ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES SCHOOL OF INFORMATION SCIENCE

EXTRINSIC HYBRID AMHARIC TEXT PLAGIARISM DETECTION FOR NEWS ARTICLE'S POSTS ON SOCIAL MEDIA

BY

BEMENET TESFAYE

GSE/8588/11

JANUARY 2025, ADDIS ABABA

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES SCHOOL OF INFORMATION SCIENCE

EXTRINSIC HYBRID AMHARIC TEXT PLAGIARISM DETECTION FOR NEWS ARTICLE'S POSTS ON SOCIAL MEDIA

A THESIS SUBMITTED TO SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE AND SYSTEMS (INFORMATION SCIENCE SPECIALIZATION)

BY : BEMENET TESFAYE

GSE/8588/11

ADVISOR: WONDWOSSEN MULUGETA (PHD)

JANUARY 2025, ADDIS ABABA

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCES SCHOOL OF INFORMATION SCIENCE

EXTRINSIC HYBRID AMHARIC TEXT PLAGIARISM DETECTION FOR NEWS ARTICLE'S POSTS ON SOCIAL MEDIA

BY : BEMENET TESFAYE

NAME AND SIGNATURE OF MEMBERS OF THE EXAMINING BOARD

WONDWOSSEN MULUGETA (PHD)

ADVISOR SIGNATURE DATE

_____ _____

SIGNATURE DATE

SOLOMON TEFERRA (PHD)

EXAMINER

_____ _____

SIGNATURE DATE

MICHAEL MELESE (PHD)

EXAMINER

_____ _____

SIGNATURE DATE

DECLARATION

This thesis has not previously been accepted for any degree and is not being concurrently submitted in candidature for any degree in any university.

I declare that this thesis entitled "**EXTRINSIC HYBRID AMHARIC TEXT PLAGIARISM DETECTION FOR NEWS ARTICLE'S POSTS ON SOCIAL MEDIA**" is a result of my own investigation, except where otherwise stated. I have undertaken the study independently with the guidance and support of my research advisor, Dr. Wondwossen Mulugeta. Other sources are acknowledged by citations giving explicit references. A list of references is appended.

Signature:_____

Bemenet Tesfaye

This thesis has been submitted for examination with my approval as university advisor.

Advisor's Signature:_____

Wondwossen Mulugeta(Ph.D.)

ACKNOWLEDGEMENTS

First and foremost, I want to express my gratitude to the Almighty God for his blessings on my research, education, and life. This work would not have been completed without the aid, encouragement, dedicated involvement, understanding, and leadership of my advisor, Dr. Wondwossen Mulugeta.

I am grateful to my Instructors and classmates in AAU for enlighten specially to my close friends Sisay Zinabu, Muluken Sholay, Getnet Mezgebu and Getaneh Damite and my office colleagues.

I am enormously grateful to my children, Lael Bemenet and Basliel Bemenet, for their love, encouragement, understanding, care, prayers, and ongoing support in helping me finish my research.

Bemenet Tesfaye

JANUARY, 2025

Addis Ababa, Ethiopia

ABSTRACT

As news posting material develops in Ethiopia, there is rising worry about plagiarism in news items on social media platforms due to potential plagiarism of articles. Plagiarism in news articles will have a significant negative impact on society by fostering a climate of mistrust, particularly regarding the authority and ownership of news organizations, as well as integrity and homogeneity reports that lack originality, which leading to violations of professional norms. In order to identify plagiarism on information that is copied and rephrased without giving credit to the original author, this research proposes a two-layer text plagiarism detection technique tailored for Amharic news items posted on social media by using different methodologies.

The study employed a methodology of purposive sampling to gather data from social media accounts belonging to government news agencies, privately owned news agencies, individuals/bloggers, and journalists, as well as international organizations like BBC Amharic. The criteria for choosing included posting daily news on a range of topics (such as politics, economics, and international news), having an Amharic-language public channel, and having a significant number of followers. Two-layer plagiarism detection has been found to be more effective at identifying semantic meaning, rephrasing, and copy pasting which the traditional detection failed to detect it.

The first layer of our approach compare candidate plagiarism detection techniques, such as fingerprint and n-gram checking, to identify potential cases of plagiarism. The second layer of the approach focuses on more advanced techniques, such as semantic LDA and fuzzy models. Then compared the performance of various mixed approaches like 1-layer fingerprint with 2-layer LDA, 1-layer fingerprint with 2-layer Fuzzy, 1-layer n-gram with 2-layer LDA, and 1-layer n-gram with 2-layer fuzzy for candidate then semantics plagiarism detection methods.

In conclusion, by incorporating both traditional like fingerprint and n-gram and advanced techniques like using LDA and fuzzy semantics, it is found that these performance metrics increase as the results of the various experiments show. The research findings indicate that merged features are better than the individual features for almost all models. Bi-gram with LDA:- Accuracy: 0.96, Recall: 0.909, Precision: 0.93, F1 Score: 0.92; Fingerprint with LDA:- Accuracy: 0.97, Recall: 0.917, Precision:

0.968, F1 Score: 0.94; Fingerprint and Fuzzy:- F1 Score: 0.66, Recall: 0.5, Precision: 1.0, Accuracy: 1.0. Based on the results of the study concluded that fingerprint with LDA outperformed in performance matrix including short time span to detect plagiarism.

Keywords:- Plagiarism, Extrinsic hybrid Amharic Text Plagiarism Detection, fingerprint and n-gram and advanced techniques like using LDA and fuzzy semantics

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AMN	Addis Media Network
APD	Amharic Plagiarism Detection
EBC	Ethiopia Broadcasting
FBC	Fana Broadcasting
IEEE	Institute of Electrical and Electronics Engineers.
PD	Plagiarism Detection
PDS	Plagiarism Detection System
SEO	Search Engine Optimization

CHAPTER ONE

1. INTRODUCTION

1.1. BACKGROUND

With the rapid advancement of technology, the internet has made information readily accessible, meeting the need for specific knowledge at the click of a button. Numerous tools have emerged in this era of digital connectivity, enabling users to acquire information from books, journals, newspapers, and more. The proliferation of smart mobile phones and declining internet costs have further increased online engagement, especially on social media platforms, where users exchange information and socialize. This shift has prompted organizations to leverage social media for promoting services, disseminating information, and capturing audience attention through network-based visibility, leading to adjustments in their business models.

While these technological advancements have enhanced information sharing, they have also made it easier to share individual work across the internet, exposing creative content to the risk of plagiarism. Plagiarism—the act of using another person's work without proper acknowledgment—is a form of theft that undermines the originality of the author. It deceives readers and misrepresents the creator's authenticity. For industries like journalism, plagiarism erodes trust, nullifies originality, and diminishes commercial value. Hence, organizations must adopt strict anti-plagiarism policies to uphold ethical standards, ensure unique information delivery, and maintain societal trust in diverse perspectives on truth.

Plagiarism, a significant ethical issue, has caused the ruin and humiliation of many creators across various fields, including literature, music, software, research papers, news, advertisements, and websites. Globally, there is a shared goal to preserve the originality of individual work by recognizing and giving credit to creators. Consequently, effective plagiarism detection (PD) has become a critical area of study. While PD tools have shown high performance with Latin-based languages like English, these tools often fail to detect plagiarism effectively in other languages due to linguistic differences.

One of the main challenges lies in understanding the unique linguistic characteristics of each language, including syntax, structure, semantics, and word morphology. High-performance models designed for languages like English are not suitable for other languages, such as Amharic, due to

these inherent differences. Understanding these linguistic features is crucial for developing language-specific plagiarism detection models. However, to the best of my knowledge, no comprehensive study has been conducted on plagiarism detection in Amharic text.

The primary objective of this thesis is to address this gap by focusing on plagiarism detection in Amharic text. This research, in my opinion, represents the first in-depth exploration of this issue for Amharic. While prior research has predominantly focused on English and European languages, Amharic, as an under-resourced language, requires specific attention. A model designed for English, for instance, cannot be applied to Amharic due to its distinct linguistic nature and structure. Therefore, developing plagiarism detection tools for Amharic is a necessity.

Detecting plagiarism in any language involves analyzing the meaning, context, and intent of text to identify similarities, even when terminology or sentence structure differs. For Amharic, semantic analysis—a sophisticated method for assessing meaning similarity—offers a promising approach to detecting plagiarism, particularly when entire documents are paraphrased. This research will focus on plagiarism detection in Amharic texts, specifically social media posts from news agencies. The data will include posts from 10 selected channels, spanning March 2020 to March 2021.

Developing an effective Amharic plagiarism detection model will require a combination of natural language processing (NLP) techniques and language-specific tools. Semantic similarity analysis provides a potential strategy for comparing text meanings and identifying plagiarism. However, further research is essential to build advanced techniques and resources tailored to the analysis of Amharic text.

The relevance of plagiarism detection research in Amharic is twofold. First, as an under-resourced language, Amharic faces significant challenges, including complex morphology, semantic ambiguities, and a lack of NLP tools and annotated data. This research aims to integrate existing knowledge and overcome these challenges to develop effective detection methods. Second, protecting the originality and commercial value of Amharic writers' work is vital for preserving the language's cultural and intellectual contributions.

Finally, in the thesis section which follows, it is intended to provide a statement of the problem, research questions, study objectives, significance of the study, the scope of the study, and organizations of the study.

1.2. STATEMENT OF THE PROBLEM

The drastic changes of technology from big computers to personal mobile technology advancements have changed our lifestyle for better through allowing us to access information at the tip of our fingerprints. Internet, one of the technological aspects, has revolutionized our lives and currently, it has become our preferred lifestyle. Nearly in everything we do and become affordable, we use it via Wi-Fi, mobile data, and ADSL. Internet assists us to find new innovations, exploring solutions for our daily information queries, and upscale our knowledge. Because of this practice, the Internet has brought a fundamental change throughout society, driving it forward from the industrial age to the networked era [5].

It has changed business, education, government, healthcare, world become small that distance doesn't matter and even the ways in which we interact with our family and is one of the key drivers. It is an easily accessible source of information for our needs and a tool for us to use when assistance is required. Our civilization is evolving into a modernized society that consumes a great deal of information. And information is coming at us from all sides like a flood [6]. Internet can facilitate an increasing number of resources for articles, journals, literature, news and many more.

Our global society is turning into a modernized culture that uses the internet for socialization using social media, and blogs to consume a large amount of information [7]. As an example, in the domain of news, journalists from all aspects of the news category, to the point where news articles written by different media writers can source other competitors', which is a common practice of sourcing articles from other news agencies to maintain their visibility, keep followers, and keep the momentum of the news outputs to inform their audiences. Without adding value to the news contents and acknowledging the source, it becomes problematic. This act of providing news enforces cultural homogenization, which may lead to news organizations' outputs becoming increasingly homogeneous [1]. Furthermore, it is a copyright violation to present someone's own work without mentioning the author/news agency. It is extremely difficult to detect such practices when it becomes a large number of news which cannot detect with manual inspection.

Because news coverage relies on public confidence, a posted content failure to acknowledge their sources raises questions on the agency’s judgment and legitimacy. News posting authors accused of plagiarism are frequently suspended/fire from their duties which we have experienced in CNN, ABC, and Fox news [10] while the accusations are investigated by the news organization which will maintain the news agency’s validity. The existence of Internet and online search engines has advanced international collaboration but at the same time it also has raised the plagiarism

opportunity.

In Ethiopia, news is shared on the internet in local languages such as Amharic as well as in other local languages. The advancement of information technology, such as social media, has created and facilitated a channel for journalists, bloggers, and authors to post Amharic contents online. As Amharic news posting content increases in Ethiopia, the technology can be misused by the writers and lead them to plagiarism. plagiarism in news items will have a substantial detrimental influencing the society by creating an atmosphere of mistrust, notably on the legitimacy and ownership of news organizations, as well as on the discipline's integrity and homogeneity reports that lucks credibility, creativity and criticality which results unlawful or unethical practices that breaches professional norms and society. According to Hiwot Molla [7], while some of the news articles are acquired from the Facebook pages of other organization's post, the news agency did not acknowledge the origins of the news stories while delivering the reports. We have never heard plagiarism in Ethiopia like we heard in BBC, CNN, or NBC [11], it doesn't imply that there is no plagiarism in Ethiopia. Rather we have no mechanism to detect it.

Therefore, methods and models capable of automated plagiarism detection become mandatory to expose legitimate work and exercise the principles of delivering ethical information to citizens. Plagiarism is becoming more prevalent in practically every aspect of modern life, including social media, journal articles, news, academy, and so on [8].

There are several types of plagiarism software, which bases on the researches, available for detecting plagiarism in other languages, as well as the characteristics of each type of software, such as which features are checked for plagiarism detection [2]. The online plagiarism detection software tools like Turnitin, EVE2, CopyCatchGold, Word Check, J Plag, PlagAware, iThenticate, PlagScan, and so on [12]. These tools detect any unlawful acts in the written works with their supporting languages, but not to Amharic. Existing researches rely on linguistic features like syntax rules or semantics that differ significantly across languages, limiting their generalizability. Because Amharic by its very nature can have meaning if the structure changes which makes unique. Example አበበ በሶ በላ፤ በሶ አበበ በላ፤ በላ አበበ በሶ. All have the same meaning even if the structure changes. And also can have many word extensions added to the original word that show multiple meanings by adding few character prefix, infix, or suffix. Example ጦውደድ፤ አልወድም፤ ወደድኩ፤ ወዘተ.

Existing plagiarism detection systems lack support for Amharic text and struggle to address the complex forms of plagiarism unique to the language. Additionally, they are not equipped to handle scenarios where a large volume of posts needs to be compared to determine if a piece of writing is original.

As a result, such uniqueness require to develop a model that can detect plagiarism at the character/text level, as well as to experiment with detection that can support plagiarism in Amharic language that can address both text and language level, by experimenting and adapting existing models from other languages and combining suitable detection models which fosters writer to do their best like in those internet independent days newspapers' written in a creative manner to present heterogeneous and creative in their original work.

1.3. RESEARCH QUESTION

The following research question will be answered in this paper for Amharic textual plagiarism detection:

- What are the challenges of detecting plagiarism in Amharic language?
- What is the best selection method for detecting plagiarism in Amharic language?
- Are these methods effective in detecting Amharic plagiarism?

1.4. OBJECTIVE

1.4.1. GENERAL OBJECTIVE

The general objective of this paper is to adapt a model that can detect plagiarism at the text level, experiment with detection to Amharic language news posts, adapting existing PD models from other languages to fit to Amharic, capable of identifying complex forms of textual similarity and ensuring broader applicability across various phrase level textual contexts, and implement optimal detection model that can foster writers to do their best in a creative manner that presents heterogeneous original work.

1.4.2. SPECIFIC OBJECTIVE

- Review the literature on plagiarism detection models for different languages and to have a better grasp of the strategies for implementation.
- Understanding of Amharic linguistic nature
- Collection of post of Amharic news digital resources from online repository

- Experiment using NLP, Amharic morphology, and different techniques on the collected data by choosing the appropriate methodology and tools
- Enhance accuracy in Amharic plagiarism detection through combining suitable detection models
- Discuss the findings in detail
- Draw conclusions and provide recommendations.

1.5. SIGNIFICANCE OF THE RESEARCH

The significance of this research lies in addressing the unique challenges posed by the Amharic language in the context of plagiarism detection. Amharic, with its complex morphology, syntax, and orthography, presents difficulties that are not encountered with many other languages. Current plagiarism detection tools, developed primarily for languages like English, are not capable of accurately detecting plagiarism in Amharic texts due linguistic variations & contextual meanings; so it executes the appropriate model, testing and improve existing model performance of detection in Amharic texts.

The research aims to bridge this gap by developing specialized plagiarism detection tools tailored specifically for Amharic. This is crucial because existing natural language processing tools and language models are limited for Amharic, making it difficult to handle the nuances of the language.

By advancing plagiarism detection technology for Amharic, the research contributes significantly to language technology development, benefiting a wide range of users including language learners, authors, and researchers. It promotes integrity, safeguards intellectual property, fosters original thought, and increases fairness by incorporating context-specific elements appropriate for Ethiopia.

The research also serves as a foundational study for further advancements in detecting plagiarism in Amharic, potentially leading to the development of the most effective detection models. The beneficiaries of this research include society at large. Institutions such as news agencies, individuals, and researchers.

1.6. SCOPE AND LIMITATION OF THE PROJECT

The scope of this project examines plagiarism detection for the Amharic language by collecting data from well-known news agency channels on social media platforms such as Telegram and Facebook. It identifies plagiarism in the context of a news agency by employing semantics for plagiarism made by the plagiarist.

The primary scope of this research is to develop an off-line Amharic textual plagiarism detection model that meets the needs of the language texts which doesn’t consider software graphs, music,

paintings, pictures, maps, technical drawings, tables, flowcharts, figure, etc. Discarding the figures and charts creates an undetectable appearance that people can manipulate and control. That means people can easily plagiarize figures and charts without current plagiarism detection models catching them [13].

But the scope of this research is to design plagiarism detection for Amharic language texts; which is monolingual not referring to cross-lingual in relatively smaller domains, selected news writers posts on social media; doesn't consider academic research papers, poems, novel writings, short stories. Furthermore, I presume that the suspicious document is just plain text or a word document not pdf or other modals like eps, tiff, etc. It doesn't include other local languages like Oromigna, Tigrigna, Guragigna etc, translations of local and foreign languages, and complex paraphrased sentences, idea/conceptual detection, online Amharic posted hyperlinked and source code detection.

Syntactic structures that have the same meaning by its nature in Amharic, due to Amharic synthetically have the same meaning SOV or SVO or VSO or VOS. For example አበበ በሶ በላ ፣ በሶ አበበ በላ ፣ በላ አበበ በሶ ፣ በላ በሶ አበበ. As seen in the example, the synthetic forms for the Amharic texts stated as SOV, convey the same notion of meaning in various forms so more attention is paid in this research to duplicate and semantic texts.

The experiment is limited to Amharic words synonym not antonym and light weight paraphrased words in time constraints. As there is no standard corpus to use, this research uses texts extracted from Amharic-Amharic dictionary. Because of resource limitations, the experimental research collects the data from an online relevant number of news entities posts.

1.7. ORGANIZATION OF THE STUDY

There are five chapters in this research. The first chapter, Introduction, discusses the history of plagiarism detection in connection to news writers and subject matter, as well as the research questions, problem statement, objectives, scope, and significance of the study. The second chapter is a literature review that gives conceptual and contextual foundation in the existing body of knowledge on Amharic plagiarism detection in news organizations. The research design and methodology employed in this study are presented in the third chapter. The data collected from

research participants is examined, interpreted, and the results are presented in chapter four. The final chapter, chapter five, will include a discussion, conclusion, and recommendations.

1.8. CHAPTER SUMMARY

The context of the research and the motivations for undertaking this research were presented in this chapter, which served as an introduction to the thesis. This was followed by a statement of the problem with the research's objectives, scope and limitation as well as its significance contribution to field. Finally, the chapter included a thesis outline as well as a summary of each chapter.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. OVERVIEW

This chapter provides a foundation for the study investigation by reviewing the literature on plagiarism detection. This section of the research looks at ways to identify plagiarism in texts written in other different languages, as well as studies on textual plagiarism that have been conducted using various techniques. To build a proper model for Amharic language Plagiarism Detection (APD), the models built by different scholar's strengths and shortcomings will be explored in detail. The chapter begins with a definition of Plagiarism Detection, followed by a discussion of its essential characteristics, as well as enabling technologies that aid in comprehending the detection phenomenon. Different plagiarism detection techniques and deployment methodologies are investigated, with the benefits and drawbacks of each will be addressed. In addition, the benefits that drive plagiarism detection adoption are addressed, as well as the drawbacks that may deter adoption.

The objective and purpose of this chapter was to evaluate the strengths and weaknesses of existing plagiarism detection approaches with the ability to combine and come up with the best solutions for Amharic plagiarism detection, as well as to analyze the shifting of trends and adopting to the models in the plagiarism detection field. Since the research done for Amharic language, we'll start with the nature of the language and move on to the methods details afterwards in exploring to identify PD.

2.2. PLAGIARISM DEFINITION

One of the problems with dealing with plagiarism is that the term "plagiarism" is often misunderstood. Plagiarism according to Oxford Learner's Dictionary, Derived from the Latin "plagiarius" which means "presenting someone else's work or ideas as your own, with or without their consent, by incorporating it into your work without full acknowledgement". In Amharic to Amharic Dictionary define as "ስርቆት (plagiarism) ማለት የሌላውን ሰው ሥራ ወይም ሀሳብ ለምንጩ ተገቢ ዕውቅና ሳይሰጡ በመውሰድ እና የራስዎ አድርጎ መጠቀም ነው።" According to literature [14]], "define plagiarism as an illegal quotation of someone else's effort, whatever effort was it (an idea, invention, writing, methodology, design, etc.), and in different ways such as copy-paste function,

by paraphrasing without exact citation."

Scholars, on the other hand, do not appear to agree on a single definition of plagiarism. This controversy arose as a result of a variety of academic wrong doing, including self-plagiarism [15]. As related to this issue, one question that may arise is how can researcher steal their own ideas? The study included 158 lecturers who were 92.2 percent Caucasian, 1.9 percent Multiethnic, 1.3 percent Asian, 1.3 percent Hispanic, and 0.7 percent Native American or Alaskan Native and the study found that instructors were still unsure if reusing self-work is considered self-plagiarism after distributing questionnaires to the selected research participants [16].

Borrowing basic sentences should not be considered plagiarism, according to Yilmaz [17], whereas duplicating some sentences that do not include an original concept is of minor value, according to Bouville [18]. It appears abstract in the sense of perfect understanding plagiarism, but it becomes extremely evident in the effort to prevent not committing.

From my perspective, there is nothing wrong with taking self-own research craft and presenting it as my own. After all, claiming someone else's work as one's own not comparable as applying to self own work. As a result, I define plagiarism as "Theft of knowledge which isn't commonly acknowledged as common knowledge or our own or purchases material written by someone else, without the author's consent, by putting it into your work without full acknowledgement". Currently, the number of detection tools has increased greatly due to the proliferation of the plagiarism problem [19].

2.3. BROAD TYPES OF PLAGIARISM

Plagiarism can be found in non-structured/Natural like written text and structured/Non-Natural like source code [20]. In relation to languages, there are two types of language that can be plagiarized.

2.3.1. Non-natural / Source code plagiarism and

2.3.2. Natural language plagiarism

2.3.1. NON-NATURAL / SOURCE CODE PLAGIARISM

This sort of plagiarism involves duplicating a software's source code by translating it from one computer language to another or using the same programming languages. A set of keywords, structures, functions (public/protected/private), meaningful variables, and so on are used in source

languages. Plagiarism in source code is a growing problem in computer science (CS) education, particularly in programming courses [21]. Since the scope of this research is not addressing source code, I will focus on the next topic.

2.3.2. NATURAL LANGUAGE PLAGIARISM

Natural language processing is a technique for detecting plagiarism in text. Plagiarism can be happen to any human languages. In the world more than 6,000 human languages [22] and very few among these languages resourced digitally. In terms of linguistic plagiarism detection, there are two methods for detecting plagiarism [23].

2.3.2.1. **Monolingual Plagiarism Detection:** Monolingual plagiarism detection deals with the automatic identification and extraction of plagiarism in a homogeneous language setting, e.g., English–English plagiarism. Most of the plagiarism detection models have been developed for monolingual detection, which is divided into two former tasks, offline and online [23]

2.3.2.2. **Cross-Lingual Plagiarism Detection:** Plagiarism detection like this is used in conjunction with text translation from one language to another. Due to the availability of language translation services such as Google Translation and variety of websites and applications, translation has become incredibly simple, allowing for quick plagiarization. It has become well-known and famous for this this type detection important. In the last several years, there has been a lot of interest in research on cross-lingual plagiarism detection [24].

2.4. TYPES OF MONOLINGUAL PLAGIARISM

Plagiarism is might include copying, paraphrasing, metaphoring, or taking appropriate ideas as a whole. Plagiarism can be categorized into 7 types in natural languages.

2.4.1. EXACT/ CLONE/ COPY AND PASTE PLAGIARISM

Exact / Clone / Copy and Paste Plagiarism without using quotation marks occurs when a student takes an exact word or modifies a phrase or paragraph from the source without citing it. The majority of the structure of a sentence/phrases/ paragraph is the same or missing some of the words in the middle in this type of plagiarism.

Copying a full text without a quotation from a website or an Electronic file (e.g. articles in word

or PDF format) and passing it off as one's own or copying sections of text from one or more sources and passing it off as one's own [25].

2.4.2. INTELLIGENT PLAGIARISM

This type of plagiarism uses the intelligent way rather than the lazy one which is direct copy and paste. Under this, there are various types of intelligent plagiarism

2.4.2.1. **Redrafting/Paraphrasing:** Paraphrase has been defined in several studies as the process of rephrasing information using linguistic characteristics (both semantic and syntactic) such as synonym substitution, word form modification, and sentence reorganization [26]. In under redrafting, there are two types of paraphrase plagiarism

2.4.2.1.1. **Simple paraphrase plagiarism:** This type of paraphrase is used to substitute synonyms, alter phrase order, or modify grammatical patterns, or when a statement has to be rewritten to be lexically and syntactically distinct while yet being semantically identical [27].

2.4.2.1.2. **Complex paraphrase plagiarism:** This type of plagiarism is made by combining basic paraphrasing with the ideas of other writers into one. It's more difficult to spot because it's a collection of thoughts rather than one's own [27].

2.4.2.2. **Metaphor/personification plagiarism:** are utilized to better and more comprehensively explain the thoughts of others [28].

2.4.2.3. **Concept/ Idea plagiarism:** is stealing someone else's idea and claim it own [27]

2.4.2.4. **Illegitimate source / 404:** Plagiarism from an unauthorized source happens when an author offers references to sources/citation that are erroneous. [27].

2.4.2.5. **Plagiarism through translation:** Stealing a text, concept, replacing synonyms, or paraphrasing from another language, such as French, Chinese, Spanish, or another language, and translate and presenting it as one's own is an example of translation plagiarism. This type of plagiarism can use the approve types of plagiarism techniques in addition to translation like paraphrasing, metaphor, or taking concepts.

2.5. PLAGIARISM DETECTION APPOACHES

There are two approaches to detect plagiarism among the documents

2.5.1. Human based Detection and

2.5.2. Technology based Detection

2.5.1. Human detection: The most basic method of detecting plagiarism in this case is to look for inconsistencies on several levels, including inconsistencies in the presented ideas, theories, hypotheses, and thoughts, inconsistencies in the text's writing style, inconsistencies in the presentation of stating the arguments, and inconsistencies in the bibliography and sources cited. The expert in that specific subject will recognize whether it is derived from a well-known author's article via experience. However, it is incredibly difficult for a person to remember the details until the concept is distinguished. In addition, if we want to detect plagiarism, we are demanded to bring experts on that specific domain to identify plagiarism that requires availability of expertise whenever we require; and the number of Amharic digital documents has also grown tremendously in recent years. This is not feasible and as a result, Automatic text technological detection is crucial as a result of this.

2.5.2. Technology detection: The technological detection programs automatically check the content of a text and compare it to either a repository of textual papers kept in the corpus or Internet content via urls, links, and online accessible libraries.

In text plagiarism detection, mainly two formal tasks are defined which are extrinsic/external detection and intrinsic/internal detection, which in turn defines the two types of PDS [29]. The suspected documents are compared against a reference source corpus in extrinsic PDS.

Unlike extrinsic PDS, a reference corpus is unavailable for intrinsic PDS. Here the suspicious document is analyzed single-handedly without being compared with any sources. The writing styles of the author, structural distributions, vocabulary richness etc. are analyzed here [23]. Thus different features are extracted for identifying these plagiarism cases.

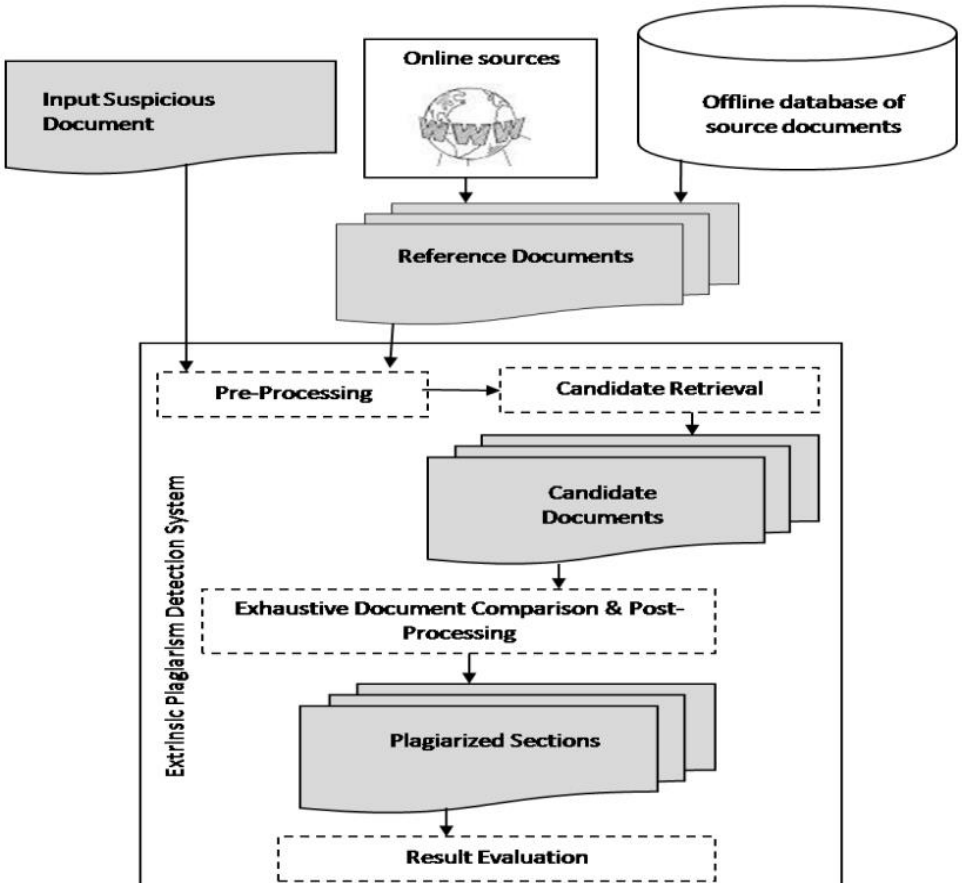

Figure 1. General Architecture of Extrinsic

2.6. TEXTUAL FEATURES FOR EXTRINSIC PLAGIARISM DETECTION

In extrinsic PDS, various detection techniques can be employed, viz., string based, vector space model (VSM) based, syntax based, semantic based, structural based and citation based techniques or a combination of these techniques called hybrid [30].

Different textual features for extrinsic plagiarism detection can be found in literature documents reviewed, which fall into one of the following methods: Lexical detection methods, Syntax-based detection methods, Semantics-based detection methods, or a combination of methods.

Lexical-based Detection: The characters works at a document n-gram or paragraph n-gram or character n-gram or word n-gram levels are the only thing that lexical detection consider. Character-based n-gram (CNG) re presentation is the simplest form whereby a document d is represented as a sequence of characters d = {(c1 , d), (c2 , d), . . . ,(cn , d)}, where (ci, d) refers to

the ith character in d, and n = d is the length of d (in characters). On the other hand, word-based n-gram (WNG) represents d as a collection of words d = {(w1 , d), (w2 , d), . . . , (wn , d)}, where (wi, d) refers to the i^{th} word in d, and n = d is the length of d (in words) with ignoring sentence and structural bounds. Simple WNGs may be constructed by using bigrams (word- 2-grams), trigrams (word-3-grams) or larger. CNG and WNG are commonly called fingerprints or shingles in text retrieval and plagiarism detection research. The process of generating fingerprints (or shingles) is called fingerprinting (or shingling). A document fingerprint can, therefore, identify the document uniquely as well as a human fingerprint does.

Syntax-based detection: algorithms took into consideration the sentence structure exhibited in part of speech (POS) of phrases and words in distinct statements using fundamental POS tags such as verbs, nouns, pronouns, adjectives, adverbs, prepositions, conjunctions, and interjections [23]. POS tagging is the task of marking up the words in a text or more precisely in a statement as corresponding to a particular POS tag. Sentence-based representation works by splitting the text into statements with the use of end-of-sentences delimiters, such as full stops, exclamation, and question marks. After splitting the text into sentences, POS and phrase structures can be constructed by using POS taggers. On the other hand, chunks is another feature that is generated by so-called windowing or sliding windows to characterize bigger text than phrases or sentences. POS could be further used in windowing to generate more expressive POS chunks. Word order, in a sentence or a chunk, could further be combined as a feature, and used as a comparison scheme between sentences. Sentence-based representation works by splitting the text into statements with the use of end-of-sentences delimiters, such as full stops, exclamation, and question marks. After splitting the text into sentences, POS and phrase structures can be constructed by using POS taggers. On the other hand, chunks is another feature that is generated by so-called windowing or sliding windows to characterize bigger text than phrases or sentences. POS could be further used in windowing to generate more expressive POS chunks. Word order, in a sentence or a chunk, could further be combined as a feature, and used as a comparison scheme between sentences.

Semantics-based detection: The meaning of phrases, paragraphs, or texts is compared using semantic-based detection algorithms [23]. Semantic features quantify the use of word classes, synonyms, antonyms, hypernyms, and hyponyms. The use of thesaurus dictionaries and lexical databases, WordNet, for instance, would significantly provide more insights into the semantic meaning of the text.

2.6.2. PLAGIARISM DETECTION METHODS

For this research to be of the highest caliber and integrity, plagiarism detection is an essential responsibility. In this review of the literature, we examine the most current developments and difficulties in computer techniques for detecting plagiarism in academic texts. We divide the present methods into three major categories: machine learning, semantic analysis, and text matching. Text-matching techniques evaluate the similarity between the suspect document and a reference corpus using lexical elements like n-grams or word hashes. Semantic analysis techniques use methods like natural language processing, information retrieval, or citation analysis to go beyond the surface level and attempt to capture the meaning and structure of the texts. Machine learning techniques develop a model that can discriminate between authentic and plagiarized texts using a variety of variables and classifiers. We discuss the strengths and limitations of each group of methods, as well as the evaluation metrics and datasets used in the literature. We conclude with some directions for future research in plagiarism detection.

Plagiarism detection has its own set of difficulties. The approaches utilized in certain text phases have an impact on detection accuracy. One of the primary issues that impacts detection accuracy and computational complexity is the curse of modification and synonym replacement. To tackle this, understanding the detection methods is mandatory and explained here below:

2.8.1.1. LEXICAL DETECTION APPROACHES

Lexical detection methods exclusively consider the characters in a text for similarity computation. The methods are best suited for identifying copy-and-paste plagiarism that exhibits little to no obfuscation. To detect obfuscated plagiarism, the lexical detection methods must be combined with more sophisticated NLP approaches [31]. Lexical detection approaches typically fall into one of the two categories we describe in the following: n-gram comparisons, and vector space models.

N-gram Comparisons. Comparing n-grams refers to determining the similarity of sequences of n consecutive entities, which are typically characters or words and less frequently phrases or sentences. n-gram comparisons are widely applied for candidate retrieval or the seeding phase of the detailed analysis stage in extrinsic monolingual and cross-language detection approaches as well as in intrinsic detection.

Approaches using n-gram comparisons first split a document into (possibly overlapping) n-grams, which they use to create a set-based representation of the document or passage (“fingerprint”). To

enable efficient retrieval, most approaches store fingerprints in index data structures. To speed up the comparison of individual fingerprints, some approaches hash or compress the n-grams that form the finger- prints. Hashing or compression reduces the lengths of the strings under comparison and allows performing computationally more efficient numerical comparisons. However, hashing introduces the risk of false positives due to hash collisions. Therefore, hashed or compressed fingerprinting is more commonly applied for the candidate retrieval stage, in which achieving high recall is more important than achieving high precision. The state of art in exhaustive detection stage of a PDS is analyzed and studied based on different techniques and methodologies presented by renowned authors [30].

Fingerprinting is widely applied extrinsic plagiarism [29]. The purpose is to reduce the size of the compared texts and speed up the comparison without missing a significant match. A document fingerprint is a list of integers resulting from hashing substrings of the document. The comparison is then performed on the fingerprint rather than the whole text [11]. The process of creating a fingerprint involves three steps:

- Chunking: the document is segmented into substrings (called chunks or minutiae). A chunk might be a sequence of letters, words or even sentences.
- Hashing: a hash function is applied to the chunks to generate list of integers.

Selection: The final fingerprint is a subsequence of the list of hashes [32].

2.8.1.2. STRING BASED DETECTION APPROACHES

This includes the simplest level of comparison where character level/ word level comparisons are made. Mainly N-gram based comparisons either character N-grams or word N-grams fall into this category [30].

Torrejón and Ramos [35] extracted the contextual and surrounding N-grams which are extended N-gram models. They used sorted word 3-grams and sorted word 1-skip-3- grams. The accuracy dropped as the paraphrasing complexity increased. Non-overlapping 250 character chunks are extracted by Kuppers and Conrad [36]. Then the word-based similarity is computed using the dice coefficient and a threshold is used to detect plagiarized fragments. The overall model performance was poor due to the extremely low recall [30].

All these detection models were effective for detecting plagiarism cases with simple copy-paste

and intelligent plagiarism cases with small random shuffling while the efficiency of detection dropped as plagiarism complexity increased. In general, N-gram based models were found to be less effective when it comes to complex obfuscation types. But the exhibition of good precision shows its potential to be combined and used in hybrid approaches [30].

2.8.1.3. VECTOR SPACE MODELS (VSM) APPROACH

This is one of the popular techniques which utilizes the lexical and syntactic features and represent the document in a vector space. Then different weighting schemes are adopted for document representations and comparisons. Mainly the two weighting schemes used are term frequency inverse document frequency (tf-idf) and term frequency inverse sentence frequency (tf-isf), where the former operates at document level and latter at sentence level. The former is used in both candidate retrieval and exhaustive analysis stage while tf-isf is mainly used in exhaustive analysis [30]. In [30], The proposed model focuses on the importance of paraphrases in detecting plagiarism, both mono-lingual and cross-lingual. To investigate the detection challenges, the authors examined the efficacy of an external plagiarism detection model based on the Vector Space Model (VSM) on the PAN-PC-2011 corpus. The model employed only 250 documents as corpus and 20 documents as suspect documents. And the outcome of Monolingual Simulated Plagdet Score (0.0524298), Recall (0.0293390), Precision (0.3780321), and Granularity (1.0541872), and when used in conjunction with any synonym addition mechanism, such as the thesaurus or dictionary or wordnet, this strategy may be more effective [33].

2.8.1.4. SEMANTICS BASED APPROACH

In semantic based techniques the meaning representation of a document is focused and is found to be efficient for paraphrased detections. Latent Dirichlet Allocation (LDA), machine learning techniques, soft computing techniques etc. fall into Semantic based Approaches [30]. Semantic similarity is used to identify the extent to which two or more terms or sentences are conceptually similar but not necessarily lexically similar [34]. In general, words from various phrases and documents are mapped, and their associations are measured in the terms of those sentences and documents, respectively, to calculate semantic similarity. Semantic similarity measure is a crucial idea in linguistics for detecting phrase or document copying.

A fuzzy based similarity approach was used in exhaustive analysis stage by Alzahrani and Salim [36] where fuzzy based semantic similarity metric computa- tions are employed. Alzahrani and

Salim [36] have introduced a statement-based plagiarism detection system for Arabic (FS-APD) using fuzzy-set information retrieval model. The degree of similarity between two statements is computed and compared to a fixed threshold value to judge whether are similar or not. This approach led to perform well on verbatim reproductions. To address the rewording, they have proposed another model named fuzzy semantic-based string similarity for extrinsic plagiarism detection (SFS-APD) [37]. This uses a shingling algorithm, Arabic WordNet lexical database [38] and Jaccard coefficient for retrieving a list of candidate documents. The suspicious document is then compared sentence by sentence with the candidate documents to compute the fuzzy degree of similarity [32]. Also, analyzed their pros and cons, and reported in a tabular form in the table below.

#	Method	Type		Lingual		Mode		Language	Recommended types of plagiarism
		Intrinsic	Extrinsic	Mono	Cross	Clone	Intelligent		
1	Character-Based	X	✓	✓	X	✓	X	Any	Copy-Paste
2	Vector-Based	X	✓	✓	X	✓	X	Any	Copy-Paste
3	Syntax-Based	X	✓	✓	X	✓	X	Specific	Copy-Paste, Sentence order switching
4	Semantic-Based LDA	X	✓	✓	X	✓	✓	Specific	Copy-Paste, structural alterations, Paraphrase
5	Fuzzy-Based	X	✓	✓	X	✓	✓	Specific	Copy-Paste, structural alterations, Paraphrase

Table 1- Plagiarism detection Methods based on features of Types, Lingual, and Mode [28].

2.7. SEMANTICS PLAGIARISM DETECTION MODELS

Semantic plagiarism detection models are algorithms intended to detect plagiarism based on the context and meaning of language rather than just the words or syntax at the most basic level. These methods examine the content of the believed plagiarized text and contrast it with a reference document to spot ideas, concepts, and themes that are the same or different from one another.

By spotting similarities that might go undetected by more conventional plagiarism detection techniques, semantic plagiarism detection models seek to increase the accuracy of plagiarism detection. This is especially crucial when the copied text has been rephrased, rewritten, or translated into another language.

1. Lexical-based approaches: This approach involves comparing the vocabulary of the suspected plagiarized material with the original content to determine similarities and differences. This method uses techniques like n-gram analysis, synonym matching, and word embeddings to detect similarities.

2. Syntactic-based approaches: This approach involves analyzing the syntax and structure of sentences to identify plagiarism. This method uses techniques such as part-of-speech tagging, dependency parsing, and sentence alignment to detect inconsistencies and similarities.

3. Semantic-based approaches: This approach involves analyzing the meaning of the content to identify plagiarism. This method uses techniques such as named entity recognition, topic modeling, and semantic similarity to identify similar themes, ideas, and concepts between the original and the suspected plagiarized content.

4. Hybrid approaches: A combination of the above approaches may be used to improve the accuracy of plagiarism detection. For instance, a combination of lexical and semantic-based approaches can provide more accurate results.

This study focuses on sentence similarity by contextual meaning in a hybrid mode, semantics is the popular detection method to implement. Text modification, translation, and concept adoption are common examples of intelligent plagiarism. This type of plagiarism may be detected using semantic, citation-based, or style-based approaches [39].

2.8. RELATED WORK

Many efforts was done to identify plagiarism that bases on the types of plagiarisms. One may detect the plagiarism to specific type but not to other type of plagiarism. The most powerful detection of disguised phrases can be done using semantic similarity measures either LSI or LDA or Fuzzy semantic popular models irrespective of their types. It can detect Verbatim/Cloning, near copy, restructuring, and paraphrasing.

Sanchez-Perez, Sidorov and Gelbukh [42] presented a tf-isf weighting scheme for the exhaustive analysis stage with cosine and dice similarity metrics. VSM approaches are also limited to detection of copy-paste and plagiarism.

In Filip Cristian [43], proposed a model to detect external plagiarism with the use of Authentic Cop, with the objective of detecting instances of plagiarism in Computer Science academic writings and purposes, the authors experimented with the cosine similarity and term frequency-inverse document frequency (tf-idf)weighting schemes on the PAN 2011 corpus and 1000 randomly selected documents, the authors performed preprocessing, removing stop words, and applying stemming, the authors performed preprocessing, removing stop words, The tested whether there is a distinction between plagiarized and non-plagiarized passages in these circumstances, i.e. if there is a threshold over which the majority of pairs with similarity greater than are plagiarized but the majority of pairs with similarity less than are not, and that are highly similar under the cosine similarity with tf-idf weighting, the prosed model cannot compare in detail the candidate selection phase will retrieve only the documents from which the current text is most likely to plagiarize; the proposed model tested N-gram sizes of 3, 4, and 5 atoms and discovered that using 4-grams produces the best results in the absence of normalization and stemming; the proposed model demonstrated its effectiveness by performing an application, yielding Precision (0.7609), Recall (0.3377), and Granularity (1.2653). The algorithms used in the candidate selection and detailed analysis phases should be further benchmarked for other combinations of threshold values, and the program should be enhanced to handle other document formats, be accessible via a user-friendly web interface, and include semantic analysis components and stemming support [43].

A method for detecting external plagiarism has been presented by integrating semantic relationships between words and their syntactic composition, this method can improve the performance of plagiarism detection by avoiding selecting source text sentences that are highly similar to suspicious text sentences but have a different meaning. This model was tested on the PAN-PC- 10 and PAN-PC-11 datasets using stop words extracted from the English language [43]. To test and compare the proposed technique's performance, we applied the approach to 200 previously used suspect documents and source documents using four different standard metrics (macro-average Precision, Recall, F-measure, and granularity). The outcome of Plagiarism Detection Using Linguistic Knowledge (PDLK) on PAN-PC-11 systems is as follows: the PDLK has a precision of (0.902), a recall of (0.702), an F-measure of (0.790. However, this result is based on just 200 documents out of 22000 documents, indicating that it does not operate on all datasets [43].

Alzahrani's [23] model propose through four main steps: (1) Pre-processing which includes tokenization and stop-word removal, (2) retrieve a list of candidate source documents for each

suspicious document using n-gram fingerprinting and Jaccard coefficient, (3) An in-depth comparison between the suspicious documents and the associated source candidate documents using k-overlapping approach, (4) Post-processing where consecutive n-grams are joined to form united plagiarized segments [32].

"The suggested model introduces an External Plagiarism Detection System (EPDS) that makes use of the Semantic Role Labeling (SRL) approach, as well as semantic and syntactic information" [44]. The suggested method is capable of detecting several types of plagiarism, including exact verbatim copying, paraphrase, phrase transformation, and word structure modification. The suggested algorithm operates on the English-language portion of the data set, analyzing 800 questionable papers and their related originals. These papers are divided into two different datasets at random (training and test dataset). The training data set has 450 papers and 350 test documents. Precision (0.921), Recall (0.622), F1 (0.743), Plagdet (0.737), and Granularity (0.737) are the results of the assessed method on the PAN-PC-11 dataset (1.011) [45]. The amount employed in testing and training is little, as the total number of English books is 22,000, and the model do not assess the effect of stop words on text relevancy [33]. Meni [46] proposed a plagiarism detection tool for Arabic documents (Aplag). Aplag is based on heuristics to compare suspect documents at different hierarchical levels to avoid unnecessary comparisons. In addition, to address the problem of rewording, Aplag replaces each word's root by the most frequent synonym extracted from Arabic WordNet [45].

Lovepreet [47] makes use of semantic information to detect duplicated material in the absence of human intervention; the suggested model makes use of an extrinsic plagiarism detection methodology that is cognitive in nature. The model determines the semantic similarity between two phrases using the Dice measure, which is backed by a lexical database such as WordNet. It also makes use of linguistic characteristics such as path similarities and depth calculation to determine the similarity between two words, which are blended using different weights. It is capable of detecting restructure, paraphrase, exact copying, and synonymized plagiarism, among other things. The PAN-PC-11 corpus was used to assess it. The proposed model is expected to produce results equivalent to or slightly better than those produced by existing models, and source articles are preprocessed using NLP features. The model preprocessed the data by normalizing the text, segmenting it into sentences, removing stop-words, and lemmatizing the words, using the English language only. The model's precision (0.934), recall (0.861), and F1-measure (0.875) values are as follows (0.875). A drawback of the proposed approach is its inability to detect complicated examples of textual plagiarism, such as input text summary information and translated

text. This method is unable to detect more intricate instances of plagiarism that are manually altered [33].

Latent Dirichlet Allocation (LDA) is a popular topic modeling technique used for plagiarism detection in natural language texts. LDA identifies similarities between documents by analyzing the frequency and co-occurrence of words within each document and across multiple documents. This helps identify duplicate or near-duplicate content, enabling more accurate plagiarism detection. According to Adnen Mahmoud and Mounir Zrigui, the PAN was an international competition that provided a corpus of 20, 611 suspect and 20, 611 source documents to assess the efficacy of plagiarism detection systems. It used random obfuscation techniques such as paraphrase (the substitution of synonyms, antonyms, and hyponyms) and sentence reorganization. For huge vocabulary, TF-IDF was slow and unable to derive semantic links between words. LDA, on the other hand, outperformed the other classical approaches (TF-IDF and LSA) with 80.7% precision, 82% recall, and 81.34% F-measure.

2.9. CHAPTER SUMMERY

This chapter discusses and reviews many studies conducted by different researchers on the issue of plagiarism detection. The studies are evaluated in accordance with the language in which they were written. Despite the fact that certain works are investigated in a few languages, the English language is central to many works on the subject. These investigations are carried out utilizing various statistical and semantic methodologies, which serve as an input to the experiment.

CHAPTER THREE

3. AMHARIC LANGUAGE

3.1. AMHARIC LANGUAGE BACKGROUND

Ethiopia is a country which is located on the north-eastern part of African continent. It is highly ethnically and linguistically diverse, with about 80 different languages spoken in its territory [48]. According to Wikipedia, Ethiopian languages have different language families which is Semitic (Tigtingna, Geez, Amharic, Guragigna,etc), Cushitic (Oromigna, Agewigna, Sidamigna, Hadigna, Solmaligna, Afarigna, etc), Omotic (Wolayttigna, Bench, Dorzigna, Hamerigna,etc), Nilo Saharan (Anuakigna, Bertigna, Gumuzigna, Majangigna, etc) and some unclassified languages.

Amharic is Ethiopia's second most widely spoken language, after Afan Oromo. When the number of people who speak Amharic as their first and second language live in all over Ethiopia. Amharic is the most widely spoken Semitic language today, second to Arabic. Amharic is an Ethiopian official language, owing with its widespread use based on second-language speakers. Amharic is one amongst those different well-spoken languages and has been spoken in most parts of Ethiopia as mother tongue or secondary. According to the 2007 census, Amharic speakers encompass 26.9% of Ethiopia's population [2]. It has estimated that more than 29 million can speak inside Ethiopia and neighboring countries. The language serves as the working and one of the amongst officially marked languages of Ethiopia. Many kinds of research have been made on the language in higher educational institutions for fulfilment of Master and PHD.

Amharic is written today using the Ge'ez alphabet called fidel, a unique tabular presentation unlike English. The Amharic alphabet as it's usually known over 30 consonantal letters with each of which has six extended versions. Each character represents a consonant+vowel sequence, however the consonant determines the fundamental form of each letter, which is changed for the vowel.

The language widely used in the education, economy, government offices, including in the public media. In concern to media, many journalists uses various mediums to post the national and international news events. Every minutes journalists post their articles to their respective organizations online media.

Amharic language is well spoken in every part of Ethiopia. Even spoken in neighboring countries.

In addition to Afan Oromo, Ethiopia's most generally spoken language, Afar, Somali, and Tigrigna, Ethiopia's federal government utilizes Amharic as a working language. It is used as a first and second language throughout the country, and it is used as a working language in many areas. The Amharic language is a Semitic language that is second most widely spoken after Arabic. Ge'ez has been updated into Amharic. Ge'ez was the first language spoken in Ethiopia [49]. In terms of writing system, the Amharic languages have their roots in Ge'ez. Amharic, unlike many other Semitic languages such as Arabic and Hebrew, has a horizontal writing style derived from the Ge'ez script and is written or read from left to right.

3.1. AMHARIC WRITING

3.1.1. ALPHABET

The Amharic language has 33 consonant symbols with seven different phonetic variants created by combining the consonants with different vowels. There are a total of 279 characters in the Amharic characters table, allowing it to have its own scripts. Fidel is the name of each character, and they each have their own distinct tone. Each syllable pattern in written Amharic reproduces the seven vowel sounds or phonetics. The basic form is the first fundamental letter; the others are derived from it by more or less regular alterations signifying the various vowels added to the fundamental letter like l for ለ, l+u for ሉ, l+i for ሊ, l+a for ላ, l+i+e for ሌ, l+e for ል, l+o for ሎ. The chart of Amharic is presented in the APPENDIX 1.

3.1.2. PUNCTUATION MARK

Every language has its own set of punctuation marks for conveying information correctly. Amharic, too, has its unique set of punctuation symbols. Amharic has around ten punctuation signals. 'Hulet Neteb' (':') is used as a word separator, and 'Arat Neteb' ('::') is used as a sentence separator. However, instead of using 'Hulet Neteb' (':') [50]. IT professionals were influenced by English and used space instead of Hulet Neteb (':'). The Amharic writing system has borrowed a few punctuation marks from foreign languages, such as the question mark '?'. The language's complete list of punctuation marks can be found in the Annex section. The dominant word order type of the languages like English uses Subject-Verb- Object order.

The explain the few basic punctuation marks in Amharic writing system are

- Hulet Neteb (two square dots arranged like colon, :, used as a word-separator like space in english and in current computer writing space in amharic) and
- Arat Netib (sentence-separator, four square dots arranged in a square pattern :: like full stop in english).

The other common type of punctuation marks in the language are:‘Netela Sereze’ (፣), an equivalent of comma, and

- ‘Derib Sereze’ (፤), which is the equivalent of semi-colon, can also be used as a list separator. All the lists of the Amharic punctuation on APPENDIX 2.

3.1.3. MORPHOLOGY

Morphology is the study of how words are built from smaller meaning-bearing elements (morphemes) [51]. A morpheme is commonly described as the smallest unit of meaning in a language. For example, ዳኞች a combination of the root word with the affixes which is either prefix, infix or suffix, ዳኝ + ች. Affixes do add additional meaning from the original. In English, Judges contains the root word which is morpheme ‘judge’ and another morpheme ‘s’. Morphology have its own distinction in Amharic unlike English langiage which might add before, in the middle, or at the end.

Affixes are characters that are appended to the root word. Prefixes, suffixes, infixes, and circumfixes are examples of affixes. Prefixes are added to the preceding character, Suffixes are added at the end of the stem word, infixes are added inside the stem, and circumfixes are words added to both the beginning and the end of the stem word, or attaching a prefix and suffix to the root word. For example, የዳኞች is circumfixes because ‘የ’added at the beginning, ዳኝ the stem word, ‘ች’ added the end of the stem word.

3.1.4. INFLECTION

The process of word construction by changing words to represent distinct grammatical categories is known as inflection. Number, definiteness, cases (accusative/objective, possessive/genitive), and gender can all be inflected in Amharic nouns [52]. On the contrary, verbs in Amharic can be inflected for any combinations of person, gender, number, case, tense/aspect and mood which results, from a single verbal root, thousands of verbs will be generated consisting a “root + vowels” merger [52]. For example, words like “ሰበረው” (säbäräw), “ሰበረች” (säbäräč), “ሰበርን” (säbärn),

“አሰበረ” (assäbärä), “ተሰበረ” (täsäbärä), “አልሰበረም” (alsäbäräm), “ካልተሰበረ” (kaltäsäbärä), “የሚሰበር” (yämisäbär), etc.

3.1.5. DERIVATION

Amharic nouns can be generated from adjectives, verbal roots by embedding vowels between consonants, stems, stem-like verbs, and nouns themselves through the derivation process [51].

In order to correctly manage Amharic writing, we must first comprehend how to get to the root word. To deal with this, we’ll need to know a lot about stemming, syntax, and semantics.

3.2. ISSUES OF AMHARIC WRITING

There are a number of issues related with the Amharic writing system making the natural language processing of Amharic documents a bit more challenging [50]. Some of these problems are:

- Redundancy of some characters
- Compound words: there is no standard way of writing Amharic compound words which enables to use space or hyphen [53].
- Spelling variation of same words (the same word can be written in various forms) like judges for ዳኛዎች, ዳኞች, ዳኘአዎች or the word ‘ቴሌቪዥን’ (‘television’) can be written as ቴሌቭዢን, ቴሌቭዥን, ቴሌቪዥን [54].
- Inconsistency of abbreviations of Amharic words [55]. All these mentioned issues pose challenges since the same word is treated in a completely different form within the process of feature preparation creating style differences between authors for a classifier

English	Character	Other form/s of the character
H	ሀ	ሐ, ኀ
S	ሰ	ሠ
A	አ	ዐ
Ts	ጸ	ፀ

Table 2 Redundancy of Characters [50]

3.3. HANDLING THE ISSUE OF AMHARIC WRITING

3.3.1. STEMMING

The process of removing the last few letters of a word, which usually results in wrong meanings and spelling, is known as stemming. It becomes more difficult in morphologically rich languages, such as Amharic, and more crucial in languages with a lot of inflectional morphology. To understand stemmer in Amharic, consider the preceding phrase. The words ‘ሰባበረ’ and ‘ሰባበሩት’ are derived from the root word ‘ሰብር’, however when we stem the word, the final few letters are removed, and the word is turned to ‘ሰባበ’ or ‘ሰባበሩ’. We will be wasting our time if we only utilize the stemmer. To achieve the right word format and contextual meanings, lemmatization is essential.

Lemmatization considers the context and converts the word to its meaningful base form, which is called Lemma. Lemmas are the base forms of all inflectional forms of words whereas a stem basically isn’t. In Amharic, these two tasks are usually used interchangeably with one serving as a replacement for the other, but the real meaning have their own differences.

3.3.2. NORMALIZATION

Normalization is the process of transforming text into a single canonical form. Because input cannot be guaranteed to be consistent before operations are performed on it, normalizing text before processing ensures that issues are separated. Compound words might be presented in a variety of ways to help us establish a logical framework. Some characters in table 1 that are redundant can be standardized to the same format. Inconsistent abbreviation can also degrade the quality of data processing. As a result, by normalizing them, we can alleviate them.

3.3.3. SEMANTIC STRUCTURE

The study of meaning is semantics, a part of linguistics. A linguistic theory that looks at the meaning of words. The meaning of a word is fully represented by its surroundings. In this case, a word’s meaning is determined by its contextual relationships the preceding or succeeding words. As a result, a difference is drawn between the meanings in participation with other words.

Lexical elements can also be categorized semantically based on whether their meanings come from single lexical elements or from their context. Lexical elements form predictable patterns of association with one another in hypernym, synonymy, and antonym, as well as homonymy, are some of the relationships between lexical terms.

3.4. CHAPTER SUMMERY

This chapter's focus is on the Amharic language's alphabet, punctuation marks, morphology and normalization. There are certain issues with Amharic textual writing that need to be addressed, such as redundancy of texts written in various characters but with the same sound and meanings, and the lack of a standard of a few words that allows for writing in space and with hyphens. Inconsistent abbreviations and spelling variances also have an influence. Therefore, this chapter concluded with on how to manage the Amahric texts to properly process in computations.

CHAPTER FOUR

4. RESEARCH DESIGN AND METHODOLOGY

4.1. OVERVIEW

Any research approach should be determined by the nature of the research problem. The term methodology refers to the overall methods and perspectives to the research process as a whole, and it is concerned with the following main topics, such as why did the researcher collect specific data? What data is collected? Where did the data come from? And how was the data collected analyzed? in related to the research question.

In this chapter, a research methodology for creating a sample as well as approaches for achieving research goals and answering the research question. Figure 1 depicts the suggested architecture, which is made up of four primary layers: input-source and destination texts, preprocessing with morphological analysis, Fingerprint matching & n-gram to detect direct or minor change, and semantics detection if any paraphrasing exists.

Research design refers to the procedures that will be utilized to acquire relevant data and the techniques that will be used to analyze them. It explains the procedures for gathering and analyzing data that aids in answering the research question.

4.2. RESEARCH DESIGN

Research design is the manner by which researchers must conduct their studies. It displays how researchers formulate their problem and objective, as well as how they present their conclusions based on data gathered during the research period. This chapter on research design and technique also shows how the data was acquired, processed, and what actions were taken to find the appropriate results that would meet the study's objectives.

As a result, this chapter discusses the research approaches used during the research process. It encompasses all aspect of the study's research methodology, from the research strategy to the dissemination of the results. To adequately meet the research experiment, this study adopted a quantitative research approach. Creswell [56] defines research technique as the selection of a research approach based on the study's nature, whether qualitative, quantitative, or mixed. Data for the study was gathered using quantitative approaches from a variety of news story sources, including telegram. As a result, in order to draw correct general conclusions, this study used a main data strategy to collect data and confirm findings from diverse data sources.

Furthermore, this study enables a researcher to assess data in a specific domain, namely news. In most cases, the study approach selects a limited news domain or a small number of news organizations in a specific period of time to detect plagiarism. The purpose of this study is to find plagiarism in Amharic. It will need to collect data from Amharic news articles and documents from Telegram online posts in order to do so.

4.3. RESEARCH APPROACH

The selection of a research approach is also based on the nature of the research problem or issue being addressed, the researchers' personal experiences, and the audiences for the study. As per the most known classification, there are mainly three research approaches: qualitative, quantitative, and mixed methods. Unquestionably, the three approaches are not as discrete as they first appear. Qualitative and quantitative approaches should not be viewed as rigid, distinct categories, polar opposites, or dichotomies.

Qualitative research is an approach for exploring and understanding the meaning individuals or groups ascribe to a social or human problem. The process of research involves emerging questions and procedures, data typically collected in the participant's setting, data analysis inductively building from particulars to general themes, and the researcher making interpretations of the meaning of the data. Quantitative research is an approach for testing objective theories by examining the relationship among variables. These variables, in turn, can be measured, typically on instruments, so that numbered data can be analyzed using statistical procedures. The final written report has a set structure consisting of introduction, literature and theory, methods, results, and discussion. Whereas mixed methods research is an approach to inquiry involving collecting both quantitative and qualitative data, integrating the two forms of data, and using distinct designs that may involve philosophical assumptions and theoretical frameworks. The core assumption of this form of inquiry is that the combination of qualitative and quantitative approaches provides a more complete understanding of a research problem than either approach alone. Therefore, a quantitative research approach is used in conducting this research.

4.4. SOURCES OF DATA

There are many ways of classifying data. A common classification is based upon who collected the data. Data that has been collected from first-hand-experience is known as primary data. Primary data has not been published yet and is more reliable, authentic and objective. Primary data has not been changed or altered by human beings; therefore its validity is greater than secondary data. Data

collected from a source that has already been published in any form is called as secondary data. Secondary data is essential, since it is impossible to conduct a new survey that can adequately capture past change and/or developments [57].This research has used primary data collected through collecting of posts on social media for this specific research.

4.5. METHODS OF DATA GATHERING

Data collection is the systematic process of acquiring and measuring information on variables of interest in order to answer research questions, test hypotheses, and evaluate outcomes [57]. Primary data is used in this study. The study used public channel News Articles uploaded specifically to social media platforms like Telegram for primary data collecting. The reason is that most of the news agencies posts the same articles on different social media outlets, like Telegram and Facebook, but the contents of that specific agency's posts are the same. The study drew on a variety of data sources for different news agencies.

The methods used in this study to obtain primary data included collecting news stories posted on telegram from various news writers such as Fana Broadcasting/ FBC, Walta, LTV, Addis Media Network / AMN, Prime Minister Office, ESAT, and Ethiopian Broadcasting / EBC, BBC Amharic, Getu Temesgen, TikVAH, and Elias Meseret . To find the similarity, more than 100 news pieces were gathered. Because the study's goals are to have a diverse public channel source of pages and a representative set of news reports in document format, the sources were carefully selected.

On a social media platform like Telegram, there are numerous sample measures or criteria that can be used to choose these public pages or blogs. The following sampling criteria and metrics were used to select a public page from the categories in this study:

- A page posts daily news stories on a variety of topics, including politics, economics, and international news.
- An Amharic-language public channel.
- The amount of followers who have an impact

Telegram is utilized in this study since it is a commonly used platform in Ethiopia that is growing in popularity. Telegram is easily accessible offline, which influences whether or not any incorrectly reported news has an impact on society.

4.6. FEATURE EXTRACTION

Amharic extrinsic hybrid plagiarism detection on first layer for candidate selection and second layer for semantic plagiarism detection involves analyzing the meaning of words and phrases in a document to detect plagiarism. Some key features of Amharic plagiarism detection are:

1. Verbatim detection: This feature involves detecting instances where a writer has used copy and paste without referring the source can be a sign of plagiarism which detected in the 1- layer.

2. Sentence structure: This feature involves analyzing sentence structures to detect instances where a writer has copied the exact sentence structure of the source material, which can indicate plagiarism which detected in the first layer.

3. Paraphrasing: This feature involves comparing the original source material with the text in question to detect any instances of paraphrasing. If the writer has used similar words or phrases, but rephrased them to avoid being detected for plagiarism or replaced with similar word/phrase, this detection feature can flag it which detected in the second layer.

4. Contextual analysis: This feature involves analyzing the surrounding context of a text to detect if writing is original or if there is a risk of plagiarism. Identifying key concepts and context provides a deeper understanding of the meaning and intent of texts which detected in the second layer.

Based on the above and in relation to this research, verbatim with sentence structure at the initial matching stage for selecting suspected contents; then to context-related problems will be examined based on the subsequent next detection stage for context analysis.

4.1. SAMPLING METHODS AND TECHNIQUES

A sample design is a method for selecting a representative sample from a population. Before any data is gathered, the sample design is determined. A census is an endeavor to collect data on every member of a population. As a result, to conduct an in-depth analysis, a census of all social media pieces from the past to the present is required, which is extremely difficult to accomplish. Purposive sampling approaches are used if the quantity of news articles is very huge. In addition, sampling is taken to manage the dataset manageability with a significant number of data that can

align to the resource I have.

Purposive sampling was used to collect data for this study based on the categories of online government news agencies, private owned news agencies, individuals/bloggers/journalists, and international organizations by using the criteria on posts daily news on a variety of topics (including politics, economics, and international news), Amharic-language public channel, and the amount of followers who have an impact based on data availability and heterogeneity. As a result, the population was all social media posts in Amharic in relation to politics, events, economy and the sampling derived for this study from social media like Telegram. The Telegram posts were chosen because the number of growth in Ethiopia increasing due to fit for small group channel to control, monitoring, easy to collect feedbacks using polls, and are simple to collect and export for these research.

4.2. CORPUS PREPARATION

After the news articles collected, a preparation process follows, which are collecting, cleaning, filtering, and consolidating data into txt files. This process used preparation tools such as notepad or others. Filtering and Cleaning the data primarily use for the next stage, which is the annotation of the post in the articles and then after used for designing a training model for Amharic plagiarism detection.

4.3. ARCHITECTURE

The three main stages of [47] method are preprocessing, computing resemblance between sentence pairs, and filtering out the plagiarized sentence pair. If the root word which is stem compared from the source with the destination document and found any similarity with more than the expected threshold, then it is plagiarized. But if in some case the plagiarist try to modify and paraphrase the message using some similar words from the original document, a way to manage context is required.

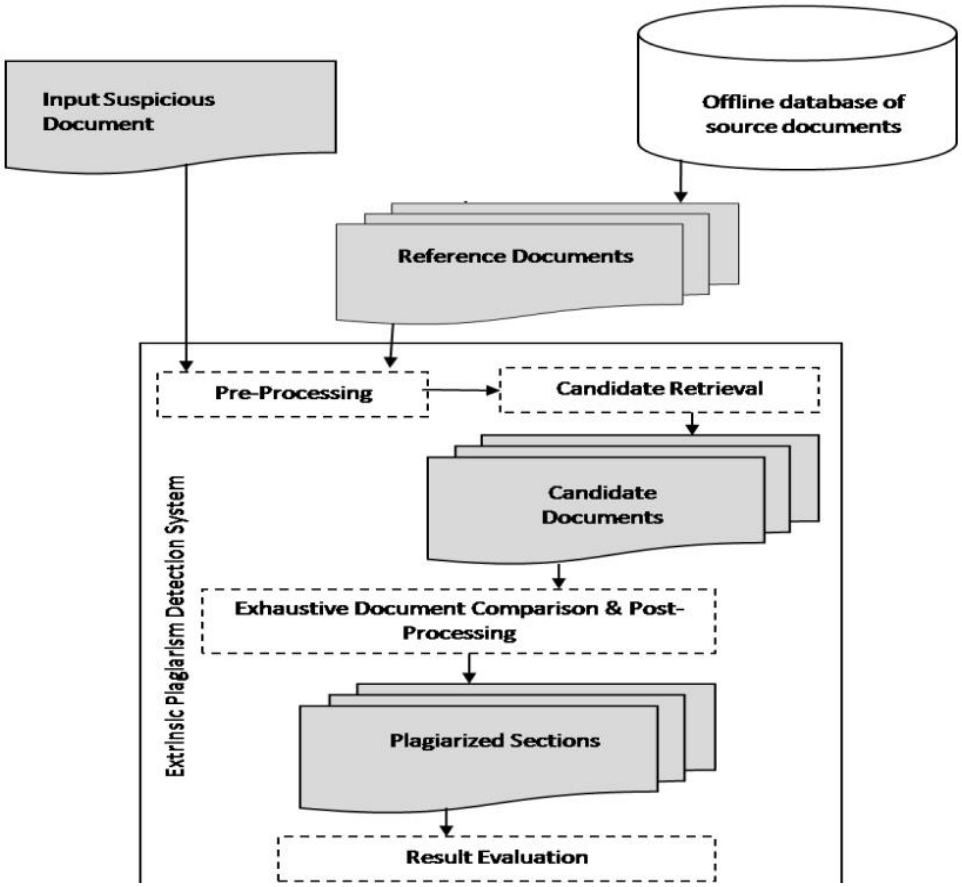

Figure 2. General workflow of the proposed model

In the proposed Architecture, suspected and source documents are preprocessed by enforcing the NLP features onto them. The proposed Architecture preprocessed the documents by normalizing the text, segmenting the text into sentences, removing the stop-words, lemmatizing, and morphologically analyze.

In an extrinsic PDS, a suspicious document is compared to a reference source document corpus or collection. The reference sources are compared to the suspect document. Reference documents may be subjected to some pre-processing when working offline and with limited resources. As a result, a heuristic retrieval process is used to find close duplicates, which are referred to as candidate documents for the suspicious document in question [58]. Candidate retrieval in the generic representation follows pre-processing. To identify blocks of text in quotation marks where possible direct reference will be ignored because implies referring the authors directly.

Normalizing the text—there can be many irrelevant characters in the text that are not used while computing the resemblance among the documents. The Amharic characters like comma (፣),

semicolon (፤), brackets {}/(), special characters, quotes “”, full stop (።), white spaces (፡), etc. need not be there while computing the resemblance. Further, the word can be present in the form of an abbreviation or different spelling that needs to be normalized.

Algorithm

```
Define Tokens
For each document File (F)
        For each document F
        Tokenize into unigram term (token) by space
        Token. Remove ()
If token t tokenized
t.getText ().replaceAll (‘’) // to remove symbols, any special character etc.
End
End
```

Segmentation—the normalized text in the suspected and the source documents is segmented into sentences. Both documents are then represented as a list of sentences. Segmentation of the text is done by using “’፤”,”።”, “?” and “!” symbols.

```
Define Segmentation: “’፤”,”።”, “?” and “!”
For each document File (F)
        For each document F
        Tokenize into unigram term (token) by Segmentation
If token t tokenized
t.getText ().writetoarray() // to remove symbols, any special character etc.
End
End
```

Eliminating stop-words—Stop-words appear very often in the text like articles, propositions, and conjunctions and these words do not bear any meaning. Upon removal of these words from the text can reduce the computation time of the model, and also improves efficiency and accuracy. The proposed model eliminates the stop-words residing in the APPENDEX 3 of Amharic stop-

```
Define stop-words
For each document File (F)
        For each term
        If token t compare term is in Stop-Words
        t.remove()
        If not
        t.getText ().writetoarray() // to remove symbols, any special character etc.
End
End
```

After pre-processing the next important stage is the document-level plagiarism detection by retrieving the verbatim sources. Usually in any practical scenario, for a detection model the suspected document has to be compared with large repositories or databases which can be some

offline databases specific to an application. In any case, the exhaustive comparison of a suspected document with all the documents in these databases will be quite time consuming. Thus to reduce this search space a document level comparison is done which retrieves the candidate sources for the given suspicious document at hand. In candidate retrieval task, the similar source documents with respect to a particular suspected document are retrieved using Jaccard Coefficient. Thus each suspicious document is associated with a source set termed as candidate set [30].

Fingerprinting Model

Fingerprint-based approaches identify plagiarism by comparing strings in papers based on common fingerprint proportions, which are character sequences found throughout the document. The purpose is to reduce the size of the compared texts and speed up the comparison without missing a significant match and set to 20% to indicate a potential intelligent plagiarism.

N-gram

N-gram based approach identify plagiarism by comparing bi-gram, tri-gram, and so on which the characters sequences found throughout the candidate document versus the actual in the corpus. The purpose is to reduce the size of the compared texts and speed up the comparison without missing a significant match and set to 20% to indicate a potential intelligent plagiarism.

Latent Dirichlet Allocation Model

If the fingerprint didn't find any plagiarism in the suspect or less than the threshold, it is forwarded to LDA model to check the semantics of the sentences. Regarding threshold, the maximum token plagiarism threshold for the whole document to mark as plagiarized if it exceeds the accepted threshold of the Latent Dirichlet Allocation similar is 80%.

Fuzzy Semantics

If the fingerprint didn't find any plagiarism in the suspect or less than the threshold, it is forwarded to Fuzzy model to check the semantics of the sentences. Regarding threshold, the maximum token plagiarism threshold for the whole document to mark as plagiarized if it exceeds the accepted threshold of the Fuzzy semantics similar is 80%.

CHAPTER FIVE

5. EXPERMENT AND EVALUATION

5.1. INTRODUCTION

This chapter addresses the experimentation and evaluation for the research work, which is Amharic plagiarism detection on news agencies to detect any verbatim or paraphrased publications that are not acknowledged. To accomplish all of this, a setup is performed on the acquired corpus, which is initially utilized as an input for plagiarism detection. The experiment setup then describes the setting in which the analysis and experiment to deliver the results will be conducted. The findings will be presented and interpreted as they are discovered.

5.2. EXPERIMENTAL SETUP

The input files are organized separately and run with different codes to get the results of the relationship among all other factors such as features and algorithm models used for extrinsic hybrid plagiarism detection. The first experiment is based on the candidate selection features which chooses n-gram and fingerprint. These features focus on the first layer candidate selection is made if the similarity between the documents above 0.2 using n-gram and fingerprint if only the verbatim obsification checked. Then after the first layer candidate selected documents that the similarity between the documents fulfill above the threshold, it goes to the next level which is checking the semantic detection. If the semantic similarity above 0.8, it highly probable that one of the document plagiarized from another one.

The two experiments are performed by using HP EliteBook 840 G1 12th Gen Intel(R) Core(TM) i7-1255U 1.70 GHz, 16.0 GB (15.7 GB usable), cores 10/ logical processor 12, 1TB SSD storage laptop. The Integrated development environment was the Windows version operating system of Windows 11. Nnltk, numpy, sklearn.feature_extraction.text, sklearn.decomposition, LatentDirichletAllocation, f1_score, recall_score, precision_score, accuracy_score, os Python 3.10.9 | packaged by Anaconda, Inc. | (main, Mar 1 2023, 18:18:15) [MSC v.1916 64 bit (AMD64)] was broadly used for these experiments. Due to grid search hyper parameter tuning and performance intensity after the fifth interactions of both experiments on two of the algorithms such as Fingerprint and Latent Dirichlet Allocation vs n-gram and Latent Dirichlet Allocation was used. The performance took 5 minute to generate the results.

For this study, 10 sources of Amharic channels articles posts with 211 days posts, 100 documents, 807 sentences, tokenized into 3,357 tokens, and of those tokens 2,275 tokens are unique keys. All

were written by different authors with different title and in different time presented in different ways with same events for politics, businesses, economy, social, and national agendas. Regarding threshold, the maximum token plagiarism threshold for the whole document to mark as plagiarized if it exceeds the accepted threshold of the Latent Dirichlet Allocation similar is 80% [60].

The tokens were given to python ferret tool to make the candidates select from the corpus which is formatted as txt. The ferret code which will help to compare similarity of texts that instructs specifically to Anaconda. "Ferret tool is incredibly simple to install and run in a Windows system that can handle the file types.txt,.rtf,.doc, and.pdf" [61]. "Ferret takes a collection of documents and turns each text into a unique trigram and reference number. A list of file-pairs with similarity scores that are ranked from the most similar pair to the least similar pair is produced after each text is compared to each other based on how many different trigrams are similar across the texts" [61].

"The number of similar trigrams in a pair of documents divided by the total number of unique trigrams in the pair is used to calculate the similarity measure. Ferret exhibits similarity scores that are precisely 0.90991" [61]. "For analytical purposes, the figures were rounded to the nearest whole number; in this instance, 0.91 was used because they are shown side by side with related portions highlighted, the method enables users to choose any pair of texts and conduct further research" [61].

The corpus consists various channels presented as follows in table

Sr no	Channel	Category	No of Documents	No of Sentences	No of Tokens	No of Unique Tokens
1.	Fana Broadcasting/ FBC	Government	20	135	1727	1036
2.	Walta	Government	13	92	877	526
3.	Addis Media Network / AMN	Government	7	65	619	371
4.	Ethiopian Broadcasting / EBC	Government	10	84	800	480
5.	BBC Amharic	International	14	97	924	554

		News agency				
6.	ESAT	Private	11	81	772	463
7.	LTV	Private	8	56	533	320
8.	Elias Meseret	Blogger	8	88	838	503
9.	Getu Temesegen	Blogger	9	109	1039	623
		Total	100	807	8129	4877
10.	Paraphrased by 5 News Authors persons (Experience from 2 Authors from Newspaper, 2 from Radio, 2 from TV)	Paraphrasers/considered as plagiarists/	10	103	981	543
		Total	110	910	9110	5420

Table 3 corpus document against the samples taken from various channels on Telegram.

CHAPTER SIX

6. RESULTS AND DISCUSSION

Introduction

The experiment process begins once all of the prerequisites have been met, as outlined in the preceding chapter. This chapter starts with a discussion and study of the characteristics of the collected corpus, which was initially utilized as an input for the entire sources that generated the corpus. The experiment setup then represents the context in which the initial corpora are physically processed and the results are analyzed. Finally, the output results are presented in forms with a brief description of the experiments' conclusions.

Data collection was one of the key tasks for the online author identification. Three online text sources were used to collect the dataset from 11 authors of Amharic text: Telegram public channel, personal blogs and online magazines. 20 Amharic authors‟ corpus dataset was collected from a previous research conducted by Baher Hussen in 2020 at the School of Information Science in Addis Ababa University [9].The experimenter collected the corpus from two sources: Kumneger Magazine and The Reporter Ethiopia Media & Communications Center. Characteristics of the authors from these sources are described in Table 5 below. Table 5 shows the name, the number of articles, the domain area, sources used, number of words and characters per article for each individual author. A closer look at the table shows that the researcher used the same domain area, journalist, and same source, magazine for all authors.

Data collection was one of the key tasks for extrinsic plagiarism detection. The news posts from the social media which was collected from 10 sources. Out of the 10, 7 of the data source where from organizations like Fana Broadcasting/ FBC, Walta, Addis Media Network / AMN, Ethiopian Broadcasting / EBC, BBC Amharic, ESAT, LTV. Out of 7, 4 of them are government owned news agencies, 2 of them are commercial media owned by private, 1 of them is an international news agency. The remaining 2 are owned by individuals as a public channel who can broadcast news to the general public using their sources to report the event. This individuals mainly focused to the information unlike company policies.

The variety of the news posters intentionally diversified by choosing frequent posts who report as a news with relation to social, health, economy, politics in addition to journalist views. This can be viewed on table 4.

To experiment on the word similarity the proposed algorithm which can collect both words from query and corpus documents can base on two word features to evaluate the detection. The first feature can check for lexical matching of both words from query and document stored using either fingerprint or n-gram algorithm. The second feature is semantic matching of both words from query and document stored using LDA or fuzzy algorithm. If the two features not recognized any plagiarism or less than the allotted threshold, the suspect against the corpus are not plagiarized. As mentioned earlier four different experiments were conducted to examine how the number of articles affects the accuracy, precision, recall and f1 score of models generated using features. These experiments included n gram with LDA, n-gram with fuzzy, fingerprint with LDA and fingerprint with fuzzy techniques.

Model Selection

According to the findings presented in table 6 this study has achieved encouraging outcomes when compared with research. This research used 2-gram, 3-gram, and 4-gram features which founds presented as follows with the format (suspect, corpus, percentage)

1. **n-gram with LDA**

n=2 and LDA
Accuracy: 0.9629629629629629
Recall: 0.9096045197740112
Precision: 0.930635838150289
F1 Score: 0.92

n=3 and LDA
Accuracy: 0.9310344827586207
Recall: 0.6479750778816199
Precision: 0.8851063829787233
F1 Score: 0.7482014388489209

n=4 and LDA
Accuracy: 0.9230769230769231

Recall: 0.8768718801996672
Precision: 0.8887015177065767
F1 Score: 0.882747068676717

n=5 and LDA
Accuracy: 0.8
Recall: 0.422680412371134
Precision: 0.7454545454545455
F1 Score: 0.5394736842105263

n=7 and LDA
Accuracy: 0.7777777777777778
Recall: 0.35233160621761656
Precision: 0.8192771084337349
F1 Score: 0.4927536231884058

n=10 and LDA
Accuracy: 0.7777777777777778
Recall: 0.4419642857142857
Precision: 0.868421052631579
F1 Score: 0.5857988165680473

2. Fingerprint and LDA

Accuracy: 0.972972972972973
Recall: 0.9175654853620955
Precision: 0.9682926829268292
F1 Score: 0.9422468354430379

3. Fingerprint and Fuzzy

F1 Score: 0.6666666666666666
Recall: 0.5
Precision: 1.0
Accuracy: 1.0

Features Comparison

The feature compared as above shows the results obtained for n-gram starting from n= 2 upto n = 10. Using bi-gram performed with LDA, Accuracy: 0.9629629629629629, Recall: 0.9096045197740112, Precision: 0.930635838150289, and F1 Score: 0.92. The performance show best results as compared to the other results with trigram up n=10. This presents in Amharic Unicode extrinsic text, the candidate selection with sentiment plagiarism detection using LDA, can have the best outcome with n=2 with LDA. The accuracy shows 96% with 100 Amharic news posts.

The feature using fingerprint with LDA results, Accuracy: 0.972972972972973, Recall: 0.9175654853620955, Precision: 0.9682926829268292, and F1 Score: 0.9422468354430379.
The performance as it is presented, outperform the bi-gram with LDA.

Another feature used to check the research is using fingerprint with fuzzy. The result found F1 Score: 0.6666666666666666, Recall: 0.5, Precision: 1.0, and Accuracy: 1.0. As compared to the bi-gram with LDA and finger print with LDA, it shows low performance. The bi-gram with LDA and fingerprint with LDA has better detection capability. The following table presents the results in table. The feature used n-gram with fuzzy kept failing after it took 6 hours to check. Therefore, cannot be able to present the results with its performance matrix. As far as the time taking to evaluate, it took more than 6 hours that cannot be applicable as compared to the other model models.

	F1	**Precision**	**Recall**	**Accuracy**	**Rank**
Bi-gram with LDA	0.92	0.930635838150289	0.9096045197740112	0.9629629629629629	2
Tri-gram with LDA	0.7482014388489209	0.8851063829787233	0.6479750778816199	0.9310344827586207	4
N=4 with LDA	0.882747068676717	0.8887015177065767	0.8768718801996672	0.9230769230769231	3
N=5 with LDA	0.5394736842105263	0.7454545454545455	0.422680412371134	0.8	7
N=7 with LDA	0.4927536231884058	0.8192771084337349	0.35233160621761656	0.7777777777777778	8
N=10 and LDA	0.5857988165680473	0.868421052631579	0.4419642857142857	0.7777777777777778	6
N-gram with fuzzy	-	-	-	-	N/A
Fingerprint and Fuzzy	1.0	1.0	0.5	0.6666666666666666	5
Fingerprint and LDA	**0.9422468354430379**	**0.9682926829268292**	**0.9175654853620955**	**0.972972972972973**	**1**

Table 4 Results of the various models against F1, Recall, Precision, and Accuracy

Graphic presentation of results

To tell if a plagiarism detection model is good against F1, recall, precision, and accuracy results, can look at the following:

The F1 score is a harmonic mean of precision and recall, which means it considers how effectively the model recognizes plagiarized material (precision) as well as how well it avoids classifying non-plagiarized text as plagiarized text (recall). A high F1 score suggests that the model is functioning well across the board.

Recall is the percentage of plagiarized content identified properly by the model. A high recall indicates that the model does not contain any plagiarized text.

Precision is the fraction of text identified as plagiarized by the model that is actually plagiarized. With a high precision, the model does not label any non-plagiarized material as plagiarized.

Accuracy is the percentage of text that the model properly classifies as plagiarized or non-plagiarized. A high level of accuracy indicates that the model is working well overall.

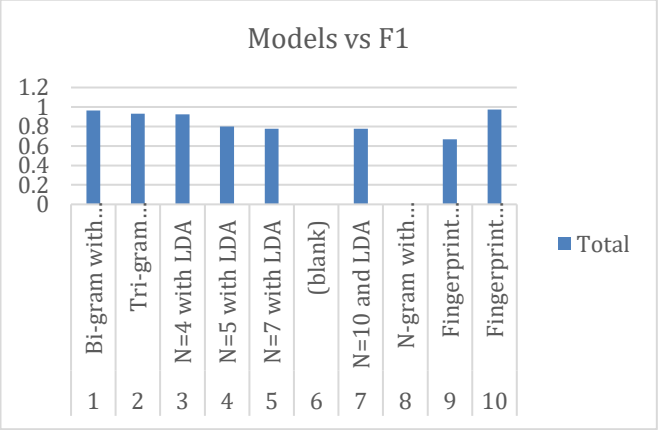

Figure 3 Models versus F1 shows 1st layer Fingerprint candidate selection and 2nd layer semantics with Fuzzy beat all

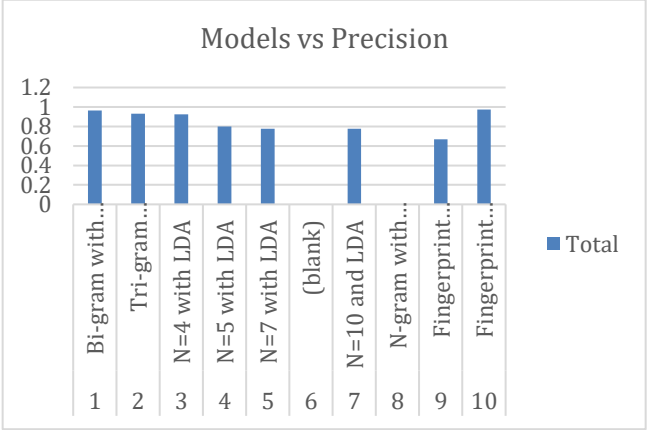

Figure 4 Models versus Precision shows 1st layer Fingerprint candidate selection and 2nd layer semantics with Fuzzy beat all

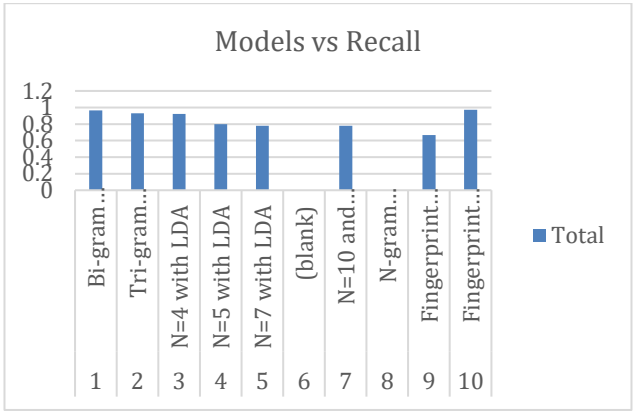

Figure 5 Models versus Recall shows 1st layer Fingerprint candidate selection and 2nd layer semantics with LDA beat all

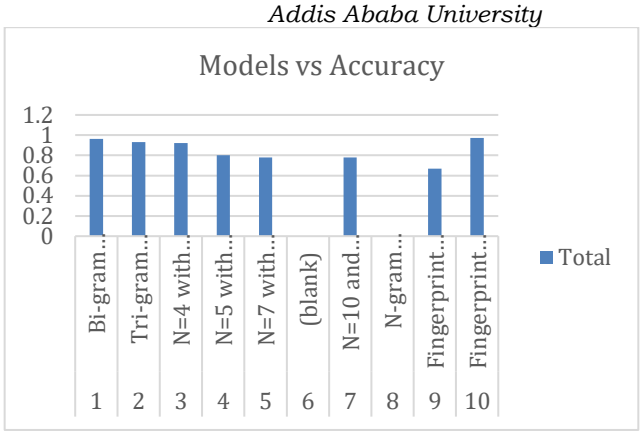

Figure 6 Models versus Accuracy shows 1st layer Fingerprint candidate selection and 2nd layer semantics with LDA beat all

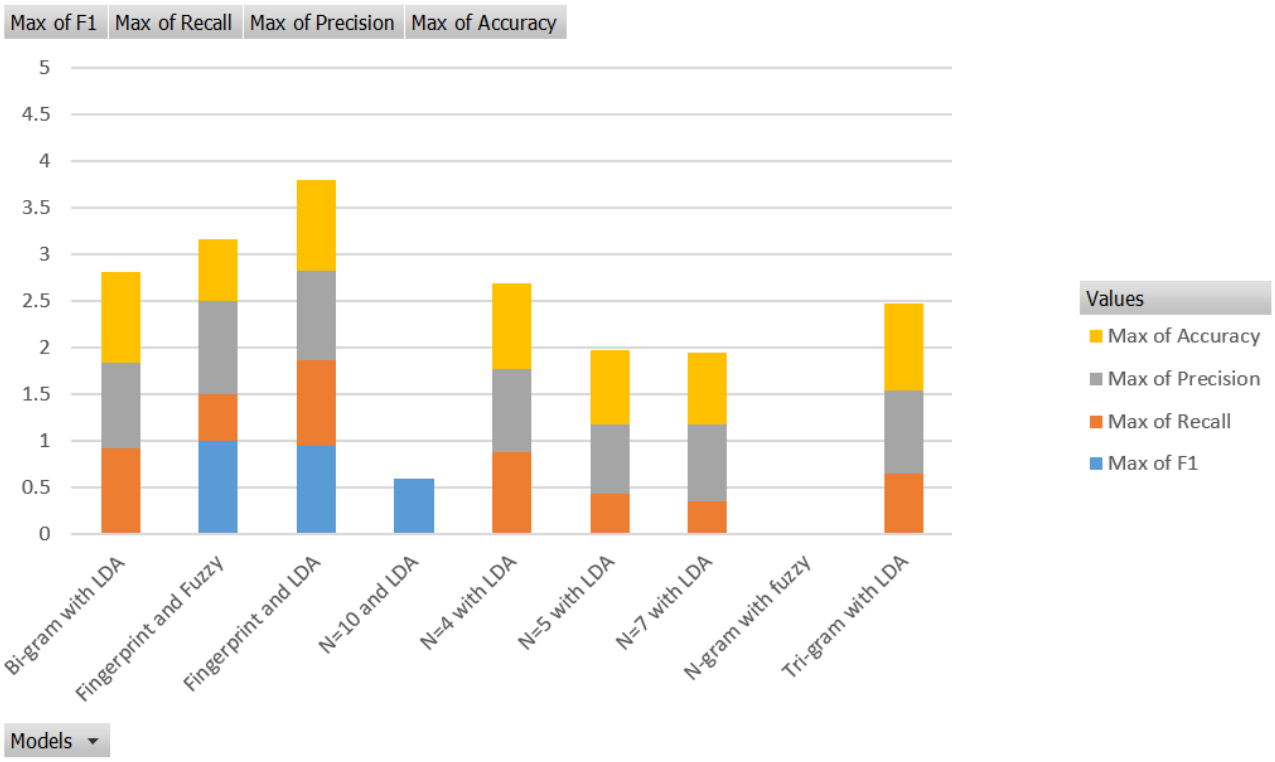

Figure 7 Models against F1, Precision, Recall, and Accuracy comparison converted to out of 5

6.3. SUMMARY

In this section the results of the different experimented models are described and interpreted with respect to the research question. All models contributed different findings which uses the Amharic texts posted on the social media as news. All the models presented different results which is presented in the above sections. The various models presented show good results except the n-gram with fuzzy plagiarism detection model.

The two-layer plagiarism detection experiment you described is interesting because it uses a combination of fingerprint-based or n-gram and LDA-based or fuzzy semantics techniques. Fingerprint-based techniques are good at identifying plagiarism, while LDA-based techniques are good at identifying fine-grained plagiarism. By combining these two techniques, the experiment was able to achieve a high overall accuracy.

One interpretation of the results of the experiment is that the two layers of the system are complementary. The fingerprint-based layer is able to quickly and efficiently identify plagiarism passages, while the LDA-based layer is able to accurately identify the semantics plagiarism strings. This suggests that the two layers can be used together to improve the overall performance of the plagiarism detection system.

Another interpretation of the results of the experiment is that the two layers of the system are synergistic. This means that the two layers work together to produce a result that is greater than the sum of its parts. For example, the fingerprint-based layer may be able to identify a plagiarism at document level, but the LDA-based layer may be able to identify the exact plagiarism strings within that document. This information can then be used to provide more detailed feedback to the user about the plagiarism.

Overall, the two-layer plagiarism detection experiment is a promising example of how different plagiarism detection techniques can be combined to improve the overall performance of the system. The experiment shows that fingerprint-based and LDA-based techniques can be used together to achieve high accuracy with all the performance matrixes in detecting Amharic plagiarism.

CHAPTER SEVEN

7.1. CONCLUSION AND FUTURE WORKS

This chapter discusses the experiments and its results on finding plagiarism detection to Amharic language, specifically to the News domain. As the internet and social media develop and distribute information between information providers and information consumers, texts misused anonymously might give an issue such as presenting someone craft work as self-own without crediting the source.

These experiment is intended to meet the objectives of the research for Amharic and detect any plagiarism issues using NLP and n-gram feature to detect verbatim and semantic plagiarism texts. The experiment presents the percentage on the obfuscation and paraphrases on the suspected document against the corpus documents. The n-gram for first candidate selection can be achieved using n-gram by experimenting n=2, n=3, n=4, n=5, n=7, and n=10 and when the candidate selection met above the threshold of 0.2 out of 1.0, it will go to the next layer which is semantic checking is done. The semantic layer uses the LatentDirichletAllocation which analyzes the semantic content of the candidate documents to determine if there is any plagiarism. This is done by considering factors such as the words that are used, the order of the words, the structure of the sentences. This approach has several advantages over a single-layer approach. First, it can detect a wider range of plagiarism cases, including cases where the plagiarized text has been paraphrased or modified. Second, it is more efficient, as the first layer can quickly identify the documents that are most likely to be plagiarized, which reduces the amount of time that the second layer needs to spend on analysis.

The experiments are systematically and scientifically arranged which is directly related to research problems. The corpus consists of 100 documents exported from Telegram from 7 news agencies channels and 3 private bloggers channel. So after each suspected queries compared against all the corpus, the output values are shown by using accuracy, precision, recall and f1-scores through all possible combination related to the n-gram.

Based on the findings, we concluded that when the number of corpus increases and the number of n-gram increases, the performance of the model goes down, because increasing the number of corpus means increasing the number of contexts which deemed the performance low. A model that gives a higher performance at a smaller number of articles per channel or bloggers serve to analyze short and dangerously illegal extrinsic texts that are usually posted on social media.

For evaluating the proposed model on detecting plagiarism, a fingerprint and an Latent Dirichlet Allocation semantics are utilized. The experiment detects plagiarism with fingerprint and Accuracy: 0.972972972972973, Recall: 0.9175654853620955, precision: 0.9682926829268292, and F1 Score: 0.9422468354430379. The study's findings are intriguing, indicating that more research into the language, given its complexity, can deliver improved performance.

7.2. CONTRIBUTIONS OF THE STUDY

The findings of this research work are:

- The problem of plagiarism detection in the field of plagiarism analysis for the Amharic language was studied, which will assist open the door for the study of other problems in the language's field.
- Adopted suitable algorithm for the model to detect plagiarism for Amharic language

7.3. FUTURE RECOMMENDATION

- Develop a user interface for the system: The two-layer plagiarism detection system is a powerful tool, but it may be difficult for non-experts to use. It would be helpful to develop a user interface that makes the system more accessible to a wider range of users.
- Offline and Online: this study is an attempt to indicate that through improvements on detecting plagiarism using the existing txt files but I would recommend offline/Extrinsic with Online Amharic plagiarism detection should be studied when the first article was posted using post time to identify.

REFERENCES

[1] B. N. J. R. L Graves, Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking, 2016.

[2] H. L. S. a. I. Gashaw, "DICTIONARY BASED AMHARIC-ARABIC CROSS LANGUAGE INFORMATION RETRIEVAL," 2016.

[3] M. G. V. S. W. K. W. Kienreich, "Plagiarism Detection in Large Sets of Press Agency News Articles," *Know-Center, Competence Center for Knowledge-Based Applications and Systems,* 2015.

[4] B. H. H. a. S. Y. E. A.-T. a. A.-T. e. al., "An academic Arabic corpus for plagiarism detection: design, construction and experimentation," *International Journal of Educational Technology in Higher Education,* 2020.

[5] Z. Dentzel, How the Internet is Chang Everyday Life, 2014.

[6] A. A. A. Salawu, LANGUAGE POLICY, IDEOLOGIES, POWER AND THE ETHIOPIAN MEDIA, 2015.

[7] H. MOLLA, A Study of Journalistic Professional Practices in Ethiopian Television News Productions: Ethiopian Broadcasting Corporation (EBC) and Life Television (LTV) in focus, International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Volume 1 Issue 7, 2020.

[8] R. I. S. A. A. J. V. P. Muna AlSallal, An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection, 9th International Conference on Developments in eSystems Engineering, 2016.

[9] z. l. lu lu, "DNAP: detection of News Article Plagiarism," 2021.

[10] F. News, "https://www.foxnews.com/politics/time-cnn-suspend-zakaria-for-plagiarism," Foxnews, 23 December 2015. [Online]. [Accessed 01 09 2022].

[11] "BBC," BBC, 7 February 2019. [Online]. Available: https://www.bbc.com/news/world-us-canada-47156917. [Accessed 10 09 2022].

[12] N. B. A Patil, "Survey on different plagiarism detection tools and Software's," 2016.

[13] S. a. M.S.Otari, "Plagiarism Detection-Different Methods and Their Analysis: Review," 2014.

[14] I. Abakush, "METHODS AND TOOLS FOR PLAGIARISM DETECTION IN ARABIC DOCUMENTS," 2016.

[15] C. &. B. D. U. Halupa, "Faculty perceptions of student self-plagiarism: an exploratory multi-university study," *Journal of Academic Ethics,,* vol. 11, no. 4, pp.

297-310, 2013.

[16] Akbar, "Defining Plagiarism: A Literature Review," *Institute Agama Islam Negeri (IAIN),* 2018.

[17] I. Yilmaz, "Plagiarism? No, we're just borrowing better English," pp. 449, 658, 2007.

[18] M. Bouville, "Plagiarism: Words and Ideas," *Science and Engineering Ethics, ,* vol. 14, pp. 311-322, 2008.

[19] D. Weber-Wul, "Test cases for plagiarism detection software," *Proc. 4th Intl. Plagiarism Conf.,* pp. 1-13, 2010.

[20] L. Sindhu, "AN INTEGRATED APPROACH FOR PLAGIARISM DETECTION IN MALAYALAM DOCUMENTS, AN INTEGRATED APPROACH FOR PLAGIARISM DETECTION IN MALAYALAM DOCUMENTS," *Department of Computer Science,* 2017.

[21] G. &. J. M. Cosma, "Towards a definition of source-code plagiarism," *IEEE Transactions on Education,* vol. 51, no. 2, , pp. 195-200, 2008.

[22] W. T. Fitch, "Unity and diversity in human language," 2011.

[23] S. M. Alzahrani, N. Salim and A. Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," *IEEE,* vol. 42, no. 2, pp. 133 - 149, 2012.

[24] C. G. M. P. C. Grozea, "Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection," *3rd PAN Workshop. Uncovering Plagiarism,* p. 10, 2009.

[25] J. S. N. S. U. R. Ruben Comas Forgas, "The "Copy and Paste" Generation Plagiarism Amongst Students, a Review of Existing Literature,," *Balearic Islands University (UIB), Spain,,* 2006.

[26] I. F. A. Badiozaman, "PARAPHRASING CHALLENGES FACED BY MALAYSIAN ESL STUDENTS," vol. 3, no. 1, 2014.

[27] D. &. P. B. Pecorari, "Plagiarism in second-language writing. Language Teaching," vol. 47, no. 3, pp. 269-302, 2014.

[28] H. A. C. a. D. K. Bhattacharyya, "Plagiarism: Taxonomy, Tools and Detection Techniques," *Dept. of CSE,* 2018.

[29] M. S. B. C. A. a. R. P. Potthast, "An evaluation framework for plagiarism detection,," *23rd Int. Conf. on Computational Linguistics,* 2010.

[30] V. K. a. D. Gupta, "Study on Extrinsic Text Plagiarism Detection Techniques and Tools," *Department of Computer Science & Engineering, Amrita School of Engineering, Amrita University, Amrita Vishwa Vidyapeetham, Bangalore, India, Department of Mathematics,* 2016.

[31] A. E. a. B. Keller, "Twitter paraphrase identification with simple overlap features and

SVMs," *9th International Workshop on Semantic Evaluation (SemEval'15),* p. 64–69, 2015.

[32] A. K. H. C. D. S. El Moatez Billah Nagoudi, "A Two-Level Plagiarism Detection System for Arabic Documents," *Cybernetics and Information Technologies,* vol. 18, no. 1, 2018.

[33] M. S. H. A.-. T. Farah K. AL-Jibory, "Hybrid System for Plagiarism Detection on A Scientific Paper," *Turkish Journal of Computer and Mathematics Education,* vol. 12, no. 13, pp. 5707-5719, 2021.

[34] G. E. E. G. P. E. AngelosHliaoutakis, "Information retvieval by semantic similarity," 2015.

[35] Torrejón and Ramos, " Using a variety of n-grams for the detection of different Kinds of Plagiarism: Notebook for PAN at CLEF," 2013.

[36] S. a. N. S. Alzahrani, "Fuzzy semantic-based string similarity for extrinsic plagiarism detection," *lab report for PAN at CLEF , Proc. of 2nd Int. Workshop PAN-10,,* 2010.

[37] F. V.-T. E. M.-y.-G. M. P. a. R. P. Sánchez-Vega, "Determining and characterizing the reused text for plagiarism detection.," *Expert Systems with Applications,* vol. 40, no. 5, pp. 1804-1813, 2013.

[38] S. a. B. M. Suchomel, "Heterogeneous Queries for Synoptic and Phrasal Search-Notebook for PAN at CLEF," *6th International Workshop PAN-14,* 2014.

[39] E. H. B. Oumaima Hourrane, "Survey of Plagiarism Detection Approaches and Big data Techniques related to Plagiarism Candidate Retrieval," *Morocco 2017 Association for Computing Machinery,* 2017.

[40] C. Boyd, "What Is Latent Semantic Indexing," *SE journal,* 2018.

[41] P. W. Foltz, "An Introduction to Latent Semantic Analysis," *Discourse Processes ,* vol. 25, pp. 259-284, 1998.

[42] S. a. G. A. Sanchez-Perez, "A Winning Approach to Text Alignment for Text Reuse Detection," in *lab report for PAN at CLEF 2014. Proc. of 6th Int. Workshop PAN-14*, 2014.

[43] M. S. H. A.-. T. Farah K. AL-Jibory, "Hybrid System for Plagiarism Detection," *A Scientific Paper a Post graduate Student,* vol. 12, no. 3, 2021.

[44] S. S. N. I. R. A. Asad A, "A linguistic treatment for automatic external plagiarismdetection," vol. 135, p. 135–146, 2017.

[45] S. E. H. R. M. A. P. V. A. P. a. C. F. William Black, "Introducing the arabic wordnet project," *In Proceedings of the third international Word- Net conference,* pp. 295-300, 2006.

[46] M. E. B. Menai., "Detection of plagiarism in Arabic documents," *International journal of information technology and computer science (IJITCS),* vol. 4, no. 10, p. 80, 2012.

[47] V. G. R. K. Lovepreet A, "A New Hybrid Technique for Detection of Plagiarism from Text Documents," *Arab. J. Sci. Eng.,* vol. 45, no. 12, p. 9939–9952, 2020.

[48] A. A. A. Salawu, "LANGUAGE POLICY, IDEOLOGIES, POWER AND THE ETHIOPIAN MEDIA," 2015.

[49] B. H. Geletu, "Authorship Attribution Model for Amharic Documents using Machine Learning," *A Thesis Submitted to the Department of Computer Science in Partial Fulfillment for the Degree of Master of Science in Computer Science,* 2020.

[50] W. Kelemework, "Automatic Amharic text news classification: A neural networks approach," vol. 6, no. 2, pp. 127-137, 2013.

[51] Daniel Jurafasky and James H.martin, "Speech and Language Processing," 2000.

[52] M. Abate and Y. Assabie, "The Development of Amharic Morphological Analyzer Using Memory Based Learning," *Ethiopia Information Communication Technology Annual Conference,* 2014.

[53] M. L. Bender, "Language in Ethiopia," *London: Oxford University Press,* 1976.

[54] T. H. GEBERMARIAM, "AMHARIC TEXT RETRIEVAL: AN EXPERIMENT USING LATENT SEMANTIC INDEXING (LSI) WITH SINGULAR VALUE DECOMPOSITION (SVD)," *D. o. C. Science, Ed., Addis Ababa University: Unpublished Masters Thesis,* 2003.

[55] Z. SINTAYEHU, "Automatic Classification of Amharic News Items," *D. o. C. Science, Ed., Addis Ababa University: Unpublished Masters Thesis,* 2001.

[56] J. Creswell, "Research design: Qualitative, quantitative and mixed methods approaches," *SAGE Publications,* 2003.

[57] "METHODS OF DATA COLLECTION," in *Basic Guidelines for Research: An Introductory Approach for All Disciplines*, Chittagong-4203, Bangladesh, Book Zone Publication, 2016, pp. 201-275.

[58] M. S. B. C. A. a. R. P. Potthast, "An evaluation framework for plagiarism detection," in *Proc. of the 23rd International Conference on Computational Linguistics*, Beijing, China, 2010.

[59] M. A. Salahli, "An Approach For Measuring Semantic Relatedness Between Words Via," vol. 14, no. 1, pp. 55-63, 2009.

[60] S. S. A.-N. Mohran H. Al-Bayed, "Intelligent Multi-Language Plagiarism Detection System," *International Journal of Academic Information Systems Research (IJAISR),* vol. 2, no. 3, pp. 19-34, 2018.

[61] M. Shahabi, "Comparing Three Plagiarism Tools (Ferret, Sherlock, and Turnitin)," *International Journal of Computational Linguistics (IJCL),* vol. 3, no. 1, 2012.

[62] T. W. S. C. a. M. K. M. Rahman, "Multilayer SOM with treestruc- tured data for efficient document retrieval and plagiarism detection," *IEEE,* vol. 20, no. 9, p. 1385–1402, 2009.

[63] H. Z. a. T. W. S. Chow, "A coarse-to-fine framework to efficiently thwart plagiarism," vol. 44, p. 471–487, 2011.

[64] F. P. COTTERELL, "Amharic word classes,," *Journal of Ethiopian Studies,* vol. 2, no. 1, pp. 33-48, 1964.

[65] G. A. D. a. M. Getachew, "Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges," 2006.

[66] M. T. a. W. Menzel, "Amharic Part-of-Speech Tagger for Factored Language Modeling," Vols. 428-433, 2009.

APPENDIX 1

```
import nltk
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
from nltk.tokenize import word_tokenize
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.metrics import f1_score, recall_score, precision_score, accuracy_score
import os


def read_unicode_files(directory_path):
    readfiles = []
    files = os.listdir(directory_path)

    for file_name in files:
        file_path = os.path.join(directory_path, file_name)
….
    return readfiles

# Define the number of topics for LDA
NUM_TOPICS = 10

# Function to perform candidate selection using simple n-gram matching
def ngram_matching(documents, n=2, threshold=0.2):
   …
    return candidates


def get_candidates(documents):
…
    return candidates

def generate_fingerprints(text, k):
…
    return fingerprints

def get_candidate_sentences(source_text, suspect_text, k, threshold=0.2):
    candidate_sentences = []
    for c in source_text:
        for s in suspect_text:
…
    return candidate_sentences
```

```
def get_semantic_topics(documents):
    # Step 2: Semantic topic modeling using LDA

    vectorizer = CountVectorizer()
    X = vectorizer.fit_transform(documents)
    #print(X)
…
    return lda.transform(X)


def get_document_topic_distribution(lda, documents):
    vectorizer = CountVectorizer()
    X = vectorizer.fit_transform(documents)
…
    return doc_topic_distr

def detect_plagiarism(candidate_documents, candidate_topics, target_documents, target_topics):
    plagiarized_count = 0
    non_plagiarized_count = 0
    tp = 0 # True Positives
    tn = 0 # True Negatives
    fp = 0 # False Positives
    fn = 0 # False Negatives
…
    accuracy = (tp+tn)/(fp+fn+tp+tn)
    print("Accuracy:", accuracy)

    recall=tp/(tp+fn)
    #print(non_plagiarized_count)
    print("Recall:", recall)

    precision=tp/(tp+fp)
    #print(precision)
    print("precision:", precision)

    f1 = 2*((precision*recall)/((precision+recall)))
    #print(f1)
    print("F1 Score:", f1)

def main():
    directory_path = 'D://corpus//'  # Replace with your directory path
    corpus= read_unicode_files(directory_path)
    #print(corpus)
    #directory_path = 'D://corpus//test//'  # Replace with your directory path
    suspect_texts= read_unicode_files(directory_path)

    # Step 1: Candidate selection using CountVectorizer
    #candidate_documents = ngram_matching(corpus)
    n=10
```

```
    candidate_documents = get_candidate_sentences(corpus,corpus,n)
    #candidate_documents = get_candidates(corpus)
    #print(candidate_documents)

# Full program
def main():
    directory_path = 'D://corpus//'  # Replace with your directory path
    documents= read_unicode_files(directory_path)
    #print(corpus)
    #directory_path = 'D://corpus//test//'  # Replace with your directory path
    suspect_texts= read_unicode_files(directory_path)# Text documents to check for plagiarism

    # Perform 1st-layer candidate selection using fingerprinting
    fingerprint_duplicates = fingerprint_matching(documents)



    # Perform 2nd-layer semantics fuzzy matching
    all_duplicates = []
    for i, j in fingerprint_duplicates:
        duplicates = fuzzy_matching([documents[i], documents[j]])
        all_duplicates.extend(duplicates)

    # Calculate metrics
    true_positives = len(all_duplicates)
    false_positives = len(fingerprint_duplicates) - len(all_duplicates)
    false_negatives = len(all_duplicates)
    total_duplicates = len(fingerprint_duplicates)

    # Calculate F1 score, recall, precision, and accuracy
    f1_score, recall, precision, accuracy = calculate_metrics(true_positives, false_positives,
                                            false_negatives, total_duplicates)

    # Display the results
    if len(all_duplicates) > 0:
….
    else:
        print("No plagiarism detected.")

    print(f"F1 Score: {f1_score}")
    print(f"Recall: {recall}")
    print(f"Precision: {precision}")
    print(f"Accuracy: {accuracy}")


if __name__ == "__main__":
    main()
```

```
import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from fuzzywuzzy import fuzz
import os


def read_unicode_files(directory_path):
    readfiles = []
    files = os.listdir(directory_path)

    for file_name in files:
        file_path = os.path.join(directory_path, file_name)

        if os.path.isfile(file_path):
            try:
                with open(file_path, 'r', encoding='utf-8') as file:
                    content = file.read()
                    readfiles.append(content)
                        …..
                continue

    return readfiles


def preprocess_text(text):
…

    return preprocessed_text



def extract_ngrams(text, n):
    ngrams = []
    words = text.split()
…
    return ngrams

def select_candidates(documents, n):
    candidates = []
```

```
    for i in range(len(documents)):
        for j in range(i+1, len(documents)):
…
    return candidates

def fuzzy_match_similarity(sentence1, sentence2):
    return fuzz.ratio(sentence1, sentence2) / 100.0
    tp = 0 # True positives (plagiarized passages correctly detected)
    fp = 0 # False positives (non-plagiarized passages incorrectly detected)
    fn = 0 # False negatives (plagiarized passages missed)
…
    return calculate_scores(tp, fp, fn)

def main():
    directory_path = 'D://corpus//'  # Replace with your directory path
…
        print("Similarity Score: ", similarity_score)
        print("\n")

if __name__ == '__main__':
    main()
```

APPENDIX 2

አንቀጽ	ሀያ	የሆኑትን	ስድስት	እስከ	ይኸኛው	-	205/	-4	8/ሀ/	/አንቀጽ 164/	(ሸ)	336)	2-55/	36//
ወይም	ሆነ	በሆኑ	ምንም	ብሎ	ከርሱ	/1/	90/	248-250)	1ኣ5	62/	481//	107/	/pubic/	356-34
እንደሆነ	በኑላ	ጀምሮ	በሆነው	ከሰላሳ	መሆኑን	1/	84/	248-22/	784/	613/3//	134/	335/3/	2)%//1/	355-34/
ላይ	በአንድ	እንደዚህ	ስለ	የሚሆኑ	ለዚያው	/	78/	248-22	784-791/	599/	479//	332)	187//ሸ/)	327/ህ
ማንም	የሆኑ	በመሆን	ይሁን	ላይም	ለዚሁ	//	200/	247/ሀ/	784-790/	///ሀ/	478/	327/ህ	/36//	.ስ
በማይበልጥ	ከአስራ	የሌለው	ከዚሁ	የሆናል	ለእነርሱም	ሰ/	189/	244/	194/	59/1/	134-153/	32-1	187//ሀ/	32-1
መሰረት	የሆነውን	በማለት	በእነዚህ	ከነዚህ	እዚሁ	180/	769/	243///3/	/1//መ/	589/	468/	3/1/(ሀ)/	184)	3/1/(ህ)/
ሁኔታ	ከላይ	ባለ	ከማናቸውም	ያህል	ቢሆንም	3/	768-770	240/	775/	588/	467/	103/1/	18/	3/
ሌላ	ሁሉ	ይህንን	ከነበረው	ከሆነና	ሀ/	/ሀ/	75/	240)	771/	158/	465/	2ኣ5	///ሀ/	284-337/
እና	መሆኑ	እንዲቆይ	በእንዳንድ	እስከ	ሰ/	66/	684/	/ሀ//	19-200/	587/	46(1)/	103/	136//	27/
ይሆናል	ሌላውን	ሌላው	በእያንዳንዱ	ለሆኑ·ት	ሐ/	179/	218/	237/	770/	583/	459/	298/	(አንቀጽ 187//ሀ/)	269-24)
ሆኖ	ከሰባት	የሚሆነው	ጊዜም	አለው	መ/	5/	671/	238-260	754/	58/3/	455//	297/	13/	26/1//2/(1)እስክ
ነው	ለሌላ	በአንዱ	እስከ	እነዚሁ	ሠ/	/ከአንቀጽ	161/	231/	750/	58-582/	129-15)	ሆን½	129-15)	
በዚህ	አለበት	አይሆንም	የሌሎች	እንደሆኑ·	ረ/	(አንቀጽ	6/	229/	187/1/	(መ)	451/	29/	123/ህ	
እስከ	ሲል	ማለት	የሚሆኑ·ት	ስለማናቸውም	ሸ/	/3/	16/	228/	/1//ሸ/	(1)/	45(2)/	ቀ/	11/	
ውስጥ	ይሆናሉ	ሲባል	ከሆነው	ስለዚሁ	-ሀ/	-2	159/	224/1/መ/	73/	58(3)/	129-131/	ሆን&	11-419	
ከአንድ	ያልሆነ	ላለ	የነበረውን	ከአንዳንድ	ሓ/	12/	156/	224/	187//ሸ/)	57/	440/	280-283/	11-154/	
በማናቸውም	በመሉ	የሆነው	ያሉ	በእነዚሁ	-ሐ/	9/	/መ//	(አንቀጽ 187//ሀ/)	ማናቸWM	57-59/	435/	284-337	1/3/	
ወር	አስራ	መሆናቸው	ከሌሎች	በእምስት	-መ/	83/	15//	224-228	728-30/	560/	127/	284-325/	9-	
ከአምስት	አስር	በዋና	አንዲት	ወይም	-ሠ/	188/	145/	22)	72/	150/	426/	284-317/	84ኣ	
በላይ	አንደ	በማቀድ	ለሌሎች	በሆኑ	-ረ/	140/	145-150/	217/	187//ሀ/	56-599	423/	1/ሸ/	8/ሰ/	
ሲሆን	ቢሆንም	ጊዜና	አንኳን	የሆኑ·በታል	ሀ.	-1	48/	213/	71/	559/	123/ህ	28/1/	728-30/	
በሆነ	አንዱ	ለዚህ	ለሆነው	ለነዚህ	ለ.	4/	49/	211-213/	70/	55/3/	(ሀ)/	28(3)/	72/	
ከዚህ	የሌላውን	ሶስተኛ	አንዳንድ	ለማንኛውም	ሐ.	208/	13/	21//	187/	55/	422-424	1/3/	ን/ኣ	
የሆነ	ከሁለት	የነገሩ	ሰኣት	አንደኛ	መ.	20/	407/	/5/	ኘ/	15/1/	420/	273/	690//	

አምስት	ሶስት	እንዲሆን	እንደነዚህ	ናቸው	ሠ.	79/	37/	98/	690/1/	549)	123/ሀ//	-133	68/4/
ማንኛውም	ካልሆነ	እነኳ	የሆኑት	አሁን	ሽ.	88/	11/	95/	690//	543/	42/	(4)/	68/
ጋር	ቢያንስ	ከሀያ	የማናቸውም	ሰባት		9ኣ	37/3/	94-95/	186/	533/	42-47/	ሀ	676/1//
አንድ	ቢሆን	ከሀምሳ	ይህንንም	እነደሆነ		183/	110(2)/	937	ረ	146-150/	123/ሀ/	አንቀጽ	66//
ልዩ	እነዚህን	ይኸው	የአንድን	እነደሆነ		/ፋርማሲስት/	34/	207/	69/	532-534/	41/	12/	656/
ከሶስት	ናቸው	ለአንድ	በመሉም	ይህቸው		131/	106/	93/	184/	53/	4010	5/	622-31
በተለይም	አንዱን	የሚችለውን	በነዚህ	ከእነዚህ		123/	40/	201-207/	680/	514-24/	123(ሀ)/	9/	58-582/
በሌላ	ሁለት	ወይም	የዚሁ	ከእነዚሁ		(ሀ)	/በአንቀጽ	9-	68/4/	511///5/	399/	20/	55/3/
ሺህ	ለዚሁ	በሚገባ	ለእያንዳንዱ	የአንቀጹ		104/	100/	9(3)	ኘ/	(ሸ))	121-128/	9ኣ	55/
ወደ	ወይዘሮ	ይህም	ስለሆነ	ወይ		148/	ሆነ&	861/	676/1/	510//	397/	79/	514-24/
ማናቸውንም	ተብሎ	እነዚህ	መሆናቸውን	የሆኑትን		-3	10/	20/)	/1//ሸ/	51(ሸ)/	397)	14/	511///5/
ከአስር	ላይሆን	ከዚያ	ማንኛውንም	የለውም		81/	27//	84/1/መ//	670/	144-149/	39/	15//	510//
የማይበልጥ	እንደሆነና	እንዲሆኑ	ሁለቱ	በሚችሉ		19/	27/	84/1/(መ)/	665/	506/3/	39-641/	16/	51(ሸ)/
ብቻ	ብሎ	ከሌላ	እንጂ	የሌላቸውን		/1//ሀ/	269-322	84/1/	180)	505-513/	117/	10/	4ኣ
እንዲሁም	ከብር	ለሆነ	ከስምንት	በሶስተኛ		68/	269-24)	2/	ቸ/	143/	379/	766/	4ኣ7/
ሌሎች	ሆኖም	በሌሎች	ሁለቱንም	በቀር		18/	266/	/36//	66/1/ሀ/	4ኣ	376/	59/	493//
ይህ	በታች	እንደሆነ	በሁለት	በኋሉ		90/3/	261/1	835/	66//	4ኣ9	111/	40/	481//
ይህን	የሌላ	እንዲህ	በእስር	የአንዱን		59/	260/	2-55/	655/	14/3/	375/	27//	48/
ከሆነ	ያላቸው	በነዚህ	በሚል	የአንዱ		155/	26/1//2/	2)%//1/	654/	4ኣ7	37)	26/1/	"ሀ"
የዚህ	ይህንኑ	በእንደዚህ	ቁጥር	ው		154/	26/1/	829/	179/·	ዚዘ	367/	254/4/	479//
ማናቸውም	ሆነው	ስምንት	ባሉ	በዚህ		54/	257/ሸ/	821/	/አንቀጽ 101/	497/	36//	25//	46(1)/
ከስድስት	በስተቀር	ሲሆንና	ከመቶ	በዚህም		14/	257)	2(ሸ)	(ሸ)	14/1	11-419	248-22/	455//
መቶ	መሆን	ምንጊዜም	እነዚህም	በዚህና		101/	256/	809/	65-687/	494-500	36/	248-22	45(2)/
ያለ	ስም	ለማናቸውም	ሲናር	ከዚህም		/4/	255)	2(ሀ)	640/	493//	በ/	243///3/	41/
መሆኑን	እንደገና	የአንድ	ሰላሳ	በሁኔታው		50/	ህ/	/3	168-162	491/	356-34	22)	42/
አንድን	የማያንስ	እነዚህኑ	ማንም	ከነዚሁ		232-237/	254/	80/	627/	488/	11-154/	21//	(Manual)
ያላቸውን	እጅግ	ሲሆኑ	ለሆኑ	ሌሎች		21/	25//	8/ሰ/	622-31	486/	355-34/	20/)	39/
ሊሆን	ግን	በሁለቱም	አለ	ይህን		98/3/	25/	2(5)	168(2)/ማዘገ	136//	35/	2/	39-641/

APPENDIX 3

Punctuation	Amharic Name
።	አንድነጥብ (anednetib)
፡	ሁለትነጥብ (huletnetib)
፡-	ሁለትነጥብከሰረዝ (huletnetibkeserez)
…	ሦስትነጥብ (sostnetib)
፣	ነጠላሰረዝ (netelaserez)
፤	ድርብሰረዝ (derebserez)
‹‹››	ትዕምርተጥቅስ (teemirteteqs)
!	ትዕምርተአንክሮ (teemirteankro)
i	ትዕምርተስላቅ (teemirteselaq)
?	የጥያቄምልክት (yetyaqemilkt)
/	እዝባር (ezbar)
()	ቅንፍ (qenef)

APPENDIX 4

	ā/ä [a]	u [u]	ī/i [i]	a [a]	ē/e [e/ɛ]	(i)/(ə) [ə]	ō [o/ɔ]
h [h]	ሀ ha	ሁ hu	ሂ hi	ሃ ha	ሄ he	ህ h(ə)	ሆ ho
l [l]	ለ le	ሉ lu	ሊ li	ላ la	ሌ le	ል l(ə)	ሎ lo
h/ḥ [h]	ሐ ha	ሑ hu	ሒ hi	ሓ ha	ሔ he	ሕ h(ə)	ሖ ho
m [m]	መ me	ሙ mu	ሚ mi	ማ ma	ሜ me	ም m(ə)	ሞ mo
s/ś [s]	ሠ se	ሡ su	ሢ si	ሣ sa	ሤ se	ሥ s(ə)	ሦ so
r [r]	ረ re	ሩ ru	ሪ ri	ራ ra	ሬ re	ር r(ə)	ሮ ro
s [s]	ሰ se	ሱ su	ሲ si	ሳ sa	ሴ se	ስ s(ə)	ሶ so
sh/š [ʃ]	ሸ ʃe	ሹ ʃu	ሺ ʃi	ሻ ʃa	ሼ ʃe	ሽ ʃ(ə)	ሾ ʃo
k'/q [k']	ቀ k'e	ቁ k'u	ቂ k'i	ቃ k'a	ቄ k'e	ቅ k'(ə)	ቆ k'o
b [b]	በ be	ቡ bu	ቢ bi	ባ ba	ቤ be	ብ b(ə)	ቦ bo
t [t]	ተ te	ቱ tu	ቲ ti	ታ ta	ቴ te	ት t(ə)	ቶ to
ch/č [ʧ]	ቸ ʧe	ቹ ʧu	ቺ ʧi	ቻ ʧa	ቼ ʧe	ች ʧ(ə)	ቾ ʧo
h/ḫ [h]	ኀ ha	ኁ hu	ኂ hi	ኃ ha	ኄ he	ኅ h(ə)	ኆ ho
n [n]	ነ ne	ኑ nu	ኒ ni	ና na	ኔ ne	ን n(ə)	ኖ no
ny/ñ [ɲ]	ኘ ɲe	ኙ ɲu	ኚ ɲi	ኛ ɲa	ኜ ɲe	ኝ ɲ(ə)	ኞ ɲo
ʾ/ʔ [ʔ]	አ (ʔ)a	ኡ (ʔ)u	ኢ (ʔ)i	ኣ (ʔ)a	ኤ (ʔ)e	እ (ʔ)(ə)	ኦ (ʔ)o
k [k]	ከ ke	ኩ ku	ኪ ki	ካ ka	ኬ ke	ክ k(ə)	ኮ ko

	ā/ä [a]	u [u]	ī/i [i]	a [a]	ē/e [e/ɛ]	(i)/(ə) [ə]	o [o/ɔ]
h/ḵ [h]	ኸ he	ኹ hu	ኺ hi	ኻ ha	ኼ he	ኽ h(ə)	ኾ ho
w [w]	ወ we	ዉ wu	ዊ wi	ዋ wa	ዌ we	ው w(ə)	ዎ wo
ʿ/ʿ [ʔ]	ዐ ʔa	ዑ ʔu	ዒ ʔi	ዓ ʔa	ዔ ʔe	ዕ ʔ(ə)	ዖ ʔo
z [z]	ዘ ze	ዙ zu	ዚ zi	ዛ za	ዜ ze	ዝ z(ə)	ዞ zo
zh/ž [ʒ]	ዠ ʒe	ዡ ʒu	ዢ ʒi	ዣ ʒa	ዤ ʒe	ዥ ʒ(ə)	ዦ ʒo
y [j]	የ je	ዩ ju	ዪ ji	ያ ja	ዬ je	ይ j(ə)	ዮ jo
d [d]	ደ de	ዱ du	ዲ di	ዳ da	ዴ de	ድ d(ə)	ዶ do
j/ǧ [ʤ]	ጀ ʤe	ጁ ʤu	ጂ ʤi	ጃ ʤa	ጄ ʤe	ጅ ʤ(ə)	ጆ ʤo
g [g]	ገ ge	ጉ gu	ጊ gi	ጋ ga	ጌ ge	ግ g(ə)	ጎ go
t'/ṭ [t']	ጠ t'e	ጡ t'u	ጢ t'i	ጣ t'a	ጤ t'e	ጥ t'(ə)	ጦ t'o
ch'/č̣ [ʧ']	ጨ ʧ'e	ጩ ʧ'u	ጪ ʧ'i	ጫ ʧ'a	ጬ ʧ'e	ጭ ʧ'(ə)	ጮ ʧ'o
p'/p̣ [p']	ጰ p'e	ጱ p'u	ጲ p'i	ጳ p'a	ጴ p'e	ጵ p'(ə)	ጶ p'o
ts'/ṣ [ts]	ጸ ts'e	ጹ ts'u	ጺ ts'i	ጻ ts'a	ጼ ts'e	ጽ ts'(ə)	ጾ ts'o
ts'/ś [ts]	ፀ ts'e	ፁ ts'u	ፂ ts'i	ፃ ts'a	ፄ ts'e	ፅ ts'(ə)	ፆ ts'o
f [f]	ፈ fe	ፉ fu	ፊ fi	ፋ fa	ፌ fe	ፍ f(ə)	ፎ fo
p [p]	ፐ pe	ፑ pu	ፒ pi	ፓ pa	ፔ pe	ፕ p(ə)	ፖ po
v [v]	ቨ ve	ቩ vu	ቪ vi	ቫ va	ቬ ve	ቭ v(ə)	ቮ vo