ	የእነ	ይ	በየ	እስከት
በእነ	ሰለ	ለእነ	እስኪ	በ	እንደየ	የሚ
የ	እነ	ሲ	ከየ	ን	እን	እስከን
እንድ	ለ	እንደ	እንድን	እንዲ	እየ	ዎቿ
ን	ያል	ላችሁ	ዎች	ዎቹን	ምና	ና
የዋ	ላቸው	ዎችን	ናም	ምናም	ም	ቸው
ባቸው	ዎችንና	ናምና	ምናን	ች	ተው	ባችሁ
ዎችንም	ምቹ	ናምን	ዩ	የው	ናንና	ህ
ዎችና	ናን	ምን	ዎ	ዎችናም	በት	ባት
ምንና	ናንም	ዎችናን	ምንም	ሽ	ነት	ዎችም
ናንን	ምንን	ዋ	ለት	ዎችምና	ምዉና	ናው
ቹ	ዎቹና	ላት	ናዉና	ምዉም	ው	ይቱ
ዎቹናም	ምውም	ቱ	ያዊ	ዎቹናን	ናዉም	ምዉን
ሁ	ኞች	ዎቹም	ዉም	ኝ	ዎቹ	ዎቹምና

Appendix V:

Amharic characters, numbers, and punctuations

Amharic characters

	ä/e	u	i	a	e	ə	o		ä/e	u	i	a	e	ə	o
	[ɛ/ə]	[u]	[i]	[a]	[e]	[i/ɨ]	[o/ɔ]		[ɛ/ə]	[u]	[i]	[a]	[e]	[i/ɨ]	[o]
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	k	ከ	ከሁ	ከሂ	ከሃ	ከሄ	ከህ	ከሆ
[h]	ha	hu	hi	ha	he	hə	ho	[k]	kä	ku	ki	ka	ke	kə	ko
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ	ገ	ኸ	ኸሁ	ኸሂ	ኸሃ	ኸሄ	ኸህ	ኸሆ
[l]	lä	lu	li	la	le	lə	lo	[h]	kä	ku	ki	ka	ke	kə	ko
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ	w	ወ	ወሁ	ወሂ	ወሃ	ወሄ	ወህ	ወሆ
[h]	ḥa	ḥu	ḥi	ḥa	ḥe	ḥə	ḥo	[w]	wä	wu	wi	wa	we	wə	wo
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ	ʾ/ʿ	ዐ	ዐሁ	ዐሂ	ዐሃ	ዐሄ	ዐህ	ዐሆ
[m]	mä	mu	mi	ma	me	mə	mo	[ʔ]	ʾa	ʾu	ʾi	ʾa	ʾe	ʾə	ʾo
ś	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ	z	ዘ	ዘሁ	ዘሂ	ዘሃ	ዘሄ	ዘህ	ዘሆ
[s]	śä	śu	śi	śa	śe	śə	śo	[z]	zä	zu	zi	za	ze	zə	zo
r	ረ	ሩ	ሪ	ራ	ራ	ሮ	ሮ	ž	ዠ	ዠሁ	ዠሂ	ዠሃ	ዠሄ	ዠህ	ዠሆ
[r]	rä	ru	ri	ra	re	rə	ro	[ʒ]	žä	žu	ži	ža	že	žə	žo
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ	y	የ	የሁ	የሂ	የሃ	የሄ	የህ	የሆ
[s]	sä	su	si	sa	se	sə	so	[j]	jä	ju	ji	ja	je	jə	jo
š	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ	d	ደ	ደሁ	ደሂ	ደሃ	ደሄ	ደህ	ደሆ
[ʃ]	šä	šu	ši	ša	še	šə	šo	[d]	dä	du	di	da	de	də	do
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ğ	ጀ	ጀሁ	ጀሂ	ጀሃ	ጀሄ	ጀህ	ጀሆ
[kʰ]	qä	qu	qi	qa	qe	qə	qo	[ɕ]	ğä	ğu	ği	ğa	ge	gə	go
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	g	ገ	ገሁ	ገሂ	ገሃ	ገሄ	ገህ	ገሆ
[b]	bä	bu	bi	ba	be	bə	bo	[g]	gä	gu	gi	ga	ge	gə	go
v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ	ʈ	ጠ	ጠሁ	ጠሂ	ጠሃ	ጠሄ	ጠህ	ጠሆ
[v]	vä	vu	vi	va	ve	və	vo	[tʰ]	tä	tu	ti	ta	te	tə	to
t	ተ	ቱ	ቲ	ታ	ቲ	ቲ	ቲ	ç	ጮ	ጮሁ	ጮሂ	ጮሃ	ጮሄ	ጮህ	ጮሆ
[t]	tä	tu	ti	ta	te	tə	to	[tʃʰ]	çä	çu	çi	ça	çe	çə	ço
č	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ	p	ጸ	ጸሁ	ጸሂ	ጸሃ	ጸሄ	ጸህ	ጸሆ
[tʃ]	čä	ču	či	ča	če	čə	čo	[pʰ]	pä	pu	pi	pa	pe	pə	po
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ	s	ሠ	ሠሁ	ሠሂ	ሠሃ	ሠሄ	ሠህ	ሠሆ
[h]	ḥa	ḥu	ḥi	ḥa	ḥe	ḥə	ḥo	[ts]	śä	śu	śi	śa	śe	śə	śo
n	ነ	ኑ	ኒ	ና	ኔ	ኑ	ኑ	ś	ሰ	ሰሁ	ሰሂ	ሰሃ	ሰሄ	ሰህ	ሰሆ
[n]	nä	nu	ni	na	ne	nə	no	[ts]	śä	śu	śi	śa	śe	śə	śo
ny/ñ	ን	ኑ	ኒ	ና	ኔ	ኑ	ኑ	f	ፈ	ፈሁ	ፈሂ	ፈሃ	ፈሄ	ፈህ	ፈሆ
[ɲ]	nyä	nyu	nyi	nya	nye	nyə	nyo	[f]	fä	fu	fi	fa	fe	fə	fo
ʾ/ʿ	አ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ	p	ፐ	ፐሁ	ፐሂ	ፐሃ	ፐሄ	ፐህ	ፐሆ
[ʔ]	ʾa	ʾu	ʾi	ʾa	ʾe	ʾə	ʾo	[p]	pä	pu	pi	pa	pe	pə	po

ሰ	ሱ	ሚ	ሚ	ሪ	ሰ	ሰ	ቁ	ቀ
[l ^w a]	[h ^w a]	[m ^w a]	[s ^w a]	[r ^w a]	[s ^w a]	[ʃ ^w a]	[k ^w ɔ]	[k ^w i]
ቋ	ቋ	ቀ	ባ	ባ	ቲ	ቲ	ኸ	ኸ
[k ^w a]	[k ^w e]	[k ^w i/ɪ]	[b ^w a]	[v ^w a]	[t ^w a]	[tʃ ^w a]	[h ^w ɔ]	[h ^w i]
ሩ	ሩ	ኸ	ሩ	ሩ	ኸ	ኸ	ኸ	ኸ
[h ^w a]	[h ^w e]	[h ^w i/ɪ]	[n ^w a]	[p ^w a]	[ʔa]	[k ^w ɔ]	[k ^w i]	[k ^w a]
ኸ	ኸ	ኸ	ኸ	ኸ	ኸ	ኸ	ኸ	ኸ
[k ^w e]	[k ^w i/ɪ]	[h ^w ɔ]	[h ^w i]	[h ^w a]	[h ^w e]	[h ^w i/ɪ]	[z ^w a]	[ʒ ^w a]
ጸ	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ
[ts ^w a]	[tʃ ^w a]	[g ^w ɔ]	[g ^w i]	[g ^w a]	[g ^w e]	[g ^w i/ɪ]	[t ^w a]	[tʃ ^w a]
ጸ	ጸ	ጸ	ጸ					
[p ^w a]	[ts ^w a]	[f ^w a]	[p ^w a]					

Amharic numbers

English	Amharic	Amharic Pronunciation	English Pronunciation	English	Amharic	Amharic Pronunciation	English Pronunciation
1	፩	አንድ	And	20	፳	ሃያ	Haya
2	፪	ሁለት	Hulet	30	፳፩	ሰላሳ	Selasa
3	፫	ሦስት	Sost	40	፳፯	አርባ	Arba
4	፬	አራት	Arat	50	፷	ሃምሳ	Hamsa
5	፭	አምስት	Amist	60	፷፮	ስልሳ	Silsa
6	፮	ስድስት	Sidist	70	፸	ሰባ	Seba
7	፯	ሰባት	Sebat	80	፸፬	ሰማንያ	Semanya
8	፰	ስምንት	Simint	90	፸፯	ዘጠና	ZeTena
9	፱	ዘጠኝ	ZeTeñ	100	፻	መቶ	Meto
10	፲	አስር	Asir	1,000	፲፪	ሺ	Shee

Amharic punctuation marks

- ※ section mark ፥ colon
- ፥ word separator ፥- preface colon
- ፥፥ full stop (period) ፥፥ question mark
- ፥፥፥ comma ፥፥፥ paragraph separator
- ፥፥፥፥ semicolon

Appendix VI:

Same sound Amharic character

Same sounded Amharic characters for normalization

Sound	Ha	Hu	Hi	Ha	He	H	Ho
Fidels	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
Sound	Se	Su	Si	Sa	Sie	S	So
Fidels	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
Sound	A	U	I	Aa	Ie	E	O
Fidels	አ	ኡ	ኢ	ኣ	ኤ	ኦ	ኦ
	ዐ	ዑ	ዒ	ዓ	ዤ	ዖ	ዘ
Sound	Tse	Tsu	Tsi	Tsa	Tsie	Ts	Tso
Fidels	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
	ፀ	ፁ	፺	፻	፼	፽	፾

Performance result of 70:30 training-testing split ratio

Performance metrics	Support Vector Machine (SVM)		Artificial Neural Network (ANN)		Naïve Bayes (NB)	
	Class label		Class label		Class label	
	non-jar	jar	non-jar	jar	non-jar	jar
Precision	91	98	93	93	96	91
Recall	98	91	92	94	90	97
F-measure	94	94	93	93	93	94
Accuracy	94.2		92.9		93.6	