



**Jimma University**  
**Jimma Institute of Technology**  
**School of Computing**

***AFAAN OROMO SENTENCE BASED PLAGIARISM DETECTION: A SEMANTIC SIMILARITY  
APPROACH***

**By: - GIZAW TADELE BEKELE**

**A Thesis Submitted to the Faculty of Computing in Partial Fulfillment for the Degree of  
Master of Science in Information Technology**

**October, 2019**  
**Jimma, Ethiopia**

**JIMMA UNIVERSITY**  
**JIMMA INSTITUTE OF TECHNOLOGY**  
**FACULTY OF COMPUTING**  
**GRADUATE PROGRAM IN INFORMATION TECHNOLOGY**

***AFAAN OROMO SENTENCE BASED PLAGIARISM DETECTION: A SEMANTIC SIMILARITY APPROACH***

***GIZAW TADELE BEKELE***

I hereby declare that this thesis on “**Afaan Oromo Sentence Based Plagiarism Detection: A Semantic Similarity Approach**” is my original work and has not been presented as a requirement for a degree in any other university, and that all sources of materials used for the thesis have been strongly referenced and acknowledged.

Name: Gizaw Tadele

Signature \_\_\_\_\_

Date: October 2019

**Approved by advisors:**

**Signature**

**date**

Advisor: Dr. Million Meshesha (PHD)



October 22, 2019

Co-Advisor: Admas Abteu (MSc)

\_\_\_\_\_

\_\_\_\_\_

**Approved by examining committee:**

**Signature**

**date**

1. \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

2. \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

3. \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

## ABSTRACT

In current day the availability of digital technology enables world community to communicate and exchange information easily. As a result of which, we are in the era of information overloading where various types of information is collected from different sources. As the amount of available digital information increases it is difficult to access information efficiently from different sources. To address this problem, machine leaning based NLP has a great contribution. In this work we focused on semantic based similarity measure for plagiarism detection from Afaan Oromo documents. To use the semantic approach, we built a sample dictionary for synonym terms representation. The study used LSI approach to decompose sentences into terms matrix for similarity calculation. We have collected 3 documents with 15 sentences, 14 sentences and 11 sentences. The documents are collected from different sources like two documents from Afaan Oromo published fiction and one document of personal bibliography from Afaan Oromo FBC. Preprocessing of text has been applied to the dataset. Java programming has been used to develop a prototype of the proposed model and SQL has been used to build sample dictionary.

The performance of the study work was tested on 10 sentences of suspicious query and 3 source documents of 275 key terms. The accuracy achieved in detecting plagiarism from suspicious query was 53.02 %.

The result gained was not high due to less dataset. In addition stemming and POS tagging has not been applied this work. The accuracy can be improved with big dataset, applying stemming and POS tagging will the recommendation for this study for future step.

**Key word: Afaan Oromo, Semantic similarity, Machine learning, LSI, Plagiarism detection**

## **DEDICATION**

I dedicate this work to my Mother **Asnakech Senbete** who dreams my success throughout her life.

## ACKNOWLEDGEMENT

First and for most I would like to thank God for his everlasting love, mercy and support throughout my life and during the work of this thesis. Then there a number of people who helped me on this work directly and indirectly. My special thanks go to **Dr. Million Meshesha** (PHD) who was my advisor and whose advice and support me starting from proposal shaping to this thesis and pass through all ordeals. Thanks for your patience, critical comment and moral encouragement for me on this thesis work. I would thanks **Admas Abtew** who was my co-advisor and who helped me on commenting proposal and he was with me until this work has been finalized.

Yusuf Hussein (Afaan Oromo dept. head, DU) you have great contribution for this work regarding to language concern.

My **mother Asnakech Senbete**, I dedicate this work for you with great love and pleasure because, you are the secret of this work success. I love you forever and may God bless your future life.

My wife **Mekdes Tadesse**, I will not forget challenge you faced during this thesis stage. I have great love and respect to you. I do not know how I can express your love. Both of my sons **Nimona** and **Moti**, I love you forever and I wish to you a bright future.

My brother **Chala** you have great contribution of moral for this thesis work because you are stayed with my families and treated them very well until I comeback.

I have special thanks to Fikadu Wayesa who was my class mate and he was with me at any challenge of this thesis work. Amanuel Aseffa, Zarihun Olana, Abuljebar Kedir, Alamisa Endebu, Magarsa Dasta, Tigist Tajebe and others whose your name was not written on this thesis, but you were with me directly or indirectly during this thesis have great thanks from me.

# TABLE OF CONTENT

Contents	
<b>DEDICATION</b> .....	iii
<b>ACKNOWLEDGEMENT</b> .....	iv
<b>TABLE OF CONTENT</b> .....	1
<b>LISTS OF FIGURES</b> .....	4
<b>LIST OF TABLES</b> .....	5
<b>LIST OF ACRONYMS</b> .....	6
<b>CHAPTER ONE</b> .....	7
<b>INTRODUCTION</b> .....	7
1.1. Background.....	7
1.2. Machine learning.....	8
1.3. Motivation.....	9
1.4. Statement of the Problem.....	10
1.5. Objective of the study.....	11
1.5.1. General objective.....	11
1.5.2. Specific objectives.....	11
1.6. Methodology of the study.....	12
1.6.1. Study design.....	12
1.6.2. Problem identification and motivation.....	12
1.6.3. Definition of the objectives for a solution.....	12
1.6.4. Data collection and analysis.....	12
1.6.5 Design and development.....	13
1.6.6. Demonstration.....	14
1.6.7. Evaluation.....	14
1.6.8. Communication.....	15
1.7. Scope and limitation of the study.....	15
1.8. Significance of the study.....	16
1.9. Thesis structure.....	16
<b>CHAPTER TWO</b> .....	17
<b>LITERATURE REVIEW</b> .....	17
2.1. Overview of Plagiarism.....	17

2.2. Plagiarism detection.....	18
2.2.1. Synonym .....	18
2.3. Approaches to plagiarism detection.....	19
2.4. Similarity measures.....	19
2.4.1. Semantic similarity .....	20
2.4.2. Syntactic similarity .....	20
2.4.3. Lexical similarity .....	21
2.4.4 Textual similarity .....	21
2.5. Methods and algorithms for similarity measures.....	21
2.5.1. Latent Semantic Indexing (LSI).....	21
2.5.2. Latent Semantic Analysis (LSA) .....	22
2.5.3. Latent Dirichlet Allocation (LDA).....	23
2.5.4. Vector Space Model (VSM) .....	23
2.5.5. Jaccard Similarity.....	24
2.6. Challenges in plagiarism detection .....	24
2.7. Evaluation techniques of sentences similarity semantically .....	25
2.8. Over view of Afaan Oromo language .....	25
2.8.1. Dialects .....	26
2.9. Related work .....	26
<b>CHAPTER THREE.....</b>	<b>29</b>
<b>METHODS AND TECHNIQUES.....</b>	<b>29</b>
3.1. Overview.....	29
3.2. The architecture of proposed model.....	29
3.2.1. Data collection .....	29
3.2.2. Text Preprocessing.....	30
3.2.3. Normalization .....	31
3.2.4. Tokenization .....	31
3.2.5. Stop word removal .....	32
3.2.7. Semantic similarity measure .....	33
3.2.8. Lexical similarity or matching .....	36
3.2.9. Semantic similarity or matching .....	37
3.2.10. Plagiarized content.....	38
<b>CHAPTER FOUR.....</b>	<b>39</b>

<b>EXPERIMENT AND EVALUATION</b> .....	39
4.1. Overview .....	39
4.2. Implementation tools .....	40
4.3. Experimental steps .....	42
4.4. Prediction process .....	46
4.5. Evaluation procedure .....	47
4.6. Experimental result .....	53
<b>CHAPTER FIVE</b> .....	59
<b>CONCLUSION AND FUTURE WORK</b> .....	59
5.1. Overview .....	59
5.2. Conclusion .....	59
5.3. Contribution of the work .....	60
5.4. Future work .....	61
<b>REFERENCES</b> .....	62
<b>APPENDIX</b> .....	66



## LISTS OF FIGURES

Figure	page
Figure 3.1. Architecture of the proposed approach.....	30
Figure 3.2. Afaan Oromo text Normalization Pseudocode.....	31
Figure 3.3. Term-document matrix [47] .....	34
Figure 3.4. Comparing lexical similarity algorithm Pseudocode .....	37
Figure 3.5. Comparing semantic similarity Pseudocode .....	38
Figure 3.6. Plagiarized content algorithm Pseudocode.....	39
Figure 4.1. Synonym representation in relational database .....	46
Figure 4.2. Proposed interface .....	53
Figure 4.3. Highlighted red color to show lexically plagiarized part .....	55
Figure 4. 4. Highlighted blue color to show semantically plagiarized part .....	56
Figure 4. 5. Highlighted part to indicate lexically plus semantically .....	57

## LIST OF TABLES

<b>Table</b>	<b>Page</b>
Table 3. 1 Sample stop word list.....	33
Table 3. 2 Sample synonym for Afaan Oromo language.....	35
Table 4.1 Result of manual work .....	49
Table 4 .1 Result of proposed model .....	50
Table 4.3 Evaluation result of proposed model .....	51

## LIST OF ACRONYMS

AOSSS	Afaan Oromo Semantic Sentences Similarity
BOW	Bag Of Words
FBC	Fana Broadcasting Corporation
FN	False Negative
FP	False Positive
IDF	Inverse Document Frequency
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
ML	Machine Learning
NLP	Natural Language Processing
POS	Part Of Speech
RDB	Relational Data Base
SQL	Structured Query Language
SSSAOPD	Semantic Sentence Similarity of AFAAN Oromo for Plagiarism Detection
STS	Semantic Textual Similarity
SVD	Singular Value Decomposition
TF	Term Frequency
TF-IDF	Term Frequency - Inverse Document Frequency
TN	True Negative
TP	True Positive
VSM	Vector Space Model

# CHAPTER ONE

## INTRODUCTION

### 1.1. Background

Natural Language Processing (NLP) is a sub-field of Artificial Intelligence that is focused on enabling computers to understand and process human languages to get computers closer to a human-level understanding of language [1][2]. Computers do not yet have the same intuitive understanding of natural language that humans do.

As a result, more research are expected to be done in NLP to enable computers communicate with human being. NLP embodies several techniques that can change the way people think, learn, and communicate with machine. One of these techniques is plagiarism detection. According to Merriam Webster's Dictionary [3], plagiarism is 'the act of using another person's words or ideas without giving credit to that person.'

Semantic plagiarism may take many forms including changing the structure of sentences (restructuring) and replacing words with their synonyms or rewriting (paraphrasing) [4], self-plagiarism that occurs when somebody submit his/her previous work, direct plagiarism that occurs word with word substitution without any quotation mark [5]. This work focus on content based semantic plagiarism detection from the Afaan Oromo text.

Plagiarism detection requires checking the similarity between two contents. Sentence similarity (or distance between sentences) is one of the central themes in proofing to what extent contents of documents are similar [6]. A similarity measure for plagiarism detection computes the degree of nearness between a pair of vectors or sentences. When sentences are treated as similar they are semantically close and describe similar concepts [7]. Hence, similarity can be used in the context of duplicate detection.

The similarity between sentences is a function of the angle between their vectors in the term vector space [1]. LSA (Latent Semantic Analysis) puts sentences together even if they don't have common words, but if the sentences share frequently co-occurring terms.

Corpus based methods assume that words with similar meaning often occur in similar contexts. Latent Semantic Analysis (LSA) represents words as compact vectors via singular value decomposition (SVD) on the corpus matrix, and reduced computational costs by training directly on the non-zero elements in corpus matrix. Vector-space model (VSM) is also used to detect semantic based similarities between two sentences [8].

Plagiarism detection with semantic similarity evaluates the similarity between concepts that are not lexicographically similar. The deep understanding of these concepts is necessary for computing semantic measures [9]. Similarity and relatedness measure can be applied to solve many problems in different applications.

The measure of similarity and relatedness can be extended to many types of entities, such as words, sentences, texts, concepts, or Ontologies depending on the requirement. Lexical semantics extracts semantic relations. Tasks such as document classification and clustering, machine translation, information retrieval, information recommendation and synonym extraction require precise measurement of semantic similarity between words [10][11].

Accurately measuring semantic similarity between words, sentences and documents present a significant challenge due to the complexity and ambiguity of natural language semantics.

## 1.2. Machine learning

Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions [12][13].

Samuel and Mitchell [13], defined Machine Learning as the field of study that gives computers the ability to learn without being explicitly programmed and computer program to learn from experience with respect to some class of tasks.

Machine learning enables analysis of massive quantities of data while it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly [12].

In today's modern technology combining machine learning with AI and cognitive technologies can make it more effective in processing large volumes of information.

Machine learning can be classified into supervised and unsupervised categories. In supervised machine learning training data includes both the input and the desired results where as in unsupervised machine learning the model is not provided with the correct results during the training and can be used to cluster the input data in classes on the basis of their statistical properties only [14].

In further steps machine learning can be applied in semantic similarity by training the machine (means computer) identify and understand patterns from training dataset. For this work computer learns the semantic meaning of terms from stored dictionary to predict the meaning for similarity measure in case of plagiarism detection from documents written in Afaan Oromo language.

### 1.3. Motivation

Sentence similarity measure has been conducted for so many languages, including English, Arabic, Indonesia, Russian and Chinese languages. However sentences similarity measuring for plagiarism detection has not been studied for most of Ethiopian languages in general and Afaan Oromo language in particular.

From the researches point of view the following field of Plagiarism can be a very motivating issues for measuring document similarity to plagiarism detection [8], such as idea plagiarism, copy-paste plagiarism, citation based plagiarism, paraphrasing plagiarism, cross-lingual plagiarism etc. In our country the existence of the issue of plagiarism of content semantically is reported in conference organized by Afaan Oromo department regarding to linguistic segmentation (qaaccessa afaanii), at Dilla University (DU), Dilla, Ethiopian on May, 2015 G.C.

Particularly, the report was presented based on two Afaan Oromo fictions written by different authors, with different title and in different duration of publication that were plagiarized from each other semantically. From the presentation the researcher observed how much it's difficult for the analyzer to manually measure the similarity of the given sentences with other sentences and given document with other documents.

Two major problems can be inferred, it's time taking, inconsistent and less accurate. So this study is inspired to explore, design a system and suggest a way to mitigate these two problems computationally.

#### 1.4. Statement of the Problem

Plagiarism of content or idea has been one of the big issues nowadays since most data are in digital form from different sources. Content of one sentence or document may be repeated in other sentence or document by missing semantic issue. Even the issue is observed and frequently raised in movie authoring and audio productions.

Particularly Afaan Oromo's two fictions are missing this concept and duplicated idea or concept exists currently. One of the examples from Afaan Oromo fiction that was written by Dhaba Wayesa called 'Godaannisa' and fiction that was written by Getachew Rabira called 'Ichima jaalalaa' has semantic similarity matching. The idea of both documents (fictions) was almost the same but the lexical structure of both documents (fictions) was almost different which means one was plagiarized from the other. It was an idea plagiarism since the concept of one document present in the other document without changing the terms and majorly representing other terms with synonyms.

Any document could be a collection of sentences and sentence could be constructed from words or terms. So before dealing with the document level semantic similarity it is better to analyze and measure semantic similarity of sentences which is the medium of words and documents. So the main target of our work in this research is to detect problem of content plagiarism lexically and semantically from collected documents.

When specified bloggers or reporters read news or post something what was new in current situation, then other individual or groups use the same concept posted by reporters already again as new idea in different time and different places [15].

International online plagiarism checkers are there in digital communication of our today's world. Those software has advantage and disadvantage to solve the problem of plagiarism detection. Our motive to model this framework was to compensate the disadvantage part of those issue for the language. Lack of security, freely unavailability of software and missing of synonyms extraction for similarity measures are some concepts observed as a gap.

Since there were no modeled design that can accept Afaan Oromo text for measuring similarity and lastly detect plagiarism based on lexical similarity and semantic similarity approach.

Not only plagiarism for fictions, news and other social media, the same thing may also happen for plagiarizing the idea of one thesis and article as a new concept in other place at different time. So, plagiarism detection is a technique to find out the theft of single message, thesis, article, scientific paper, literary works, source code and others [16].

Plagiarism detection systems (many of which are commercial based) are designed to detect word co-occurrences and light modifications, but are unable to detect severe semantic and structural alterations [17].

The main aim of this study is therefore to measure similarities of new provided query or sentences with collected documents written in Afaan Oromo language to detect plagiarism. To this end, the current study attempts to investigate and answer the following research questions.

- How to represent sentences for similarity measure?
- Which similarity measure is the best for sentences level plagiarism detection?
- To what extent the proposed approach works in detecting plagiarism?

## 1.5. Objective of the study

### 1.5.1. General objective

The general objective of the study is to build semantic based sentence similarity for plagiarism detection in Afaan Oromo text.

### 1.5.2. Specific objectives

To accomplish the general objective of the study the following specific objectives are formulated.

- To review related works so as to identify approaches and techniques used in the study
- To select suitable approach for sentence level semantic similarity measure
- To study Afaan Oromo text deeply and identify synonyms of the word
- To design architecture of system for the semantic based sentence similarity checking for plagiarism detection.
- To test and evaluate the performance of the designed prototype



## 1.6. Methodology of the study

To explore the problem stated in a research there must be a need to follow sequential procedure or techniques which is called research methodology. Methodology guides the scholars for evaluation, validity and reliability of the overall work. For this work we collected data from two Afaan Oromo fictions and one document from FBC Afaan Oromo.

### 1.6.1. Study design

For this study design science research methodology has been selected because it is an outcome based designing solution, which offers specific guidelines for evaluation and iteration within research procedures. The overall procedure and techniques occur in step by step from problem identification and motivation of semantic sentence similarity to communication stage. Design science research is typically applied to categories of artifacts including algorithms, human or computer interfaces, design methodologies (including process models) and languages [18].

### 1.6.2. Problem identification and motivation

Motivation is the moral inspiration to find solution for problem identified in specified area. Before that problem statement must be clearly stated by reviewing related work done to model the new architecture. Before we have insight of the area, others literature reviews are explored and work done related to the study has been reviewed.

### 1.6.3. Definition of the objectives for a solution

The objective of this work is to develop Afaan Oromo Sentence Based Plagiarism detection with semantic similarity approach. The model of the study developed for unsupervised text to plagiarism detection.

### 1.6.4. Data collection and analysis

In this stage, we collected the data based on the design science research procedure for process model. To collect the necessary data set for training and testing the system designed we collect document written in Afaan Oromo text from different sources, such as two different fictions of Afaan Oromo published in different time and one document from Afaan Oromo FBC.

In case of data collection sample terms must be organized means organizing of structured from unstructured text. In case of Afaan Oromo preparing structured sentences from unstructured is very vast task since it need strongly language expert for structure of language script, to identify more synonyms for single word because Afaan Oromo single word may have more synonyms which is case for plagiarism in the language.

In addition to that identifying the similar sentences those are going to be measured semantically from Afaan Oromo sentences needs additional technique or knowledge expert's skill for the language to clearly know more synonym terms or words. As far as his/her knowledge concerned Afaan Oromo language expert must know synonyms words spoken throughout the whole Oromia regional state and in other place where Afaan Oromo where spoken. The big problem was all synonym words are not known in all Oromia regional state. So the model we are going to design can solve the mentioned problem somehow.

Relational database or dictionary is used to solve the problem of semantic similarity by representing synonym terms together. In relational database all synonym terms are clustered into same group by having similar group and different lexical architecture by SQL server or MYSQL database tools.

#### 1.6.5 Design and development

Design and development stage clearly define the framework of implementation tools and environmental architecture. Edraw max 7.9 is used for designing architecture. Java Programming Language (NetBeans 8.2) is used to process and implement the LSI algorithm that process semantic similarity analysis purpose. Structured query language (SQL 2012) for grouping synonym term together for semantic similarity measure.

Since we didn't get organized softcopy of the document we change hardcopy into softcopy some content of the documents (fictions) for this work for analysis and evaluation purposes. So the corpus was prepared manually as a sample to evaluate the work. The collection of documents we used was encountered as training dataset for this study.

To detect plagiarism sentences semantic similarity is measured using Latent Semantic Indexing (LSI). LSI is to find and fit a useful model of the relationships between terms, sentences and documents.

LSI examines the words used in sentences and looks for their relationships with other words. A truncated singular value decomposition (SVD) is used to estimate the structure in word usage across documents [19]. Retrieval is then performed using the database of singular values and vectors obtained from the truncated SVD.

Semantic similarity is the complex task for unexplored language previously because there were no annotated data input. Since our work is done based on clustering or grouping synonym terms together by relational database, the approach we follow in this study is machine learning approach because the algorithm learn from the synonym based clustered or grouped terms (Relational database) to calculate the similarity of the sentences.

#### 1.6.6. Demonstration

The proposed model has been demonstrated with the dataset to analyze the performance and it's efficiencies as per stated research problem. The system has been demonstrated for user with its parameter to evaluate the model with evaluation method steps.

#### 1.6.7. Evaluation

It is the stage of comparing the objectives of proposed solution corresponding to actual observed results from the demonstration. Result can be evaluated via performance metrics and analysis techniques throughout the process. IR evaluation techniques such as recall, precision and F-score are considered for semantic sentence similarity measure. Precision is a fraction of retrieved sentences that are relevant, whereas recall is a fraction of relevant sentences that are retrieved. F-measures combines the harmonic mean of precision and recall. The overall performance of the model is evaluated manually with human judgment to evaluate this work with LSI algorithm on specified corpus.

To evaluate the performance of the proposed work, it has been proposed to use Afaan Oromo's sentences randomly that we are going to measure their similarity to identify whether semantically they are similar or not. The performance of the study could be measured from two perspectives and the summary would be generalized. Sample sentences we used for evaluation purpose was the testing dataset for this study work. Synsets are interlinked by means of conceptual-semantic and lexical relations. In other words synset represents a group of words, in which all words have a similar meaning.

The first perspective of performance was human judgement and the second was model evaluation which could be performed as human judgements to verify the model detect plagiarism from similarity measure encounters.

#### 1.6.8. Communication

The overall procedure and techniques of this work must be clear for other researchers and other relevant audiences. Even finding of the semantic sentences similarity for plagiarism detection must be presentable to others for literature and future works. The model of the plagiarism detection, finding of the study and procedures has been presented and understood by other scholars.

#### 1.7. Scope and limitation of the study

It is important to mention at the outset what is the scope of our work. This study attempts to develop a prototype plagiarism detection system at sentences level using semantic similarity measures. For experiment two documents are collected from Afaan Oromo published fictions and one document from Afaan Oromo FBC.

The approach of this work is for a given new query to compare all terms of in a query with all terms in the collected documents to detect how much it is plagiarized by all documents. The approach applied LSI as data representation and comparison model.

This work is limited to Afaan Oromo textual content to detect plagiarism. Plagiarism detection can be possible in image, video and audio but for this work they are out of scope. Similarity measure for Plagiarism detection approach for multi languages is also not the scope of this work.

The main limitation while processing the study is the absence of readily available annotated corpus and dictionary for semantic extraction during natural language processing from Afaan Oromo language.

### 1.8. Significance of the study

Similarity measure for plagiarism detection is an important tool in almost all NLP application areas. Plagiarism detection software serves as an important preprocessing tool for tasks such as information extraction, information retrieval and other text processing applications.

Semantic based sentences similarity measure modeler in case of plagiarism detection for Afaan Oromo can serve as an input for other works like topic modeling, text summarization, information retrieval and recommendation system. In addition semantic based sentences similarity measurement has great significance in semantic searches, plagiarism detection, and automatic technical surveys.

This work can be used as a reference for research that will be conducted for plagiarism detection in other local languages. This thesis work has been designated for sentences based semantic similarity measure for Afaan Oromo, with lexically matching and representing synonym words on dictionary for semantic extraction during matching time to get plagiarized parts of source documents suspicious query.

### 1.9. Thesis structure

This thesis contains five chapters. Chapter 2 discuss about literature review about similarity measure, overview about Afaan Oromo language and related work done in the area of similarity measure for plagiarism detection. Chapter 3 crucial part of this research that focus on design and architecture of the proposed model system. Chapter 4 detail about experimental procedure, evaluation and discussion of this research. Chapter 5 concerns conclusion of idea achieved from experimented proposed problem and show direction future work.

## **CHAPTER TWO**

### **LITERATURE REVIEW**

In this chapter, reviews of literatures that are relevant to this research have been done. Obviously, plagiarism is a problem or an issue that can affect knowledge sharing and innovation/invention, whether it was content plagiarism, idea plagiarism or any other type of plagiarisms.

This chapter presents first review of concepts related to plagiarism detection, what, why and how. The conceptual review is then followed by related works review where different works related to this study are presented.

#### **2.1. Overview of Plagiarism**

The Merriam Webster dictionary[3] defines that, the act of plagiarism is: "to steal and pass off the ideas or words of another as one's own", Plagiarism is simply taking other people's words and/or ideas, using them, and then without giving credit to the person who thought of them or originality, pretending that those words/ideas belong to others.

In an instructional setting, plagiarism occurs when a writer deliberately uses someone else's language, ideas, or other original (not common-knowledge) material without acknowledging its source [20]. The reason why People plagiarize are [21], not knowing any better, pressure, competition, lack of confidence, work perceived as too difficult, lack of consequences, lack of interest.

The consequence of plagiarism are [21], unethical act because it is intellectual theft. It shows disrespect for the rights of the original author, it tarnishes the Universities reputation, and diminishes the value of once qualification, it casts suspicion on the honest work of the students. As noted by Grace [21], there are five types of plagiarism observed, as discussed below.

Copy and Paste Plagiarism or Direct Plagiarism.

This type of plagiarism happens when somebody copy a sentence, phrase, or paragraph word by word, but do not quote the source for acknowledgement.

Word Switch Plagiarism.

This is a plagiarism observed in a situation where the writer rephrases a person's work and insert it into his/her/their own work without acknowledging its original source. If someone take a sentence from a source and change a few words without acknowledging the source, it is still plagiarism.

### Blending Plagiarism

In this type of plagiarism, the writer may mix words or ideas from an unacknowledged source with his/her/their own words or ideas. The other option is the writer may mix together words and ideas taken from uncited several sources into a single work. Blending plagiarism may happen also when an attempt is made to use together properly cited source with uncited one.

### Insufficient Acknowledgement.

This is a type of plagiarism which is frequently happening in writing. Because most writers are correctly cite other's source once, but continue to use the author's work without giving additional proper citation.

### Self-Plagiarism.

This type of plagiarism happens when the writer or scholar attempts to publish one paper or report multiple times. For instance, self-plagiarism happens when a student uses an assignment completed for one class to satisfy the assignment for a different class. Even if you modify a previous paper or assignment, you must get permission from your professor/ instructor and correctly cite your previous paper.

## 2.2. Plagiarism detection

Plagiarism detection is the process of locating instances of plagiarism within a work or sentences. To detect plagiarism of any form, it is essential to have broad knowledge of its possible forms and classes, and existence of various tools and systems for its detection. Based on impact or severity of damages, plagiarism may occur in an article or in any production in a number of ways [22].

### 2.2.1. Synonym

Synonym is a word or phrase that means exactly or nearly the same as another word or phrase in the same language. Words that are synonyms are said to be synonymous if different words have

the same meaning. For example, words begin, start, commence, and initiate are all synonyms of one another.

Synonymy is a lexical relation between word forms that could handle by WordNet or grouping by relational database. In WordNet the important relationship could be represented in meaning. WordNet take into consideration the semantic areas of the word so that there is not only a text matching for similarity but in addition looking for word meanings as well [23].

The impact of synonymy is that if a semantic consists of synonym word, then the other synonym of the word usually is not used in the same sentences. We have to use the same word in expressing same meaning [7].

### 2.3. Approaches to plagiarism detection

Hiremath and Otari [24] discussed that there are mainly two types of plagiarism that occurs most frequently. Textual plagiarism and source code plagiarism are the two frequently occurring plagiarism as per their discussion. There are many plagiarism detection approach and techniques such as [24].

- textual based plagiarism that delivers satisfying results if the plagiarized text is copied (copy and paste),
- citation based plagiarism that compares the occurrences of citations in order to identify similarities, and
- shape based plagiarism for flowchart that detecting flow chart figure plagiarism based on shape based image processing.

### 2.4. Similarity measures

Similarity measure is an idea of evaluation to know the result of how much two or more concepts are equal or not. Similarity measure can be considered for textual plagiarism like copy and paste of documents, sentences and even word. Similarity means that finding relevant meaning of the given text and computing the accuracy between them [25].

A similarity measure and evaluation can be used to calculate similarity between two documents, two queries and one document with one query. The main objective of similarity is to identify the extent to which the given query in one document is repeated in another document.



Similarity can be measured at different levels, such as semantic similarity, syntactic similarity and lexical similarity [5].

#### 2.4.1. Semantic similarity

Semantic similarity is used to identify the extent to which two or more terms or sentences are conceptually similar but not necessarily lexically similar [26].

Basically, semantic similarity is computed by mapping terms from different sentences, documents and by measuring their relationships in the terms of that sentences and documents respectively. In linguistic Semantic similarity measure is an important concept to identify plagiarism content of sentences or documents. Plagiarized content of sentences or documents which cannot handled by lexical similarity can be detected by semantic similarity measure with learning the synonyms of given terms from dictionary or WordNet to detect plagiarism.

Measures of relatedness or similarity are used in a variety of applications, such as information retrieval, automatic indexing, word sense disambiguation and automatic text correction [7]. These terms however, are not identical. Semantic similarity is a special case of relatedness and takes into consideration only hyponymy/hypernymy relations.

The relatedness measures may use a combination of the relationships existing between words depending on the context or their importance. To illustrate the difference between similarity and relatedness, Reznik [2] provides the widely used example of car and gasoline. These terms are not very similar; they have only few features in common. Semantic similarity metrics determine the extent of the similarity of both concepts of the sentences and documents semantically.

#### 2.4.2. Syntactic similarity

Syntactic similarity is an important activity in the area of high field of text, sentences, documents, data mining, natural language processing and information retrieval [25]. Syntactic similarity is how similar are two words or terms with respect to their syntactic function or role.

In the field of data mining syntactic similarity is exploited in application like cleansing data for mining and warehousing, duplicate detection and mining knowledge from text [25].

If syntactic similarity of words or terms are measured just we could mis-semantic case because synonym terms may exist in other sentences which are lexicographically different but have same meaning.

#### 2.4.3. Lexical similarity

In linguistics, lexical similarity is a measure of the degree to which the word sets of two given languages are similar. Lexical similarity of content may exist in sentences and documents.

When physical existences of one query's term appear lexically in other sentences and documents lexical similarity of the query's terms and terms of other sentences or documents can 100 percent which is direct copy. Generally a lexical similarity of 100% mean a total overlap between the sentences or documents, whereas 0% means there are no common words.

#### 2.4.4 Textual similarity

The main goal of text similarity is to compute how two piece of texts are close to each other. The closeness of text can be surface and meaning. Surface closeness is lexical similarity and meaning based is semantic similarity of texts.

Text similarity measures play an increasingly important role in text related research and applications in tasks such as information retrieval, text classification, document clustering, topic detection, topic tracking, questions generation, question answering, essay scoring, short answer scoring, machine translation, text summarization and others [27].

### 2.5. Methods and algorithms for similarity measures

With a great motive, NLP concept can be integrated with machine learning algorithms. There are methods used to measure sentences similarity semantically to detect plagiarism, as discussed below [12].

#### 2.5.1. Latent Semantic Indexing (LSI)

It is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts

contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings.

A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts.

LSI is used to find and fit a useful model of the relationships between terms and sentences and examines the words used in a sentences and looks for their relationships with other words.

The main goal of LSI is to enhance the accuracy of information retrieval. LSI use a technique called singular value decomposition (SDV) to extract unstructured data within documents and identify relationships between the concepts contained therein [14], it finds the hidden (latent) relationships between words (semantics) in order to improve information understanding (indexing).

So the SVD concept is decompose the documents and queries into terms and compare the relationships of both documents and queries terms matrix. In matrix calculations single term of query will compare with each terms of documents until all terms are compared lexically and semantically to get the result of similarity measure. The Singular Value Decomposition is a highlight of linear algebra. A is any m by n matrix, square or rectangular. There are two vectors U and V to decompose documents into rows and column. Those v's and u's account for the row space and column space of A.

$$A = U\Sigma V^T \dots\dots\dots (2.1)$$

Where U is an m×m orthogonal matrix<sup>1</sup> whose columns are the eigenvectors of AA<sup>T</sup>, V is an n×n orthogonal matrix whose columns are the eigenvectors of A<sup>T</sup>A, Σ is an m × n diagonal matrix of the form [28]. Where ‘m’ represent terms and ‘n’ represent documents.

### 2.5.2. Latent Semantic Analysis (LSA)

Almost it is the same with latent semantic indexing (LSI). It is the technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of sentences and the terms they contain by producing a set of concepts related to the sentences and terms.

To analyze word meaning LSA produces measures of word-word, word-passage and passage-passage relations that are well correlated with several human cognitive phenomena involving association or semantic similarity [13].

LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of rows while preserving the similarity structure among columns.

### 2.5.3. Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data such as text corpora [29]. LDA model enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments [29][30]. LDA is a Bayesian network statistical approach that is effective on building latent topics by relating words contextually for huge documents [18].

### 2.5.4. Vector Space Model (VSM)

The similarity between two documents or sentences are computed according to the Vector Space Model (VSM) [31], as the cosine of the inner product between their documents or sentences vectors using the following formula [32].

$$Sim(q, d) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2} \sqrt{\sum_i d_i^2}} \dots\dots\dots (2.2)$$

Where  $q_i$  and  $d_i$  are the weights in the two vector representations. Given a query, all sentences are ranked according to their similarity with the query. This model is also known as the bag of words model for sentences retrieval. Vector Space Model (VSM) supposing that the words are independent, and each sentences is expressed in a space vector, each word is a dimension of the space vector, that simplify the complexity relationship of the words and convert the computing of the sentences similarity into the computing of the angle between vectors[33][34].

The lack of common terms in two sentences does not necessarily mean that both sentences are unrelated. Semantically similar concepts may be expressed in different words in the sentences and the queries, and direct comparison by word-based VSM is not effective. For example, VSM will not recognize synonyms or semantically similar terms.

### 2.5.5. Jaccard Similarity

The Jaccard similarity index (sometimes called the Jaccard similarity *coefficient*) compares members for two sets or sentences to see which members are shared and which are distinct.

It's a measure of similarity for the two sets of data, with a range from 0% to 100%. The higher the percentage, the more similar the two sets or sentences [35].

Although it's easy to interpret, it is extremely sensitive to small samples sizes and may give erroneous results, especially with very small samples or data sets with missing observations. To calculate Jaccard index the following measurement will takes place [35].

1. Count the number of members or terms which are shared between both sentences.
2. Count the total number of members or terms in both sentences (shared and un-shared).
3. Divide the number of shared members (1) by the total number of members (2).
4. Multiply the number found in (3) by 100.

Jaccard similarity can be calculated by the following formula for sentence similarity [36].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \dots\dots\dots (2.3)$$

Where A is number of member of set i or sentence i and B is number of member in set j or sentence j. Jaccard similarity result is divide intersection (commonly shared terms) of both sets and sentences by union of both sets or sentences.

### 2.6. Challenges in plagiarism detection

Plagiarism detection for the texts that contain significant changes in syntax and in meaning but mostly inadequate and inefficient [37].

Its representation has biggest challenge in the detection of these changes of meaning with synonym terms, because it requires analysis of texts that carry similar meanings and making a decision whether there is a plagiarism or not.

Semantic similarity is a confidence score that reflects the semantic relation between the meanings of two terms [38], it is difficult to gain a high accuracy score because the exact semantic meanings are completely understood only in a particular context.

Paraphrase plagiarism is one of the difficult challenges facing plagiarism detection systems. Paraphrasing occur when texts are lexically or syntactically altered to look different, but retain their original meaning [39].

Most plagiarism detection systems (many of which are commercial based) are designed to detect word co-occurrences and light modifications, but are unable to detect severe semantic and structural alterations such as what is seen in many academic documents [23]. Hence many paraphrase plagiarism cases go undetected.

## 2.7. Evaluation techniques of sentences similarity semantically

WordNet [40] is an on-line lexical reference system developed at Princeton University. WordNet attempts to model the lexical knowledge of a native speaker of English. WordNet can also be seen as an ontology for natural language terms. It contains around 100,000 terms, organized into taxonomic hierarchies [26], Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synset).

Relational database is also used to solve the problem of semantic similarity by representing synonym terms in the same group together. So relational database is the approach we across through to represent and evaluate semantic similarity by considering synonym terms in addition to lexical similarity.

## 2.8. Over view of Afaan Oromo language

Afaan Oromo is a mother tongue for Oromo. Currently Afaan Oromo is the official language of the regional state of Oromia (the largest regional state in Ethiopia) being used as a working language in offices, educational language for all non-language subjects in junior-secondary schools (1-8 grades).

The Oromo people constitute the single largest ethnic group in Ethiopia, where the Oromia region contains a third of Ethiopia's land area and population [41][42]. The Oromo language, also known as Afaan Oromo, which is spoken as a first language by 87% of Oromia's 27 million people. Many others (as yet unquantified) speak it as a second language.

Oromo (Afan Oromo / ኣረቓኛ) is a Cushitic language spoken by about 30 million people in Ethiopia, Kenya, Somalia and Egypt. It is the third largest language in Africa. The Oromo people are the largest ethnic group in Ethiopia and account for more than 40% of the population.

### 2.8.1. Dialects

The main dialects of Oromo are Wellega (spoken in the West Wellega, East Wellega, Illubabor, and Jima zones), Tulama (in the North, West, and East Shewa zones), Wello (in Northern Shewa and Southern Amhara), Arsi (in the Arsi and Bale zones), Harar (in the West and East Harerge Zones), and Borena (in the southern-most zone by the same name) [42].

This classification scheme is general, as there is no official division of Oromo dialects, and many dialects go by multiple names (Wellega is also called Mecha).

Additionally, more isolated dialects are spoken by small Oromo communities that remain in eastern Amhara and southern Tigray. The major differences in dialects are in the form of pronouns, certain verb conjugations, and colloquial lexicon. For example 'She' in Wellega Oromo is “isheen”, while other regions in Oromo would use “isiin”, “ishiin”, or “iseen”.

Those all pronouns are similar or synonym but lexically they are different due to dialects of the Afaan Oromo language in different zone of Oromia region.

### 2.9. Related work

Due to the increase of web based information and number of internet users', it is difficult to find the relevant documents for users to particular needs. In semantic similarity measure LSI or LSA was the popular model identified [8].

Man Yan Miranda [43] suggested using a semantic-based approach for plagiarism detection, by combining an information retrieval model based on tf-idf with latent semantic indexing (LSI) and they identified as bag-of-words approach at the document level can represent the documents better as the feature is not limited by the sentence boundaries.

They have discussed the way and direction to detect plagiarism by textual similarity measure and contribute as machine learning bring benefits to the plagiarism detection framework.

They have been discussed as N-gram string matching was the great parameter based on suspicious text and source documents. Means as documents size increased n-gram size must be increased to get plagiarized part of source documents. An n-gram represents n number of consecutive words. During the experimental evaluation they have achieved different accuracy value as their experimental results. For example overall accuracy of 93.7% with small size of corpus, 84.2 % with four-class classification of overall accuracy.

Lucia D. Krisnawati [44] proposed LSA algorithm to detect plagiarism with similarity measure for Indonesia texts. As per author's motivations each measure takes account on different aspects of object properties, they will result in different values, even if they are applied to the same objects for machine. Semantic concept included in the work used WordNet Bahasa was a WordNet version for Malay language which covers Indonesian and Malaysian.

They have identified as the granularity of n-grams plays a great role in increasing the plagiarism detection accuracy. They have achieved 100 % with 7-gram of their corpus whereas the lowest detection rate, 80 % with 4-gram of their corpus.

Recently one year ago Belyy A. V. and Dubova M. A [45] have been proposed a new approach for advanced plagiarism detection in Russian language.

The study was focused on sentence similarity measure for plagiarism detection for the Russian language. They have used supervised sentence embedding to identify semantic case in their work.

Means if sequential sentences  $u$  and  $v$  are semantically related, then the angle between vectors  $f(u)$  and  $f(v)$  is close to 0 and cosine distance of  $f(u)$  and  $f(v)$  is close to 1.

King Abdulaziz [46] and Khalid Shams [16] has introduced as latent semantic indexing (LSI) or latent semantic analysis (LSA) is a technique in natural language processing to detect plagiarism. Khalid Shams have not test their program on a lot of data and they cannot tell the accuracy of their work. They are promised as they are going to release a prototype very soon including words from the WordNet to enhance the accuracy level of their work.



They explored as few lexicographers argue that there cannot be a synonym for any word because every word is different by its phonetic, origin and uses.

But they realized as this arguments was not considered because plagiarism has done widely by using synonym. They have kept synonym terms on database as wordlist for semantic extraction during matching.

To enhance the performance of IR system Anita R., Subalalitha C. N., Abhilash Dorle and Karthick Venkatesh [47] have been applied by combining semantic search using latent semantic indexing (LSI) and WordNet. They have been compared the result of term based information retrieval and LSI plus WordNet based information retrieval system in their work. They have been keeps the steps of LSI concept to decompose document into matrix by using SVD to decompose into term-documents algebraic techniques and WordNet to overcome the problem of synonym. They observed from their evaluation result as LSI and WordNet based information system has high performance than term based information retrieval system. The maxim accuracy they have got from LSI and WordNet based was 96 % whereas 90% from term based information retrieval system.

The aim of this study is to develop sentence based semantic similarity for plagiarism detection for Afaan Oromo. There was no model designed of semantic similarity for plagiarism detection for Afaan Oromo yet. The model we designed for this study can serve as base or input for other researcher's those who are interesting in the area of information retrieval, text summarization and recommendation system.

## CHAPTER THREE

### METHODS AND TECHNIQUES

#### 3.1. Overview

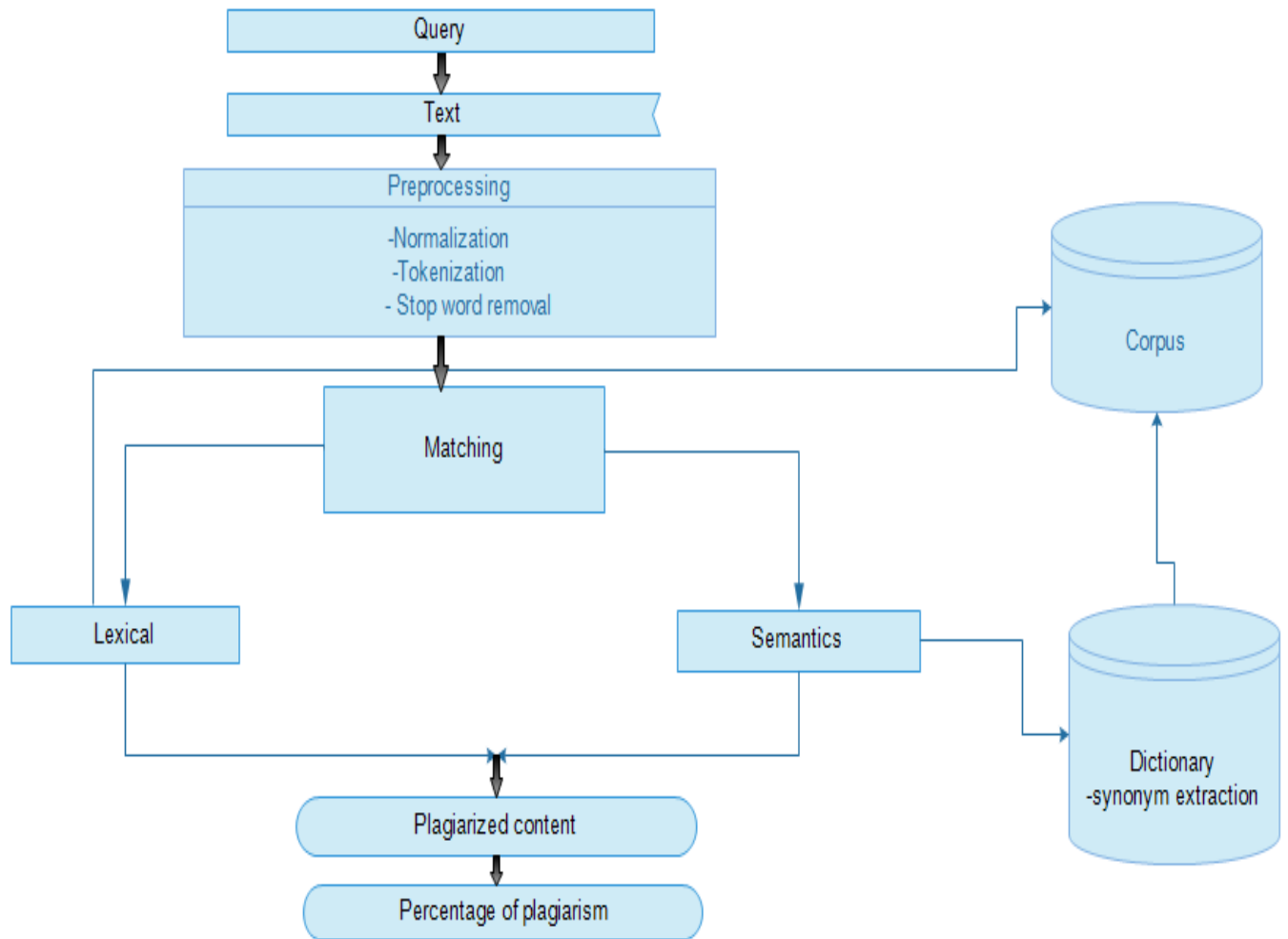
The aim of this study is to design plagiarism detection model for Afaan Oromo documents. To this end, we come up with architecture of the system, based on which different techniques and methods are identified that are followed for developing semantic sentences similarity for plagiarism detection with LSI algorithm and customize the system. Collecting sample documents we used from Afaan Oromo, implementation of the system to measure sentences similarity for plagiarism detection and evaluate the output of the systems are discussed.

#### 3.2. The architecture of proposed model

Figure 3.1 shows the proposed architecture in this study of semantic sentence similarity for plagiarism detection. The model explore the proposed system from feeding plain text of query to decision point of similarity at the end. The terms normalization, tokenization, stop word removal are performed on both query's terms and documents collected to decide whether they similar or not. There is sample dictionary for measuring semantic similarity.

##### 3.2.1. Data collection

We have collected three documents as training data with total of 40 sentences and 10 sentences for testing as testing data. Two documents are collected from manually published Afaan Oromo fiction of different title. One document was collected from Afaan Oromo FBC which concerned about bibliographic history of one women. We included the third document into our collected documents simply to analyze similarity measures (lexical and semantic) of different documents has been almost nearest to zero unless few terms are randomly found in the documents. But the first two documents are suspected documents to be plagiarized when testing sentences could be from Doc0 or Doc1.



*Figure 3.1. Architecture of the proposed approach*

### 3.2.2. Text Preprocessing

The preprocessing subsystem includes stop-word removal, stemming and parsing (breaking the input document in to a collection of sentences). We are collecting the documents from Afaan Oromo published fictions (Godaannisa and Ichima jaalalaa) manually. Generally Preprocessing text is called tokenization or text normalization which pass through the LSI concept for plagiarism detection direction.

### 3.2.3. Normalization

Normalizing text before storing or processing is used for separation of concerns, because input is guaranteed to be consistent before operations are performed on it. In this stage all words or terms are converted to lowercase.

Number (1, 2, I, II) punctuation mark or symbols (? : " ! | ? @ # \* ~ \$ % ^ & ( ) { } <> [ ] \_ + = - , " ... \ ; - \_ + £) are removed from query and documents before any task of NLP can be incorporated on texts.

We applied all necessary normalization step to our work. So number, symbols, any punctuation mark will not considered in our work to measure similarity for plagiarism detection.

Prototype for normalization

```
For each document collection C
  For each document D
    Tokenize into unigram term (token)
    Token. Lowercase( )
    If token t tokenized
      t.getText ().replaceAll ("(^a-zA-Z) // to remove symbols, any special character etc.
    End
  End
End
```

Figure 3.2. Afaan Oromo text Normalization Pseudocode

### 3.2.4. Tokenization

Tokenization is splitting into stream of tokens by using java code or divides a text into tokens, which are fragments selected as useful units for semantic parsing. Tokenization process is an integral part of IR systems, involves pre-processing of given sentences and generates respective tokens. In tokenization techniques count of token were used to establish a value “Word Count or Token Count” which can be used as indexing/ranking process to identify tokens independently.

Tokenizing the sentence or document into tokens simplifies the task of information extraction, text summarization, information retrieval and similarity measure.

For this study tokenization of sentences provided as a query and tokenization of documents stored into separate tokens are used for similarity measure of both query and documents to detect plagiarism.

Generally tokenization in similarity measure is used to form the matrix to compare single token (uni-gram) of provided query with single token (uni-gram) of documents one by one until each specified tokens are compared independently.

### 3.2.5. Stop word removal

It is the process of removing highly frequent term from sentences. Sometimes, some extremely common words which would appear to be of little value in helping select sentences matching a user need are excluded from the vocabulary entirely. These words are called stop words.

The general purpose for determining a stop list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a stop list.

Stop word list are a list of words that should not be stemmed as they are non-content bearing words. Even POS tagging is not necessary for stop word list in case of plagiarism detection since they are not considered as a parameter for similarity measure.

As there were stop words in English language like “the”, “since”, “and” there are stop words in Afaan Oromo those add noise on documents rather than contribute description on documents.

On stop word removal procedure Fiseha and Debela [48][49] respectively, describe that as to remove stop words from the input document after accepting list of words from tokenize module and removes stop words from segmented sentence by finding the match with list of stop word stored in stop word knowledge base.

He identified some stop words of Afaan Oromo in his work like waan (because), garuu (but), sun (that), utuu (if), akka (such as, like) as sample stop word in his study.

The techniques to prepare stop words list Eyob [50] identifying stop word is building list of stop words manually containing set of content.

After all stop words are removed from document collected the left terms are key terms those are represent the respective documents. So all key words are indexed or stored as bag of words (BOW) to be retrieved for comparison purpose later on for this study.

*Table 3. 1 Sample stop word list*

Number	Words
1	Sun
2	isaan
3	Ol
4	Yoo
5	Fi
6	kee
7	kun
8	koo
9	As
10	garuu

Table 3.1 shows stop word must be removed from query that going to extract the stored documents and documents stored and going to extracted.

### 3.2.7. Semantic similarity measure

In this study to detect plagiarism sentence-level similarity measure approach is used including semantic similarity measure. To accomplish this work Latent Semantic Indexing (LSI) is applied as algorithm of semantic similarity measure for plagiarism detection. LSI is to find and fit a useful model of the relationships between terms and sentences. LSI examines the words used in a sentences and looks for their relationships with other words.

Thus LSI examines the words used in a sentences and looks for their relationships with other words whereas LDA can be used in the semantic analysis of long documents with complex mathematics.

LSI is the application of SVD which is algebraic mathematical expression of matrix in terms of query to terms of document for similarity measure. The matrix is terms of query (t) to terms of documents (d). Using LSI in sentences based semantic similarity has two advantages [51].

The first one is reducing the original  $t \times d$  matrix into  $k \times d$  matrix  $S_k \cdot D_k$ , where  $k$  is orders of magnitude smaller than  $t$  and the second one is due to the solid mathematical background of the used transformation, is that it can also be viewed as a noise reduction process.

A truncated singular value decomposition (SVD) is used to estimate the structure in word usage across sentences [19]. Retrieval is then performed using the database of singular values and vectors obtained from the truncated SVD. SVD is an algebraic mathematical used to decompose documents and queries in column and rows matrix for similarity calculation [52]. Column is represent documents and rows represent unique terms (query). The decomposition of document for similarity calculation in term-documents matrix could be obtained due to SVD algebraic mathematical calculation.

$$A = U \Sigma V^T \dots\dots\dots (3.1)$$

Whereas  $U$  and  $V$  are orthogonal matrices and  $\Sigma$  diagonal matrix.

Then preprocessing has been applied to suspicious query and source documents. The term-documents matrix would be applied to identify relevant documents.

	D1	D2	D3	D4	D5	D6	D7
binary	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
computer	0.0198	0.0000	0.0198	0.0198	0.0000	0.0000	0.0000
computer system	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
engineering	0.1138	0.0000	0.1138	0.1138	0.0000	0.0000	0.0000
eps	0.1733	0.0000	0.1733	0.1733	0.0000	0.0000	0.0000
generation	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
graph	0.0000	0.0559	0.0000	0.0000	0.0559	0.0559	0.0559
human	0.1336	0.0000	0.1336	0.1336	0.0000	0.0000	0.0000
interface	0.0198	0.0000	0.0198	0.0198	0.0000	0.0000	0.0000
intersection	0.0000	0.0105	0.0000	0.0000	0.0105	0.0105	0.0105
machine	0.0198	0.0000	0.0198	0.0198	0.0000	0.0000	0.0000
management	0.0595	0.0000	0.0595	0.0595	0.0000	0.0000	0.0000
minors	0.0000	0.0454	0.0000	0.0000	0.0454	0.0454	0.0454
opinion	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
ordered	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
random	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
response	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
survey	0.0000	0.1859	0.0000	0.0000	0.1859	0.1859	0.1859
system	0.2871	0.0000	0.2871	0.2871	0.0000	0.0000	0.0000
testing	0.1138	0.0000	0.1138	0.1138	0.0000	0.0000	0.0000
time	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
user	0.0000	0.1405	0.0000	0.0000	0.1405	0.1405	0.1405
user interface	0.0595	0.0000	0.0595	0.0595	0.0000	0.0000	0.0000

Figure 3.3. Term-document matrix [47]

As we can observed from figure 3.3 LSI has been used application of SVD to decompose document into terms and compare specified term with documents term. As indicated on figure 3.3 term-document management used to identify relevant document easily.

Table 3.2 Sample synonym for Afaan Oromo language

Word	Synonyms
gaarii/good	dansaa, baroo, hosee, mishaa
baay'ee/ a lot	hedduu, danuu, bacaa
ishee/ she	isee, isii, ishii
ariitiin/ quickly	saffisaan, daddaffiin, hatattamaan
soba/ false	kijiba, dhara
rooba/rain	Bokkaa
hiyyeesa/ poor	Dhabaa
qoricha/ medicine	dawaa, qorsa
qorra/ cold	Diilalla
rifeensa/ hair	Dabbasaa
kabaja/ respect	ulfina, tabaroo
waraabessa/ hyena	hobolaa, gulloo
asi/ here	addana
misiraachoo/ amazing	aaga
Citaa/ grass	Gaalala, marga

Table 3.2 contains sample Afaan Oromo word that contain one and more than one synonyms

The other phase is sentences representation process after preprocessing or normalization to represent synonyms terms by using grouping relational database. We have prepared or built sample grouped terms to represent Afaan Oromo synonyms words with corresponding meaning for all terms after preprocessing those have synonyms. Means representing all synonyms terms together as semantically they have the same meaning throughout the documents. This means generally we prepared sample dictionary for synonym representation which helps the user for semantic handling managements.

For example represent Baayyee/Danuu/Hedduu/Bacaa together as they have same meaning semantically but lexically they are different.



Since our approach will be machine learning this synset representation system support our work as the algorithm learn from the sample dictionary whether the retrieved texts are from the same synset or not to decide whether they are similar or not semantically as result of similarity measure. Because all synonym and related terms or words are represented with their respective group.

The next phase is retrieving independent terms of sentences from entire collection of represented documents that has all terms of each sentences. Then there was LSI algorithm concept that process and analyzes the similarity of both query and documents by learning from clustered relational database whether semantically they similar or not. There is lexical similarity in addition to semantic case to improve the similarity measure of this study.

### 3.2.8. Lexical similarity or matching

Lexical similarity is when two sets of different sources are equal. Means when both sets are overlap each other physically or copy and paste similarity. The method of comparison of sequences of lexically chained, and similarity is computed by using LSI on the lexical chains and the term frequency-inverse document frequency (tf-idf) of weighted keywords.

As dictated in his work Chong [43], lexical changes involve the addition, deletion or replacement of words in the text and detection would require the analysis of lexical information throughout the text.

For example if one term of query found physically in collected documents

Then the term in document is highlighted with red color and plagiarism ration is calculated as 1 out of all terms of specific document specifically to identify plagiarized part.

Prototype to compare lexical similarity

```
For each document in collection C
  For each term in document D
    If document term equals with query term // physically present in D
      Count plagiarism term
    End
  End
End
```

*Figure 3.4. Comparing lexical similarity algorithm Pseudocode*

### 3.2.9. Semantic similarity or matching

Measuring Semantic Textual Similarity between words or terms, sentences, paragraph and document plays an important task in information technology and computational linguistic due to information overload in today's digital world. Language is the founder or source of information as input for the technology to accomplish problem proposed particularly with machine.

Like that of lexical similarity semantic similarity deals with determining how similar two pieces of texts are. But semantic similarity is meaning based (synonym) rather than physical availability of terms in both sources. Semantic matching or similarity happens when terms of query available has synonym terms in dictionary.

Prototype to compare semantic similarity

```
For each document in collection C
  For each term in document D
    If document term not equals with query term
      else
    For each synonym terms in dictionary Dict // for semantic extraction
      If D term equals with query term
        Count plagiarism terms (catch plagiarism)
      End
    End
  End
End
```

Figure 3.5. Comparing semantic similarity Pseudocode

3.2.10. Plagiarized content

Plagiarizing is the crucial case in information extraction and information retrieval because content can be plagiarized as image, audio, video and text. In this study we consider textual plagiarism detection. Plagiarized content of texts is word or terms taken from others sources word, sentences, paragraph and document. Copying contents from somebody’s work is dishonest activity in anywhere. Texts similarities can be measured in different features to detect plagiarism very well from sentences or documents.

The features are lexical similarity, syntactic similarity and semantic similarity of text. We have used both lexical similarity for physical presence of both terms (copy and paste) and semantic similarity for matching synonym (meaning based) terms of both sources (query and document collected) in study. For our case plagiarized content is the summary result of both lexical similarity measure and semantic similarity measure for plagiarism detection.

Prototype for plagiarized content

```

for each document in collection C
    For each term in document D
        if document term equals with query term
            count plagiarism term
        else
            For each synonym terms in dictionary Dict // for semantic extraction
                If document term equal with query term
                    Count plagiarized term
            End
    End
End

```

*Figure 3.6. Plagiarized content algorithm Pseudocode*

## CHAPTER FOUR

### EXPERIMENT AND EVALUATION

#### 4.1. Overview

This chapter contain the experiment and evaluation conducted for this study work means similarity measure for plagiarism detection approach. The approach of the study has been across with the

concept of LSI algorithm to measure similarity of sentences. To show direction of plagiarism detection proposed solution must be at least partially simulated rather it become suspended wish.

## 4.2. Implementation tools

For implementation development of tool used in LSI model for this study is java NetBeans version 8.2 and SQL server 2012 as backend to represent synonym terms in relational database. We have select java NetBeans 8.2 because java has more sophisticated library and package like hashmap, array List and soon those can assist to implement similarity measure for plagiarism detection.

We have used computer dell laptop with one Terabyte (1 Terabyte) of hard disk (HDD), 500 giga bite (500 GB) of random access memory (RAM), window 10 operating system with 64 bits for implementation of proposed model.

There is no officially annotated, authorized and publicly available tagged Afaan Oromo text for work of NLP. For this study we are collecting two documents from Afaan Oromo published fictions manually and one document from Afaan Oromo FBC. The corpus has 40 sentences. Each sentences are tokenized into 357 tokens and of those tokens 275 tokens are unique keys those represent all three collected documents. Both fictions are written by different authors with different title and in different time.

“Godaannisa” is one fiction title and “Ichima jaalalaa” is the second fiction title. “Godaannisa” what is known as “scar”. The scar can be comes from different background of problem. “Ichima jaalalaa” in English is “scar of love”.

According to this context the cause of pain is love. So Godaannisa is more general than Ichima jaalalaa which more specific at concept level. But at concept level both titles are semantically similar. We can generalize as Godaannisa fiction and Ichima jaalalaa fiction has semantic similarity at title level.

For semantic similarity handling we collect seventy eight (78) synonym words from different domain of Afaan Oromo Daily communication.

We collect those words randomly for our sample work from more than seven (7) different Oromia Zone like Bale zone, Arsi zone, Borena zone, Guji zone, Wallaga zone, Shewa zone, Jimma zone and Harerge zone.

For example wild animal in English “Hyena” is named in Afaan Oromo as “hobolaa” in Bale and Arsi zone but in all Shoa zone and Wallaga zone it named as “Waraabessa”.

The word “False” in English is named in Afaan Oromo as “dhara” in Harerge and Bale zone. But it is known as “kijiba” in Arsi zone, East Shewa zone and in some part of Bale zone woreda. The same word is communicated “soba” in Wallaga zone, total Shewa zone except some part of east Shewa, Jimma zone and Illubabor zone.

The English word “a lot” in Afaan Oromo communicated as “bacia” in Guji zone and some part of Borena zone, but “hedduu” in Harerge zone, Bale, Arsi and East Shewa zone. The same word is used with the same meaning as “baayyee” in Wallaga zone, Shewa zone and central part of the region. Jimma zone, Illubabor zone and some part of Wallaga are use the same word with the same meaning as “danuu”.

Each terms was first tokenized according to tokenization principle consistently. By using java NetBeans tools we put all terms on hashMap (dictionary) and on array list for comparison. HashMap is a Map based collection class that is used for storing Key and value pairs, it is denoted as Hashmap<Key, Value> or HashMap<K, V>. It does not sort the stored keys and Values. It must need to import java.util.HashMap or its super class in order to use the HashMap class and methods.

From stored list the comparison for similarity will takes place between all words or terms of query provided with all words or terms sentences of document collection stored on array string.

Means all words of query must be compared with all words of document collection correspondingly for similarity measuring.

In the process of word matching the proposed algorithm compare both words from query provided and document collected based on two word features. The first feature is physical or lexical matching of both words from query and document stored.

The second feature is synonym matching of word from query with word of document collected for semantic similarity. This information can be retrieved from relational database that contain synonym term representation with their corresponding group.

From the above prototype two lists those are used to handle terms of query and terms of document as a bag of words (BOW) are defined as  $l_1$  and  $l_2$ . We used java Array List function to store terms temporarily for further process of adding into dictionaries.

We have define two dictionary as  $dict_1$  and  $dict_2$  to collect terms or words from query provided and document collected. For comparison purpose whether lexically or semantically the terms from both query and documents are cross checked in the dictionaries. The above sample prototype is only matrix of similarity of terms from query to documents. But to be sure about plagiarism between two sentences bi-similarity matrix is more average and better.

The above sample prototype implies that for comparison if the term from query is not available in document collected lexically or physically it must be checked from the groups of relational database for semantic similarity.

### 4.3. Experimental steps

The experimental procedure of the model is carried out step by step starting from accepting plain text as query and go further up to decision of whether the content of query provided and collected documents are plagiarized or not. Java programming language has been used for model implementation and SQL server for synonym representation for relational database (sample dictionary).

These synonym terms are equal in degree of similarity in Afaan Oromo language. So if calculate similarity of both sentences lexicographically only, the similarity of both sentence is very less or plagiarism can be happen easily.

If calculate similarity of both sentence semantically only, the similarity of both sentence is less again or plagiarism can happen in the language. In natural language processing (NLP) context to measure similarity of two sentences applying lexical similarity plus semantic similarity reduce plagiarism problem or the overall similarity of both sentences can determine the similarity of both sentence whether they are similar or not.

We used accuracy to measure the performance of sentence similarity semantically for Afaan Oromo. Open semantic sentence similarity java version open source has been selected as a tool to develop the system by adopting the code to our work.

The description of figure 3.1 contain the following detail.

**Step 1.** Any texts from different sentences can be provided as a query to retrieve relevant terms from entire collection of documents those are represented as BOW by representative terms (key terms).

**Step 2.** The queries are going to be normalized to key word that represent the queries to retrieve the relevant terms from stored documents. Here the preprocessing phase will takes place means tokenization (splitting the whole text into single term or word).

These words are considered as individual tokens. Stop word removal (remove highly frequented term or word or more frequently repeated terms).

Stop words are the words which are having less importance and repeating frequently. Some Afaan Oromo stop word lists are ‘osoo’, ‘as’, ‘dha’, ‘hin’, ‘ni’, ‘kan’, ‘kuunnoo’, ‘koo’, ‘yoo’, ‘ol’, ‘oli’, ‘gadi’, ‘irra’, ‘achi’, ‘utuu’, ‘kee’, ‘fi’, ‘sun’, ‘kun’ etc. Proper care is taken so than accuracy remains unaffected without increasing recall.

Stop words are highlighted by black color as configured in figure 4.3 and figure 4.4. For this work we have used black color to hide stop words from retrieved documents and retrieved with black color (stop word) does not have value on sentences similarity measure. Means on plagiarism ratio calculation stop word (black part) of retrieved document has no any value (ignored).

Stemming was the technique of reduce a word to its grammatical root to reduce inflectional problem that may cause difference between one words meaning for machine learning due to morphological analysis like suffix for Afaan Oromo. Consider those three words ‘nyaate’, ‘nyaatte’ and ‘nyaatan’. All these words have same root form as ‘nyaat-’. But those words are different terms for machine (computer) when stemming concept missed.



If we do not perform the term stemming, then the relational value between nyaate, nyaatte and nyaatan will reduce and it also affects the overall similarity value. It improves the efficiency of the method by increasing recall.

**Step 3.** The words or terms matching will takes place to calculate the similarity of the query and stored document's (corpus) terms based on content or lexical similarity measure. Here word-to-matching will takes place.

To make as plagiarized part of documents are clearly visible to the user physically or lexically plagiarized or copy and pasted part of document's terms are highlighted with red color.

**Step 4.** In addition to lexical similarity to enhance accuracy of similarity to detect plagiarism semantic similarity is also important. Semantic similarity can be calculated by look up synonyms terms from relational data base of mismatched words from input text (query) and stored Documents (corpus).

For this work we have try to show by highlighting blue color semantically plagiarized part of stored documents (D0 and D1). There is also similarity ratio for both document independently to indicate amount of plagiarism.

**Step 5.** Calculate percentage of matching terms we made a value between queries provided and collected documents over the similarity and number of words or terms.

If there were mismatched terms from both query and corpus after calculating percentage of matching terms or similarity lexically, they must be checked from grouped relational database of synonym representation to know whether the terms are semantically similar or not.

As the both figure 4.3 and figure 4.4 shows the last result of similarity was represented in percentage as plagiarized ratio from both lexical similarity and semantic similarity together.

For this study area we have limited and set the threshold value if there were similarity from percentage ratio. Threshold value is used for decision making purpose of amount of plagiarized contents in the documents in percent from similarity measure summary.

**Step 6.** Lastly based on the above technique of procedure the system can give an opinion whether the terms of query and document stored as corpus should be checked for plagiarism or not [16].

Our sequence of work has follow the above procedure sequentially to calculate similarity of the sentences lexically and semantically to detect plagiarism.

In this step the last result of similarity measure to detect plagiarism of query's term with collected document's terms can be represented in percentage.

For this work we have collected two documents from Afaan Oromo fictions and one document which was history of one women from FBC Afaan Oromo. The documents are document 0 (D0), document 1 (D1) and document 2 (D2) sequentially as sample corpus for this work of study.

D0 has stored with 15 sentences and 101 tokens, D1 has stored with 14 sentences and 115 tokens and D2 has stored with 11 sentences and 141 tokens. To know the plagiarized content of documents there must be query (Q) with unlimited tokens but less than source documents terms. This is used to test our work how much the documents content has been plagiarized.

In this study percentage of plagiarized content of the documents by query provided can be separately measured for all documents D0, D1 and D2 in percentage independently. The single document (D0) percentage of similarity measure is calculated from lexical similarity plus semantic similarity of query and document terms to get the ratio of similarity for identification of plagiarized part of stored documents.

In addition to put plagiarized content or part of documents with ratio in present, we have try to show plagiarized part documents by highlighting the color. For clarity of this work we have used two colors for highlighting the plagiarized content of documents. Our similarity ratio has derived from two features of similarity measuring techniques (lexical and semantic) similarity.

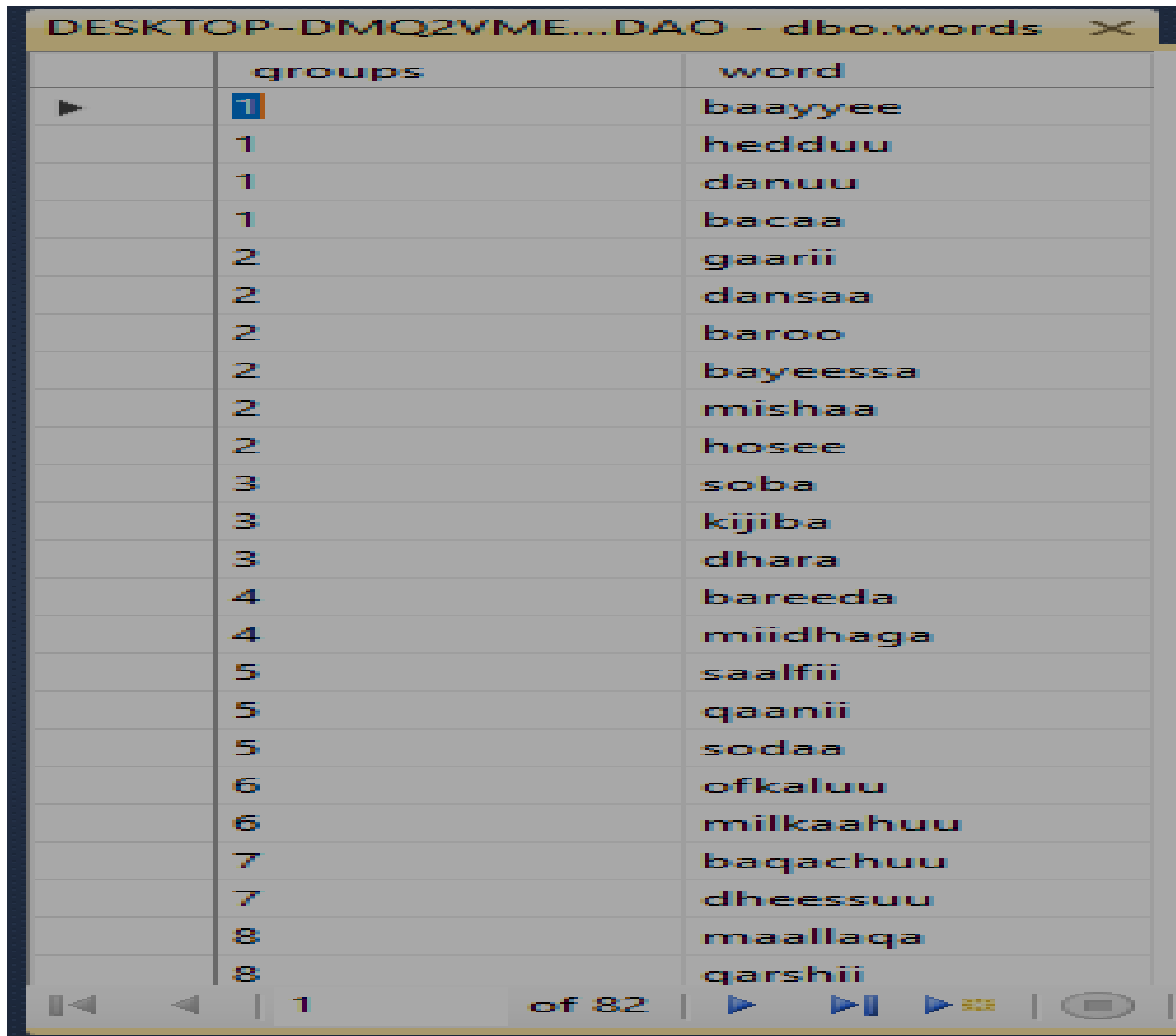
For highlighting document's part of plagiarized we used two colors red and blue for lexically plagiarized source documents and semantically plagiarized source documents respectively.

So terms of documents (D0, D1 and D2) highlighted with red color is lexically plagiarized documents terms with query's terms or copy and paste part of plagiarism and documents (D0 and D1) highlighted with blue color is semantically plagiarized with query's terms.

#### 4.4. Prediction process

The prediction process works on the plain text sentences entered by a user (query). First, the plain text or sentence is pre-processed and collected documents are also preprocessed in the same manner.

The following representation shows term representation with their synonym in the database.



groups	word
1	baayyee
1	hedduu
1	danuu
1	bacaa
2	gaarii
2	dansaa
2	baroo
2	bayeessa
2	mishaa
2	hosee
3	soba
3	kijiba
3	dhara
4	bareeda
4	miidhaga
5	saalfii
5	qaanii
5	sodaa
6	ofkaluu
6	milkaahuu
7	baqachuu
7	dheessuu
8	maallaqa
8	qarshii

Figure 4.1. Synonym representation in relational database

We can understand from figure 4.1 how synonym terms are classified into the same group with their respective meaning semantically in relational database.

For testing sentences are randomly taken from Afaan Oromo daily spoken communication from the society and Afaan Oromo written documents (fictions) those have semantically plagiarized.

We prepared sample dictionary that contain synset or synonym term in one group for our work that stated in figure 4.1. Means our algorism, LSI learns from relational database as term has synonym or not during similarity measure for checking semantic similarity.

For example the English term “false” by Afaan Oromo “**soba**” word has two synonyms “**dhara**” and “**kijiba**” as our sample relational database representation and the other words or terms has their own synonyms as we have tried to figure out sample terms in figure 4.1. We have used Structured Query Language (SQL) server to grouping the words with their corresponding groups in relational database.

We adopted the LSI algorism concept to our work of design to solve the problem or the gap we proposed. The semantic sentence similarity measure can be applied by checking the synonym terms from grouped relational database for semantic measurement.

From these point of view with natural language processing concept the algorism we adopt to our work learn the synonym words from relational database we have been designed. When similarity of two terms, sentences paragraphs and documents are going to be measured, the meaning or synonym of words has been checked from sample synonym representation.

The sample synset we designed for synonym representation can serve as WordNet for our work in Afaan Oromo language. So this was the technique to represent synonym words or terms to detect plagiarism by similarity measure as much as possible in our work for the language.

#### 4.5. Evaluation procedure

To evaluate the performance of the proposed work, the study has been proposed to use Afaan Oromo collected synonym terms without any domain specification for semantic extraction. We are going to measure sentence based similarity whether semantically they are similar or not in addition to lexical similarity for plagiarism detection.

The performance of test will evaluating manually and by applying IR system performance evaluation techniques those are recall and precision measure techniques. Precision is fraction of retrieved sentences that are relevant whereas recall is fraction of relevant sentences that retrieved.

The reason why we propose this framework is that in practical situation people modify word by using synonym when they plagiarize intentionally.

To overcome the problem of plagiarism by using synonym term our proposed frame work may contribute great solution for Afaan Oromo language.

Thus it can analyze semantic features of the sentences at word level to enhance the Afaan Oromo Information retrieval system. Because of these characteristics we believe that it is possible to capture the meaning of word in sentences to compare the original and plagiarized sentences.

We have evaluating query (texts) and stored two three documents with semantic and lexical similarity manually. As per our evaluation the result we get is vary from that of our proposed methodology. Means we distributed 10 queries at sentence level as testing data and three documents D0 with 15 sentences of 101 words, D1 with 14 sentences of 115 words, D2 with 11 sentences of 141 words as training data for evaluation purpose manually. The respondents are 18 in numbers and all are Afaan Oromo fluent speakers including Afaan Oromo department head of Dilla University as language expert.

We classified our dataset into training data and test data. The classification ratio of our dataset was 90 % training data and 10 % testing data. All terms of individual document for all three documents are counted as training data whereas all terms of query are considered as testing data for our work. The number of collected document's sentences are 15, 14 and 11 for D0, D1 and D2 respectively whereas averagely query sentences has one (1) sentence for one query. The same size can estimated at term level.

If the query's term available in the source documents physically the similarity can be calculated as lexical similarity was 1 whereas query's term was not found in source documents lexically similarity is calculated as 0. But when semantic similarity approach has been applied for this scenario there may be probability to the term have synonym. When the hybrid case happens, their average similarity is between 0 and 1.

All respondents are requested to compare all sentences (queries) and all documents (D0, D1, D2) as we have tried to indicate on **ANNEX** of our work part. Almost most of respondent's results vary when they put the result of both queries' and collected document's similarity due to semantic issue matters for communication.

In our opinion we conclude their result variation may be due to environmental location because one of our respondents was from North shoa, Oromia and two of them were from Guji, Oromia.

The one from North shoa didn't know the word 'Bacaa' instead he know as 'Hedduu' which is exactly the synonym of 'Bacaa'. The one from Guji didn't know the word 'Katabe' instead they know as 'Barreesse' which is again exactly the synonym of 'Katabe'.

But all of our respondents are from Oromia region, they are Oromo and all terms or words of our sentences for evaluations are Afaan Oromo text. Due to the above problem the manual sentences similarities performance is less than our system's performance.

Not only these, some people use representative synonym of word in different purpose, especially for plagiarism. Due to this also accuracy of sentence similarity become less for reader, listener and generally for audience.

*Table 3 Result of manual work*

NO	Query	Document 0	Document 1	Document 2
1	Q1	64.8	59.16	7.67
2	Q2	20.12	36.67	6.2
3	Q3	31.67	80	16.67
4	Q4	9.33	22.67	71.83
5	Q5	26.17	47.33	3.67
6	Q6	25.83	44.16	5.3
7	Q7	100	78.67	4.4
8	Q8	44.33	47.5	0
9	Q9	0	0	2
10	Q10	0	0	0

Table 4.1 shows result gained from manual evaluation we have been done.

So for the first pair of query 1 (Q1) terms with all D0 terms averagely all respondent put 64.8% query 1 was plagiarized from D0 and our model work evaluate both similarity as 27.1 %, Q1 terms with all D1 terms averagely all respondents result was 59.16 % Q1 was plagiarized from D1, but model proposed for this work evaluate both similarity as 25.6 % and Q1 terms with all D2 averagely all respondents result was 7.67 % Q1 has been plagiarized from D2, but our model evaluate both similarity as 15.1 % .

The same procedure has been followed up to comparison of Q10 with all terms of D0, D1, and D2 respectively as shown in table 4.1 and figure 4.2.

The proposed model result was less than human judgment as observed on both figures.

Human can judge based on his/her perspective similarity of given documents and queries. Since we collect manually the judgements of peoples their response as similarity may fluctuate due their dialect background.

*Table 4.2. Result of proposed model*

*Table 4 Result of proposed model*

NO	Query	Document 0	Document 1	Document2
1	Query 1	27.1	25.6	15.1
2	Query 2	19.5	21.1	16.1
3	Query 3	28	26	17
4	Query 4	23.1	20.4	33.2
5	Query 5	31.67	25.5	16.3
6	Query 6	19.5	19.1	115.4
7	Query 7	22.6	22.5	0
8	Query 8	21	21	15.1
9	Query 9	0	10.4	17.4
10	Q1uery 10	0	0	15.1

Table 4.2 shows proposed model result from all source documents and query provided.

The proposed algorithm evaluates the comparison of terms similarity task as query’s term per individual documents term’s for similarity measure lexically and semantically by extracting synonym terms from dictionary.

We have used the IR evaluation metrics formula to calculate precision, recall and F-measure. Where TP = True positive, FP = False positive, TN = True negative, FN= False negative

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \dots\dots\dots 4.1$$

Precision represents relevant and retrieved terms divided by number of suspicious query’s term.

Recall =  $TP / (TP + FN)$  ..... 4.2

Recall represents relevant and retrieved terms divided by number of source document's term

F-measure =  $2 * (P * R) / (P + R)$ .....4.3

F-measure was the harmonic mean of precision and recall calculated from equation

Accuracy =  $(PT + TN) / (P + N)$ ..... 4.4

Accuracy represents number of all correct predictions divided by the total number of the dataset.

*Table 5 Evaluation result of proposed model*

Query per document	Precision	Recall	F-measure	Accuracy
Query1 with Doc0	68.33	51	58.4	68
Query1 with Doc1	68.33	49	57.2	67.3
Query1 with Doc2	43	34	38	52
Query2 with Doc0	55	48	51.3	48.12
Query2 with Doc1	46	47	46.5	44.89
Query2 with Doc2	46	35	38	32.8



Query3 with Doc0	67.1	51	58	52.81
Query3 with Doc1	100	53	69.3	52.53
Query3 with Doc2	44.3	36	40	44.6
Query4 with Doc0	36.25	39	37.6	78.62
Query4 with Doc1	36.25	35	35.6	78.1
Query4 with Doc2	72.5	57	64	51.8
Query5 with Doc0	43.33	43	43.2	41
Query5 with Doc1	76.67	49	60	37.89
Query5 with Doc2	46.67	36	41	44.8
Average result	56.65	44.2	49.2	53.02

Table 4.3 Evaluation result in precision, recall, F- measure accuracy.

Table 4.3 indicate that average precision 56.65 %, recall 44.2 %, F-measure 49.2 % and accuracy 53.02 % achieved from the model proposed.

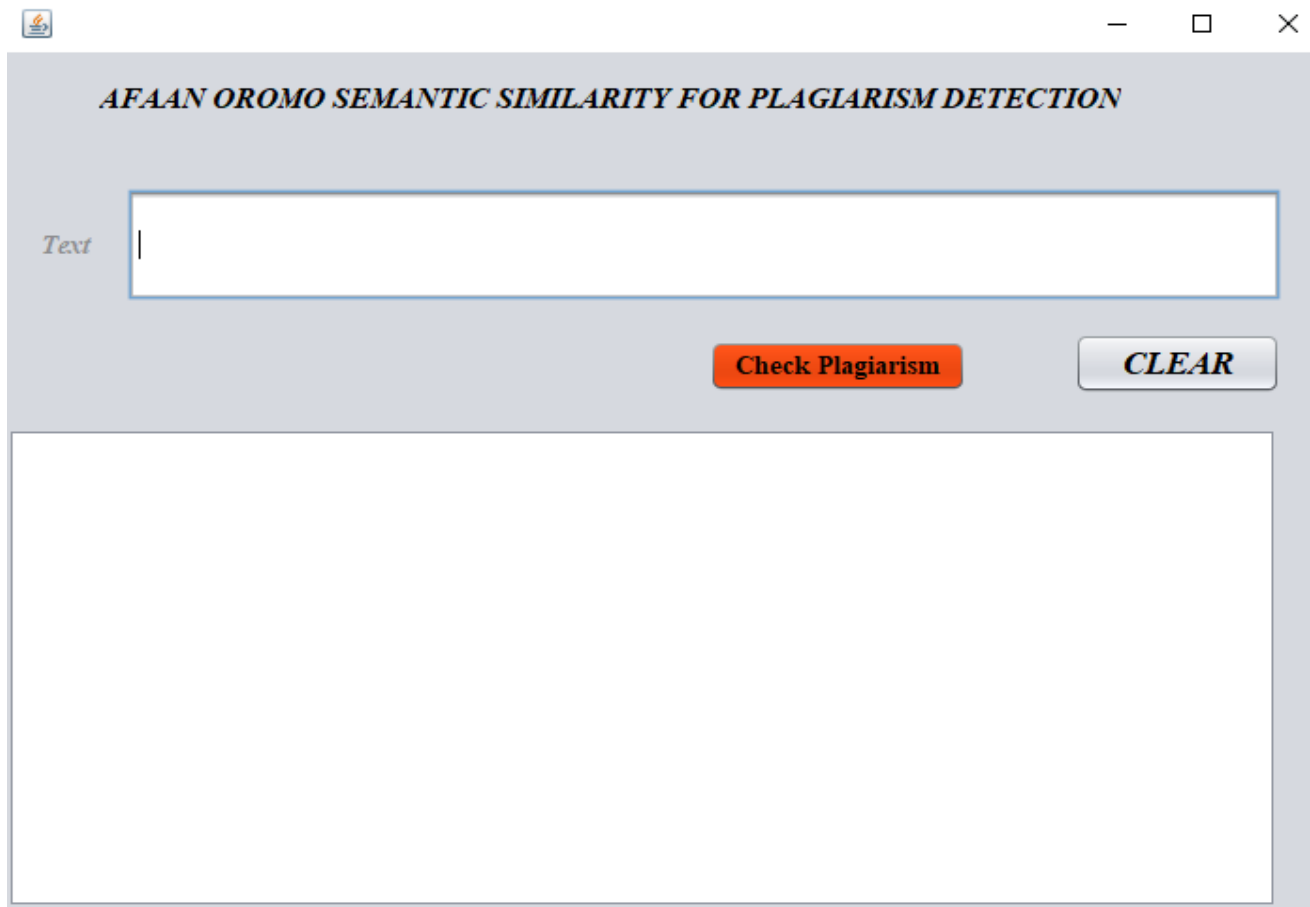
As the result average recall gained was less value because we tested one sentence of query with more sentence of documents. The same was true for average precision value and F-measure average value achieved. The overall accuracy of this work can be increased if the size of query’s terms (suspicious) sentences or document increase.

As per the similarity results from the table 4.1 we got the manual evaluation result based on our respondents answer. We have distributed 10 sample queries with three documents for 18 respondents as they put the similarity result of query per documents one by one.

As per our request all of our respondents are volunteer and they have filled similarity of both all queries with all documents manually. We take their result as single manual evaluation on the table 4.1 by adding all answer of the first paired terms of query with documents and divide the sum by number of respondents (18). So the single value for first similarity was an average result of all respondents.

#### 4.6. Experimental result

Suspicious text or query could be provided as ‘Text’ on the figure 4.2 and extraction has been takes place from source or collected documents.



*Figure 4.2. Proposed interface*

This was the proposed interface that has been created with java NetBeans 8.2 to show the general interface.

Suspicious text or query could be provided as ‘Text’ on the figure 4.2 and extraction has been takes place from source or collected documents. Then when we click on ‘check plagiarism’ button of figure 4.2 similarity matching of suspicious text provided as ‘Text’ with source documents will be calculated separately for all collected source documents. Overall matching can be calculated from lexical matching and semantic matching.

In addition to percentage of similarity measure as plagiarism ratio to detect plagiarized content of documents we have used highlighting technique to show plagiarized part. Highlight with color the part of plagiarism is help to clearly visualize plagiarized part of document with great focus.

Highlighting technique is an additional features of similarity ratio that show plagiarized part of document with specified color background for this study.

The highlighted with red color observed on figure 4.3 shows terms of query provided plagiarized directly from both documents (D0 and D1) or copy and paste of the terms. For our work we randomly select red color to shade lexically plagiarized part, so it is not mandatory for other works. We have used only two documents with one sentence of query to show lexicographically plagiarized part of D0 and D1 in below figure and we include D2 in the evaluation parts.

## AFAAN OROMO SEMANTIC SIMILARITY FOR PLAGIARISM DETECTION

Text yeroo dhiyootti akka oromoo qaqqabu abdiin qaba. ulfina namaaf hin kennu

Check Plagiarism

CLEAR

doc 0 ifaajjii yeroo dhiyootti akka ummata oromoo akka qaqqabu fedhii abjuu natti tahee Wayaa ilaali tokkollee ija namaatti tolu jiru. haalli jireenya baayyee ulfaataa Hawwiin soba dubbattee dhugaa awwaalte. Ati saalfii beektaa? Sammuun wal falmuu jalqabe

The plagiarism ratio with doc 0=0.14705882352941177

Average Plagiarized

doc 1 Hawwii guddaan tattaaffii hojii hawaasa bira akka naaf qaqqabu Huccuu ati uffattu hunduu namatti tolan. kijiba dubbachuun hawaasa keessatti ulfina namaaf kennu. Mindaan xiqqaa waan ta'eef jireenyi hedduu rakkisa. Qaanii beektaa? Yaadni ofumaa waliin wal mormuu eegale.

The plagiarism ratio with doc 1=0.1388888888888889

Average Plagiarized

Figure 4.3. Highlighted red color to show lexically plagiarized part

The highlighted with blue color observed on figure 4.4 shows terms of query plagiarized semantically from both documents (D0 and D1) that the meaning of terms are extracted from sample dictionary we prepared for synonym representation. This semantic similarity value is added on lexical similarity value to increase the overall accuracy of similarity measure to detect plagiarism from the collected documents. We have used blue color to shade semantically plagiarized part of document to differ it from lexically plagiarized part of documents which is red color. The blue color is not mandatory parameter, simply it was randomly selected to highlight semantic parts in this work. We have used only two documents with one sentence of query to show semantically plagiarized part of D0 and D1 in below figure.

***AFAAN OROMO SEMANTIC SIMILARITY FOR PLAGIARISM DETECTION***

**Text** dhara dubbatanii jiraachuun aadaa bacaa jibbisiisaa dha.

Check Plagiarism
CLEAR

doc 0 ifaajjii yeroo dhiyootti akka ummata oromoo akka qaqqabu fedhii abjuu natti tahee Wayaa ilaali tokkollee ija namaatti tolu jiru. haalli jireenya baayyee ulfaataa Hawwiin soba dubbattee dhugaa awwaalte. Ati saalfii beektaa? Sammuun wal falmuu jalqabe

The plagiarism ratio with doc 0=0.1176470588235294  
Average Plagiarized

doc 1 Hawwii guddaan tattaaffii hojii hawaasa bira akka naaf qaqqabu Huccuu ati uffattu hunduu namatti tolan. qjiba dubbachuun hawaasa keessatti ulfina namaaf kennu. Mindaan xiqqaa waan ta'eef jireenyi hedduu rakkisa. Qaanii beektaa? Yaadni ofumaa waliin wal mormuu eegale.

The plagiarism ratio with doc 1=0.0555555555555555  
Plagiarism Free

Figure 4. 4. Highlighted blue color to show semantically plagiarized part

From figure 4.5 there were query's those are lexically plagiarized directly and semantically from documents (D0 and D1) by synonym term representation.

Red color highlighted and blue color highlighted are there to indicate lexically plagiarized part of documents and semantically plagiarized part of documents respectively for all three documents independently. But black color highlighted parts indicates stop words of the documents retrieved by query request. Black color highlighted or stop words has no effect on similarity calculation to get result of ratio plagiarism because the algorithm exclude stop words from any task of calculation. We have used only two documents with one sentence of query to show lexically and semantically plagiarized part of D0 and D1 in below figure. But we have used all three documents collected in the evaluation part of this study.

**AFAAN OROMO SEMANTIC SIMILARITY FOR PLAGIARISM DETECTION**

**Text** fedhii koo guddaan tattaaffii hojiikootii ummata bira qaqqabsiisuu dha.

Check Plagiarism
CLEAR

doc 0 ifaajji yeroo dhiyootti akka ummata oromoo akka qaqqabu fedhii abjuu natti tahee Wayaa ilaali tokkollee ija namaatti tolu jiru. haalli jireenya baayyee ulfaataa hawwiin soba dubbattee dhugaa awwaalte. Ati saalfii beektaa? Sammuun wal aimuu alqabe

The plagiarism ratio with doc 0=0.20588235294117646  
Average Plagiarized

doc 1 hawwi guddaan tattaaffii hojii hawaasa bira akka naaf qaqqabu Huccuu ati uffattu hunduu namatti tolan. kijiba dubbachuun hawaasa keessatti ulfina namaaf kennu. Mindaan xiqqaa waan ta'eef jireenyi hedduu rakkisa. Qaanii beektaa? Yaadni ofumaa waliin wal normuu segale.

The plagiarism ratio with doc 1=0.19444444444444448  
Average Plagiarized

Figure 4. 5. Highlighted part to indicate lexically plus semantically

For example term “tattaaffii” which is provided in the text (query) has available directly as it is in D1 which copy and pasted in query provided from D1. That is why it was highlighted with red color in figure 4.5 to show as it was plagiarized part of D1 in this work. For plagiarism ratio calculation “tattaaffii” term has been counter as plagiarized term of D1. But “tattaaffii” was not physically present in D0 where as its synonym term “ifaajjii” is there in D0.

Threshold based retrieval system is better if the text of query submitted in case of similarity measure for plagiarism detection [53]. The author suggest that as plagiarism is not allowed anywhere but to negotiate the decision for plagiarism threshold value is better like if 10 % and above of document part was plagiarized decide as there was plagiarism case for the documents.

Based on the concept of [53]we have set threshold values from plagiarism ratio result of query provided per all documents collected as dataset.

Threshold value helps to decide whether given text (query) is plagiarized or not based on plagiarism ratio value calculated from sentences similarity measure (lexical and semantic). For this study threshold value was set to the following statements.

If plagiarism ratio is greater than 40 % mostly plagiarized,

If plagiarism ratio is less than 40 % and greater than 10 % average plagiarized and

If plagiarism ratio is less than 10 % plagiarism free.

The above rule based value used to make boundary of plagiarism for this work in three stage as we have seen in the above rule. It is difficult to decide as 'mostly plagiarized' when most of Q1 terms are available in one specific stored document (D0) because the document may have huge data. So when sentence similarity of mostly copied terms of Q1 from the document (D0) and document (D0) with huge data is calculated the result of plagiarism ratio may be less. Matrix comparison of a few terms with huge terms result can be less than matrix comparison of a few terms with few terms result as textual similarity measure. Hence good decision of plagiarism can be better on huge dataset with good accuracy.

In logic of AO language one word may have one or more synonym terms. But there is a word that doesn't has synonyms.

## CHAPTER FIVE

### CONCLUSION AND FUTURE WORK

#### 5.1. Overview

The availability of NLP discipline has great contribution to enabled digital machines (computers) to understand human languages and process them for further steps. IR and similarity measures are major tasks incorporated in digital information area. IR was for searching by query and retrieving the response from documents whereas similarity measure can be comparison of query with query of sentences, paragraph and documents for plagiarism detection, summarization, and recommendation system.

In this work we have developed sentence based semantic similarity (SSS) system, called Afaan Oromo sentence based semantic similarity (*AOSSS*) measure for plagiarism detection. Before developing *AOSSS*, we have studied some of the sentence based semantic similarity systems developed for English language and other surveying work related to this study work.

Sentence based semantic similarity measure approaches have also been studied to select an approach that can give the best performance for the given constraints of Afaan Oromo language. The nature, structure of Afaan Oromo language has also been studied before developing the system to investigate the problem.

#### 5.2. Conclusion

This study proposed and designed a system called *AOSSS* measure to solve Afaan Oromo plagiarism problem. The system was designed based on machine learning approach. We have implemented the machine learning features by using LSI algorithm concept to decompose the sentence into term for vector representation for query provided and documents stored. The LSI algorithm concepts was implemented for semantic similarity measure for the language. LSI algorithm was used to index the term with its value in java hashmap and adopt the model for similarity measure.

For evaluation target we first considered human judgment manually with different respondents of language speakers and plagiarized ration calculated from the adopted LSI algorithm to this study.



The result obtained from both point can be counted as accuracy of the system. Hence we have got accuracy of calculated ratio of 53.02 %. This value was very small because our queries size was very smaller than source documents size. Increasing size of suspicious query's term to extract terms of source documents can enhance the accuracy values.

The proposed problems are almost implemented as main work to answer the research questions of this study. Synonym terms represented together in dictionary for semantic extraction, NLP disciples like normalization, tokenization, stop word removal and little rule based stemming features are implements to this work. There was no labeled data for Afaan Oromo for evaluation purpose so we have implemented the adopted algorism concept on small dataset which has impact on performance. So for further research performance can be enhanced if implemented on huge dataset.

### 5.3. Contribution of the work

We have tried to solve the problem of plagiarism semantically by representing word with their corresponding group by sample dictionary in relational database for Afaan Oromo language. From the previous work of scholar we identified the benefit of why tokenization to word, stemming, part of speech tagging, stop word removal are necessary.

We collect sample synset or synonym terms of Afaan Oromo together and group them with their corresponding meaning or synonym. When collecting and organizing those terms we have contact Afaan Oromo experts to get more words those have synonyms in the language.

Since one term of Afaan Oromo language may have more than two synonym that could be spoken in different part of Oromia region and other place we prepared sample dictionary for synonyms terms representation.

So the adopted algorism learn the meaning of word or term of sentences going to be measured from the sample synset we developed and meaning of the word will be known or retrieved from the relational database we designed as a sample dictionary.

The method we applied to handle sentence based semantic similarity measure for plagiarism detection can be contribution for the language.

#### 5.4. Future work

Sentence based or documents based similarity measure was very complex for under resourced languages like Afaan Oromo language. Therefore we will recommend as previously implemented and evaluated work will be free for next new research like stemming and POS tagging for Afaan Oromo and other languages.

Dataset size was the critical parameter for evaluation to achieve better results. In our case we have used small dataset with medium result. For future work we have plan to enhance the performance of this work by increasing the size of our dataset. N-gram word matching parameter is also a best plagiarism detector technique we have planned for this study and recommend for other local languages.

Standardize the sample dictionary we prepared for synonym term as synset by relational database. Means this sample dictionary must be standardized as WordNet for English throughout further step for Afaan Oromo and other local language with respective rule if not.

Preparing annotated dataset or corpus in the language for semantic sentence similarity even for semantic document similarity.

We have plan to enhance the performance of the current work in the future steps by deeply focusing on stemming and POS tagging since we didn't apply POS tagging in this work. Even we didn't fully apply stemming concept for this work rather we tried to manage the algorism concept for stemming little bit. So if this study work will evaluated with fully integrated stemming concept the achieved result will be better.

So others researchers and experts can repeat the same study with the same method and approach for plagiarism detection at sentences or document level for other local languages.

## REFERENCES

- [1] T. H. Dan Jurafsky, "Document Similarity in Information Retrieval," in *Document Similarity in Information Retrieval*, 2012, p. 81.
- [2] P. Resnik, "Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research*, vol. 11, p. 95–130, 1999.
- [3] merriam, "dictionary/plagiarism," merriam-webster, 4 september 2016. [Online]. Available: <http://www.merriam-webster.com/dictionary/plagiarism>. [Accessed 12 june 2019].
- [4] M. P. J. P. a. V. S. Hussein Soori, "Semantic and Similarity Measure Methods for Plagiarism Detection of Students' Assignments," *Ostrava - Poruba*, vol. 2, no. 3, pp. 45-62, 2018.
- [5] G. H. Alexander Budanitsky, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," 8 July 2004.
- [6] P. 1. k. Tomas Ptacek, "Advanced Methods for Sentence Semantic Similarity," *Masters thesis*, 2012.
- [7] M. A. Salahli, "An Approach For Measuring Semantic Relatedness Between Words Via," *Mathematical and Computational Applications*, vol. 14, no. 1, pp. 55-63, 2009.
- [8] K. K. Rajkumar Kundu, "Contextual Plagiarism Detection Using Latent Semantic," *ISSN (Online)*, pp. 2455-9024, 2010.
- [9] A. J. A. Muftah, "Document Plagiarism Detection Algorithm Using Semantic Networks," *Universiti Teknologi Malaysia*, November 2009.
- [10] C. K. L. Sumathy, "A Hybrid Approach for Measuring Semantic Similarity between Documents and its Application in Mining the Knowledge Repositories," (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 8, 2016.
- [11] I. S. i. T. David Kirk Evans, "Multi-Lingual Analysis for Summarization", *COLUMBIA UNIVERSITY*, 2005.
- [12] M. Naghibzadeh, "Semantic similarity assessment of words using weighted WordNet," *Springer-Verlag Berlin Heidelberg*, 2012.
- [13] P. W. Foltz, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259-284, 1998.
- [14] C. Boyd, "What Is Latent Semantic Indexing," *SE journal*, March 6, 2018.
- [15] M. C. Malik Muhammad Saad Missen, "Comparing Semantic Associations in Sentences and Paragraphs for Opinion Detection in Blogs," *Lyon, France.*, 2009.

- [16] K. Shams, "PLAGIARISM DETECTION USING SEMANTIC ANALYSIS," *Masters thesis*, April 2010.
- [17] V. U. T. a. C. Bowerman, "Methods for Detecting Paraphrase Plagiarism," 2011.
- [18] M. T. N. B. M. V. M. P. R. a. S. C. N. Sukhija, "Topic Modeling and Visualization for Big Data in Social Sciences," *Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress*, 2016.
- [19] N. N. a. R. B. Vasile Rus, "Similarity Measures Based on Latent Dirichlet Allocation," *Springer-Verlag Berlin Heidelberg*, vol. 1, no. LNCS 7816, p. 459–470, 2013.
- [20] wpacouncil.org, 22 september 2016. [Online]. Available: <http://wpacouncil.org/positions/WPAplagiarism.pdf>. [Accessed 16 July 2019].
- [21] www.indiana.edu, indiana, 21 march 2018. [Online]. Available: <https://www.indiana.edu/~istd/examples.html>. [Accessed 26 May 2019].
- [22] H. A. C. a. D. K. Bhattacharyya, "Plagiarism: Taxonomy, Tools and Detection Techniques," *Dept. of CSE, Tezpur University*.
- [23] Y. J.-Y. K. H.-R. Chen C-Y, "Plagiarism Detection using ROUGE and WordNet," *Journal of Computing*, vol. 2(3), p. 34–44., 2010.
- [24] S. a. M.S.Otari, "Plagiarism Detection-Different Methods and Their Analysis: Review," *International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN*, vol. 1, no. 7, pp. 2349-2163, August 2014.
- [25] A. Kaur, "A Novel Approach for Syntactic Similarity between Two Short Text," vol. 4, no. 06, pp. ISSN 2277-8616, JUNE 2015.
- [26] G. E. E. G. P. E. [22] AngelosHliaoutakis, "Information retrieval by semantic similarity," in *Technical University of Crete (TUC)*, Canada, 2015.
- [27] W. H. G. A. A. Fahmy, "A Survey of Text Similarity Approaches," *International Journal of Computer Applications*, vol. 68, no. No.13, p. 0975 – 8887), April 2013.
- [28] Y.-B. Jia, "Singular Value Decomposition," Com S 477/577 Notes, Sep 13, 2018.
- [29] A. Y. N. a. M. I. J. David M. Blei, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
- [30] Y. S. H. W. Y. H. Junsheng Zhang, "Calculating statistical similarity between sentences," *Journal of Convergence Information Technology*, vol. 6, no. 2, February 2011.
- [31] G. Salton, "Automatic Text Processing," in *The Transformation Analysis and Retrieval of Information by Computer*, AddisonWesley, 1989.

- [32] T. a. M. F. Rahutomo, "Semantic Cosine Similarity," *International Journal of Advanced Computer Science*, 2007.
- [33] B. Liu, "Web Data Mining," *Springer Berlin Heidelberg*, 2011.
- [34] S. S. Q. W. LiHong Xu, "Text Similarity Algorithm Based On Semantic Vector Space Model," *Dingfuzhuang Street (E)*, vol. 1, 2009.
- [35] Stephanie, "How to Calculate the Jaccard Index," December 2, 2016.
- [36] L. Zahrotun, "Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method," *Computer Engineering and Applications*, vol. 5, no. 1, pp. 2252-5459, February 2016.
- [37] E. A.-S. HadeelQasemGheni, "Plagiarism Detection using Semantic Analysis," *Indian Journal of Science and Technology*, vol. Vol 9(1), p. DOI: 10.17485/ijst/2016/v9i1/84235, January 2016.
- [38] T. S. Thanh Dao, "WordNet-based semantic similarity measurement," 28 Apr 2016.
- [39] Z. W. a. M. Palmer, "Verb Semantic and Lexical Selection," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistic*, p. 133–138, 1994.
- [40] R. C. C. S. C. Mihalcea, "Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity," *In: American Association for Artificial Intelligence (AAAI'06)*, Boston (2006).
- [41] T. Guya, "CaasLuga," in *Afaan Oromoo Jildii-1, Gumii Qormaata Afaan OromootiinKomishinii*, Finfinnee, Aadaa fi TurizimiiOromiyaa, 2003.
- [42] wikibooks, "en.wikibooks.org," wiki/Afaan\_Oromo, 6 december 2018. [Online]. Available: [https://en.wikibooks.org/wiki/Afaan\\_Oromo/Introduction](https://en.wikibooks.org/wiki/Afaan_Oromo/Introduction). [Accessed 23 may 2019].
- [43] M. Y. M. Chong, "A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques," degree of Doctor of Philosophy, University of Wolverhampton, 2013.
- [44] L. D. Krisnawati, "Plagiarism Detection for Indonesian Texts," Masters thesis, August 29, 2016.
- [45] B. A. V. a. D. M. A, "FRAMEWORK FOR RUSSIAN PLAGIARISM DETECTION USING SENTENCE EMBEDDING SIMILARITY AND NEGATIVE SAMPLING," *Computational Linguistics and Intellectual Technologies, Proceedings of the International Conference*, 30, June 2, 2018.
- [46] S. A. King Abdulaziz University, "A Framework For Plagiarism Detection In Arabic Documents," DOI : 10.5121/csit.2015.50201, 2010.
- [47] S. C. N. A. D. a. K. V. Anita R., "Semantic Search Using Latent Semantic Indexing And Wordnet," *ARPN Journal of Engineering and Applied Sciences*, vol. 12, no. 2, pp. ISSN 1819-6608, 2017.

- [48] F. B. Tesema, "Afaan Oromo Automatic News Text Summarizer Based on Sentence Selection Function," *Masters thesis*, 2013.
- [49] D. Tesfaye, "Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach," *Masters thesis*, 2010.
- [50] E. N. Alemu, "Afaan Oromo –Amharic Cross Lingual Information Retrieval: A corpus Based Approach," *Masters thesis*, 2013.
- [51] R. R. e. r̃ek, "Semantic-based plagiarism," January 10, 2007.
- [52] P. Z. A. Liqiang Pan, "Similarity Calculation Method of Chinese Short Text Based on Semantic Feature Space," (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, 2015.
- [53] G. Cosma, "Evaluating the Performance of LSA for Source-code Plagiarism Detection," *Informatica*, vol. 36, p. 409–424, 2012.

## APPENDIX

### SAMPLE COLLECTED DOCUMENTS

#### Documents 0 (D0)

Aadde Mabiraat Walda Sillaasee kanan harkaan barreesse komputeraan barreessanii kan naa qindeessan galata na biraa qabu. ifaajjii koo yeroo dhiyootti akka ummata oromoo qaqqabu fedhii koo isa abjuu natti tahee dha. Wayaa kee ilaali tokkollee kan ija namaatti tolu hin jiru. haalli jireenya isaanii baayyee ulfaataa dha. Hawwiin soba dubbattee dhugaa awwaalte. Ati garuu saalfii beektaa? Sammuun isaa wal falmuu jalqabe. Ni bareeda mitii adaraa? Akka nama dhoqqee dhiituu narra deeme. Eessa akkan bule, enyu akka bule waliin hin beeku. Ganama sireerratti of arge. dinqii siif! Enyu waliin akka bulte hin beektu? Ati hooma of hin beektu. Ija guutuun ilaalu sodaate irraa garagale.

#### Document 1 (D1)

Kanan ani harkaan barreesse makiinaa fi komputeraan kan naaf Katabe Aadde Yeshii Abbayyaa fi Aadde Elsaabeet Gurmeessaa yoom iyyuu akkan galatoomfadhetti. Namoota yaadan nafaana turan hunda galata narraa qabu. Hawwii koo guddaan tattaaffii hojii kootii hawaasa koo bira akka naaf qaqqabu dha. Huccuu ati uffattu hunduu namatti hin tolan. kijiba dubbachuun hawaasa keessatti ulfina namaaf hin kennu. Ni miidhaga mitii sirritti hubattee garuu? Mindaan isaanii xiqqaa waan taheef jireenyi hedduu isaan rakkisa. Na tuffate yaada koo irra deeme. Qaanii beektaa? Yaadni isaa ofumaa isaa waliin wal mormuu eegale. Akkamitti akkan buleefi eenyu wajjin akkan bule himuu haa hafuuti ofii koo iyyuu hin beeku. Ajaahiba! aboo kan waliin bulte illee hin beektuu? Ilaaluu dadhabeen achi irraa garagale.

## **Document 2 (D2)**

Ilmaan qonnaan bulaa irraa dhalattee hojii daldalaatiin dursitee mullatte, kutannoo fi murannoon jiraannaan bakka yaadan gahuun akka dandahamu agarsiiftee daandii haaraa saaqxe.

Hin dandahamuu dandeessee, harka dheeraa osoo hin taane kaayyoo dheeraa qabatee imala milkii eegalte. Shiroy fi paastaa gurguruun daandii jireenyaa kan eegalte harra haadha mana nyaataa filatamaa tahuu dandeesse. Aadde Maammituun dhalatanii kan guddatan Lixa Oromiyaa Magaalaa Ayiraa yommuu tahu, hojii daldala kan itti eegalan garuu Handhuura Oromiyaa Finfinnetti. Gaafa cidha namaa deemanii hojii nama gargaaruu hedduu jaaalatu turan. Sababa kanaafis namoonni hedduun gaafa qophii wayii qabaatan namni isaan jalqaba waammatan Maammituu ture. Fedhii hojii ittiin horatan nyaata xixiqqaa akka shiroy fi paastaa itti hojjachuu eegalan. Abjuun dubartii cimtuu tanaas karaa walakkaa imaltee gara isaanitti dhihaatte. Fagoonis dhihaattee hin taatuunis taatee argamtee milkii biraa akeekte. Haaluma Kanaan jireenya baayyee gaarii jiraachaa jiru.