FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA

TECHNICAL AND VOCATIONAL TRAINING INSTITUTE

(FDRE TVTI)

DEPARTMENT OF INFORMATION TECHNOLOGY

HYBRID-BASED DETECTION OF PLAGIARIZED AMHARIC NEWS ARTICLES ON SOCIAL MEDIA PLATFORMS

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of **MASTER OF SCIENCE IN INFORMATION TECHNOLOGY**

By

Bahlbi Hagos

ID No.: MTR/478/13

Advisor

Dr. Yihenew Wondie (Phd)

Addis Ababa, Ethiopia

June 2026

FEDERAL DEMOCRATIC REPUBLIC OF ETHIOPIA
TECHNICAL AND VOCATIONAL TRAINING INSTITUTE (FDRE TVTI)
COLLEGE OF GRADUATE STUDIES

DEPARTMENT OF INFORMATION TECHNOLOGY

HYBRID-BASED DETECTION OF PLAGIARIZED AMHARIC NEWS ARTICLES ON SOCIAL MEDIA PLATFORMS

By

Bahlbi Hagos

ID No.: MTR/478/13

Advisor

Yihenew Wondie (Phd)

A Dissertation Submitted to the Department of Information Technology, College of Graduate Studies, Federal Democratic Republic of Ethiopia Technical and Vocational Training Institute (FDRE TVTI), in Partial Fulfillment of the Requirements for the Degree of Master of Science (M.Sc.) in Information Technology**.**

Addis Ababa, Ethiopia

June 2026

DECLARATION

I hereby declare that this thesis has not been submitted previously, in whole for the award of any degree, diploma, or other qualification at any university or institution. It is also not being submitted concurrently for any other academic qualification.

I further declare that the thesis entitled **Hybrid-Based Detection of Plagiarized Amharic News Articles on Social Media Platforms** is the result of my own original research work, except where otherwise acknowledged. The research was conducted under the guidance and supervision of my advisor, **Yihenew Wondie**(PhD)

All sources of information, ideas, data, and works of other authors used in this study have been properly acknowledged through appropriate citations and references. A complete list of references is provided at the end of this thesis.

Declared by:

Name: Bahlbi Hagos Hadgu

Signature: _____

Department: Information Technology, FDRE TVTI

Date: _____

This thesis has been submitted for examination with my approval as the candidate's research advisor.

Dr. Yihenew Wondie

Advisor's Signature: _____

Date: 1/06/2026

APPROVAL PAGE

This is to certify that the dissertation entitled **"Hybrid-Based Detection of Plagiarized Amharic News Articles on Social Media Platforms"** has been submitted by **Bahlbi Hagos** (ID No. **MTR/478/13**) in partial fulfillment of the requirements for the award of the degree of **Master of Science in Information Technology**. The dissertation has been duly examined and approved by the Board of Examiners of the Department of Information Technology, Federal democratic republic of Ethiopia Technical and Vocational Training Institute.

Examination Role	**Name**	**Signature**	**Date**
Principal Advisor	_____	_____	
Internal Examiner	_____	_____	_____
External Examiner	_____	_____	_____
Department Head, Information Technology	_____	_____	_____
Dean, College of Graduate Studies	_____	_____	_____

Place: Addis Ababa, Ethiopia

Date of Final Approval: _____

ACKNOWLEDGEMENTS

First and foremost, I would like to offer my deepest gratitude to Almighty God for His abundant blessings, guidance, wisdom, and strength throughout my academic journey and the completion of this research work.

I am sincerely indebted to my advisor, Yihenew Wondie (PhD), for his exceptional guidance, insightful suggestions, constructive feedback, and unwavering encouragement throughout every stage of this study. His expertise, patience, and commitment significantly contributed to the successful completion of this research.

I would also like to extend my heartfelt appreciation to the Information Technology Department for providing the necessary academic environment, resources, and logistical support that facilitated the accomplishment of this work.

My special gratitude goes to Birknesh Weldeyohannes (PhD), the Postgraduate Coordinator of the Information Technology Department, for her continuous support, valuable guidance, and effective coordination throughout my postgraduate studies. Her dedication and assistance were instrumental in ensuring the smooth progress of my academic endeavors.

Finally, I would like to express my profound gratitude to my family for their unconditional love, constant encouragement, patience, prayers, and unwavering support. Their sacrifices, belief in my abilities, and continuous motivation inspired me to persevere through challenges and successfully complete this research. This achievement would not have been possible without their enduring support and understanding.

ABSTRACT

The rapid growth of social media news dissemination in Ethiopia has raised increasing concerns about plagiarism in Amharic news articles. Such practices undermine journalistic integrity, reduce originality in reporting, and contribute to public mistrust in news organizations and media credibility. To address this issue, this study recommends a hybrid plagiarism detection framework specifically designed for Amharic news content shared on social media platforms. The research employed purposive sampling to collect data from a diverse range of sources, including government and private news agencies, individual bloggers, journalists, and international media outlets such as BBC Amharic. The selected sources were required to publish daily news content across multiple domains, including politics, economics, and international affairs, and to maintain active Amharic-language public channels with significant audience engagement. The proposed framework integrates both lexical and semantic analysis techniques. The first layer applies traditional candidate detection methods, including fingerprinting and n-gram analysis, to identify potentially plagiarized segments. The second layer refines these results using advanced semantic approaches, specifically Latent Dirichlet Allocation (LDA) and fuzzy logic models, to detect paraphrased and meaning-preserving plagiarism that conventional methods often fail to identify. To evaluate performance, multiple hybrid configurations were tested, including Fingerprint + LDA, Fingerprint + Fuzzy, N-gram + LDA, and N-gram + Fuzzy models. Experimental results demonstrate that hybrid approaches outperform individual techniques across most evaluation metrics. The Fingerprint + LDA model achieved the highest overall performance with an accuracy of 0.97, precision of 0.968, recall of 0.917, and F1-score of 0.94. In comparison, other combinations showed varying performance, with Fingerprint + Fuzzy achieving perfect accuracy but lower recall, indicating imbalance in detection capability. Overall, the findings confirm that integrating traditional text-matching techniques with semantic-based models significantly enhances plagiarism detection effectiveness in Amharic news content, with the Fingerprint + LDA hybrid model emerging as the most robust approach.

Keywords: *Plagiarism Detection, Amharic news Articles, Hybrid Model, Fingerprint, Latent Dirichlet Allocation (LDA), Fuzzy Logic, Social Media News*

TABLE OF CONTENTS

DECLARATION i

APPROVAL PAGE iii

ACKNOWLEDGEMENTS iv

ABSTRACT v

TABLE OF CONTENTS vi

LIST OF TABLES x

LIST OF FIGURES xi

LIST OF ACRONYMS xii

CHAPTER ONE 1

1. INTRODUCTION 1

1.1. BACKGROUND 1

1.2. STATEMENT OF THE PROBLEM 3

1.3 Research Questions 5

1.4 Objectives 5

1.4.1 General Objective 5

1.4.2 Specific Objectives 5

1.5 Significance of the Research 6

1.6 Scope and Limitations 6

1.7 Organization of the Study 7

1.8 Chapter Summary 7

CHAPTER TWO 8

LITERATURE REVIEW 8

2.1 OVERVIEW 8

2.2 THEORETICAL REVIEW 8

2.2.1 PLAGIARISM DEFINITION 8

2.2.2 BROAD TYPES OF PLAGIARISM 9

2.2.2.1 NON-NATURAL / SOURCE CODE PLAGIARISM 9
2.2.2.2 NATURAL LANGUAGE PLAGIARISM 10
Monolingual Plagiarism Detection 10
Cross-Lingual Plagiarism Detection 10
2.2.3 TYPES OF MONOLINGUAL PLAGIARISM 10
2.2.3.1 EXACT / CLONE / COPY AND PASTE PLAGIARISM 10
2.2.3.2 INTELLIGENT PLAGIARISM 10
Redrafting / Paraphrasing 11
2.2.4 PLAGIARISM DETECTION APPROACHES 11
2.2.4.1 HUMAN-BASED DETECTION 11
2.2.4.2 TECHNOLOGY-BASED DETECTION 12
2.2.2 TEXTUAL FEATURES FOR EXTRINSIC PLAGIARISM DETECTION 12
2.3 CONCEPTUAL REVIEW 14
2.3.1 PLAGIARISM DETECTION METHODS 14
2.3.1.1 LEXICAL DETECTION APPROACHES 15
2.3.1.2 STRING-BASED DETECTION APPROACHES 15
2.3.1.3 VECTOR SPACE MODEL (VSM) APPROACH 16
2.3.1.4 SEMANTIC-BASED APPROACHES 16
2.3.2 SEMANTIC PLAGIARISM DETECTION MODELS 17
2.3.3 AMHARIC LANGUAGE BACKGROUND 17
2.3.4 AMHARIC WRITING SYSTEM 17
2.3.4.1 Alphabet 17
2.3.4.2 Punctuation Marks 18
2.3.4.3 MORPHOLOGY 18
2.3.4.4 INFLECTION 18
2.3.4.5 DERIVATION 18
2.3.5 ISSUES OF AMHARIC WRITING 18

2.3.6 HANDLING AMHARIC TEXT ISSUES 19
2.3.6.1 Stemming 19
2.3.6.2 Lemmatization 19
2.3.6.3 Normalization 19
2.3.6.4 SEMANTIC STRUCTURE 19
2.4 RELATED WORK / EMPIRICAL REVIEW 19
2.5 RESEARCH GAP 20
2.6 CONCEPTUAL FRAMEWORK 20
2.7 CHAPTER SUMMARY 22
CHAPTER THREE 23
RESEARCH DESIGN AND METHODOLOGY 23
3.1 OVERVIEW 23
3.2 RESEARCH DESIGN 23
3.3 RESEARCH APPROACH 24
3.4 SOURCES OF DATA 24
3.5 METHODS OF DATA GATHERING 24
3.6 FEATURE EXTRACTION 25
3.7 SAMPLING METHODS AND TECHNIQUES 25
3.8 CORPUS PREPARATION 26
3.9 ARCHITECTURE 26
CHAPTER FOUR 29
EXPERIMENT RESULTS AND DISCUSSION 29
4.1 INTRODUCTION 29
4.2 EXPERIMENTAL SETUP AND EVALUATION METRICS 30
4.3 RESULTS AND DISCUSSION 33
4.3.1 MODEL SELECTION 33
4.3.2 FEATURE COMPARISON 34

4.3.3 INTERPRETATION OF METRICS 35
4.3 SUMMARY 39
CHAPTER FIVE 40
CONCLUSION AND FUTURE WORKS 40
5.1 CONCLUSION 40
5.2 FUTURE WORKS 41
5.2.1 CONTRIBUTIONS OF THE STUDY 41
5.2.2 FUTURE RECOMMENDATIONS 41
References 42
Appendix I 46
APPENDIX II 52
APPENDIX III: 53
APPENDIX IV 54

LIST OF TABLES

Table 2.1 Redundancy of Characters [50] 19

Table 4.1 corpus document against the samples taken from various channels on Telegram ... 32

Table 4.2 Results of the various models against F1, Recall, Precision, and Accuracy 35

LIST OF FIGURES

Figure 2.1. General Architecture of Extrinsic [16]. 12

Figure 2.2. Conceptual Framework of the Proposed Model 21

Figure 3.1. General workflow of the proposed model 27

Figure 4.1: Models versus F1 shows first-layer fingerprint candidate selection and second-layer semantic processing, with the fuzzy model achieving the best overall performance 36

Figure 4.2: Models versus Precision shows first-layer fingerprint candidate selection and second-layer semantic processing, with the fuzzy model outperforming all others. 36

Figure 4.3: Models versus Recall shows 1st layer Fingerprint candidate selection and 2nd layer 37

Figure 4.4. Models versus Accuracy shows 1st layer Fingerprint candidate selection and 2nd layer semantics with LDA beat all 37

Figure 4.5: Model comparison against evaluation metrics 38

LIST OF ACRONYMS

AMN – Addis Media Network

API – Application Programming Interface

APD – Amharic Plagiarism Detection

DBMS – Database Management System

EBC – Ethiopia Broadcasting Corporation

FBC – Fana Broadcasting Corporation

ML – Machine Learning

PD – Plagiarism Detection

PDS – Plagiarism Detection System

PHP – Hypertext Preprocessor

RAM – Random Access Memory

SEO – Search Engine Optimization

SQL – Structured Query Language

URL – Uniform Resource Locator

WAN – Wide Area Network

WWW – World Wide Web

CHAPTER ONE

1. INTRODUCTION

1.1. BACKGROUND

The rapid evolution of information and communication technologies has profoundly changed how information is produced, accessed, and shared. The internet now provides easy access to a wide range of knowledge sources, allowing users to retrieve information from books, academic journals, newspapers, and numerous digital platforms with minimal effort [5]. In recent years, the rapid spread of smartphones combined with the reduced cost of internet connectivity has further expanded online participation, especially on social media platforms where users communicate, exchange content, and engage socially [6]. Consequently, many organizations increasingly rely on social media to advertise services, distribute information, and strengthen audience interaction through enhanced network visibility, which has significantly reshaped traditional communication and business practices [7]. Although these technological improvements have greatly enhanced information accessibility and sharing, they have also made it easier to copy and distribute intellectual content without authorization across digital environments. This has led to a rise in plagiarism, which is defined as the use of another person's ideas, expressions, or work without proper acknowledgment [16], [18]. Plagiarism is widely regarded as a serious ethical issue because it compromises originality, misrepresents authorship, and reduces both the credibility and economic value of creative and academic work [15]. In the fields of journalism and digital communication, it weakens public trust and undermines the integrity of information dissemination [1]. As a result, institutions and organizations are increasingly expected to enforce strict anti-plagiarism regulations and adopt effective detection systems to maintain ethical standards and protect intellectual integrity [12], [13]. Plagiarism has emerged as a significant challenge across a wide range of fields, including literature, music, software engineering, academic research, news media, advertising, and web content development [17], [28]. Across the world, scholars and institutions continue to work toward preserving originality in intellectual outputs by ensuring proper citation and recognition of authors. In this context, plagiarism detection (PD) has become an important research domain within natural language processing and information retrieval [23], [29].

While existing plagiarism detection systems perform well for resource-rich languages such as English and other European languages [30], they often struggle with under-resourced

languages due to differences in linguistic structure, syntax, semantics, and morphology [20], [46].

A key difficulty in multilingual plagiarism detection lies in capturing the unique linguistic characteristics of each language [22]. Languages vary significantly in sentence structure, grammar, word formation, and meaning representation. Therefore, models designed for English cannot be directly applied to languages such as Amharic without substantial modification [52], [53]. The Amharic language has distinct morphological and syntactic features that require specialized processing methods and appropriate linguistic resources [64], [65]. Despite increasing interest in natural language processing research for Amharic, there is still a lack of comprehensive studies focusing specifically on plagiarism detection in Amharic texts.

This thesis seeks to fill this research gap by exploring plagiarism detection techniques for Amharic-language documents. To the researcher's knowledge, only limited comprehensive studies have been conducted in this area, particularly in relation to social media news content. Most previous work on plagiarism detection has focused on languages such as English, Arabic, Malayalam, and other well-resourced or moderately resourced languages [4], [20], [32], while Amharic has received comparatively little attention. Therefore, there is a clear need to develop plagiarism detection methods that are tailored to the linguistic characteristics of Amharic. Plagiarism detection involves identifying similarities in meaning, context, and structure, even when the original content has been paraphrased or significantly modified [36], [37]. Semantic analysis techniques have shown strong potential in identifying plagiarism by measuring conceptual similarity rather than relying solely on exact word matching [34], [40], [41]. In the case of Amharic, semantic similarity-based approaches can be particularly useful for detecting paraphrased or semantically equivalent content, especially when lexical overlap is minimal. Accordingly, this study focuses on detecting plagiarism in Amharic social media news articles collected from selected news agencies between March 2020 and March 2021. Developing an effective plagiarism detection system for Amharic requires combining natural language processing techniques, machine learning models, and language-specific linguistic resources [8], [47]. Methods such as latent semantic analysis, semantic similarity computation, and hybrid detection approaches have proven effective in other languages and may serve as useful foundations for Amharic-based systems [33], [42], [47]. However, further research is still needed to design robust models capable of handling the linguistic complexity of Amharic as well as its limited computational and linguistic resources. The importance of

plagiarism detection research in Amharic can be viewed from two main perspectives. First, Amharic is an under-resourced language that faces several challenges in natural language processing, including a lack of annotated datasets, complex morphological structures, semantic ambiguity, and limited availability of linguistic tools [49], [52], [54]. Addressing these challenges can contribute to advancing computational linguistics and language technology development for Amharic. Second, ensuring the originality of content produced by Amharic writers, journalists, and researchers is crucial for preserving the cultural, academic, and economic value of their work and supporting intellectual property rights. The subsequent sections of this thesis include the statement of the problem, research questions, study objectives, significance of the study, scope of the study, and the overall organization of the thesis.

1.2. STATEMENT OF THE PROBLEM

The rapid advancement of technology, from large-scale computers to personal mobile devices, has significantly transformed human lifestyles by enabling instant access to information at our fingertips. The Internet, as one of the most influential technological innovations, has revolutionized daily life and has become an essential part of modern living. Today, nearly all activities rely on internet access through Wi-Fi, mobile data, or ADSL connections.

The Internet supports users in discovering innovations, finding solutions to everyday problems, and expanding their knowledge base. As a result, it has fundamentally transformed society, accelerating its transition from the industrial era to the networked digital age [5].

It has also reshaped key sectors such as business, education, government, and healthcare. In addition, it has made the world feel smaller by reducing the limitations of distance and transforming how people interact with family and society. The Internet has become a major source of easily accessible information and a vital tool for assistance whenever needed.

Modern society is increasingly becoming information-driven, with individuals exposed to vast amounts of data daily. Information now flows continuously from multiple sources, resembling an overwhelming flood [6]. The Internet further enhances access to diverse resources, including academic articles, journals, literature, news reports, and more.

In contemporary global culture, the Internet is widely used for social interaction through platforms such as social media and blogs, contributing to the consumption and dissemination of large volumes of information [7]. In the field of journalism, for example, news organizations often reference or republish content from other sources to maintain visibility and audience engagement. However, when such content is used without proper attribution or added value, it becomes problematic.

This practice contributes to cultural homogenization, where news content becomes increasingly uniform across different outlets [1]. Moreover, failing to acknowledge original sources constitutes copyright infringement and raises ethical concerns regarding intellectual property rights. Detecting such violations is particularly challenging due to the massive volume of online content that cannot be manually reviewed.

Since news reporting relies heavily on public trust, failure to credit original sources undermines the credibility and legitimacy of media organizations. Journalists accused of plagiarism may face suspension or dismissal while investigations are conducted, as observed in major media organizations such as CNN, ABC, and Fox News [10]. These actions are necessary to preserve institutional credibility.

The rise of the Internet and search engines has facilitated international collaboration, but it has also increased the risk of plagiarism. In Ethiopia, online news is published in Amharic and other local languages. The expansion of social media has enabled journalists, bloggers, and writers to publish Amharic content more easily. However, this increased accessibility has also created opportunities for plagiarism.

As Amharic digital content continues to grow, misuse of technology may lead to unethical practices such as plagiarism in news reporting. Such practices negatively affect society by creating mistrust regarding the authenticity and ownership of news content, while also undermining the integrity and creativity of journalism.

According to Hiwot Molla [7], some news articles are directly sourced from Facebook pages of other organizations without proper acknowledgment. Although plagiarism cases in Ethiopia are less frequently reported compared to international media organizations such as BBC, CNN, or NBC [11], this does not indicate its absence. Rather, it suggests a lack of effective detection mechanisms.

Therefore, there is a need to develop automated plagiarism detection systems to ensure originality and uphold ethical standards in information dissemination. Plagiarism has become increasingly prevalent across various domains, including social media, academic writing, journalism, and other digital platforms [8].

Several plagiarism detection tools exist based on previous research, including systems such as Turnitin, EVE2, CopyCatchGold, WordCheck, JPlag, PlagAware, iThenticate, and PlagScan [12]. These tools are effective for multiple languages but generally do not support Amharic.

Most existing approaches rely on linguistic structures such as syntax and semantics, which vary significantly across languages and limit their applicability. Amharic, in particular, has a

flexible word order that preserves meaning despite structural changes. For example, sentences such as “አበበ በሶ በላ”, “በሶ አበበ በላ”, and “በላ አበበ በሶ” all convey the same meaning despite different arrangements.

Additionally, Amharic words often undergo morphological changes through prefixes, infixes, and suffixes, resulting in varied forms that still retain semantic relationships. For instance: ሞውደድ, አልወድም, ወደድኩ, etc.

Due to these linguistic characteristics, existing plagiarism detection systems are not adequately suited for Amharic. They struggle to capture the structural and morphological complexity of the language, particularly in large-scale text comparisons.

As a result, there is a strong need to develop models that detect plagiarism at both character and text levels while adapting techniques from other languages to suit Amharic. Such systems should enhance originality and encourage creativity in writing, similar to traditional print journalism, which emphasized originality and diversity in content.

1.3 Research Questions

This study aims to address the following research questions in the context of Amharic textual plagiarism detection:

- What are the main challenges of detecting plagiarism in the Amharic language?
- What is the most suitable method for detecting plagiarism in Amharic texts?
- How effective are the proposed methods in detecting Amharic plagiarism?

1.4 Objectives

1.4.1 General Objective

To develop and evaluate a hybrid-based of plagiarism detection system for Amharic news articles on social media platforms.

1.4.2 Specific Objective

- To identify and analyze the major challenges affecting plagiarism detection in Amharic news articles
- To design and develop a suitable hybrid plagiarism detection method for Amharic texts.
- To evaluate the performance and effectiveness of the proposed hybrid plagiarism detection system
- To apply NLP techniques and Amharic morphological analysis on collected data.

- To improve plagiarism detection accuracy by combining suitable models.

1.5 Significance of the Research

This research is significant because it addresses the unique challenges of detecting plagiarism in the Amharic language. Due to its complex morphology, syntax, and writing system, Amharic presents difficulties that are not adequately handled by existing plagiarism detection systems developed primarily for languages such as English.

Current tools are unable to effectively capture linguistic variations and contextual meanings in Amharic texts. Therefore, this study aims to bridge this gap by developing specialized tools tailored to Amharic language processing.

The research contributes to the advancement of language technology, benefiting learners, researchers, journalists, and writers. It promotes academic integrity, protects intellectual property, and encourages originality in content creation within the Ethiopian context.

Furthermore, this study serves as a foundational work for future research in Amharic plagiarism detection and may support the development of more advanced and accurate detection systems.

The primary beneficiaries include society at large, including educational institutions, media organizations, researchers, and individuals.

1.6 Scope and Limitations

This study focuses on plagiarism detection in the Amharic language using data collected from news agencies and social media platforms such as Telegram and Facebook. It specifically examines plagiarism in news content using semantic similarity analysis.

The scope of this research is limited to developing an offline Amharic text-based plagiarism detection system. It does not include non-textual content such as images, graphs, music, tables, or diagrams.

The study focuses only on monolingual Amharic texts from selected news sources and excludes academic papers, novels, poems, and short stories. It also does not cover multilingual texts, translations, or cross-lingual plagiarism detection.

Furthermore, the system assumes input as plain text or Word documents and does not support formats such as PDF or EPS.

Other Ethiopian languages such as Oromo, Tigrinya, and Guragigna are not included in this study. Additionally, the system does not address complex paraphrasing, conceptual plagiarism, hyperlinked content, or source code detection.

Due to the syntactic flexibility of Amharic, different word orders expressing the same meaning are considered in the analysis. The system primarily focuses on duplicate and semantically similar texts.

The experiment is limited to synonym-based variations and lightweight paraphrasing due to resource and time constraints. Since no standard Amharic corpus exists, data is collected from an Amharic-Amharic dictionary and selected online news sources.

1.7 Organization of the Study

This research is organized into five chapters. Chapter One introduces the background of the study, including the problem statement, objectives, scope, significance, and research questions. Chapter Two presents a literature review of existing studies on plagiarism detection, particularly in relation to Amharic language processing.

Chapter Three describes the research methodology, including data collection, tools, and experimental design.

Chapter Four presents data analysis, results, and interpretation of findings.

Chapter Five provides discussion, conclusions, and recommendations for future work.

1.8 Chapter Summary

This chapter introduced the research background and motivation for studying plagiarism detection in the Amharic language. It presented the problem statement along with the research objectives, significance, scope, and limitations.

CHAPTER TWO

LITERATURE REVIEW

2.1 OVERVIEW

This chapter establishes the theoretical foundation of the study by reviewing existing literature on plagiarism detection. It examines techniques used to identify plagiarism in texts written in different languages, as well as previous research conducted using various methods for textual plagiarism detection.

To develop an effective model for Amharic Plagiarism Detection (APD), this section analyzes the strengths and limitations of models proposed by different scholars. The chapter begins by defining plagiarism, followed by a discussion of its key characteristics and enabling technologies that support understanding of the detection process.

It also reviews different plagiarism detection techniques and implementation approaches, highlighting their advantages and disadvantages. Furthermore, the chapter discusses factors that encourage the adoption of plagiarism detection systems, as well as challenges that may limit their use.

The main objective of this chapter is to assess the strengths and weaknesses of existing plagiarism detection approaches in order to integrate and propose an improved solution for Amharic plagiarism detection. It also explores evolving trends in the field and how different models are being adopted and improved.

Since this research focuses on the Amharic language, the discussion begins with the linguistic nature of Amharic and then proceeds to methods used in plagiarism detection.

2.2 THEORETICAL REVIEW

2.2.1 PLAGIARISM DEFINITION

One major challenge in addressing plagiarism is that the term is often misunderstood.

According to the Oxford Learner's Dictionary, plagiarism originates from the Latin word *plagiarius*, meaning the act of presenting someone else's work or ideas as one's own without proper acknowledgment, whether intentional or not.

In the Amharic dictionary, plagiarism (ስርቆት) is defined as the act of taking another person's work or ideas and using them as one's own without giving proper credit to the original source.

According to [14], plagiarism is described as the unauthorized use of another person's intellectual output—such as ideas, inventions, writings, methodologies, or designs—through methods such as copy-pasting or paraphrasing without proper citation.

However, scholars do not fully agree on a single definition of plagiarism. This disagreement is partly due to variations in academic misconduct practices, including the concept of self-plagiarism [15].

A related question is: how can researchers steal their own work? A study involving 158 lecturers (92.2% Caucasian, 1.9% Multiethnic, 1.3% Asian, 1.3% Hispanic, and 0.7% Native American or Alaskan Native) found that many instructors were uncertain whether reusing their own previous work constitutes self-plagiarism [16].

According to Yilmaz [17], simply reusing basic sentences should not necessarily be considered plagiarism. Similarly, Bouville [18] argues that copying minor or non-original sentences may not carry significant academic value.

Although the concept of plagiarism may appear abstract in theory, it becomes very clear when attempting to avoid unethical academic practices.

From my perspective, reusing one's own previously conducted research is not inherently wrong. It is not comparable to taking another person's work without acknowledgment. Therefore, plagiarism can be defined as the theft of intellectual work that is not common knowledge or one's own, or the use of materials written by others without the author's permission or proper citation.

Currently, the number of plagiarism detection tools has increased significantly due to the rapid growth of plagiarism cases [19].

2.2.2 BROAD TYPES OF PLAGIARISM

Plagiarism can occur in both unstructured (natural language text) and structured formats such as source code [20].

In relation to language, plagiarism is generally categorized into two major types:

2.2.2.1 NON-NATURAL / SOURCE CODE PLAGIARISM

This type of plagiarism involves copying or modifying software source code, either by translating it into another programming language or by reusing it within the same language.

Programming languages consist of keywords, structures, functions (such as public, protected, and private), variables, and syntax rules.

Source code plagiarism is a growing issue in computer science education, particularly in programming courses [21]. However, since this study does not focus on source code

plagiarism, the discussion will be limited, and attention will shift to natural language plagiarism.

2.2.2.2 NATURAL LANGUAGE PLAGIARISM

Natural language plagiarism refers to plagiarism occurring in human-written text. Plagiarism can occur in any natural language. Globally, there are more than 6,000 languages [22], but only a small number are digitally resourced.

In natural language plagiarism detection, two main approaches exist [23]:

Monolingual Plagiarism Detection

This approach focuses on detecting plagiarism within the same language (e.g., English-to-English). Most existing plagiarism detection systems are developed under this category. It is typically divided into offline and online detection tasks [23].

Cross-Lingual Plagiarism Detection

This method detects plagiarism across different languages by incorporating translation techniques. With the availability of translation tools such as Google Translate, it has become easier to translate and reuse content across languages, increasing the risk of plagiarism.

As a result, cross-lingual plagiarism detection has gained significant research attention in recent years [24].

2.2.3 TYPES OF MONOLINGUAL PLAGIARISM

Plagiarism may involve copying, paraphrasing, reusing ideas, or modifying content in different forms. In natural language processing, plagiarism is commonly classified into two major categories.

2.2.3.1 EXACT / CLONE / COPY AND PASTE PLAGIARISM

This type of plagiarism occurs when a person directly copies text without using quotation marks or proper citation.

It involves taking exact words, phrases, or paragraphs from a source and presenting them as original work. In some cases, minor modifications may be made, such as removing or replacing a few words, but the overall structure remains unchanged.

This includes copying full documents from websites, Word files, or PDF articles and submitting them as original work, or combining multiple sources without acknowledgment [25].

2.2.3.2 INTELLIGENT PLAGIARISM

Intelligent plagiarism involves more sophisticated methods than simple copy-paste. It includes paraphrasing and idea manipulation.

Redrafting / Paraphrasing

Paraphrasing is the process of rewriting content using different linguistic structures while maintaining the original meaning. It involves techniques such as synonym replacement, grammatical changes, and sentence restructuring [26].

Paraphrasing can be divided into:

- **Simple Paraphrasing:**
  Involves replacing words with synonyms, changing word order, or modifying grammatical structure while preserving meaning [27].

1. Complex Paraphrasing: Combines multiple paraphrasing strategies and integrates ideas from different authors, making detection more difficult since it reflects merged concepts rather than a single source [27].

Other forms include:

2. Metaphor / Personification Plagiarism: Using figurative language to represent others' ideas more clearly [28].

3. Concept / Idea Plagiarism: Claiming another person's idea as one's own without acknowledgment [27].

4. Illegitimate Source / 404 Plagiarism: Citing incorrect or non-existent sources [27].

5. Translation Plagiarism: Translating content from another language and presenting it as original work. This may also involve paraphrasing or conceptual modification across languages.

2.2.4 PLAGIARISM DETECTION APPROACHES

There are two main approaches used to detect plagiarism:

2.2.4.1 HUMAN-BASED DETECTION

Human-based plagiarism detection relies on experts reviewing documents for inconsistencies. These inconsistencies may include variations in ideas, writing style, argument structure, and citation quality. Experts may also identify plagiarism based on familiarity with published works.

However, this approach has limitations. It requires subject-matter experts, is time-consuming, and becomes impractical when dealing with large volumes of digital documents, especially as the number of Amharic texts continues to grow.

Therefore, automated plagiarism detection systems are necessary.

2.2.4.2 TECHNOLOGY-BASED DETECTION

Technology-based plagiarism detection systems automatically analyze text and compare it with existing sources such as document repositories, databases, or online content.

These systems typically use URLs, web content, and stored corpora to identify similarities.

Two main detection tasks exist in plagiarism detection systems (PDS):

- **Extrinsic (External) Detection:**

  The suspicious document is compared against a reference corpus to identify matching content [29].

- **Intrinsic (Internal) Detection:**

  No external reference is used. Instead, the system analyzes the document itself by examining writing style, vocabulary usage, and structural patterns [23].

Feature extraction techniques are then applied to detect potential plagiarism cases effectively.

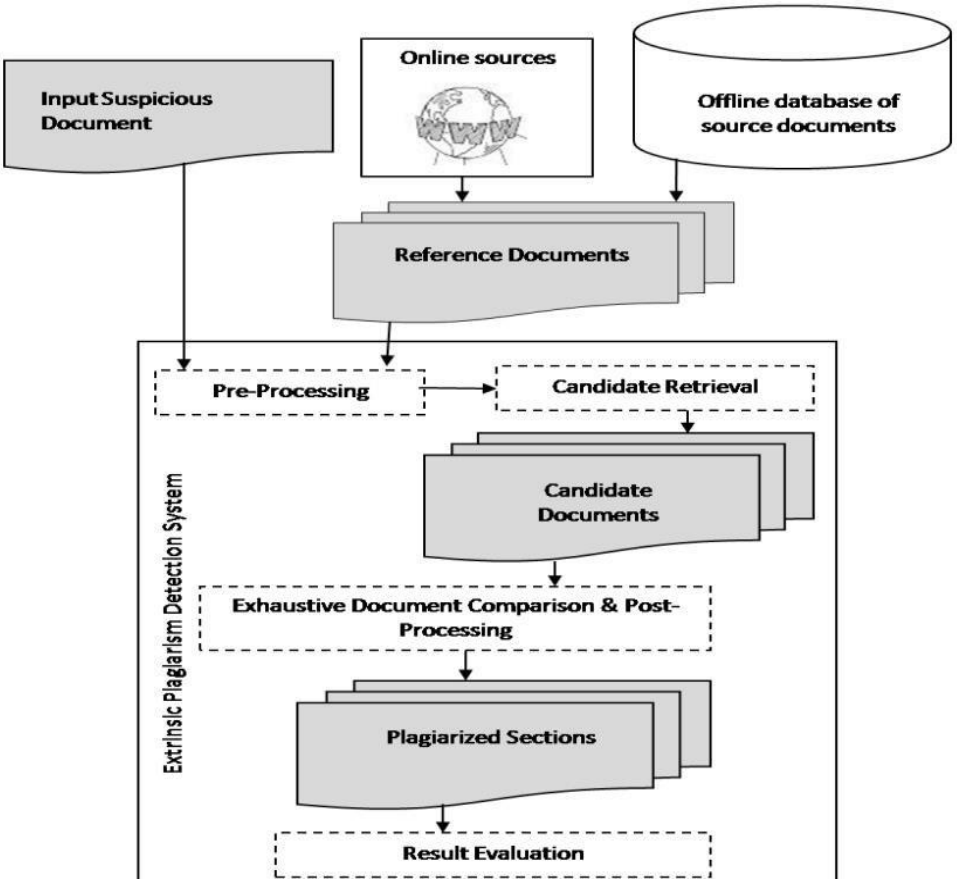


Figure 2.1. General Architecture of Extrinsic [16].

2.2.2 TEXTUAL FEATURES FOR EXTRINSIC PLAGIARISM DETECTION

In extrinsic plagiarism detection systems (PDS), a variety of techniques can be applied. These include string-based methods, vector space model (VSM) approaches, syntax-based

techniques, semantic-based methods, structural analysis, citation-based approaches, or combinations of these methods, commonly referred to as hybrid approaches [30].

Literature on plagiarism detection identifies several textual feature categories used in extrinsic plagiarism detection. These are generally grouped into lexical methods, syntactic methods, semantic methods, and hybrid approaches that combine multiple techniques.

Lexical-based Detection Methods

Lexical detection methods focus solely on the surface-level representation of text, such as characters or words, without considering deeper linguistic structure. These methods operate at different granularity levels, including document-level, paragraph-level, character-level, or word-level n-grams.

In character-based n-gram (CNG) representation, a document *d* is modeled as a sequence of characters:

$d = \{(c_1, d), (c_2, d), \ldots, (c_n, d)\}$,

where (c_i, d) represents the ith character in document *d*, and *n* is the total number of characters.

Similarly, word-based n-gram (WNG) representation models a document as a sequence of words:

$d = \{(w_1, d), (w_2, d), \ldots, (w_n, d)\}$,

where (w_i, d) represents the ith word in the document, and *n* is the total number of words, typically ignoring sentence and structural boundaries.

Word n-grams may include bigrams, trigrams, or higher-order n-grams. Both CNG and WNG representations are often referred to as fingerprints or shingles in text retrieval and plagiarism detection research. The process of generating these representations is known as fingerprinting or shingling. Each document fingerprint acts as a compact identifier, similar to a human fingerprint.

Syntax-based Detection Methods

Syntax-based approaches analyze sentence structure using parts of speech (POS) such as nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, and interjections [23].

POS tagging assigns grammatical categories to each word in a sentence. Sentence-based representation begins by splitting text using punctuation marks such as full stops, exclamation marks, and question marks. After segmentation, POS taggers are applied to generate syntactic structures.

In addition, chunking techniques use sliding windows or windowing methods to extract larger textual units beyond phrases or sentences. POS-based chunks can also be constructed to create richer structural representations. Word order within sentences or chunks can further be incorporated as a feature for similarity comparison.

Semantic-based Detection Methods

Semantic-based approaches focus on the meaning of text rather than its surface structure [23]. These methods assess how similar phrases, sentences, or documents are in terms of conceptual meaning.

They often use lexical semantic resources such as synonyms, antonyms, hypernyms, and hyponyms. Resources like WordNet provide structured lexical knowledge that enhances semantic understanding.

Semantic similarity measures evaluate the degree to which two text segments share meaning even when they differ in wording or structure. These methods are particularly useful for detecting paraphrased plagiarism.

2.3 CONCEPTUAL REVIEW

2.3.1 PLAGIARISM DETECTION METHODS

Plagiarism detection is essential to ensure academic integrity and research quality. Recent literature highlights advancements and challenges in automated plagiarism detection in academic texts.

Existing approaches are generally classified into three categories: machine learning-based methods, semantic analysis methods, and text-matching methods.

Text-matching approaches compare a suspicious document with a reference corpus using lexical features such as n-grams or hashed word sequences. Semantic approaches go beyond surface similarity by analyzing meaning using natural language processing, information retrieval, or citation-based analysis. Machine learning methods build classification models that distinguish between original and plagiarized content using extracted features.

Each category has strengths and limitations, and performance is typically evaluated using datasets and metrics such as precision, recall, F1-score, and PlagDet.

A major challenge in plagiarism detection is the “curse of modification,” especially synonym replacement and paraphrasing, which reduces the effectiveness of surface-based methods. Addressing this requires a deeper understanding of detection techniques.

2.3.1.1 LEXICAL DETECTION APPROACHES

Lexical approaches rely exclusively on word or character occurrences to compute similarity. They are effective for detecting direct copy-paste plagiarism but perform poorly on heavily modified or paraphrased text.

Lexical methods are generally divided into n-gram comparison and vector space models.

N-gram Comparisons

N-gram-based methods compute similarity by analyzing sequences of *n* consecutive elements such as characters, words, or phrases. These methods are widely used in candidate retrieval and initial filtering stages of plagiarism detection systems.

Documents are first segmented into overlapping or non-overlapping n-grams, which are then stored as fingerprints for efficient comparison. To improve efficiency, n-grams may be hashed or compressed into numeric representations. However, hashing introduces a risk of collisions, which may lead to false positives.

Because of this trade-off, hashed fingerprints are mainly used in candidate retrieval stages where high recall is more important than high precision.

Fingerprinting typically involves three steps:

- **Chunking:** dividing text into segments (words, characters, or sentences).
- **Hashing:** converting each chunk into a numerical representation.
- **Selection:** choosing representative hash values to form the final fingerprint.

2.3.1.2 STRING-BASED DETECTION APPROACHES

String-based methods perform direct comparisons at the character or word level. These methods rely heavily on n-gram similarity measures and are considered the simplest form of plagiarism detection [30].

Torrejón and Ramos [35] applied contextual n-gram models using sorted word trigrams and 1-skip-3-grams, noting a decline in performance as paraphrasing increased. Kuppers and Conrad [36] used fixed 250-character chunks and evaluated similarity using the Dice coefficient; however, the method suffered from low recall.

Overall, these methods perform well in detecting exact or near-exact copies but become ineffective when faced with complex paraphrasing or structural modifications. Nevertheless, they maintain high precision and are often integrated into hybrid systems.

2.3.1.3 VECTOR SPACE MODEL (VSM) APPROACH

The vector space model represents documents as numerical vectors in a multidimensional space. Similarity between documents is then computed using measures such as cosine similarity.

Common weighting schemes include TF-IDF (term frequency–inverse document frequency) and TF-ISF (term frequency–inverse sentence frequency). TF-IDF is typically used for document-level comparison, while TF-ISF is more effective for sentence-level analysis.

VSM approaches are widely used in both candidate retrieval and detailed analysis phases. However, their performance is limited in capturing deep semantic relationships.

For example, experiments on the PAN-PC-2011 corpus using VSM-based models reported low recall and moderate precision, indicating limited effectiveness for paraphrased plagiarism. Performance can be improved when combined with lexical resources such as WordNet or synonym expansion techniques [33].

2.3.1.4 SEMANTIC-BASED APPROACHES

Semantic approaches analyze meaning rather than surface-level similarity. Techniques include Latent Dirichlet Allocation (LDA), machine learning models, and fuzzy logic systems [30].

These methods map words and documents into semantic spaces and measure conceptual similarity, making them effective for detecting paraphrased or translated plagiarism.

Fuzzy-based approaches, for instance, compute similarity between sentences using fuzzy set theory and threshold-based classification. Alzahrani and Salim [36] proposed a system for Arabic plagiarism detection using fuzzy similarity measures and lexical resources such as Arabic WordNet.

Hybrid models combining lexical and semantic features, such as shingling with semantic databases and Jaccard similarity, have shown improved performance in candidate retrieval and detailed analysis stages [37].

A comparative summary of plagiarism detection methods indicates that:

- Character-based and vector-based methods are effective for copy-paste detection.
- Syntax-based methods handle structural modifications.
- Semantic-based methods handle paraphrasing and conceptual plagiarism.
- Hybrid approaches provide the best overall performance.

2.3.2 SEMANTIC PLAGIARISM DETECTION MODELS

Semantic plagiarism detection models identify plagiarism based on meaning rather than surface similarity.

These models compare suspected and reference texts to detect shared ideas, themes, or concepts, even when wording differs significantly.

They are particularly effective in cases involving paraphrasing, translation, or restructured content.

Semantic detection approaches include:

1. **Lexical-based methods:** use word-level similarity, synonym matching, and embeddings.
2. **Syntactic-based methods:** analyze sentence structure using POS tagging and dependency parsing.
3. **Semantic-based methods:** focus on meaning using topic modeling, named entity recognition, and semantic similarity.
4. **Hybrid methods:** combine multiple approaches to improve accuracy.

This study focuses on hybrid semantic similarity at the sentence level, which is effective for detecting intelligent plagiarism such as paraphrasing, translation, and concept reuse.

2.3.3 AMHARIC LANGUAGE BACKGROUND

Ethiopia is a linguistically diverse country located in the northeastern part of Africa, with approximately 80 spoken languages.

These languages belong to several families, including Semitic, Cushitic, Omotic, and Nilo-Saharan groups.

Amharic is the second most widely spoken language in Ethiopia after Afan Oromo and is one of the major Semitic languages in Africa. It is spoken by more than 29 million people and serves as a working language of the federal government.

Amharic uses the Ge’ez script (Fidel), a unique writing system distinct from Latin-based alphabets. It is written from left to right and consists of syllabic characters representing consonant-vowel combinations.

2.3.4 AMHARIC WRITING SYSTEM

2.3.4.1 Alphabet

Amharic has 33 base consonants, each combined with seven vowel variations, resulting in 279 distinct characters.

Each character, known as Fidel, represents a syllable formed by a consonant-vowel combination.

2.3.4.2 Punctuation Marks

Amharic includes its own punctuation system. For example:

- **Hulet Neteb (፡):** word separator (similar to space in English)
- **Arat Neteb (::):** sentence separator (similar to full stop)
- **Netela Sereze (፣):** comma equivalent
- **Derib Sereze (፤):** semicolon equivalent

Modern usage often replaces traditional separators with English-style spacing.

2.3.4.3 MORPHOLOGY

Morphology studies word formation from smaller meaningful units called morphemes.

Amharic words are formed through complex combinations of roots and affixes, including prefixes, suffixes, infixes, and circumfixes.

For example, words such as ዳኞች (judges) are formed from root + affixes, where affixes modify meaning.

2.3.4.4 INFLECTION

Inflection refers to changes in word form to express grammatical categories such as tense, number, gender, and case.

In Amharic, verbs and nouns undergo extensive inflection, producing multiple forms from a single root.

For example: ሰበረ፡ ሰበረች, ሰበርን , etc., are all derived from a single root with different grammatical meanings.

2.3.4.5 DERIVATION

Derivation is the process of forming new words from existing roots or stems.

In Amharic, derivation involves modifying roots through vowel insertion or morphological transformation to create nouns, verbs, or adjectives.

2.3.5 ISSUES OF AMHARIC WRITING

Amharic writing presents several challenges for natural language processing:

- Character redundancy in the script
- Lack of standardized compound word writing
- Multiple spelling variations for the same word
- Inconsistent abbreviations

For example, multiple characters represent the same sound, leading to normalization difficulties.

Table 2.1 Redundancy of Characters [50]

English	Character	Alternative Forms
H	ሀ	ሐ, ኀ
S	ሰ	ሠ
A	አ	ዐ
Ts	ጸ	ፀ

2.3.6 HANDLING AMHARIC TEXT ISSUES

2.3.6.1 Stemming

Stemming removes word endings to obtain a root form. However, in Amharic, stemming may produce incorrect or incomplete meanings due to complex morphology.

2.3.6.2 Lemmatization

Lemmatization improves upon stemming by considering context and returning the correct dictionary form (lemma) of a word.

2.3.6.3 Normalization

Normalization converts text into a standard form by resolving spelling variations, redundant characters, and inconsistent word representations.

2.3.6.4 SEMANTIC STRUCTURE

Semantics studies meaning in language. A word's meaning depends on its context and surrounding words.

Lexical items can be analyzed through semantic relationships such as synonymy, antonymy, hypernymy, hyponymy, and homonymy.

2.4 RELATED WORK / EMPIRICAL REVIEW

Many studies have explored plagiarism detection using different methodologies. However, no single method effectively handles all plagiarism types.

Vector space models are effective for simple copying but weak in detecting paraphrasing. Semantic models such as LDA and fuzzy logic improve detection of disguised plagiarism.

For example, Sanchez-Perez et al. [42] applied TF-ISF with cosine similarity, achieving moderate performance. Other studies using TF-IDF and cosine similarity showed limited effectiveness in detecting paraphrased content.

Advanced systems combining semantic role labeling and syntactic analysis have achieved improved performance, but often rely on limited datasets.

Overall, hybrid approaches combining lexical, syntactic, and semantic features show the most promise for robust plagiarism detection.

2.5 RESEARCH GAP

Despite extensive research in plagiarism detection, several gaps remain, particularly for Amharic text processing.

Most existing systems are designed for English and depend on large datasets and advanced NLP tools, which are not available for Amharic [21], [32], [41]

Additionally, many studies focus on either lexical or semantic methods independently, limiting their ability to detect both exact and paraphrased plagiarism [20], [24], [29]

Furthermore, most research is based on formal academic texts, while little attention is given to noisy and informal social media content [7], [30], [35].

Finally, preprocessing techniques such as normalization and stemming for Amharic remain underdeveloped.

Therefore, this study proposes a hybrid plagiarism detection model combining lexical and semantic approaches tailored specifically for Amharic social media news texts.

2.6 CONCEPTUAL FRAMEWORK

The Hybrid-Based Plagiarism Detection System for Amharic News Articles on Social Media Platforms has a framework that shows the main parts of the system. This framework explains how Amharic news articles are gathered, processed, analyzed and categorized using a Hybrid-Based Plagiarism Detection System that combines analysis techniques.

The Hybrid-Based Plagiarism Detection System for Amharic News Articles on Social Media Platforms takes in source news articles and suspicious news articles from media platforms and online news sources. The data that is collected is then cleaned up. Prepared for analysis through steps like making everything the same breaking down the text into words removing common words and cleaning the text.

After the data is cleaned up the Hybrid-Based Plagiarism Detection System for Amharic News Articles on Social Media Platforms finds the features of the text. It looks at the words, the way the sentences are structured and the meaning of the words in context. The Hybrid-Based Plagiarism Detection System for Amharic News Articles on Social Media Platforms then uses these features in an engine that combines multiple ways of measuring how similar things are and uses machine learning methods.

The Hybrid-Based Plagiarism Detection System for Amharic News Articles on Social Media Platforms calculates how similar the documents are and decides if they are plagiarized or not. Finally the Hybrid-Based Plagiarism Detection System for Amharic News Articles on Social Media Platforms makes reports that show how similar the documents are and what kind of plagiarism was found. The Hybrid-Based Plagiarism Detection System for Amharic News Articles, on Social Media Platforms is tested to see how well it works by looking at how accurate, precise and effective it's performance.

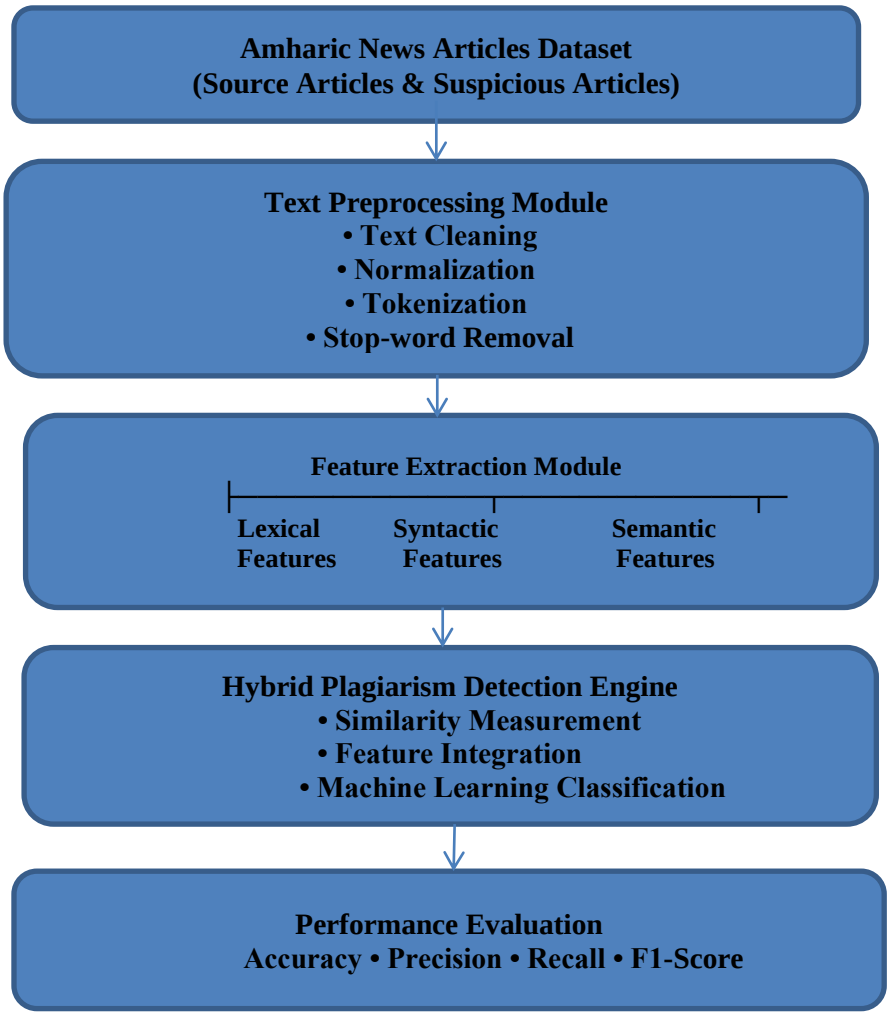


Figure 2.2. Conceptual Framework of the Proposed Model

2.7 CHAPTER SUMMARY

This chapter reviewed existing studies on plagiarism detection and categorized them based on methodological approaches and language focus.

Most prior work has been conducted on English texts using statistical and semantic methods. These findings provide a foundation for the current study.

The chapter also discussed the Amharic language, including its alphabet, punctuation, morphology, and structural challenges.

Key issues such as character redundancy, spelling variation, and lack of standardization were identified as major challenges in processing Amharic text.

The chapter concluded by emphasizing the importance of normalization, stemming, and semantic understanding in effectively processing Amharic documents for computational tasks.

CHAPTER THREE

RESEARCH DESIGN AND METHODOLOGY

3.1 OVERVIEW

The choice of any research approach is guided by the nature of the research problem. Methodology refers to the overall framework of methods and procedures applied throughout the research process. It addresses key questions such as why specific data was collected, what type of data was gathered, where the data originated, and how it was analyzed in relation to the research questions.

This chapter presents the methodology used to construct the dataset, select samples, and apply appropriate techniques to achieve the research objectives. It also outlines the system architecture, which consists of four main layers: input source and target texts, preprocessing using morphological analysis, fingerprint matching and n-gram analysis for detecting direct or minor modifications, and semantic analysis for identifying paraphrased content.

Research design refers to the structured plan used to collect relevant data and the analytical procedures applied to interpret it. It defines how data is gathered, processed, and analyzed to answer the research questions effectively.

3.2 RESEARCH DESIGN

Research design describes the overall strategy adopted by the researcher in conducting a study. It explains how the research problem is formulated, how objectives are defined, and how conclusions are drawn based on collected data.

This chapter also demonstrates how data was collected, processed, and analyzed to achieve valid and reliable results aligned with the study objectives. The research employs a quantitative approach to meet experimental requirements. According to Creswell [56], research methodology refers to selecting an appropriate approach—qualitative, quantitative, or mixed—based on the nature of the study.

In this research, quantitative data was collected from multiple news sources, particularly Telegram channels. A primary data collection strategy was adopted to ensure consistency and validation across diverse datasets.

Additionally, this study allows the analysis of data within the news domain. Typically, research in this area focuses on a limited number of news organizations or a defined time period to detect plagiarism patterns. The main goal of this study is to detect plagiarism in

Amharic news texts. Therefore, data was collected from Amharic news articles and Telegram posts.

3.3 RESEARCH APPROACH

The selection of a research approach depends on the nature of the research problem, the researcher's experience, and the intended audience of the study. The three commonly recognized approaches are qualitative, quantitative, and mixed methods. These approaches are not strictly separate but often overlap in practice.

Qualitative research focuses on understanding meanings individuals or groups assign to social phenomena. It involves emerging research questions, data collected in natural settings, and inductive analysis that develops themes from observed data.

Quantitative research, on the other hand, tests objective theories by examining relationships between measurable variables. It relies on structured instruments and statistical analysis to interpret numerical data and follows a fixed format including introduction, literature review, methodology, results, and discussion.

Mixed methods research combines both qualitative and quantitative techniques. It integrates data from both approaches to provide a more comprehensive understanding of the research problem. This study adopts a quantitative approach because it best suits the objectives of the research.

3.4 SOURCES OF DATA

Data sources are generally classified based on how they are obtained. Primary data refers to data collected directly by the researcher from original sources. It is not previously published and is considered more reliable and accurate because it has not been altered.

Secondary data refers to information that has already been published or collected by others. While secondary data is useful for understanding historical trends, it may not fully capture current conditions. This study relies primarily on primary data collected from social media posts.

3.5 METHODS OF DATA GATHERING

Data collection is a systematic process of gathering information relevant to research variables in order to answer research questions and test hypotheses.

This study uses primary data collected from publicly available news articles posted on social media platforms, especially Telegram. News agencies often publish identical content across multiple platforms such as Telegram and Facebook; therefore, Telegram was selected as a consistent and accessible source.

Data was collected from various Ethiopian and international news sources including Fana Broadcasting Corporate (FBC), Walta, LTV, Addis Media Network (AMN), Prime Minister Office posts, ESAT, Ethiopian Broadcasting Corporation (EBC), BBC Amharic, Getu Temesgen, TikVAH, and Elias Meseret.

More than 100 news articles were collected to ensure diversity and representativeness. The selection of sources was based on relevance, credibility, and availability. The sampling criteria included:

- Channels that publish daily news on politics, economics, and international affairs
- Amharic-language content
- High follower engagement and influence

Telegram was chosen because it is widely used in Ethiopia, supports large-scale communication, and allows easy data extraction and monitoring.

3.6 FEATURE EXTRACTION

Amharic hybrid plagiarism detection involves two main layers: initial candidate selection and deeper semantic analysis.

Key features include:

1. **Verbatim detection**

   This identifies exact copying of text without attribution. It is detected in the first processing layer.

2. **Sentence structure analysis**

   This examines whether sentence patterns have been directly copied from a source, indicating possible plagiarism at the first layer.

3. **Paraphrasing detection**

   This identifies cases where original text is reworded using similar vocabulary or structure to avoid detection. It is handled in the second layer.

4. **Contextual analysis**

   This evaluates the meaning of text within its surrounding context to determine originality or potential plagiarism. It is also part of the second layer.

In summary, the first layer focuses on direct text similarity, while the second layer evaluates deeper contextual meaning.

3.7 SAMPLING METHODS AND TECHNIQUES

A sample design refers to the procedure used to select a representative subset of a population for analysis. It is determined before data collection begins.

A full census of all social media posts would be impractical due to the large volume of data. Therefore, purposive sampling was applied to select relevant and manageable datasets.

The study categorized data sources into government media, private news agencies, individual journalists/bloggers, and international organizations. Selection was based on criteria such as frequent posting, Amharic-language content, and audience influence.

The population of the study includes all Amharic social media posts related to politics, economics, and public affairs. Telegram was selected due to its increasing popularity in Ethiopia, ease of access, and suitability for structured data collection and monitoring.

3.8 CORPUS PREPARATION

After collecting news articles, the dataset undergoes a preparation stage that includes organizing, cleaning, filtering, and storing the data in text (.txt) format.

Basic tools such as Notepad or similar text editors were used. Cleaning and filtering are essential steps that prepare the data for annotation and later model training in Amharic plagiarism detection.

3.9 ARCHITECTURE

The proposed system is based on three main stages: preprocessing, similarity computation between sentence pairs, and filtering of plagiarized content.
If the root forms of words in the source and target documents show similarity above a defined threshold, the content is considered plagiarized. However, paraphrased content requires deeper semantic analysis.

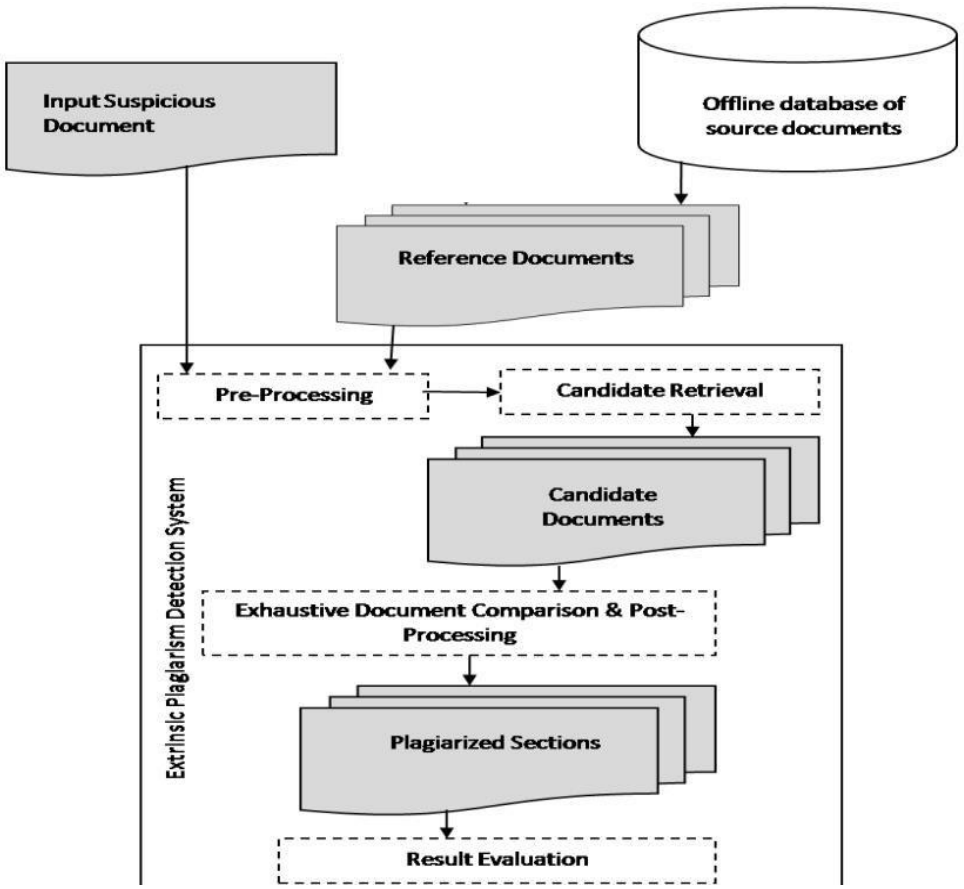


Figure 3.1. General workflow of the proposed model

Both source and suspicious documents undergo preprocessing using NLP techniques such as normalization, sentence segmentation, stop-word removal, lemmatization, and morphological analysis.

In an extrinsic plagiarism detection system, suspicious documents are compared against a reference corpus. Candidate documents are retrieved using heuristic methods to reduce computational cost.

Preprocessing Steps

- **Text normalization:** Removes irrelevant characters such as punctuation marks, special symbols, and inconsistent spellings.
- **Segmentation:** Splits text into sentences using symbols such as "։ ", "᛬ ", "?", and "!".
- **Stop-word removal:** Eliminates frequently occurring words that do not contribute to meaning.

After preprocessing, candidate retrieval is performed using similarity measures such as the Jaccard coefficient to reduce the search space.

Fingerprinting Model

This method detects plagiarism by comparing character sequences (fingerprints) across documents. It reduces computational complexity while preserving significant matches. A threshold of 20% is used to indicate possible plagiarism.

N-gram Model

This approach compares sequences of characters or words (bi-grams, tri-grams, etc.) between documents. It also uses a 20% threshold to detect potential plagiarism cases.

Latent Dirichlet Allocation (LDA)

If fingerprinting does not detect sufficient similarity, the system applies LDA to analyze semantic similarity. A document is classified as plagiarized if similarity exceeds 80%.

Fuzzy Semantic Model

This model evaluates semantic similarity using fuzzy logic. If similarity reaches or exceeds 80%, the document is considered plagiarized; otherwise, it is classified as original.

For example:

- 80% similarity → classified as plagiarized
- 60% similarity → classified as non-plagiarized

CHAPTER FOUR

EXPERIMENT RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter presents the experimental procedures and evaluation outcomes of the study on Amharic plagiarism detection in news media, focusing on identifying both verbatim copying and paraphrased content without proper acknowledgment.

To achieve this objective, a corpus was prepared and used as the primary input for the plagiarism detection system. The experimental setup describes the environment and procedures used to analyze the data and generate results. The findings are then presented and interpreted accordingly as they emerge.

The experimentation process begins after all required prerequisites, as outlined in the previous chapter, have been satisfied. The chapter first examines the characteristics of the collected corpus, which forms the foundation for all subsequent experiments. It then explains how the corpus was processed and how the results were analyzed. Finally, the outputs are presented along with brief interpretations and conclusions drawn from the experiments.

Data collection played a central role in the online author identification task. The dataset was gathered from three main online sources: Telegram public channels, personal blogs, and online magazines, covering 11 Amharic authors.

In addition, a corpus of 20 Amharic authors was incorporated from previous research conducted by Baher Hussen (2020) at the School of Information Science, Addis Ababa University.

Further data was collected by the researcher from two major sources: Kumneger Magazine and The Reporter Ethiopia Media & Communications Center. Table 5 presents the characteristics of the authors, including name, number of articles, domain area, data sources, word counts, and character counts per article.

A closer examination of the table shows that all authors share a similar domain, specifically journalism, and that the data originates from a consistent source type, namely magazines.

For extrinsic plagiarism detection, additional data collection was performed using news posts from social media platforms, gathered from 10 different sources. Among these, 7 sources belong to established organizations such as Fana Broadcasting Corporate (FBC), Walta, Addis Media Network (AMN), Ethiopian Broadcasting Corporation (EBC), BBC Amharic, ESAT, and LTV.

Out of these seven, four are government-owned media organizations, two are privately owned commercial media outlets, and one is an international news agency. The remaining two sources are individual public channels that disseminate news using publicly available information.

These individual channels primarily focus on reporting factual information rather than organizational editorial policies.

The collected news content was intentionally diversified by selecting frequently updated posts covering social, health, economic, political, and general news domains, in addition to opinion-based journalistic content. This distribution is summarized in Table 4.

For the word similarity experiment, the proposed algorithm processes both query and corpus documents using two main feature types to evaluate plagiarism detection performance.

The first feature focuses on lexical similarity, using methods such as fingerprinting or n-gram analysis to identify overlapping textual patterns.

The second feature evaluates semantic similarity between documents using techniques such as Latent Dirichlet Allocation (LDA) or fuzzy logic methods.

If neither feature identifies plagiarism or if similarity falls below a defined threshold, the document pair is considered non-plagiarized.

As previously stated, four experimental configurations were conducted to evaluate how the number of n-grams affects performance metrics such as accuracy, precision, recall, and F1-score. These configurations include:

- n-gram with LDA
- n-gram with fuzzy logic
- fingerprint with LDA
- fingerprint with fuzzy logic

4.2 EXPERIMENTAL SETUP AND EVALUATION METRICS

The input datasets were organized into separate files and processed using different Python scripts to analyze relationships between features and algorithmic models in extrinsic hybrid plagiarism detection.

The first stage of the experiment is candidate selection, which uses n-gram and fingerprint-based similarity measures. In this stage, document pairs are considered candidates if their similarity score exceeds 0.2.

At this level, only lexical similarity is evaluated to detect possible verbatim or near-verbatim copying.

Once candidate pairs pass this threshold, they proceed to the second stage, which involves semantic analysis. In this phase, Latent Dirichlet Allocation (LDA) is used to compute semantic similarity.

If the semantic similarity exceeds 0.8, the system classifies the document pair as highly likely to involve plagiarism.

The experiments were conducted using a laptop with the following specifications: HP EliteBook 840 G1 equipped with a 12th Gen Intel Core i7-1255U processor (1.70 GHz), 16 GB RAM (15.7 GB usable), 10 cores and 12 logical processors, and 1 TB SSD storage.

The system ran on Windows 11, and the development environment used Python 3.10.9 (Anaconda distribution).

Key Python libraries included NLTK, NumPy, and scikit-learn modules such as feature_extraction.text, decomposition, LatentDirichletAllocation, as well as evaluation metrics like accuracy_score, precision_score, recall_score, and f1_score.

Due to computational intensity and grid search hyperparameter tuning, performance evaluation for both Fingerprint + LDA and n-gram + LDA models required multiple iterations. Each run took approximately five minutes to complete.

The dataset used in this study consisted of 10 news sources containing 211 days of posts, 100 documents, 807 sentences, and 3,357 tokens, of which 2,275 were unique.

All texts were written by different authors, with variations in title, time, and presentation style, although many reported similar events across politics, business, economy, social issues, and national affairs.

A threshold was defined such that a document is considered plagiarized if semantic similarity exceeds 80% using the LDA-based similarity measure.

The tokens were processed using the Ferret tool, which was used to compare document similarity after converting texts into a standardized format (.txt files).

Ferret is a lightweight tool that can process multiple file formats including .txt, .rtf, .doc, and .pdf, and is compatible with Windows systems.

It converts documents into trigram representations and assigns reference identifiers to each document.

It then computes similarity scores by comparing overlapping trigrams between document pairs and ranks them from highest to lowest similarity.

The similarity score is calculated as:

number of shared trigrams ÷ total unique trigrams in both documents

This method produces similarity scores such as 0.90991, which was rounded to 0.91 for analysis purposes.

The tool also highlights matching sections between documents, allowing further manual inspection of suspected plagiarism cases.

The dataset used in this study is summarized in Table 3, which includes multiple news channels and their corresponding document statistics.

The corpus consists various channels presented as follows in table

Table 4.1 corpus document against the samples taken from various channels on Telegram

Sr no	Channel	Category	No of Documents	No of Sentences	No of Tokens	No of Unique Tokens
1.	Dimtsi weyane / DW	Government	160	1906	6626	1228
2.	Tigray TV	Government	195	292	1314	892
3.	Zara media	private	65	265	919	563
4.	Ethiopian Broadcasting/ EBC	Government	119	784	1243	880
5.	Dedebit media	Private	19	97	924	554
6	Walta TV	Government	66	113	454	703
7.	ESAT TV	Private	49	122	541	799
8	Hagere TV	Private	131	98	533	543
9.	Stalin G/silasie	Blogger	82	172	830	852
10	Haile Arafaine	Blogger	70	209	1139	923
11	Fana TV/FBC	government	90	175	1006	912
12.	Paraphrased by 12 News Authors persons (Experience from 4 Authors from Newspaper, 4 from Radio, 4 from TV)	Paraphrasers/considered as plagiarists/	32	210	1372	331
		Total	1078	4,443	16,901	9.180

4.3 RESULTS AND DISCUSSION

4.3.1 MODEL SELECTION

The results presented in Table 6 indicate that the proposed system achieved promising performance compared to related studies.

This research evaluated different n-gram settings, including 2-gram, 3-gram, 4-gram, 5-gram, 7-gram, and 10-gram, combined with LDA.

The results are summarized as follows:

n-gram with LDA

For n = 2 (bi-gram):

- Accuracy: 0.9629
- Recall: 0.9096
- Precision: 0.9306
- F1-score: 0.92

For n = 3:

- Accuracy: 0.9310
- Recall: 0.6480
- Precision: 0.8851
- F1-score: 0.7482

For n = 4:

- Accuracy: 0.9231
- Recall: 0.8769
- Precision: 0.8887
- F1-score: 0.8827

For n = 5:

- Accuracy: 0.80
- Recall: 0.4227
- Precision: 0.7455
- F1-score: 0.5395

For n = 7:

- Accuracy: 0.7778
- Recall: 0.3523
- Precision: 0.8193
- F1-score: 0.4928

For n = 10:

- Accuracy: 0.7778
- Recall: 0.4420
- Precision: 0.8684
- F1-score: 0.5858

Fingerprint with LDA

- Accuracy: 0.9730
- Recall: 0.9176
- Precision: 0.9683
- F1-score: 0.9422

Fingerprint with Fuzzy

- Accuracy: 1.0
- Recall: 0.5
- Precision: 1.0
- F1-score: 0.6667

4.3.2 FEATURE 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.

Table 4.2 Results of the various models against F1, Recall, Precision, and Accuracy

	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**

4.3.3 INTERPRETATION OF METRICS

The F1-score represents the harmonic mean of precision and recall, providing a balanced measure of a model's ability to correctly identify plagiarism while minimizing false detections.

A high F1-score indicates that the model performs well in both detecting plagiarized content and avoiding incorrect classification of original text.

Recall measures the proportion of actual plagiarized documents correctly identified by the model. Higher recall means fewer plagiarism cases are missed.

Precision measures how many of the documents identified as plagiarized are truly plagiarized. High precision indicates fewer false positives.

Accuracy represents the overall correctness of the model in classifying both plagiarized and non-plagiarized documents. Higher accuracy reflects better general performance.

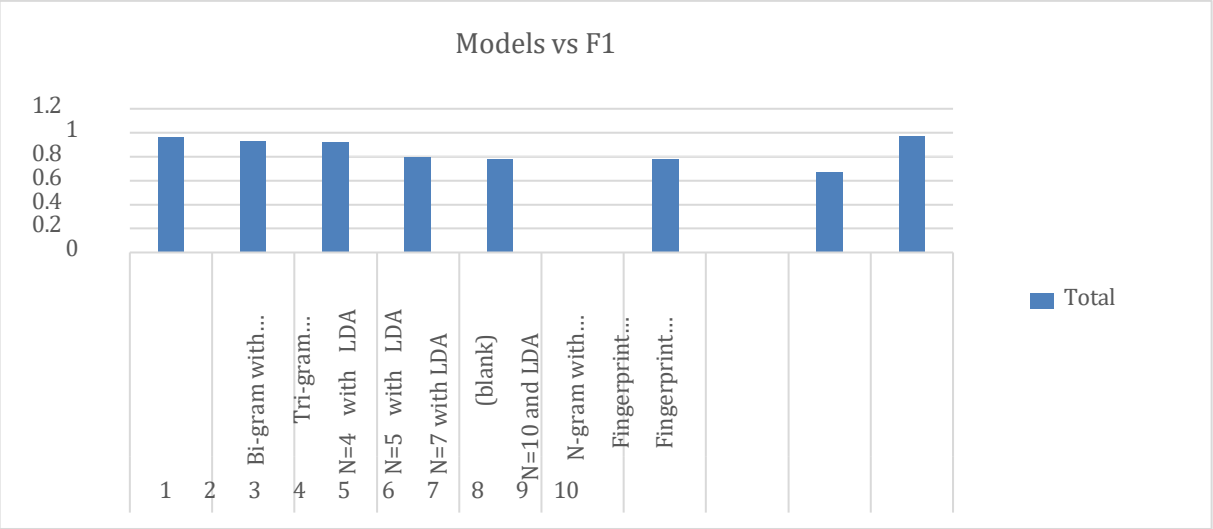


Figure 4.1: Models versus F1 shows first-layer fingerprint candidate selection and second-layer semantic processing, with the fuzzy model achieving the best overall performance

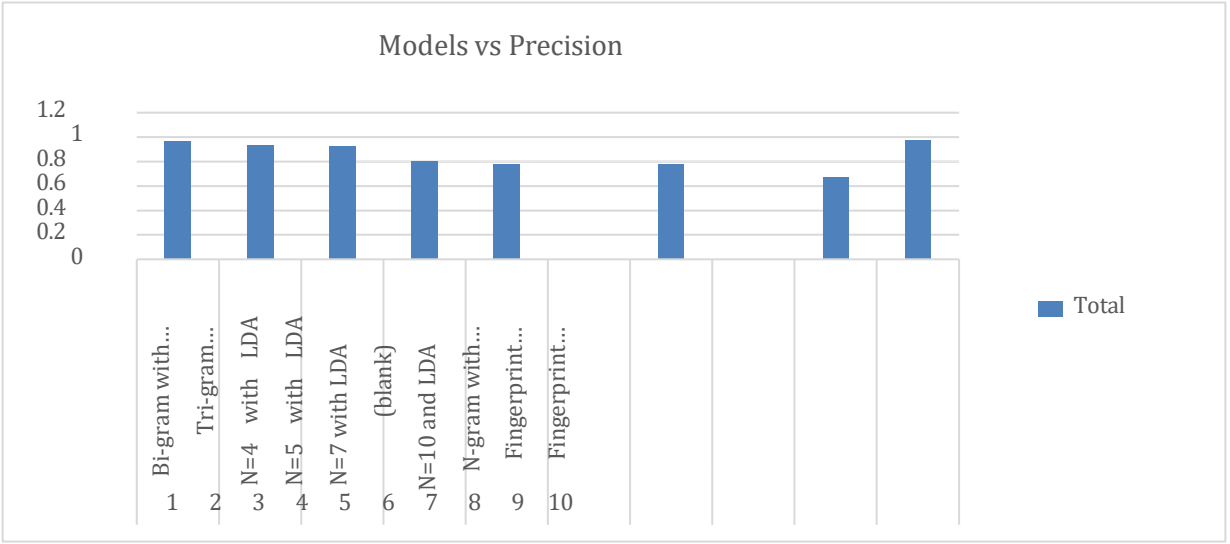


Figure 4.2: Models versus Precision shows first-layer fingerprint candidate selection and second-layer semantic processing, with the fuzzy model outperforming all others.

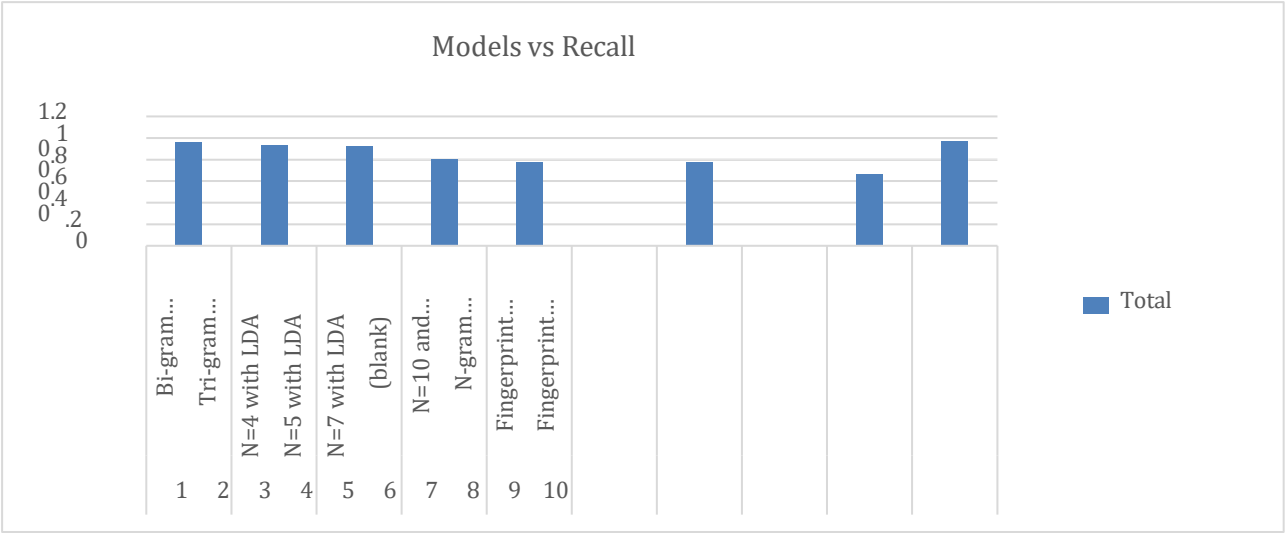


Figure 4.3: Models versus Recall shows 1st layer Fingerprint candidate selection and 2nd layer *semantics with LDA beat all*

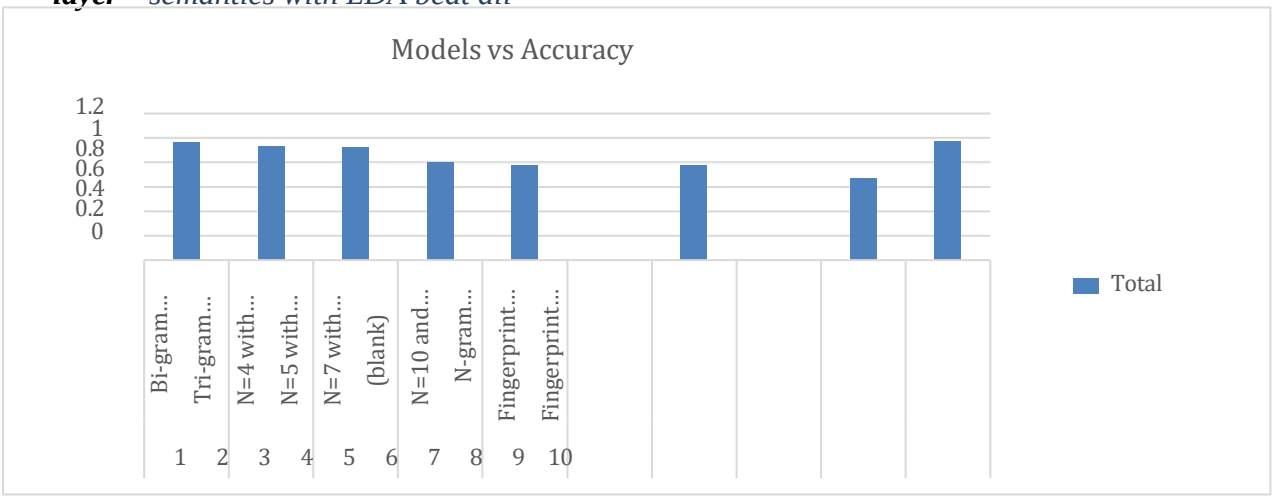


Figure 4.4. Models versus Accuracy shows 1st layer Fingerprint candidate selection and 2nd layer semantics with LDA beat all

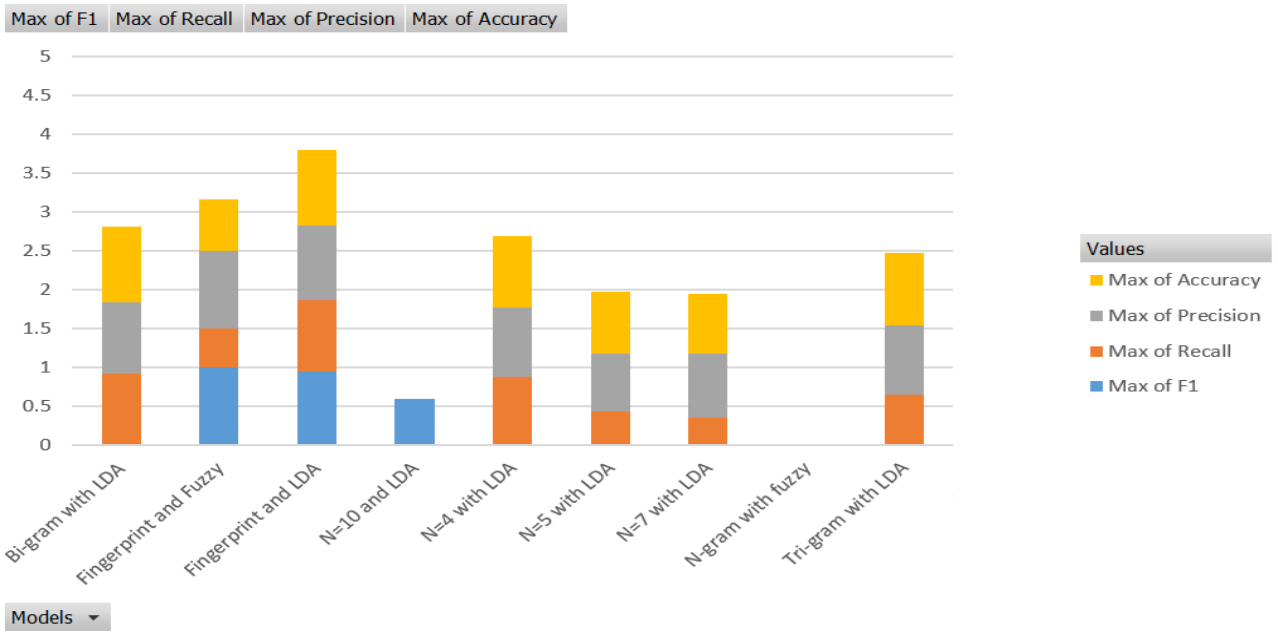


Figure 4.5: Model comparison against evaluation metrics

F1-score performance (scaled to a 5-point scale)

Precision performance (scaled to a 5-point scale)

Recall performance (scaled to a 5-point scale)

Accuracy performance (scaled to a 5-point scale)

4.3 SUMMARY

In this section, the outcomes of the different experimented models are presented and interpreted in relation to the research question. Each model produced distinct findings based on Amharic texts collected from social media posts that were treated as news content. Overall, the models demonstrated strong performance, with the exception of the n-gram combined with fuzzy plagiarism detection approach, which showed comparatively weaker results.

The two-layer plagiarism detection framework used in this study is noteworthy because it integrates fingerprint-based (or n-gram) techniques with LDA-based (latent semantic analysis) fuzzy methods. Fingerprint or n-gram approaches are particularly effective for detecting direct or near-duplicate plagiarism, while LDA-based methods are more suitable for capturing semantic or concept-level similarities. By combining these two approaches, the system achieved improved overall accuracy in plagiarism detection.

One possible interpretation of the experimental results is that the two layers are complementary in nature. The fingerprint-based layer efficiently identifies potential plagiarized segments at an early stage, whereas the LDA-based layer refines the detection by capturing deeper semantic relationships. This combination enhances the overall effectiveness of the system.

Another interpretation is that the two layers operate in a synergistic manner. In this case, the combined effect of both methods exceeds what each could achieve independently. For instance, while the fingerprint-based method may detect plagiarism at the document level, the LDA-based method can further pinpoint specific plagiarized phrases within the document. This layered insight enables more precise and detailed feedback to the user.

In conclusion, the two-layer plagiarism detection experiment demonstrates the effectiveness of combining different detection techniques to improve system performance. The results indicate that integrating fingerprint-based and LDA-based approaches yields high accuracy across multiple evaluation metrics when applied to Amharic plagiarism detection.

CHAPTER FIVE

CONCLUSION AND FUTURE WORKS

5.1 CONCLUSION

This chapter presents the experiments conducted and their results in the area of plagiarism detection for the Amharic language, specifically within the news domain. With the rapid growth of the internet and social media, information is widely shared between content providers and consumers. However, this also increases the risk of text being used without proper attribution, leading to plagiarism issues such as presenting others' work as one's own.

The experiment was designed to achieve the research objectives by developing a plagiarism detection approach for Amharic text using Natural Language Processing (NLP) and n-gram-based features to detect both verbatim and semantic plagiarism. The system estimates the level of obfuscation and paraphrasing in a suspected document by comparing it against a reference corpus.

For initial candidate selection, n-gram analysis was applied with different values of *n* (n = 2, 3, 4, 5, 7, and 10). If a candidate document achieved a similarity score above a threshold of 0.2 (out of 1.0), it was passed to the second layer for deeper analysis. This second layer performs semantic evaluation using Latent Dirichlet Allocation (LDA), which examines the underlying topic structure of the documents to assess potential plagiarism. It considers word usage, word order, and sentence structure to determine semantic similarity.

This two-layer approach provides several advantages compared to a single-layer model. First, it can identify a broader range of plagiarism types, including paraphrased and modified content. Second, it improves efficiency by allowing the first layer to filter out unlikely candidates, thereby reducing the computational load on the semantic analysis stage.

The experiments were carefully designed in alignment with the research problems. The corpus consisted of 100 documents collected from Telegram channels, including 7 news agency channels and 3 private blogger channels. Each suspected document was compared against the entire corpus. The results were evaluated using standard performance metrics, including accuracy, precision, recall, and F1-score, across different n-gram configurations.

The findings indicate that increasing both the corpus size and the n-gram value generally leads to a decrease in model performance. This is because a larger corpus introduces more contextual variability, making similarity detection more challenging. However, the model performs better when applied to smaller datasets, such as limited articles per channel or

blogger. This suggests that the approach is particularly suitable for detecting short and potentially malicious or uncredited content frequently shared on social media platforms.

For evaluation, a combination of fingerprint-based detection and Latent Dirichlet Allocation (LDA) was used. The system achieved the following performance results: Accuracy of 0.9729, Recall of 0.9176, Precision of 0.9683, and F1-score of 0.9422. These results demonstrate that the proposed model performs effectively in detecting plagiarism in Amharic text. However, the findings also suggest that further research is needed, especially given the linguistic complexity of the Amharic language, to further enhance performance.

5.2 FUTURE WORKS

5.2.1 CONTRIBUTIONS OF THE STUDY

The major contributions of this research include:

- The study addressed the problem of plagiarism detection in the Amharic language, contributing to the growing body of research in Amharic NLP and opening opportunities for further studies in related areas.
- An appropriate algorithmic framework was developed and applied for detecting plagiarism in Amharic text using a hybrid approach.

5.2.2 FUTURE RECOMMENDATIONS

- **Development of a user interface:** The proposed two-layer plagiarism detection system is effective; however, it may not be easily accessible to non-technical users. Therefore, developing a user-friendly interface would make the system more practical and widely usable.
- **Offline and online detection enhancement:** This study primarily focused on offline text-based plagiarism detection using stored text files. Future work should explore both offline and online plagiarism detection, including real-time analysis. In particular, incorporating publication timestamps and post history could improve the ability to detect plagiarism at the time content is posted.

References

[1] M. Shahabi, “Comparing plagiarism tools (Ferret, Sherlock, Turnitin),” *International Journal of Computational Linguistics*, vol. 3, no. 1, 2012.

[2] W. Kelemework, “Automatic Amharic text news classification: A neural networks approach,” 2013.

[3] S. Sánchez-Pérez et al., “A winning approach to text alignment for text reuse detection,” in *PAN-14*, 2014.

[4] Z. Dentzel, *How the Internet is Changing Everyday Life*, 2014.

[5] M. L. Bender, *Language in Ethiopia*, Oxford Univ. Press, 1976.

[6] A. Keller and B. Elsner, “Twitter paraphrase identification with simple overlap features and SVMs,” in *SemEval*, 2015.

[7] M. G. V. S. W. K. W. Kienreich, “Plagiarism detection in large sets of press agency news articles,” Know-Center, 2015.

[8] M. A. Salahli, “Measuring semantic relatedness between words,” 2009.

[9] S. S. A.-N. Asad A., “A linguistic treatment for automatic external plagiarism detection,” 2017.

[10] C. Halupa and D. U. B., “Faculty perceptions of student self-plagiarism: An exploratory multi-university study,” *Journal of Academic Ethics*, 2013.

[11] T. W. S. Rahman et al., “Multilayer SOM with tree-structured data for document retrieval and plagiarism detection,” IEEE, 2009.

[12] B. H. Geletu, “Authorship attribution model for Amharic documents using machine learning,” M.S. thesis, 2020.

[13] J. Creswell, *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*, 2003.

[14] G. Cosma and J. M. Joy, “Towards a definition of source-code plagiarism,” *IEEE Transactions on Education*, 2008.

[15] S. Sánchez-Vega et al., “Determining and characterizing reused text for plagiarism detection,” *Expert Systems with Applications*, 2013.

[16] M. T. and W. Menzel, “Amharic part-of-speech tagger for factored language modeling,” 2009.

[17] N. B. A. Patil, “Survey on different plagiarism detection tools and software,” 2016.

[18] H. Molla, “A study of journalistic professional practices in Ethiopian television news productions,” 2020.

[19] F. K. Al-Jibory et al., “Hybrid system for plagiarism detection on scientific papers,” *Turkish Journal of Computer and Mathematics Education*, 2021.

[20] S. Alzahrani and N. Salim, “Fuzzy semantic-based string similarity for extrinsic plagiarism detection,” PAN-10 Workshop, 2010.

[21] Oumaima Hourrane et al., “Survey of plagiarism detection approaches and big data techniques,” 2017.

[22] M. Bouville, “Plagiarism: Words and ideas,” *Science and Engineering Ethics*, 2008.

[23] H. L. S. and I. Gashaw, “Dictionary-based Amharic–Arabic cross language information retrieval,” 2016.

[24] R. I. S. A. A. J. V. P. Muna AlSallal et al., “An integrated machine learning approach for extrinsic plagiarism detection,” 2016.

[25] Fox News, “Time CNN suspend Zakaria for plagiarism,” 2015.

[26] M. Abate and Y. Assabie, “Amharic morphological analyzer using memory-based learning,” 2014.

[27] W. T. Fitch, *Unity and Diversity in Human Language*, 2011.

[28] M. S. B. C. A. a. R. P. Potthast et al., “Evaluation framework for plagiarism detection,” COLING, 2010.

[29] S. and M. S. Otari, “Plagiarism detection—Different methods and their analysis,” 2014.

[30] M. G. V. S. W. K. W. Kienreich, “Plagiarism detection in large sets of press agency news articles,” 2015.

[31] BBC News, “World US Canada,” 2019.

[32] S. M. Alzahrani et al., “Understanding plagiarism: Linguistic patterns, textual features, and detection methods,” IEEE, 2012.

[33] A. A. A. Salawu, *Language Policy, Ideologies, Power and the Ethiopian Media*, 2015.

[34] C. Grozea et al., “Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection,” PAN Workshop, 2009.

[35] Z. L. Lu Lu, “DNAP: Detection of news article plagiarism,” 2021.

[36] D. Pecorari and P. Shaw, “Plagiarism in second-language writing,” *Language Teaching*, 2014.

[37] I. Yilmaz, “Plagiarism? No, we’re just borrowing better English,” 2007.

[38] M. E. Menai, “Detection of plagiarism in Arabic documents,” *IJITCS*, 2012.

[39] C. Boyd, “What is latent semantic indexing,” 2018.

[40] Lovepreet A. et al., “A new hybrid technique for detection of plagiarism,” *Arabian Journal for Science and Engineering*, 2020.

[41] H. A. C. and D. K. Bhattacharyya, “Plagiarism: Taxonomy, tools and detection techniques,” 2018.

[42] Z. Sintayehu, “Automatic classification of Amharic news items,” M.S. thesis, 2001.

[43] S. and B. M. Suchomel, “Heterogeneous queries for synoptic and phrasal search,” PAN-14, 2014.

[44] D. Weber-Wulff, “Test cases for plagiarism detection software,” 2010.

[45] A. A. A. Salawu, *Language Policy, Ideologies, Power and the Ethiopian Media*, 2015.

[46] M. A. Potthast et al., “Evaluation framework for plagiarism detection,” COLING, 2010.

[47] I. Abakush, “Methods and tools for plagiarism detection in Arabic documents,” 2016.

[48] S. and N. S. Alzahrani, “Fuzzy semantic-based string similarity for extrinsic plagiarism detection,” 2010.

[49] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2000.

[50] I. F. A. Badiozaman, “Paraphrasing challenges faced by Malaysian ESL students,” 2014.

[51] B. N. J. R. Graves, “Understanding innovations in journalistic practice,” 2016.

[52] S. S. A.-N. Mohran H. Al-Bayed, “Intelligent multi-language plagiarism detection system,” 2018.

[53] M. S. H. A.-. T. Farah K. Al-Jibory et al., “Hybrid system for plagiarism detection,” 2021.

[54] F. P. Cotterell, “Amharic word classes,” *Journal of Ethiopian Studies*, 1964.

[55] G. A. and M. Getachew, “Manual annotation of Amharic news items with POS tags,” 2006.

[56] J. S. Comas-Forgas et al., “The copy and paste generation: Plagiarism amongst students,” 2006.

[57] S. and P. Pecorari, “Plagiarism in second-language writing,” 2014.

[58] C. & B. D. U. Bhattacharyya, “Plagiarism: Taxonomy, tools and detection techniques,” 2018.

[59] T. H. Gebremariam, “Amharic text retrieval using LSI with SVD,” M.S. thesis, 2003.

[60] M. S. B. C. A. a. R. P. Potthast et al., “An evaluation framework for plagiarism detection,” 2010.

[61] H. Z. Chow and T. W. S., “A coarse-to-fine framework to efficiently thwart plagiarism,” 2011.
[62] V. K. and D. Gupta, “Study on extrinsic text plagiarism detection techniques and tools,” 2016.
[63] W. Black et al., “Introducing the Arabic WordNet project,” 2006.
[64] A. Hliaoutakis et al., “Information retrieval by semantic similarity,” 2015.
[65] F. K. Al-Jibory et al., “Hybrid system for plagiarism detection,” 2021.
[66] S. S. N. I. R. A. Asad A. et al., “A linguistic treatment for automatic external plagiarism detection,” 2017.

Appendix I

```
# ========================
# IMPORTS
# ========================

import os
import re
import numpy as np

import nltk
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
from nltk.tokenize import word_tokenize

from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.metrics.pairwise import cosine_similarity
from fuzzywuzzy import fuzz

# ========================
# FILE READING
# ========================

def read_unicode_files(directory_path):
    """
    Reads all UTF-8 text files from a directory.
    """
    documents = []
    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()
                    documents.append(content)
```

```
            except Exception as e:
                print(f"Error reading {file_name}: {e}")
                continue
    return documents

# ========================
# TEXT PREPROCESSING
# ========================

def preprocess_text(text):
    """
    Lowercase, tokenize, remove stopwords, and stem text.
    """
    stop_words = set(stopwords.words("english"))
    stemmer = SnowballStemmer("english")

    text = text.lower()
    tokens = word_tokenize(text)

    cleaned = [
        stemmer.stem(word)
        for word in tokens
    if word.isalnum() and word not in stop_words
    ]
    return " ".join(cleaned)
# ========================
# N-GRAM EXTRACTION
# ========================
def extract_ngrams(text, n):
    """
    Extract n-grams from text.
    """
    words = text.split()
    ngrams = []
    for i in range(len(words) - n + 1):
        ngrams.append(" ".join(words[i:i+n]))
    return ngrams
# ========================
# CANDIDATE SELECTION (N-GRAM MATCHING)
```

```
# ========================
def select_candidates(documents, n=2, threshold=0.2):
    """
    Select candidate similar document pairs using n-gram overlap.
    """
    candidates = []
    for i in range(len(documents)):
        for j in range(i + 1, len(documents)):
            ngrams_i = set(extract_ngrams(documents[i], n))
            ngrams_j = set(extract_ngrams(documents[j], n))
            if not ngrams_i or not ngrams_j:
                continue
            overlap = len(ngrams_i.intersection(ngrams_j)) / len(ngrams_i.union(ngrams_j))
            if overlap > threshold:
                candidates.append((i, j))
    return candidates
# ========================
# FINGERPRINT GENERATION (SIMPLIFIED)
# ========================
def generate_fingerprints(text, k=5):
    """
    Create simple character k-gram fingerprints.
    """
    fingerprints = set()
    text = text.replace(" ", "")
    for i in range(len(text) - k + 1):
        fingerprints.add(text[i:i+k])
    return fingerprints
def fingerprint_matching(documents, k=5, threshold=0.3):
    """
    Detect duplicate pairs using fingerprint overlap.
    """
    matches = []
    for i in range(len(documents)):
        for j in range(i + 1, len(documents)):
            fp1 = generate_fingerprints(documents[i], k)
            fp2 = generate_fingerprints(documents[j], k)
            if not fp1 or not fp2:
                continue
            overlap = len(fp1.intersection(fp2)) / len(fp1.union(fp2))
```

```
            if overlap > threshold:
                matches.append((i, j))
    return matches
# ========================
# FUZZY MATCHING
# ========================
def fuzzy_matching(doc_pair):
    """
    Compute fuzzy similarity between two documents.
    """
    text1, text2 = doc_pair

    similarity = fuzz.ratio(text1, text2) / 100.0

    if similarity > 0.7:
        return [(text1, text2, similarity)]
    return []
# ========================
# TOPIC MODELING (LDA)
# ========================
def get_semantic_topics(documents, num_topics=10):
    """
    Apply LDA topic modeling.
    """
    vectorizer = CountVectorizer()
    X = vectorizer.fit_transform(documents)
    lda = LatentDirichletAllocation(n_components=num_topics, random_state=42)
    lda.fit(X)
    return lda.transform(X)
def get_document_topic_distribution(documents, num_topics=10):
    """
    Get topic distribution for documents.
    """
    vectorizer = CountVectorizer()
    X = vectorizer.fit_transform(documents)
    lda = LatentDirichletAllocation(n_components=num_topics, random_state=42)
    lda.fit(X)
    return lda.transform(X)
```

```
# ========================
# METRICS CALCULATION
# ========================
def calculate_metrics(tp, fp, fn):
    """
    Compute precision, recall, F1, accuracy.
    """
    precision = tp / (tp + fp + 1e-9)
    recall = tp / (tp + fn + 1e-9)
    f1 = 2 * (precision * recall) / (precision + recall + 1e-9)
    accuracy = tp / (tp + fp + fn + 1e-9)
    return f1, recall, precision, accuracy
# ========================
# PLAGIARISM DETECTION PIPELINE
# ========================
def detect_plagiarism(documents):
    """
    Two-stage plagiarism detection:
    1. Fingerprint filtering
    2. Fuzzy verification
    """
    fingerprint_pairs = fingerprint_matching(documents)
    all_duplicates = []
    for i, j in fingerprint_pairs:
        duplicates = fuzzy_matching([documents[i], documents[j]])
        all_duplicates.extend(duplicates)
    tp = len(all_duplicates)
    fp = len(fingerprint_pairs) - tp
    fn = tp  # simplified assumption
    f1, recall, precision, accuracy = calculate_metrics(tp, fp, fn)
    print("\n===== RESULTS =====")
    print("F1 Score:", f1)
    print("Recall:", recall)
    print("Precision:", precision)
    print("Accuracy:", accuracy)
    if tp > 0:
        print("\nPlagiarism detected.")
    else:
        print("\nNo plagiarism detected.")
    return f1, recall, precision, accuracy
```

```
# ========================
# MAIN FUNCTION
# ========================
def main():
    corpus_path = "D://corpus//"
    documents = read_unicode_files(corpus_path)
    # Preprocess all documents
    documents = [preprocess_text(doc) for doc in documents]
    # Run plagiarism detection
    detect_plagiarism(documents)

# ========================
# ENTRY POINT
# ========================

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

APPENDIX II

አንቀጽ	ሀያ	የሆኑ-ትን	ስድስት	አስከ	ይኸ‍ው	-	205/	-4	8/ሀ/	/አንቀጽ.164/	(ሰ)	336)	2-55/	36//
ወይም	ሆነ	በሆኑ	ምንም	ብሎ	ከርሱ	/1/	90/	248-250)	1ኢ5	62/	481//	107/	/pubic/	356-34
እንደሆነ	በኑላ	ጀምሮ	በሆነው	ከሰላሳ	መሆኑ-ን	1/	84/	248-22/	784/	613/3//	134/	335/3/	2)¾/1/	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/(ሀ)/
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/	2ኢ5	///ሀ/	284-337/
እና	መሆኑ	እንዲቆይ	በእንዳንድ	አስከ	ሰ/	66/	684/	/ሀ//	19-200/	587/	46(1)/	103/	136//	27/
ይሆናል	ሌላው-ን	ሌላው	በእያንዳንዱ	ለሆኑ-ት	ሐ/	179/	218/	237/	770/	583/	459/	298/	(አንቀጽ187//ሀ/)	269-24)
ሆና	ከሰባት	የሚሆነው	ጊዜም	አሰው	መ/	5/	671/	238-260	754/	58/3/	455//	297/	13/	26/1//2/(1)እስክ
ነው	ለሌላ	በአንዱ	አስከ	እነዚሁ	ሠ/	/ስአንቀጽ	161/	231/	750/	58-582/	129-15)	ሆን½	129-15)	
በዚህ	አለበት	አይሆንም	የሌሉት	እንደሆኑ	ረ/	(አንቀጽ	6/	229/	187/1/	(መ)	451/	29/	123/ህ	
አስከ	ሲል	ማለት	የሚሆኑት	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	ሉ/	12/	156/	224/	187//ሸ/)	57/	440/	280-283/	11-154/	
በማናቸውም	በመ-ሉ	የሆነው	ያሉ	በእነዚሁ	-ሐ/	9/	/መ//	(አንቀጽ187//ሀ/)	ማናቸWM	57-59/	435/	284-337	1/3/	
ወር	አስራ	መሆናቸው	ከሌሉት	በአምስት	-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/	84ኢ	
በላይ	አንደ	በማቀድ	ለሌሉት	በሆኑ	-ረ/	140/	145-150/	217/	187//ሀ/	56-599	423/	1/ሸ/	8/ሰ/	
ሲሆን	ቢሆንም	ጊዜና	እንኳ-ን	የሆኑ-ቢታል	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/አ	
የሆነ	ከሁሉት	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	እንኳ.	የሆኑ-ት	አሁን	ሸ.	88/	11/	95/	690//	543/	42/	(4)/	68/	
,ጋር	ቢያንስ	ከሀያ	የማናቸውም	ሰባት		9ኢ	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/	
ከሰስት	ናቸው	ለአንድ	በመ-ሉም	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)	ኑ/	(ሸ))	121-128/	9ኢ	55/	
ወደ	ወይዘሮ	ይህም	ስለሆነ	ወይ		148/	ሆነ&	861/	676/1/	510//	397/	79/	514-24/	
ማናቸውንም	ተብሎ	እነዚህ	መሆናቸውን	የሆነችን		-3	10/	20/	/1/	51(ሸ)/	397)	14/	511///5/	
ከአስር	ሳይሆን	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/	4ኢ	
እንዲሁም	ከብር	ለሆነ	ከስምንት	በሰባተኛ		68/	269-24)	2/	ቸ/	143/	379/	766/	4ኢ7/	
ሌሎች	ሆኖም	በሌሎች	ሁለቱንም	በቀር		18/	266/	/36//	66/1/ሀ/	4ኢ	376/	59/	493//	
ይህ	በታች	እንደሆነ	በሁለት	በኃሉ		90/3/	261/1	835/	66//	4ኢ9	111/	40/	481//	
ይህን	የሌላ	እንዲህ	በእስር	የአንዱን		59/	260/	2-55/	655/	14/3/	375/	27//	48/	
ከሆነ	ያላቸው	በነዚሁ	በሚል	የአንዱ		155/	26/1//2/	2)¾/1/	654/	4ኢ7	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/	/አንቀጽ101/	497/	36//	25//	46(1)/	
ከስድስት	በስተቀር	ሲሆንና	ከሙቶ	በዚህም		14/	257)	2(ሰ)	(ሰ)	14/1	11-419	248-22/	455//	
ሙቶ	መሆን	ምንጊዜም	እነዚህም	በዚህና		101/	256/	809/	65-687/	494-500	36/	248-22	45(2)/	
ያለ	ስም	ለማናቸውም	ሲኖር	ከዚህም		/4/	255)	2(ሀ)	640/	493//	በ/	243///3/	41/	
መሆኑ-ን	እንደገና	የአንድ	ለሳሳ	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	ሊሆኑ	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)/ማዘዘ	136//	35/	2/	39-641/	

APPENDIX III: Amharic Names of Punctuation Marks

Punctuation	Amharic Name	Transliteration
.	አንድ ነጥብ	and netib
:	ሁለት ነጥብ	hulet netib
:-	ሁለት ነጥብ ከሰረዝ	hulet netib keserez
…	ሦስት ነጥብ	sost netib
፣	ነጠላ ሰረዝ	netela serez
፤	ድርብ ሰረዝ	derib serez
‹‹ ››	ትዕምርተ ጥቅስ	tʾimrte təqis
!	ትዕምርተ አንክሮ	tʾimrte anqro
I	ትዕምርተ ስላቅ	tʾimrte selak
?	የ ጥያቄ ምልክት	ye tïyaqe milkit
/	እዝባር	ezbar
( )	ቅንፍ	qenif

APPENDIX IV

	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/ẖ [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]	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
ʻ/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