ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

AUTHOR IDENTIFICATION OF AMHARIC ONLINE TEXT USING STYLOMETRY AND N-GRAM FEATURES AND DIFFERENT CLASSIFICATION TECHNIQUES

BY

SISAY ZINABU GESIT

ADVISOR: SOLOMON TEFERRA ABATE(PhD)

June, 2021

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL SCIENCES

SCHOOL OF INFORMATION SCIENCE

AUTHOR IDENTIFICATION OF AMHARIC ONLINE TEXT USING STYLOMETRY AND N-GRAM FEATURES AND DIFFERENT CLASSIFICATION TECHNIQUES

BY

SISAY ZINABU GESIT

A thesis submitted to the College of Natural and Computational Sciences of Addis Ababa University in partial fulfillment of the requirements for the degree of Master of Science in Information Science and Systems (Information Science Track)

Name and Signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
Solomon Teferra (PhD)	Advisor:	_____	_____
Martha Yifiru (PhD)	Examiner:	_____	_____
Melkamu Beyene(PhD)	Examiner:	_____	_____

ABSTRACT

Users in cyberspace generated a vast amount of text data by hiding their identity. Those anonymous online text writers are distributing misinformation throughout the world. In Ethiopia also, the number of anonymous writers who are hiding their identity increases from time to time. Such writers use different languages and different social media accounts. Amharic is one of more than 80 Ethiopian languages in which misinformation are spread by anonymous online writers.
Author identification is a scientific method of identifying the author of anonymous texts by recognizing and extracting features of the author's writing style. To our knowledge, there is no authors identification model or published work to identify anonymous writers for Amharic so as to take the necessary measures. This thesis, therefore, aims at exploring the development of model for identifying Amharic text authors using stylometry, n-gram or both features and three classification algorithms: support vector machine, Naive Bayesian and Neural Network multilayer perceptron. In addition, the research investigates the effects of number of articles per author and number of authors on the performance of the author identification model. To achieve the aim of the study, experimental research methodology was followed. The necessary data (Amharic online texts) to train the model is collected and pre-processed, features are extracted and selected. The effects of increasing the number of authors and number of articles per authors are investigated in two experiments. The discrimination capability of the features and models was then tested using an anonymous Amharic online text from a suspected list. From the first experiment, the number of authors is inversely proportional with accuracy, precision, recall and f1-scores. On the other hand, these performance metrics increase as the number of articles per author increases, as the results of the two experiments show. The research findings indicate that merged features are better than the individual features for almost all models. NN-MLP-logistics has 90.47% accuracy and 90% model performance score for merged features and 27 authors. SVM Linear has 97.52% accuracy and 98% model performance score for merged features and 100 articles per author.Based on the results of the study we conclude that the Neural Network models are preferred to other classification models for small number of online text per authors to authorship identification and also the results are stable and show the best identification capability throughout number of suspects. We have conducted the experiments with limited number of authors; we recommend that further study can be conducted for more number of Amharic online text authors.

ACKNOWLEDGEMENT

First of all I'd like to thank God for the strength and wisdom He granted me in finishing this thesis.

First and foremost, I'd like to thank Solomon Teferra Abate (PhD) of Addis Ababa University's School of Information Science, College of Natural and Computational Sciences for his assistance with my thesis. His continued monitoring, encouragement, and critical support at every stage of my writing from beginning to end. Even his guiding in the proper direction starting from redesigning this thesis and in all process highly contributes to complete this thesis as my own work.

I would like to thank very much Befekadu Goraw Habteyes (PhD) research specialist/Data Analyst of Arizona State University. His support was with me to help me analyze experiment's results in a simple and graphical way.

Mulunesh Guttema, my mother, deserves special recognition for guiding me through life and shaping me into the man I am today without my father who died while I was a kid. In addition, all of my brothers and sisters have been supportive and encouraging during my years of education as well as the research and preparation of this thesis.

I'd like to express my gratitude to my wife, Tigist Goraw Habteyes, for suffering with me through all of my hardships, absences, and frustration. She also supported the family during my graduate studies. I would not have been able to complete my graduate studies without her encouragement and support. In addition to her, I'd like to thank my two sons, Aklile-Mariam and Fikir. They missed most of their refreshment time with dad during my graduate studies.

TABLE OF CONTENTS

Contents

LIST OF TABLETS

LIST OF FIGURES

CHAPTER ONE
INTRODUCTION

1.1.Background

The exponential growth of internet technology in cyberspace makes human life more sophisticated, builds productive systems, creates easily accessible communication channels and makes the world one society. Fast and easy communication is enabled via the Internet, social media, blogs, emails, and other applications. They've also given the gift of anonymity, which is much more valuable.

Because of anonymous writers, most of the world's society doesn't feel secure because of improper or illegal online users. Illegal online texts can be distributed rapidly and might attack the peaceful relationship among people at national or organizational level, child pornography, ethnic racism, spam and religious affairs. Nowadays, most of the terrorist groups use social media as a cyber-weapon and as one of their major communication channels to recruit their members. These activities have produced the concept of "cybercrime" [**1**]. Since 2020, the global pandemic has expanded cybercrime and made it the biggest threat that causes huge crises all over the world.

In Ethiopia, social media play a vital role for organizations to communicate with people and most of them have their own account on social media. The mainstream media are also copying their program broadcast through these social media. The social media became the biggest platform to advertise products most organizations. Nowadays, most of the people in the country prefer social media as a primary source of information [**2**]. In contrast, some social media and online text in Amharic and other ethnic groups' language in the country are being misused and people don't feel secure [**2**]. It is difficult to identify illegal social media writers and take legal actions because they have an anonymous nature. This has been a challenge in most parts of the world and some scholars have been working to find ways of identifying the true writer of social media disinformation. Author identification is a sub part of author analysis application of text mining which helps to identify the true writer of anonymous text using writing style features and machine learning techniques [**3**].

This research focuses on identifying writing style features (n-gram and stylometric), classification techniques which are effective for author identification of specific Amharic online

text from suspected lists. The technique also explores the effect of author identification accuracy and performance when the number of documents per author increases. Amharic is Ethiopian national and working language, which has about thirty-two alphabets and geez numbers. More than ninety million people speak this language as their native or their second language [**4**]. This research has been identified based on writing-style features such as stylometry (lexical, syntactic) and n-gram features; and classification algorithms that include Naïve Bayesian (NB), Neural Network (NN) and Support Vector Machine (SVM). It used experimental methods and evaluation metrics such as accuracy, precision, recall, and F-score to identify efficient classification algorithms and evaluate each group of features of Amharic online text.

1.2. Statement of the problem

Ethiopia is currently in a situation of national level violence, racism and politically related instabilities due to misinformation distributed very rapidly using social media, magazines and newspapers and the illegal writers are increasing exponentially [**4**][**1**]. Amharic social media, newspapers and magazines writers write false narratives and directly attack the country at a national level targeting at protected and sacred characteristics of the society like race, ethnicity, national origin, place of birth, religion and gender identity [**5**][**1**].The anonymous nature of social media invites the misuse of online text and makes Amharic illegal text writers the one who can't be identified [**6**]. The misuse of social media and illegal Amharic online text users is rapidly increasing from time to time in the country because of the lack of identity tracing forensic professionals [**7**].

Every language has its own characteristic writing style, which is likewise distinct for each author. The writing style of Amharic differs from many other languages in many ways. From the character set used in the language's writing system to sentence structure, from the set of punctuation used that is unique to the language to morphological complexity and depth. Additionally, the language's distinct grammatical rules and other qualities enable authors who write in it to create a wide range of styles that are unique not just to the author but also to the language. Similarly to how each author has a distinct style each language has a distinct style [**8**]. When an author writes in a given language, he or she must adhere to the language's grammar and other standards. Stylometry is one of the writing style features which help to identify authors of a text [**9**] confirmed that English and Chinese languages using Stylometry features recorded that

different results based on their result in English language outperformed than Chinese language. Likewise, [**10**] discovered that 1-grams in Chinese varied significantly from 1-grams in English, implying that n-gram characteristics vary by language. Stylometry is divided into four groups Lexical (character-based and word-based), Syntactic, Structural and Content specific features each of them described in literature review part. Based on the above and also different studies on authorship identification shows that writing style features like n-gram and stylometry features are different from language to language[**11**]. Even though the writing style features are efficient to author identification of online text, there is not enough research conducted on Amharic language yet. According to a recent (2020) [**12**], a model of author attribution of Amharic language was designed using machine learning algorithms such as SVM and NB and n-gram features. However, no research was found related to combining and comparing both stylometric and n-gram features for author identification of Amharic online texts. These methods have been used and were effective in some of the other global languages online texts, but the accuracy and discrimination capability different from language to language such as [**9**]. they were reported that Chinese and English language totally different in accuracy and discrimination capability[**13**][**6**][**9**] [**14**] [**15**].

Different studies on resourceful languages like English reported that using n-gram character based features are better for author identification of online short text than stylometric features such as lexical and synthetic. At Halmstad University in Sweden, three professionals found that n-gram character based perform better than stylometric on English short online text [**6**]. On the other hand, [**9**] obtained that stylometric features perform better for online short messages in both English and Chinese languages, but the result is different: English outperformed Chinese [**9**]. The other scholars, in the Department of Computer Science at University Campus of Amravati, India, used the online Enron corpus email dataset and analyzed it using C50 and SVM classification algorithms. The word based and character based uni-gram author discriminating capability was best for short email texts [**16**].But in resource scarce languages like Amharic, there is no enough empirical research for the language. Moreover, Amharic language does not have Natural Language Processing (NLP) tools like extracting Stylometric features from Amharic text and data or corpus to perform related research of Amharic author identification tasks which mean many collected author's text.

Findings from previous researches on author identification for other languages show that when the number of authors or suspects increases, the author identification classification

techniques discrimination capability and accuracy decreases or fluctuate the performance [**6**][**16**][**9**]. But no research performed to see the effect of increasing number of authors in Amharic language through hyper parameter tuning of Neural Network, support vector machine.

The number of texts per author directly affects the performance of author identification, which means classification algorithms need enough datasets to develop the best prediction model, but no research done how much the number of text per author affect the author identification performance [**17**] [**13**] [**18**] [**6**]. Thus, to create best prediction model with small number of texts per author which is important identifying an effective classification algorithm from SVM, NB and NN related to Amharic text corpus.

1.3.Research question

The thesis answers the following research questions.

- What type of writing style features is successful in the Amharic text Author identification?
- What is the relation between the performance of Amharic Author identification and the number of authors or suspects?
- What is the relation between the performance of Amharic Author identification and the number of articles (text) per author?
- From SVM, NB and NN which classification algorithms are more effective for Amharic online text Author identification?

1.4. Objective of the study

1.4.1. General Objective

The general objective of this research is to explore the impact of using different features, classification algorithms, number of authors and articles per author in the development of a model that identifies the author of anonymous Amharic online text from the suspected list.

1.4.2. Specific Objectives

In order to achieve the general objective the following specific objectives are drawn:-

- Conduct critical literature review on the author identification of online text of any language.
- Investigate various effective writing-style features for author identification on Amharic online texts.
- Collect Amharic online text corpus from personal blogs, telegram channels, online magazines, and online newspapers.
- Design data preprocessing, feature extraction and classification module of the model.
- Extract writing style features from online Amharic texts.
- Develop the prototype model based on the design using python3 programing language and Linux operating system.
- Doing two experimental analyses on the developed model.
- Identify best classification models that perform best on author identification of online Amharic text.
- Compare and contrast n-gram, stylometry and both combined for Amharic online text and find out which is relatively best.
- Calculate post-analysis evaluation metrics such as precision, recall and f-score with respect to features, algorithms, number of texts per author and number of authors.
- Report the result and make appropriate conclusions and recommendations for future research.

1.5. Scope and Limitation of the Study

The scope of this research focuses only on the identification of the author of anonymous Amharic online texts. Stylometry and n-gram features are used for appropriate author identification. Most of the previous study reported that these two features are out performed and frequently used for author analysis tasks, but frequent pattern mining not often used. Only three classification algorithms are also identified and used that include Naïve Bayesian, SVM and NN. These three algorithms are very popular and powerful for text classification tasks the author identification somehow similar with text classification problem. These classification algorithms are highly affected by the number of classes. The research, therefore, identified and included a specific number of authors/suspects only on Amharic online texts. Other Ethiopian languages are not included in this research. Only three online data sources such as telegram, magazines and personal blogs are used.

The proposed research has the following limitations

- To identify authors efficiently, a limited number of authors may be preferable. Based on the previous study, the number of authors for classification techniques directly affects the performance of algorithms.
- Incremental evaluations such as features extraction and classification algorithms need high performance computers and resources. Especially Neural network multi-layer perceptron and SVM hyper parameter tuning are performance intensive. So it needs high performance computer.
- Time and cost limitations to access more resources like collecting more authors' online texts corpus.

1.6. Significance of the study

The study has a paramount significance and contribution in the development of testing tools and initial study for programs developed for the national security of the country as author identification model of Amharic text. The findings suggest that the best algorithm and features to evaluate the performance of models that are developed to Author analysis related tasks. This work helps to identify the best model that would detect online Amharic text in cyberspace. Moreover, security institutions can save a huge capital and resources of the country if they can use such evaluation systems as additional for detecting the illegal online Amharic text from

suspected list and take predictive and prescriptive measures. These measures could protect the country from social crises and unfortunate conflicts due to false information before it is too late. Some additional benefits include:

- Forensic professionals of Ethiopia could use these models to evaluate the performance of their model that helps to trace the problem of identifying authors of online Amharic texts.
- This research also suggests some writing styles features which are effective for Amharic author identification tasks.
- Illegal online messages are also posted in other Ethiopian languages such as Afan Oromo, Tigrigna, and Gurgegna, etc. Thus, this research gives direction to these ethnic languages by adopting features related to those language and classification algorithms.
- No research is done before related to Amharic author identification using stylometry and n-gram together, so this research will be the first to study the behavior of Amharic online texts with both stylometry and n-gram features. Therefore, it will serve as a reference for further studies by researchers who will study author analysis for Amharic texts.

1.7. Organization of the thesis

This thesis organized as follows, chapter one presented background information of the study, statement of the problem, research questions, objective of the thesis, scope and limitation of the study, and significance of the study. Chapter two contains the literature review related to this thesis and direct related work chronologically. Chapter three described design the architecture with detailed of flowchart of the developed model and methodology of the thesis. Chapter four presented collected corpus analysis and detailed experimental setup of the models. Chapter five describes result and discussion in tabular, graphical and narrative way accordingly. Chapter six finally presented conclusion and recommendation of the thesis.

CHAPTER TWO
LITERATURE REVIEW AND RELATED WORK

2.1. **Introduction**

In this section we are trying to strongly address, understand and analyze related concepts, literatures, and methodologies that are important to this research problem. Important points that should be reviewed are Author analysis, Author identification, Nature of social media and other online data, stylometry features, n-gram features, feature extraction, machine learning and comparing text classification with author identification. It is also important to see the chronological findings related to this research and assess the research gap of the topic.

Stylometry and n-gram writing features are used for identifying writers like fingerprints are used to identify the identity of individuals. This section is organized in eight sub-sections and the first two discuss author analysis and author identification. The third and the fourth subsection discuss the two features, stylometry and n-gram. The mathematical and theoretical concepts of classification, more specific text classification were discussed in the subsequent two subsections. The last subsections discussed text extraction and pre-processing of Amharic online text.

2.2. **Author Analysis**

Author Analysis is the scientific process of analyzing the features of a written document to confirm its authorship. The reporter reviewed authorship analysis as a statistical study of linguistic and computational characteristics of the written documents of individuals [**19**]. This implies that there are various methods for authorship identification and analysis for a set of provided texts. The interest of authorship analysis and identification research has increased for over a decade. The researchers [**20**]is one of the researchers who presents a survey of the studies and the techniques used in the field of authorship analysis. She also shows the process of examining the characteristics of a piece of work in order to draw conclusions on its authorship [**21**].

Part of author categorization is one of the problems in “author analysis” [**22**]. Author analysis is a collection of fields which includes author characterization and similarity detection. The profile or characteristics of the authors are pieces of documents that determine authorship

characterization. Most preferable and sophisticated machine learning techniques have been applied for authorship analysis. In [**22**]identified four principal aspects of authorship analysis that can be applied to software forensics. These include author identification, author discrimination, author characterization and author intent determination.

This study grouped the definition of authorship analysis into three main fields. The first group is similarity detection or author verification. Similarity detection detects plagiarism by analyzing two texts [**23**]. However, the detection of similarities between two published pieces is greatly different from the identification of the author. The second group of authorship analysis is author identification or author attribution and it is a subset of author analysis that deals with discriminating some anonymous writers of a text from the suspected list of authors. Author attribution uses text categorization or text classification by extracting features represented in the feature vector [**20**]. Anonymity of cyberspace hides the identity of writers and makes author identification of online messages challenging [**18**].

The third category of authorship analysis is author characterization or Authorship profiling. Authorship characterization summarizes the characteristics of an author and generates the author profile based on his or her writings. Elements of those characteristics include gender, cultural and educational background, and also first or second language.

An examination of the relationship between authorship identification and Authorship characterization was written by Middleton Thomas and some others [**24**].

2.2.1. Author Identification

Author identification is the science of discriminating the writer of a piece of text from a list of authors. It is also called author attribution by linguistic professionals [**9**]. It is a process of exploring the behavior of a piece of writing to conclude the authorship. A linguistic study area called stylometry is the basis for author identification. Stylometry refers to analysis of writing style. Thus, author identification can be identified as the task of inferring characteristics of a document's author from the textual characteristics of the document. It is an attempt to infer the authorial characteristics of a piece of linguistic data. In other words, author identification is the method of identifying who wrote a particular text [**25**]. This is a very useful technique that could be applied to resolve arguments between two or more people as ownership of a written document.

In [**26**]used only a few more than thirty books recommender systems (RSs) surveyed to take the actual text of books into account. Another category of book RSs considers stylometry features. They are learned from an author's writing, and can help in authorship identification. Reporters of [**27**] Divided stylometry features into five parts such as lexical, syntactic, semantic, application-specific and character-based.

After proper cleaning, feature extraction and normalization of data, there are two important roles of author identification. The first step is to establish different features that can be used to differentiate authors' writing styles and the second task is to find the best classification algorithm that finds the most likely author of a given text [**19**]. Stylometric and n-gram features determine the values of features calculated on feature vectors where suspect names are class labels. Consequently, the dataset is divided into training and testing dataset to construct and validate the classification models to predict the class labels. Both the features and the form of [**28**]addressed some of the problems of authorship identification in a number of ways. According to another study by [**29**], text analytics and natural language processing concepts are used in conventional techniques to describe an author's writing style. More profound paradigms have been used to address this topic in recent years. Some features, such as semantic and syntactic features are not based on simple statistical analyses [**30**]. Therefore, it is important to discuss some of the most common author identification features in this section below.

2.2.2. Stylometry Features

It is a group of writing-style linguistic features that is based on the observation of a distinctive writing style of different writers. Based on the previous study by Zeng et al. 2006 [**13**], five types of stylometric features are identified to capture elements that are unique to the author. These are lexical, syntactic, content-specific, structural and idiosyncratic features.

Most studies exclusively concentrate on stylometric features in order to identify authors. To cope with the challenges of deciding the best set of features, the majority of researchers combine two or more forms of stylometric features[**31**]. For example, some may combine lexical with syntactic, syntactic, and other features, such as punctuation with word-length distributions, and others combine lexical with syntactic, syntactic, and other features [**31**][**32**].This research combined lexical and syntactic features because these features have shown the best discriminating power for languages with these features [**13**].

Lexical features (**F1**) are used to learn about the preferred use of characters and words of an individual. These features include frequency of individual alphabets, frequency of special characters, total number of uppercase letters, capital letters used in the beginning of sentences, average number of characters per word, average number of characters per sentence [**20**]. Amharic language has characters that are completely different from English language. However, this research adopted these groups of features to Amharic language because the features have the benefit of being able to be extended to any corpus in any language with no additional criteria other than the existence of a tokenizer[**20**].

Syntactic features (**F2**) are part of stylometric features in sentence level which include punctuations, function words, and part of speech. Syntactic features have a discriminating power which has different writing styles of sentences [**13**].

Content-specific features are domain dependent features which have the discriminating power of a person with a specific writing domain [**13**].Structural features are the organized structure and layout of all individual pieces of documents written and examine sentence and paragraph structures within documents. Structure features were suggested to e-mail message author identification [**33**].Idiosyncratic features capture elements that are unique to the author. Misspelled words, abbreviations, the use of emojis, and other special characters are examples of such features. These four Stylometric features are illustrated on the following table 1. For this research adopt those Lexical character-based and syntactic features, because those two group of stylometric features easy to develop and the previous researchers reported that relatively good performance for English and Chines [**9**].

Table 1 Stylometric features for English Source from [9].

Features	Description
Lexical features	
Character-based features	
1. Total number of characters(C)	
2. Total number of alphabetic characters/C	
3. 3 Total number of upper-case characters/C	
4. Total number of digit characters/C	
5. Total number of white-space characters/C	
6. Total number of tab spaces/C	
7. [7–32]. Frequency of letters (26 features) A–Z	
33. [33–53] Frequency of special characters (21 features)	
Word-based features	
54. Total number of words (M)	
55. Total number of short words (less than four characters)/M	e.g., and, or
56. Total number of characters in words/C	
57. Average word length	
58. Average sentence length in terms of character	
59. Average sentence length in terms of word	
60. Total different words/M	
61. Hapaxlegomena*	Frequency of once-occurring words
62. Hapaxdislegomena*	Frequency of twice-occurring words
63. Yule's K measure*	A vocabulary richness measure defined by Yule
64. Simpson's D measure*	A vocabulary richness measure defined by Simpson
65. Sichel's S measure*	A vocabulary richness measure defined by Sichele
66. Brunet's W measure*	A vocabulary richness measure defined by Brune
67. Honore's R measure*	A vocabulary richness measure defined by Honore
68–87. Word length frequency distribution /M (20 features)	Frequency of words in different length
Syntactic Features	
88–95 Frequency of punctuations (8 features)	",", ".", "?", "!", ":", ";", " ' ", " " "
96–245 Frequency of function words (303 features)	The whole list of function words is in the appendix.
Structural Features	
245. Total number of lines	
246. Total number of sentences	
247. Total number of paragraphs	
248. Number of sentences per paragraph	
249. Number of characters per paragraph	
250. Number of words per paragraph	
251. Has a greeting	
252. Has separators between paragraphs	
253. Has quoted content	Cite original message as part of replying message
254. Position of quoted content	Quoted content is below or above the replying body
255. Indentation of paragraph	Has indentation before each paragraph
256. Use telephone as signature	
257. Use url as signature	
258. Use email as signature	
Content-specific Features	
271 – 281. Frequency of content specific keywords (11 features) "deal", "obo", "sale", "wtb", "thx", "paypal", "check", "windows", "software", "offer", "Microsoft"	

2.2.3. N-gram features

N-gram is a contiguous sequence of n items from a given sample of text or speech in the fields of computational linguistics and probability. Depending on the application, the objects may be phonemes, syllables, letters, words, or base pairs. Usually, n-grams are derived from a text or speech corpus. N-grams are also known as shingles when the items are words [**34**].N-gram represents a sequence of n elements in a text next to one another. The elements may be of any form, such as a set of characters, phrases, symbols, syllables, etc[**6**]. Example of a character based 3-gram of "ለአውሮፕላኑ ጥራትና ብቃት ወሳኝ ነው።" is ለአው,አውሮ,ውሮፕ etc. and word based of 3-gram is ለአውሮፕላኑ ጥራትና ብቃት,ጥራትና ብቃት ወሳኝ and ብቃት ወሳኝ ነው።

The selected classification algorithms don't understand the collected texts in our case so we have to convert those collected text corpus into a vector using n-gram character based or word-based. So there are steps to convert text into a vector space model. Numbers are used as inputs in classification algorithms. This necessitates the translation of the texts into numerical vectors [**35**].This procedure is referred to as tokenization or vectorization. Tokenization divides the texts into words or smaller sub-texts so that the relationship between the texts and the class can be better generalized. The "vocabulary" of the dataset is defined in this way (set of unique tokens present in the data). Vectorization defines a useful numerical metric for describing these texts.The subsequent paragraphs discussed how to tokenize and vectorize n-gram models. Feature selection and normalization techniques are also discussed.

Text is defined as a set of specific n-grams in an n-gram vector: list of n adjacent tokens (typically, words). For example, in the text "The mouse ran up the clock.", the word 1-gram or unigram (n = 1) are ['the', 'mouse', 'ran', 'up', 'clock']; the word 2-gram or bigrams (n = 2) are ['the mouse', 'mouse ran', 'ran up', 'up the', 'the clock'], and so on. Tokenizing into word unigrams or bigrams offers reasonable accuracy while having less computation time.Vectorization transforms these n-grams into numerical vectors using classification algorithm models after tokenization where the text samples split into n-grams [**35**]. The indexes assigned to the unigrams and bigrams created for two texts are shown in the example below.

Texts: 'The mouse ran up the clock' and 'The mouse ran down'

Index assigned for each token: {'the': 7, 'mouse': 2, 'ran': 4, 'up': 10, 'clock': 0, 'the mouse': 9, 'mouse ran': 3, 'ran up': 6, 'up the': 11, 'the clock': 8, 'down': 1, 'ran down': 5}

Count encoding is one of the vectorizing options of texts after assigning indexes to the n-grams. Each sample text is represented as a vector that counts the number of tokens in the text. For example, since the word "the" appears twice in the text, the element corresponding to the unigram "the" is now represented as 2. 'The mouse ran up the clock' = [1, 0, 1, 1, 1, 0, 1, 2, 1, 1, 1, and 1].

A similar procedure was applied to the Amharic text. For example, "አንተ እዚህ ለብቻህ ቤትህ ውስጥ ሆነህ አንተ ነህ አሉ" = [2, 1, 1, 1, 1, 1, 1]. The problem with the above two methods is that common terms such as "a", "the" and others aren't penalized. This results in higher token count for common terms that are not useful for author identification[**35**] [**36**].Therefore, tf-idfVectorization is used to resolve these problems. Vectorizing the above example using tf-idf would be: 'The mouse ran up the clock' = [0.33, 0, 0.23, 0.23, 0.23, 0, 0.33, 0.47, 0.33, 0.23, 0.33, 0.33].

There are a number of other vector representations, but the count and tf-idf encoding mentioned above are the most famous ones. This research implemented the tf-idf character-based vector representation.

2.2.4. Classification

Categorical (discrete, unordered) class labels are predicted by classification models, called classifiers. For instance, to categorize bank loan applications as either safe or risky, we can construct a classification model to understand the knowledge at large [**37**]. Classification has many uses, including the detection of fraud, target marketing, prediction of results, development, and medical diagnosis. Researchers in machine learning, pattern recognition, and statistics have suggested several methods of classification.

Before a classification method is applied, the data has to be pre-processed and organized properly. This first step is the learning process (or training phase) in which the classifier is constructed on a training dataset by a classification algorithm. Analytically, for example, a tuple X is represented by an n-dimensional attribute vector, X D={x1, x2, …,xn} where n measurements made on the tuple from n database attributes, A1, A2, …, An, respectively. Correspondingly, another database attribute called the class label is discrete-valued and unordered. It is categorical (or nominal) in that each value acts as a class or group. The training sets are sampled randomly from the analyzed database [**37**]. This step is also known as

supervised learning because it is pre-defined to which class each training tuple's class label belongs to.

The classification method at this step is considered as learning a mapping or function, $Y_D=f(X)$, which can predict the associated class mark Y of a given tuple X. After formulating the mathematical relationship between the class label and the independent variables, the classification model was implemented using different algorithms.

The last step is to calculate the predictive accuracy of the classifiers [**37**]. A test set for both independent variables and class labels is used to validate and calculate the accuracy of the classifiers. The percentage or k-fold cross-validation affects the accuracy of a classifier on a given test set. There are lists of the classification algorithms. However, three classification algorithms such as Support Vector Machine, Naive Bayes and Neural Network, are found to be effective for author identification.

Support vector machines (SVM) are a group of supervised learning techniques used for classification, regression, and identification of outliers [**38**]. There is nothing more (or less) than dual-form linear learning machines that map their input vectors by the use of kernels to a feature space and compute the optimal hyperplane [**39**]. The hyperplane is used to optimize the difference in space between points. This allows the model designed for new instances to predict the target class. SVM overcomes both linear and non-linear issues. They are effective in high dimensional spaces, memory efficient and versatile. However, probability estimates are not given directly and over-fitting could be a major problem when the number of features is much higher than the number of samples.

Naive Bayes methods are a series of algorithms for supervised or classified learning based on the application of Bayes' theorem with the "naive" assumption of conditional independence given the value of the class variable between each pair of characteristics in a learning problem. Naïve Bayes classifiers are highly scalable, requiring a number of linear parameters. Instead of costly iterative approximation as used for many other types of classifiers, maximum-likelihood training can be done by evaluating a closed-form expression that takes linear time [**40**].

The Bayes theorem states the following relation, provided the class variable y and the dependent function vector x_1 through x_n[**40**].

$$P(y|x1,\ldots,xn)=P(y)P(x1,\ldots,xn \mid y) \div P(x1,\ldots,xn) \quad (1)$$

Using the naive conditional independence assumption that

$$P(xi|y,x1,\ldots,xi-1,xi+1,\ldots,xn)=P(xi|y), \quad (2)$$

for all i, this relationship is simplified to

$$P(y|x1,\ldots,xn)=P(y)\prod_{i=1}^{n} P(xi \mid y) \div P(x1,\ldots,xn) \quad (3)$$

Since P(x1,…,xn) is constant given the input, we can use the following classification rule:

$$P(y|x1,\ldots,xn) \propto P(y)\prod_{i=1}^{n} P(xi|y)$$

$$y\hat{} = \arg_{y} \text{Max} P(y)\prod_{i=1}^{n} P(xi|y), \quad (4)$$

and to measure P(y) and P(xi|y), use Maximal A Posteriori (MAP) estimation. P(y) is then the relative frequency of classY in the training set. Despite their somewhat over-simplified assumptions, in many real-world scenarios, including document classification and spam filtering, Naive Bayes classifiers have performed very well [39]. Naive Bayesian is advantageous because the algorithm is very simple and needs a small amount of training data. As a result, velocity is favored over higher precision in some situations. Besides, It works well with high-dimensional data including classification of text and identification of email spam [40]. However, these classifiers could be disadvantageous with respect to the assumptions made because all characteristics are not typically independent in real life. Simplicity and rapidity compromises precision [**41**].

Neural Networks reflect the actions of the human brain, enabling computer programs in the fields of AI, machine learning and deep learning to identify patterns and solve common problems [**42**].Artificial neural networks (ANNs) and simulated neural networks (SNNs) are subsets of machine learning that are at the core of deep learning algorithms. The human brain inspired their name and form, which simulates how biological neurons interact with one another [**42**].Artificial neural networks (ANNs) consist of a layer of nodes comprising an input layer, an output layer, and one or more hidden layers. Each node, or artificial neuron, is connected to the next and has a weight and threshold associated with it. If a node's performance reaches a certain threshold, the node is enabled, and data is transmitted to the next layer of the network. Otherwise, no data is passed on to the network's next layer [**42**].To learn and improve their accuracy over time, neural networks use training data. However, once these learning algorithms have been fine-tuned for precision, they become powerful tools in computer science and artificial intelligence, enabling us to rapidly classify and cluster data. In contrast to manual detection by human experts, tasks in speech recognition or image recognition will take minutes compared to hours. Google's search algorithm is one of the most well-known neural networks [**42**].To illustrate how the neural network algorithm works, consider each node as a separate linear

regression model, with input data, weights, a bias (or threshold), and an output. This is what the formula will look like:

$$\sum_i^m \quad wxi + bias = w1x1 + w2x2 + w3x3 + bias \tag{5}$$

$$\text{Output} = f(x) = \begin{matrix} 1 if \sum \quad w1x1 + bias \geq 0 \\ 0 \; if \sum \quad w1x1 + bias < 0 \end{matrix} \tag{6}$$

Weights are allocated until the input layer is calculated. These weights are used to evaluate the value of each variable, with larger ones contributing more to the output than smaller ones. Both inputs are first multiplied and then aggregated by their respective weights. The output is then passed through an activation function, which decides the output. If the output exceeds the threshold defined, the node is fired (or activated), passing data to the next layer of the network. This results in one node's output being the next node's input [**42**].

2.2.5. Text Classification

A classical problem in natural language processing (NLP) is text classification, also known as text categorization, which seeks to assign labels or tags to textual units, such as sentences, queries, paragraphs and documents[**43**]. It has a broad range of applications, including answering questions, detecting spam, analyzing sentiment, categorizing news, classifying user intentions, moderating content, and so on [**43**]. To name a few, text data can come from numerous sources, including web data, emails, chats, social media, tickets, insurance claims, user feedback, and customer service questions and answers. Text is an amazingly rich source of data. However, due to its unstructured nature, extracting insights from text can be difficult and time-consuming. Author identification could be a kind of text classification problem but it is different from text classification, because in author identification the writing style is also vital besides the content of the text.

2.2.6. Text preprocessing

As we understand, in numeric form, machine learning requires data. To translate text into a numeric vector, we used encoding techniques such as BagOfWord, Bi-gram, n-gram, TF-IDF, and Word2Vec. But before encoding, we first need to clean the text data and this process is called text preprocessing and is the very first step to solve the NLP issues [**44**].

In authorship attribution, text preprocessing is a very significant and tedious phase. The quality of the feature extraction and classification stages is determined by this preprocessing phase [24]. Tokenization, normalization, stop word removal and lemmatization are some of a variety of text preprocessing tasks. Tokenization is one of the text preprocessing tasks that transform a given text by breaking the input text into pieces or tokens or morphemes. By finding word boundaries in texts, the input texts are tokenized into a series of tokens. As boundary markers, white space and punctuation marks are used in some NLP works. Punctuation marks are thought to be powerful indicators of an author's writing style in authorship attribution studies. The pattern and punctuation marks that an author chooses to use in writing a particular text have been carefully considered to reflect the author's writing style. Amharic has its own punctuation marks, such as aratnetib (::), deribserez (፤), huletnetib (:), timihiirteankiro(!) and so on. When it comes to encodings, the fact that a given language has several encodings for the same character sets is a significant issue for tokenization.

Normalization is the method of converting previously uncanonical text into a single canonical form. It allows for separation of concerns by normalizing text before storing or processing it, because input is guaranteed to be consistent before operations are carried out on it. Text normalization involves understanding what form of text is to be normalized and how it is to be handled afterwards; there is no protocol for all-purpose normalization [**45**]. For example, in Amharic text "ይህን" and "ይሕን" should be normalized to one form "ህ" otherwise the two words considered as different words.

Stop word removal is commonly done on articles to delete some of the terms in a given study that are not really required. Stop-Words elimination is a standard filtering activity using prepositions, conjunctions, etc., that often appear in a text without any detail about the content [**44**]. In Author identification, stop words have the best discriminating power and they are one of the stylometric features sub category, syntactic features[**13**].Appendix A lists some of the Amharic stop words. The task of lemmatization entails morphological analysis of words, which

entails gathering together the different inflected forms of a word so that they can be analyzed as a single object. The method aims to map verb root forms to infinite tense while nouns are mapped to a single form [**46**].

2.2.7. Feature extraction

For further study, the writing styles of the authors must be extracted from the unstructured text. Human beings can extract features with high accuracy when given a predefined feature set, stylometric and n-gram, in vector space models. However, due to the large number of online messages and the large number of writing-style features, manual feature extraction is too labor intensive and time consuming[**13**].Furthermore, different feature extraction procedures are needed due to language differences. Automatic word extraction is more difficult in Chinese, for example, since there are no word boundaries. In Amharic huletnetib(:) and space together help to separate words and aratnetib(::) to separate sentences. This problem can be solved by extracting characters from a text and reassembling words from the extracted characters [**13**].

This research developed and extracted stylometric features by counting and calculating each category (lexical and syntactic) of stylometric features. However, character based n-gram features were automatically extracted using the sklearn python library.

2.3. **Amharic Language**

Amharic is written using a special script derived from the Ge'ez alphabet. Geʿez (/ˈgiːɛz/-ግዕዝ) is the liturgical language of the Ethiopian Orthodox Tewahedo Church [**47**] It is an ancient south semitic language of the Ethiopic branch and the root language for Amharic language. The language comes from the horn of Africa, more specifically from southern Eritrea and northern Ethiopia areas. Ge'ez has evolved linguistically to develop Tigrigna and Amharic languages [**48**]. Therefore, Amharic is a semitic language, together with other semitic subfamilies and modern languages, Arabic and Hebrew. The Amharic language, also known as "Amaregna", is one of Ethiopia's most widely spoken languages. It is primarily spoken in the country's central highlands. It began as a language used by soldiers for security reasons, but it quickly spread across Ethiopia and became the country's official language since Emperor Menilik's reign [**48**].

Like many other semitic languages, the Amharic language has a richness in all parts of speech(POS) which is based on triconsonantal roots with vowel variants describing modifications to, or supplementary detail of the root form [**49**]. Compound words are represented in the Amharic writing system in a variety of ways, and there is no agreed-upon spelling pattern for compounds. As a result of this reason and the country's size, widespread linguistic dispersion, lexical variation and homophony are very common in the language [**49**]. Additionally, the language has a lot of challenges. Certain phonemes with various symbols have meaning in Geez but not in Amharic. When used in Amharic, different symbols have the same phonemes. For example, ሀ (hä) with ሐ,ኸ ,ሓ,ሃ,ኻ ,ኀ and ኅ. Such problems create inconsistency or confusion in writing the same word with different Amharic spelling and more is discussed in the following session.

2.3.1. Amharic writing system

The Amharic writing system has a number of problems that make natural language processing of Amharic documents more difficult [**50**].Every basic syllable pattern in modern Ethiopic script comes in seven different orders, representing the seven vowel sounds. The first order from seven is the most basic; subsequent orders are derived from it by less frequent modifications that indicate the various vowels. There are 33 basic symbols or forms, which result in alphabets or in Amharic "fidels"(ፊደሎች) patterns of 7*33=231 syllables [**49**]. Appendix B provides more detail on the seven orders of the Amharic full alphabet, horizontally the 33 alphabets vertically.

Consonants of various forms are one of the problems in the Amharic language. As mentioned above, all of the Amharic's scripts are derived from Geez but it did not select those symbols that are only needed for the phonemes from the Geez alphabet [53]. As a result, some phonemes with different symbols have meaning in Geez but are unknown in Amharic. As Getachew [53] says, there is no comprehensive study and standard dictionary to use and consult these symbols. For example, it is unclear whether or not to compose like "Tsähäy" (sun) as ፀኃይ፣ፀሐይ፣ፀሀይ፣ጸኃይ፣ ጸሐይ and ጸሀይ with different "hä" and "Tsä" sound and with six possibilities of one word. Wolf Leslau[**51**] observed the problem of dictionary preparation in his Concise Amharic Dictionary, as expressed in his own words:

> *"As is well known, the Amharic alphabet has various letters that are identical in the pronunciation. This is the case of ኦ and ዐ, both letters being vowel carriers and no longer consonants; ሀ, ሐ and ኀ pronounced h; ሰ and ሠ pronounced s; ጸ and ፀ pronounced ts'. As a result of the merger of these letters, there is considerable lack of consistency in the Amharic spelling."*

Another issue in Amharic is that some orders are interchangeably used. In 1982, BeletuReda[**52**] discussed the confusion between the first and fourth orders of some consonants in her study. For instance, there are six different spellings of the same word "Häyl" power in English which are ሀይል, ሃይል, ሐይል, ሓይል, ኃይል and ኀይል. Similarly, the word "ayn", meaning eye, can be written in four different ways: ዐይን, ዓይን, አይን and ኣይን. There is also no consistency in the use of the consonant "ወ" in the second order, "ዉ", and its sixth order, "ዎ".

The word, for example, dog can be found spelled ውሻ and ዉሻ can be pronounced as "wshä" and "wusha", respectively.

There are twenty characters in the Amharic numbering system. They represent the numbers one to ten, as well as multiples of ten (twenty to ninety), and hundreds and thousands. Because there is no representation for the zero (0) symbol, no place value, no comma, and no decimal point, the Amharic numbering system is unsuitable for arithmetic calculations [**53**].

The Amharic writing system has about ten punctuation marks [**54**].However, few of them are used in practice, particularly in computer systems [**55**]. In the Amharic writing system, the simple punctuation marks are HuletNetib (:) and AratNeteb (::). HuletNetib has two colon-like square dots used to separate two words but not used most of the time. Modern Amharic uses space as a word separator. AratNetib has four square dots arranged in a square pattern and serves as a sentence separator. The language also has other common types of punctuation marks like lists in Amharic texts are separated by NetelaSereze(፣), a comma-like character tailed by an American Standard Code for Information Interchange (ASCII) space and the equivalent of a semi-colon. DeribSereze (፤) may also be used to separate lists. The question mark '?' is one of the few punctuation marks borrowed from foreign languages by the Amharic writing system. List of all Amharic punctuation marks are located under Appendix C.

2.3.2. Amharic Morphology

Amharic is one of the most morphologically complex languages, with a morphological phenomenon known as root pattern morphology. Patterns consisting of a set of vowels that are inserted among the consonants of a root to form a stem. In the Amharic script, various affixes (prefix, infix, suffix, and circumfix) are used to create inflectional and derivational morphemes, adding to the morphological characteristics [**56**].

Looking at the word formation process through inflection and derivation helps to grasp the language's morphological complexity. Inflection is the process of word formation by changing words to express different grammatical categories. Inflections can be applied to nouns, adjectives, pronouns, and prepositions. Number, definiteness, cases like accusative/objective or possessive/genitive, and gender can all be inflected in Amharic nouns [59]. Verbs can be inflected for any number of factors including person, gender, number, case, tense/aspect, and

mood. Thousands of verbs (in surface forms) will be produced from a single verbal root. Verbal stems are formed by combining "root + vowels + template" [**56**].For example, the stem “seber” means ‘broke’ is made up of the root verb "ስብር” (sbr) + ee + CVCVC[**57**].It is possible to construct verbal stems such as “ሰብር” (säbr), “ሰበር” (säbär), “ሰባብር” (säbabr) and ተሰባብር(täsäbabr), “ሰበረው” (säbäräw), “ሰበረች” (säbäräč), “ሰበርን” (säbärn), “አሰበረ” (assäbärä), “ተሰበረ” (täsäbärä), “አልሰበረም” (alsäbäräm), “ካልተሰበረ” (kaltäsäbärä), “የሚሰበር” (yämisäbär), etc.

Amharic nouns can be derived from adjectives, verbal roots by embedding vowels between consonants, stems, stem-like verbs, and nouns themselves through the derivation process[**56**]. There are only a few primary adjectives in the language. Adjectives may be derived from nouns, stems, compound words, and verbal roots. Adjectives can also be created by intercalating vocalic components into roots or adding a postfix to bound stems. Different verbal stems can be used to create Amharic verbs. Due to the need for precise and quick information access, the development of Amharic language processing resources as well as digital information access and storage facilities is receiving increased attention[**58**].

Based on BayeYimam[**59**], the subject of the verb is indicated by subject suffix pronouns, as in “ገደልኩ”, meaning ‘I killed’; Optionally, the verb's direct object is labeled, as in “ገደለኝ”, meaning ‘he killed me’; Optionally, certain prepositional phrase complements are marked on the noun, for example, “እስኪገድለኝ” which means ‘until he killed me’. Bound morphemes are attached to the verb and include functional elements including negation signs, conjunctions, and certain auxiliary verbs. For example, in “አልገደልኩም” ‘I did not kill’, the negation is indicated by the prefix term “አል”. Additional morphology in Amharic verbs indicates the person, number, and gender of the verb's object (second and third person singular).

Another problem with Amharic compound words was discussed by Bender and Ferguson[**53**] in 1976. For instance, it's ambiguous either “ሽንት ቤት” "shäntbet" or “ሽንት ቤት” " shänt bet " is the correct spelling for toilet. The phrase "sämtoäl" which means ‘he has heard’ is another good example. “ሰምቶአል”, “ሰምቷል” or “ሰምትዋል” are all possible spellings. In fact, such a problem exists in various languages with words written in various ways. In English, for example, the words "recognize" and "recognise" are two different spellings of the same word.

Another issue that [**60**] mentions is the translation of foreign words into Amharic. As a result, any automatic Amharic text processing should take into account the aforementioned issues.

There is also no consistency in the spelling of abbreviations, as evidenced by the documents examined. When abbreviating the word ዓመተ ምህረት (in the year AD), for example, one can find ዓ.ም or ዓም as possible abbreviations. As a result, these types of phrases should be incorporated into a common phrase. Furthermore, the use of hyphens is often contradictory.

2.3.3. Amharic Grammar

Grammar is a collection of structural rules that regulate how sentences, clauses, phrases, and terms are placed together in a natural language. These principles govern the order in which words should be combined to form sentences. Basic problems in Amharic grammar include word order and morphological agreements. In the language, sentences are made up of verb phrases and noun phrases. Because of the numerous prefixes and suffixes, Amharic sentences are short in terms of number of words [**61**] [**62**]. Take the following sentence as an example, “Abebe ate his lunch.” “አበበምሳውንበላ::” “ው” represents the word "his" and “ው” is written as a suffix in the noun ምሳ, which means "lunch.". As a result, we can conclude that an Amharic sentence has a limited number of words relative to the English language.

In formal Amharic documents, the subject-object-verb (SOV) word order is used, rather than the subject-verb-object (SVO) sequence used in English. An example of a simple Amharic sentence:“አበበምሳውንበላ::” This means “Abebe ate his lunch” and the sentence ended with a verb “በላ”(ate). While there may be OSV sequences in some Amharic texts like “ልጁንአባቱመከረው::” that is to say “The boy is advised by his father”, where the object is suffixed by the object marker “ን” in this case in formal Amharic texts, however, this word order is seldom used [**63**].

The meaning of a sentence may be changed depending on where words are put in a sentence unless the word (sentence object) has the object marker ‘-ን'/'-n'. For example, ‘አንበሳ ጅብ ይበላል’ this means “Lion eats hyena” and ’ጅብ አንበሳ ይበላል’ this means “hyena eats Lion” the words used in both sentences have the same, but they have different meanings. ‘አንበሳ’ and ‘ጅብ’ are subject of the first and second sentences, respectively. There are no subject markers or

morphemes in Amharic nouns (affix)[**59**]. A subject, on the other hand, can be defined by its location in a sentence. ‘-ን'/'-n' is an Amharic suffix that is used as a symbol for objects. An object can easily be marked by a noun with this suffix. The Amharic words must agree with each other in addition to the order and location of words in a sentence. The verb, for example, has a subject and an object (in number, gender and person), for a noun adjective (in number and gender), adverb in conjunction with a verb (in time), etc. [**64**][**59**].

2.3.4. Amharic Parts of Speech

Like the English language and others, the parts of speech in Amharic are different. Identification of a word's function in the form of the sentence in which it appears is a part of speech tagging[**65**]. Part of speech tagging is achieved with the aid of an Amharic POS tagger, which generates part of speech from Amharic words in a sentence. The tokenized text from the tokenization component is used as the input to this component. The part of speech n-gram feature extraction component in the feature extraction stage uses the output of this component, which are words with their corresponding parts of speech[**66**].

Amharic parts of speech include nouns, pronouns, adjectives, adverbs, and prepositions. Nouns are a type of parts of speech that can be used to describe a place, a name, or a number of objects or items[**67**]. Many factors influence the development of nouns, including gender (masculine/feminine), plural or singular, informal or formal, possession, and so on. Suffixes are applied to the root of the nouns to create the various combinations. Different suffixes are used for different genders[**67**].Take, for example, the English phrase "how are you?”. In English, this sentence is the same for both men and women, but in Amharic, it is different “dehnane**h**” ”ደህናህ” and “dehnane**sh**” “ደህናሽ?.When speaking to a female, the end of the word should sound “sh”, for male, the word ends in "h" sound. If they are referring to a male or female, they form plurals by adding “woč” or “oč” to the noun (depending on whether the word ends in a vowel or a consonant)[**68**].

Pronouns are words that are used in place of nouns and can do all of the functions that nouns do[**24**]. A pronoun can function as a subject, direct object, indirect object, object, or a combination of these functions[**68**].There are ten personal pronouns in Amharic, which are “እኔ”(I),”አንተ”(you for male),”አንቺ”(you for female)

,"እሱ"(He),"እሷ"(she),"እኛ"(we),"እናንተ"(you for plural form),"እነሱ"(they),"እርስዎ"(you for respect) and "እርሳቸው"(he/she for respect)[**24**][**68**].

Prepositions are limited-number words that are commonly used before nouns to indicate their relationship to another part of a clause. Prepositions such as "የ"(ye) , "ከ"(ke) , "እንደ" (ende) are some of the examples. Verbs are words that explain an action. A collection of (usually three) consonants makes up an Amharic verb root. To build verb forms, one or more suffixes and prefixes are typically added, and they always agree with their subjects. Individual, number, and gender are all identified in verbs, and a verb form may also agree with the verb's direct or indirect object. The person, number, and second and third person singular, gender, and the object of the verb are often indicated by additional morphology in Amharic verbs[**67**]. Some examples of Amharic verbs are "መከረኝ" meaning in English "he advised me","መከረሽ" means "he advised you: feminine" and "መከረ" meaning in English 'he advised".

There are only a few primary adjectives in Amharic. Adjectives may be used on their own, but they are most widely used as a suffix to a noun. Adjectives, in general, are words that define or modify another person or object in a sentence. For instance, in the case of ቀይእስክሪቢቶ (a red pen), the adjective ቀይ (red) describes the noun እስክሪቢቶ (pen) [**62**].A part of speech that adds more detail to a noun, adjective, another adverb, a word, a clause, or a sentence is called an adverb [**62**].Some examples of Amharic adverbs are ትንሽ (a little bit), ብዙ(a lot), ዳግም(also), ሁሌጊዜ (always) etc.

Conjunction is a word that connects words, phrases, clauses, sentences, and other items. They come in a small number and can be combined with verbs, nouns, and adjectives. A few Amharic conjunction examples are እና (and), ስለዚህ(because) and ስለሆነም (Therefore). In this research we are using most of the character based features both in n-gram and stylometry. Function words are used in the syntactic stylometry feature. Lack of a well-developed Amharic POS tagger was one of the challenges in using every word of the online texts author identification purposes.

2.4. **RELATED WORK**

2.4.1. Author identification using n-gram

In [**69**] experimented with language independent profiles for authorship identification using feature extraction characters of byte level n-gram. As a corpus, the experimenters collected modern Greek weekly newspapers called TOBHMA and books from different authors of English, Chinese and Greek. The writer used dissimilarity measures as a technique and showed relatively the best result, around 83%.

The researchers [**70**]performed research on feature selection for author identification using variable length character based n-gram. The writers reported 50 different authors of English Reuters corpus. This number was a relatively large number of authors used in the report and also they used SVM classification algorithms.

Recently,[**12**]designed a model of the author attribution for Amharic document using machine learning SVM and NB. The writer used the writing style features of n-gram (character based, word based, part of speech tag, space and punctuation) with a python sklearn library of tf and tf-idf. The study also used a dimensionality reduction technique called PCA (Principal Component Analysis) to reduce irrelevant features and enhance the performance of the designed model. The writer used 10-fold cross-validation to split the training and testing sets of the collected dataset to deploy the two classification models. The results of the designed model were evaluated using the most common way of evaluation metrics called precision, recall, and f-score.

The dataset used for the training and testing of the model by this author was collected from two major sources. The first category was from “Kumneger” magazines and Reporter newspapers. These were collected for twenty (20) authors from 2013 to 2019, with different topics and more than 2000 documents. The second category was collected from poems of Bewketu Seyoum and Gebrekristos Desta, and constitutes more than 120 poems.

His SVM results achieved 86.77% accuracy with the first collected data set, the combination of writing style features character based 3-gram and word plus_pos 4-gram. The average precision, recall and f-score were 88%, 87% and 87%, respectively. The second dataset of the two authors resulted in 96% accuracy[**12**]. This research has also used the same metrics and added accuracy to get more justifiable comparison among the features and models. Using Stylometry and n-gram features and also additionally using Neural Network multilayer perceptron makes this research different from the previous Baher Hussen[**12**].works and also in

addition to this evaluating the features and the three classification algorithm in the way of by increasing number of authors and number of articles per authors.

2.4.2. Author identification using stylometry

In [**15**] have used stylometry features and studied the effects of a number, the type of feature sets, size of text, and the number of text documents per author on the author identification accuracy for both emails and text documents .The analyzed result for five authors reported 84% of accuracy, on average. In another study, [**14**] extracted the most appropriate features that represent the style of an author using Stylometry linguistic style and suggested a strong insight to define unique identity features as a fingerprint. Based on their analysis, they obtained 75.1% accuracy using SVM on 10 different authors of Portuguese language. In [**71**] conducted a similar approach to identify changes in the writing style of seven authors of English but used another supervised learning algorithm called logistic regression in addition to SVM. In [**17**] used English articles, short stories and emails of 15 maximum authors. The experimenters used SVM, A multilayer perceptron (MLP) and Decision Tree from classification algorithms and obtained 84% accuracy based on MLP.

In [**72**] analyzed SMS authors incrementally starting from 5 to 70 English SMS suspects. The analysis used uni-gram features with statistical methods of cosine similarity and Euclidean distance. The authors suggested that ten authors' dataset gives the best accuracy estimate. They also found that the cosine similarity analysis method gives 33.5% to 40% accuracy if all the 70 authors' dataset is used.

2.4.3. Author identification using stylometry and n-gram

In [**16**]analyzed and identified generalized features and computational author identification. As a computational method, they used SVM and performed better, resulting in 88% accuracy with generalized features which are stylometry and n-gram. The experimenters used a corpus of Enron corpus email dataset C50 of 50 authors. Based on their experiment, word-based and character based uni-gram features have better discriminating capability for short email texts. Analyzed author identification using statistical techniques such as Euclidean, cosine and Manhattan distance with stylometry and n-gram and experimented on 40 authors of the English tweet dataset [**6**]. The result obtained through the experiment was 92 - 98.5% accuracy and the report said that n-gram based analysis showed relatively better results.

2.4.4. Summary of related work in table

Table 2 Summary of related studies with different languages, features and algorithm techniques

Author and Reference	**Year**	**Features**	**Techniques**	**Number of Authors**	**Language**	**Accuracy**
Anderson et [**15**]	2001	Stylometry	SVMs	5	English	84%
Vlado et al.[**69**]	2003	n-gram	dissimilarity measure	9	English, Chinese and Greek	83%
EfstathiosStamatatos and John Houvardas[**73**]	2006	n-gram	SVM	50	English	74.04%
Daniel et al. [**14**]	2007	Stylometry	SVM	10	Portuguese	75.1%
Ragel, R et al. [**72**]	2013	n-gram(unigram)	cosine similarity and the Euclidean distance	70	English	39%
Smita et al. [**16**]	2015	Stylometric & n-gram	SVM	50	English	88%
Sujata et al. [**17**]	2018	Stylometry	SVM, MLP and decision tree	15	English	84%
Grigori et al. [**71**]	2018	Stylometry	SVM and logistic regression	7	English	88.9%
BaherHussen[**12**]	2020	n-gram	SVM and Naïve Bayesian	20	Amharic	87%
Nicole Mariah Sharon Belvie et al. [**6**].	2020	Stylometric and n-gram	Cosine, Euclidean and Manhattan distance	40	English	92 -98

2.4.5. Summary of Related work

Some studies, books, websites and papers about the problem and conceptual frame of author identification are reviewed and discussed in this chapter. The related work section is specially reviewed based on author identification features and techniques in different perspectives. The reviews are analyzed based on the number of suspects, the features that were used, the techniques used, the accuracy, and the language that suspects used. There are studies on different languages, but the English language constitutes the largest number of studies related to author identification. Resource scarce languages like Amharic have limited or no studies till these days.

There is only one thesis done by [**12**] related to this problem on Amharic language. However, his work focused on developing a machine learning model of author identification using only n-gram, not using stylometry. Moreover, he did not explore the effect of the number of authors and number of texts per author model performance. Such a problem affects the performance from language to language, from features to features and from techniques to techniques. Stylometry is one of the writings style features which help to identify authors of a text [**9**] confirmed that English and Chinese languages using Stylometry features recorded that different result based on their result English language outperformed than Chinese language. Likewise, Stewart Yang et .al discovered that 1-grams in Chinese varied significantly from 1-grams in English, implying that n-gram characteristics vary by language [**10**]. So no one studied how much the number of authors and the number of texts per author affects the performance in Amharic language through hyper parameter tuning. No work was done to compare the performance with Stylometry and n-gram. This thesis, therefore, addressed and compared the effect of author identification performance using character based n-gram, stylometric, and both features. It also explored the responsiveness on performance of the models as the number of authors and the number of text per authors increase or decrease in hyper parameter tuning of classification algorithms.

CHAPTER THREE
DESIGN AND METHODOLOGY

3.1. Introduction

This chapter describes in detail the Methodology and preparation of the experimental environment of the prototype system to meet the research objectives. Features are very important for authorship identification and this chapter illustrates the conceptual framework of the features described as the architecture of author identification using flowcharts and pseudo codes. First the Methodology of the research described Second, author identification architecture is described in detail with data collection, stylometry and n-gram features and sub components of Author identification architectures presented in charts and pseudo codes of text pre-processing and feature extraction as subsections.

3.2. Methodology

To answer the research questions experimental methodology is selected and it is the best alternative, because it is easy to see the models performance of cause and effect through hyper-parameter tuning of each research questions in experiment. The two experiments prepared in systematic way to investigate the effects of increasing the number of authors and number of articles per authors. The analysis and evaluation processes of the experiment were conducted using python3 programming language using an important python library such as sklearn, because most of classification algorithm and evaluation matrix integrated as built-in library. Accuracy, Precision, recall, and f-score were used as evaluation metrics to measure the performance of the features and algorithms in author identification. The supervised machine learning classification techniques to investigate the effect of changes in writing style features, classification algorithms, and number of authors on the performance of author identification on Amharic online texts. In other words, sensitivity analysis was conducted to observe the relative change in author identification performance.

3.2.1.Evaluation

Writing-style features such as n-gram and stylometric are evaluated using incremental way for Amharic text using n-gram character based features only, stylometric features only, Stylometric and n-gram (1-gram and 2-gram character-base), respectively.

The effect of text document size per suspect through hyper parameter tuning is also evaluated like inserting text document incrementally with constant value. The results were analyzed using evaluation metrics such as precision, recall, and f-score of the Amharic online text from the above incremental perspectives. This research identifies the best features that accurately identify authors of specific Amharic online text. The study also identifies the best performing classification algorithm by evaluating three classification techniques. These include SVM, NB, and NN simultaneously which means evaluate the result at the same time. Accuracy score is another metric used to compare the algorithms in addition to the three evaluation metrics.

3.3.**Author Identification Architecture**

Design high level architecture to prepare prototype classification model of a system increase the visibility through development. It is, therefore, important to prepare some Amharic author identification prototypes models. However, as one of the resource scarce languages, Amharic language doesn't have a well-developed ready-made NLP tools and dataset related to Stylometric feature extraction [**75**],but in this reseach Amharic text preprocessing and Amharic Stylometric feature extraction developed. As a result, this research attempts to show the architecture of Author identification and full descriptions in the subsequent subsections.

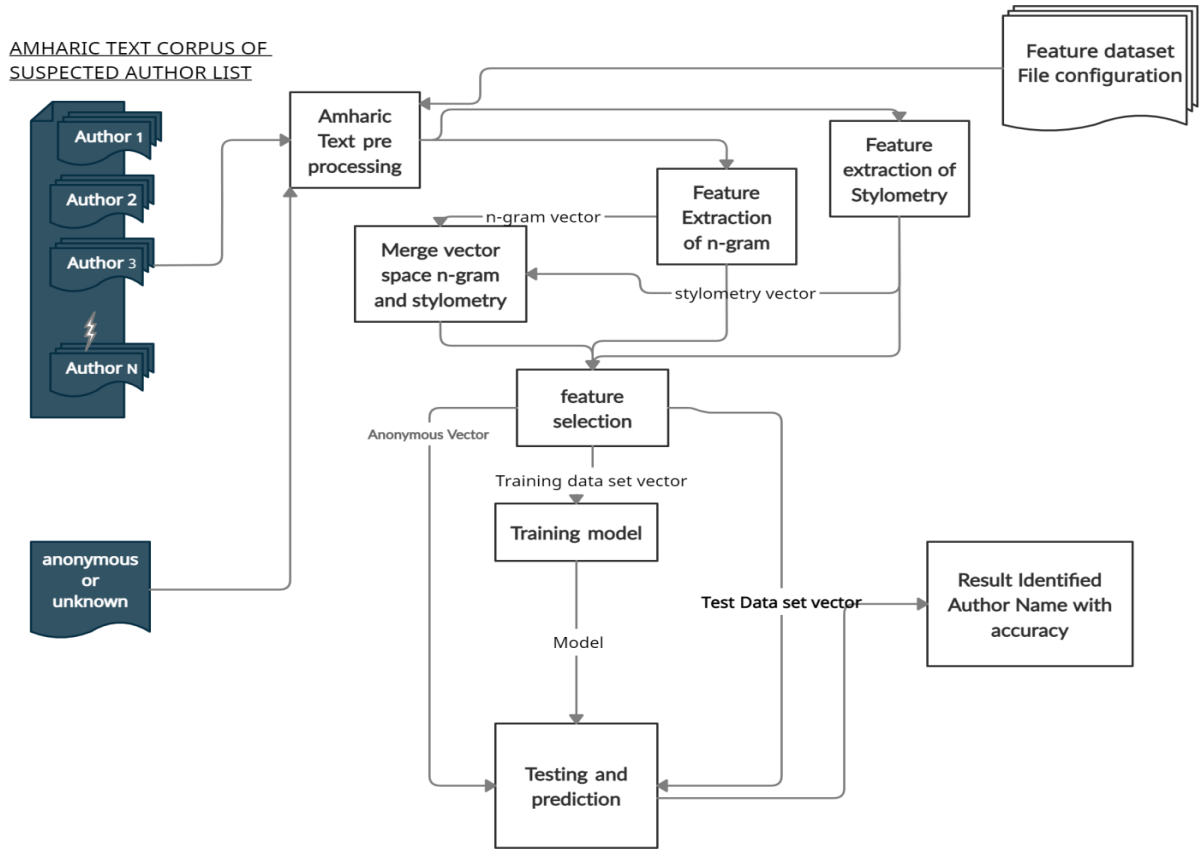

Figure 1. Architecture of author identification for Amharic online text

3.3.1. Design Process

The research used online creately.com design platform for drown a design and architecture, because this platform contain many appropriate features for design architecture ,flowcharts and which is easy to collaborate.Python3 programming language to develop a prototype model based on the design that evaluates to the performance of author identification models. sklearn python3 library was one of the tools used for feature extraction, feature selection and classification algorithms on the training and testing datasets of online text that contain built in methods and easy to import each method in the code. Python programing language is now a day's powerful and preferable to NLP related tasks.

3.3.2.Data Collection

Data collection was one of the key tasks for the online Amharic text corpus. Three online text sources were used to collect the data: Four authors corpus from magazines and personal blogs were collected manually and 20 authors' corpus collected from previous researcher [**12**].Seven authors of Amharic telegram text data was automatically exported in json format from public channel of each seven authors. The steps of export was done by opening that target public channel of telegram application then click the top right side three dot and select **export chat history** in json format. The exported Json format pre-processed using simple python code which is open source from github convert json format to text format. Form this collected corpus by using 5 fold cross validation methods hyper parameter tuning of gridsearchcv training and testing data used automatically. Why gridsearchcv applied here to answer the research questions of hyper parameter tuning applied through this method easily rather than splitting techniques.

3.3.3.Features of stylometry and n-gram

Understanding about each features of this study is very important this research dataset or corpus analysis described in detail at chapter Four section 4.2 in tabular form. In this section described the Stylometric features which are selected to this research and n-gram features.

Stylometric, n-gram and the combination of the two features might be effective for author identification. Table 3 below describes Amharic stylometric features prepared on the configuration file without putting hard code in python of this experiment prototype. Table 3 Amharic Stylometric part adopted from Zhenge table 1 Stylometric features [**9**] .

Table 3. Description of the source dataset features for Amharic language

Label	Features	Description
F1	Lexical features character-based ● Total number of Amharic alphabet/ Total number of characters ● Total number Amharic or Geez number characters/ Total number of characters ● Frequency of an Amharic alphabets(287 features) each alphabet assign as a feature whether exit or not in any text. ● Frequency of special characters (28 features)	This is a set of (ሀ ሁ ሂ ሃ ሄ ህ ሆ …. ፐ ፑ ፒ ፓ ፔ ፕ ፖ ፗ) The whole list Amharic alphabets in the Appendix C. # $ % & ' ( ) * + , - . / ; < = > @ [\ ] ^ _ ` { | } ~
F2	Syntactic Features ● Frequency of Amharic punctuations (10features).Amharic punctuation and English punctuation are different for example ፧and ፤are only Amharic punctuations. ● Frequency of Amharic function words(268 features)	:? ! :: ※፣፤፥፦-፧፨ እንደ ሆነ በኡላ ጀምሮ … The whole list Amharic function words in the Appendix A [**74**]
N-gram	Those n-gram character based (unigram and bigram character based) features extracted using python methods of class feature extraction scikit-learn library.	Forexample “ለአውሮፕላኑ ጥራትናብቃትወሳኝነው።” bigram {ለአ,አው,ውሮ,ሮፕ,ፕላ,ላኑ …} are features of bigram TfidfVectorizer...and..CountVectorizerpython functions used to extract such features.

3.3.4.Text preprocessing

In any related text classification tasks, text preprocessing tasks are mandatory because text preprocessing in machine learning plays a significant effect on the accuracy of the model. Preprocessing methods are important to extract features and stylometry and n-gram use different preprocessing components.

According to Figure 2 below, all collected suspects/authors are stored as a list of folders by authors' name and one unknown author in one database corpus file. Each folder contains the online Amharic text documents that belong to each individual author and the unidentified text within the unknown author folder.

For Stylometric preprocessing component, the list of features is not hard coded and the whole lists of features as described in the Table 3 above are inserted in a configuration text filename of "feature_Amha_alpa_speci_panc.txt" with space delimiter. These features help to, first, tokenize with space and stop word list features and then normalize into some normal form using python3. For example,ሲሉንና is one of the Amharic stop word features, syntactic features. It will be normalized in preprocessing components as ሊሆንና. This helps to count each stop word from the Amharic text with the same form. Lexical features of Amharic alphabet are not normalized, rather count is used for author identification without normalization.

In an n-gram preprocessing component, two lists are prepared in python3. The first list contains all Amhacric text in each sub directory of the corpus including unknown author text and the second list contains their corresponding Authors name from the corpus sub directory. Using these folders and text files as initial input data, the n-gram sicik-learn python library is used to extract character based n-gram features. The sicik-learn n-gram tfidfVectorizer and CountVectorizer methods convert this list of Amharic text to multi-dimensional vector space.

The flowchart in Figure 2 below, therefore, shows the step-by step preprocessing tasks of stylometry and n-gram components. The flow charts were drawn using an online tool called creately[**76**]. A closer look at the flowchart shows that **F1** represents lexical features of stylometry and **F2** represents syntactic features.

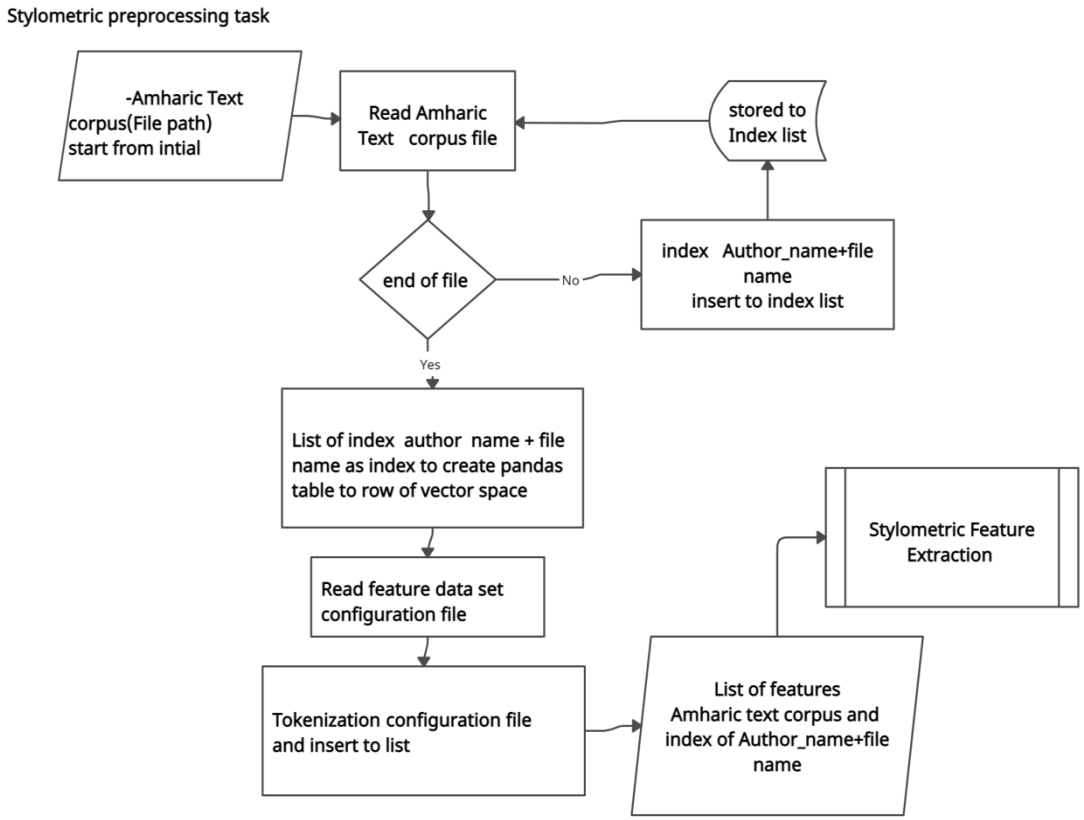

Figure 2 Preprocessing of Amharic Stylometric and n-gram features

This flow chart shows that as input configuration file and Amharic corpus and finally the output moves as input for feature extraction. The final block of diagram shows that predefined process of feature extraction.

The following Pseudo code shows the Stylometric preprocessing and directly maps with the flowchart above. The output or return from Stylometric preprocessing is an input to Stylometric feature extraction. The following code is developed on python3 on Linux notepad editor IDE.

1. **Stylometric preprocessing pseudo code**

*# **This is sudo code of stylometric preprocessing***

***Parameters**:- Amharic Text corpus(file path),Stylometric configuration file path(F1 and F2)*

***Method:-**Stylom Preprocessing*

Initialize

Index list of table

Total list of stylometric features(F1 and F2)

for each file subdirectory name(the author name) and file name

if total file not traversed

subdirectory name(Author name) concatenation with file name append to index list of table

end if

end for

open Stylometric configuration file path(F1 and F2)

Tokenize with space opened Stylometric configuration file path(F1 and F2)

Normalize only words from Tokenized Stylometric configuration file path(F1 and F2)

and insert to Total list of stylometric features(F1 and F2)

***returntype**:-index list of table*

Total list of stylometric features(F1 and F2)

and Amharic Text corpus(file path)

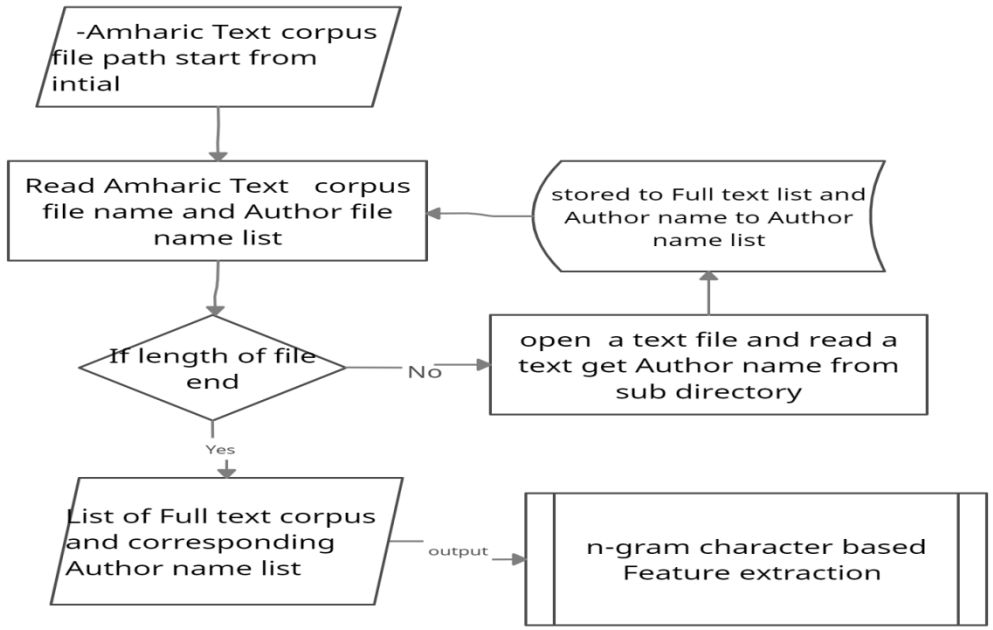

Figure 3 Preprocessing of n-gram features

Based on the above flowchart figure 3 the next pseudo code describes how to prepare n-gram feature preprocessing tasks. Like the stylometry feature, the return values of the preprocessing of n-gram serves as input for n-gram feature extraction. Using the same IDE, python3 programing language, Ubuntu 18.04 LTS linux platform and notepad++ editor, the following pseudo code was developed for n-gram preprocessing component

2. N-gram preprocessing pseudo code

 # This is pseudocode of N-gram preprocessing

Parameters:-Amharic corpus file Path

Method:-N_gram_preprocessing

Initialize

Author name list

Full list of Amharic corpus text

For each file in Amharic corpus file Path

if not end Amharic corpus file Path in main directory

assign a file path to read in subdirectory

for each file in subdirectory

open file

read file

append read text to Full list of Amharic corpus text

append the corresponding file author to Author name list

Endfor

Endif

Endfor

Return Type:-Author name list and Full list of Amharic corpus text

3.3.5.Feature Extraction

Extracting the right collection of features to reflect an author's writing style is the most important function in the identification phase. Feature extraction is a critical component in an author identification mission. The features extracted, the identification methods used, and the dataset in which the research is conducted distinguish author recognition studies from one another. After preprocessing, the feature extraction tasks are performed based on stylometric and n-gram feature extraction methods. This stage changes each Amharic text corpus into multi-dimensional vector space and authors names into a class label, which is suitable for proposed classification algorithms.

Due to lack of Amharic stylometric feature extraction tools, two groups of stylometric features such as lexical(**F1**) and syntactic(**F2**)features are used for this research purpose. Both Lexical (character based) and syntactic (list of Amharic function word and Amharic punctuation) extracted based on Figure 4 flowchart. Figure 4 shows a flowchart of Amharic stylometric feature extraction process using the list of features that are stored in a configuration file with space delimiter.

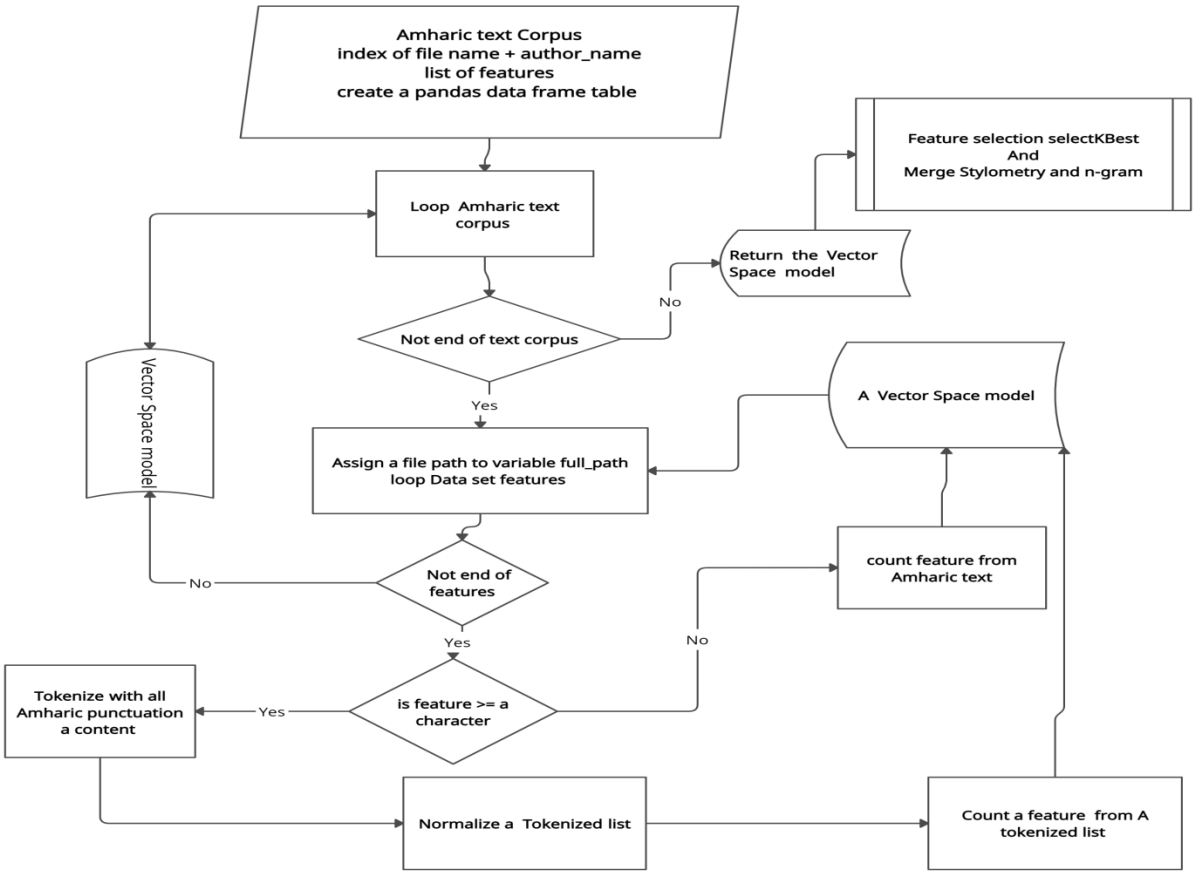

Figure 4. Extraction of Amharic Stylometric feature

In addition to the flowchart of stylometric feature extraction, pseudo codes are described at a high level to easily understand the development of Amharic stylometric feature extraction. Count Amharic feature extraction is one of the stylometric feature extraction that will be described in the next section. Amharic feature extraction method receives a specific feature and a text as a parameter and returns the integer value as described in the flowchart above and pseudo code below. In Amharic stylometric feature extraction, most of the Amharic lexical features are characters and feature characters do not need tokenization with punctuation and do not need to be normalized to a common alphabet. But if the feature is a word, it is

important to tokenize and normalize to some common syntax of Amharic spelling. For example, ሲሐንና need to be normalized into ሢሆንና, because Amharic words written in different spelling like this example so to count a function word the feature part function word and a text function word should be the same form. The outputs of Stylometric feature extraction pass as input to feature selection and merges stylometry with n-gram features for decomposition of the features.

3. Stylometric feature extraction pseudo code

 # This is sudo code of stylometric feature extraction Method
Parameters:-Amharic corpus file Path,list author name concatenate with file name as index and list of full features(F1 and F2)
Method:-Stylo_Feature_extr
Initialize
Amharic corpus file path
list of full features(F1 and F2)
create a pandas table DataFrame using as index list author name concatenate with file name and
as column name list of full features(F1 and F2)
For each file in Amharic corpus file Path
IF not end Amharic corpus file Path in main directory
assign a file path to Full path variable
FOR each Features list of full features(F1 and F2)
IF not end of list of full features(F1 and F2)
Assign result of a Amharic Features extract method to a DataFrame table address of index with column value// at this point Amharic Stylometric feature extract method called
Endif
Endfor
Assign the corresponding file author name to a DataFrame class value
Endif
Endfor
Return Type:-Final DataFrame table result

This is pseudo code of Amharic stylometric feature extraction

The above feature extraction method call another method which is Amharic_stylo_extractor to count each character or words and assigned the return value to panda's table corresponding cell which is column name counted feature with row name the corresponding text.

4. Amharic stylometric feature extraction pseudo code

Parameters:-A feature and A specific text content
Method:-Amharic_stylo_extractor
Read Amharic text
IF length of character greater than one
Tokenize text content with Amharic punctuation
Normalize the tokenized list
Count the feature from normalized list
Assign the count on count variable
else
Count feature from text content, because the feature is character
Return type: Integer count

In this research the n-gram feature extractor simply used built-in methods of python sklearn [**77**] [**78**]by just passed the preprocessed list of full text and return back multidimensional vector space of extracted features. So n-gram feature extraction methods or tasks receive input from n-gram preprocessing tasks. Term frequency (tf) and term frequency inverse document frequency (tfidf) feature extraction methods are used to extract and weigh n-gram character based tasks. CountVectorizer and TfidfVecorizer are the two feature extraction classes to perform tf and tfidf within python sklearn functions, respectively. The full text corpus list passes as input parameters into CountVectorizer and TfidfVecorizer functions [**77**] [**78**].The outputs of these methods are multidimensional vector space which means a list of text corpus changed to a matrix vector. The outputs of the two methods are very large matrices due to the large number of features. The outputs of CountVectorizer and TfidfVecorizer will serve as feature selection of the merged method that decomposes Stylometric and n-gram features.

3.3.6.Feature selection

Feature selection is the method of translating larger-dimensional data into smaller-dimensional data while maintaining the same features by removing irrelevant ones. It is a method of selecting the features in your data that contribute the most to the prediction variable or performance that you're interested in. Many models' accuracy can be negatively affected by having unnecessary features in their data [**79**]. Not all columns or features have an impact on author identification tasks. As discussed in the feature extraction section of author identification, there are many features that create a sparse matrix with a very high multi-dimensional space, especially n-gram. The most relevant features should be selected by using machine learning feature selection techniques. The feature selection component of this research uses selectKBest of the scikit-learn python library [**80**]. This feature selection function receives a stylometric feature vector, n-gram feature vectors and merge of the two vectors as input. The outputs of the three vectors are the most important selected features which have an impact on the author identification task. For example, the sparse matrix of 2500 rows and 3420 columns might be changed to a vector with 2500 rows and 410 columns using selectKbest applied. From the column values, only 410 have an impact on the classification model.

3.3.7.Classification Models

It is necessary to establish a representation of an author's writing style as a fingerprint pixel in order to identify an anonymous text in the problem of authorship identification.
Features are already pre-processed, extracted and finally datasets of vector space passed through feature selection tasks for effective creation of the classification model. Classification model is the process of finding a model that describes and distinguishes data classes and concepts. This research proposed three proposed algorithms such as Multinomial Naïve Bayesian, SVM and Multi-layer Perceptron Neural Network developed using scikit-learn gridsearchcv parameter tuning python library to answer the research questions[**81**].

In this experiment, the training and test dataset is splitted using 80/20 percent simple partitioning techniques. Moreover, SVM and Neural Network use grid search cross-validation of model selection to evaluate different parameters of the two algorithms [**82**]. The training data is used to create a classification model and the test data set helps to evaluate the accuracy of the classification model. The score method is used to evaluate the accuracy of the model of each algorithm developed using the training dataset. Precision, recall, and f1-score the three metrics to calculate score and they are measured using scikit-learn matrix class python library method of classification_report. The calculation uses predicted author test class and original author test value after model creation. Predict method also used to identify the anonymous text writer by using anonymous text vector as input.

Figure 5 below shows the classification model flowchart. It describes how to use the scikit-learn library for classification models.

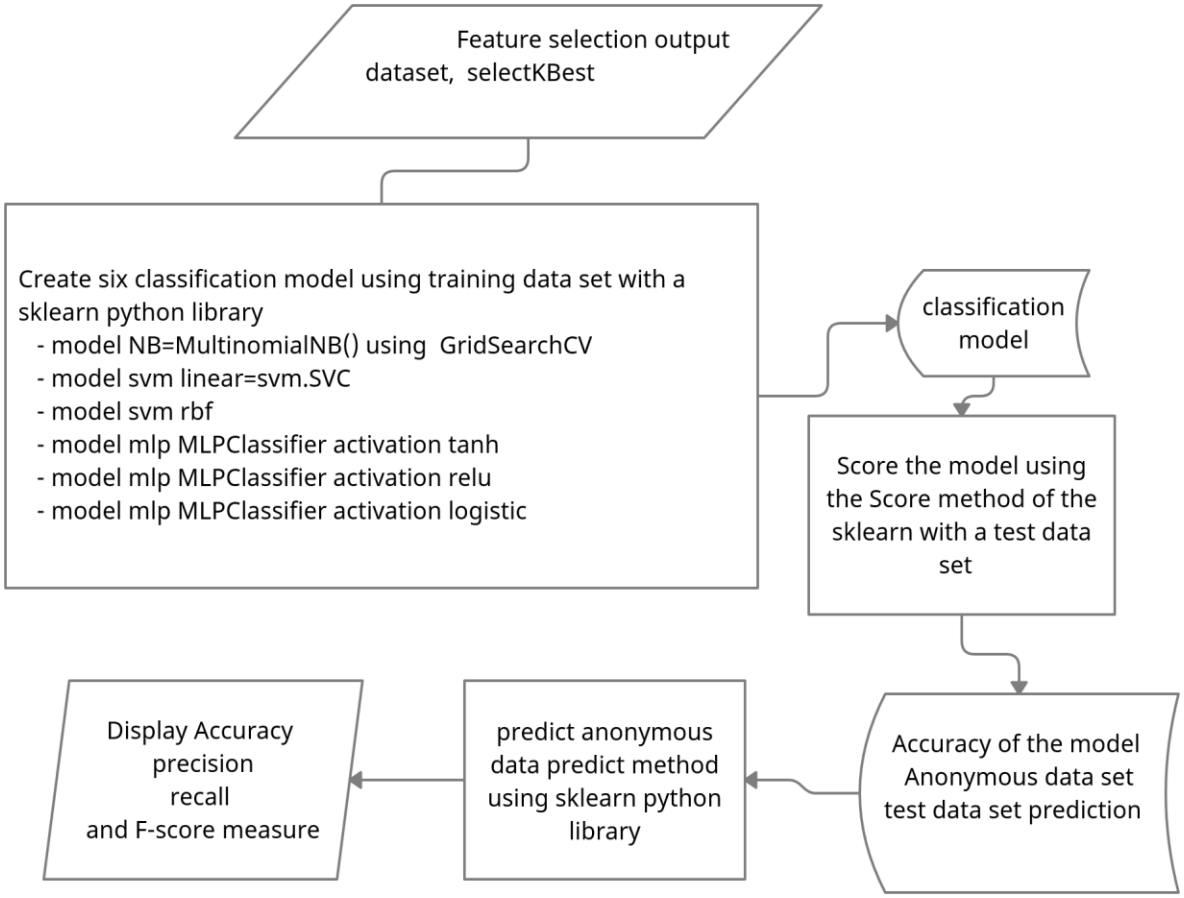

Figure 5 Classification model flowchart

CHAPTER FOUR
CORPUS ANALYSIS AND EXPERIMENT SETUP

4.1. Introduction

The experiment process starts after all the requirements are set up as discussed in the previous chapter. This chapter begins with the discussion and analysis of the characteristics of the collected corpus initially used as an input for the whole author identification. Then, the experiment setup depicts the environment where the initial corpuses are physically processed and results were analyzed. Finally, the output results are presented in tabular and graphical forms with a brief discussion of the findings from the experiments.

4.1.1. Corpus Analysis

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

This research adds 11 more authors from diverse sources and domain areas and each author is described in detail in Table 4 below. Activists, religion preachers, poets, politicians and psychologists are the domain areas used in addition to journalists. This will help to observe the application of this research experimental procedure for a diverse group of authors. The last two columns of both tables show the range of the number of words and characters of the online Amharic texts or articles posted by each author. This range is analyzed by the Linux command of *"wc -w -m *"* where "-w" is a command that generates the number of words and "-m" delivers the number of characters in the articles.

Table 4 Corpus analysis of 11 writers

R.No	Name	**No Article's**	**Domain Area**	**Source**	**No words per article's**	**No characters per article's**
1	Bewketu Seyoum	100	Poet	Telegram	537- 64659	5000-500000
2	Professor Mesfin Woldemariam	100	Political	shegerblogs.com ,ethiopiazare.com	150-1594	744-8556
3	Deacon Daniel Kibret	100	Religious and Political	Personal blog	96-3890	533-21715
4	Deacon Yaregal Abegaz	100	Religious	Amharic Religious Document	250-311	917-1500
5	Sisay Sahilu	75	Journalist	Reporter Magazines	250-311	716-1500
6	Natnael Mekonnen	100	Activist	Telegram	3721-6194	40000
7	Memeher DrZebene Lemma	100	Religious	Telegram	44-21470	212-111642
8	ZemedkunBekele	100	Political and Religious	Telegram	808-41286	2613-212196
9	Seyoum Teshome	100	Activist	Telegram	63-107	440-500
10	Dr. Deacon Rodass Taddese	100	Religious	Telegram	932-1006	4238-5000
11	Nahusenay Tsedalu	100	Psychologist	Telegram	50-1446	467-8000

Source: Author

Table 5 Corpus analysis of 20 writers

R. No	Name of the authors	Number of Articles	Domain Area	Source	Number of words per article	Number of characters per article
1	Berhanu Fekade	100	Journalist	Newspaper	207-1728	1161-6191
2	Biruk Abdu	80	Journalist	Newspaper	70-1272	344-6955
3	Dawit Endeshaw	75	Journalist	Newspaper	57-1869	315-10083
4	Dawit Taye	106	Journalist	Newspaper	64-2579	361-13918
5	Dawit Tolossa	105	Journalist	Newspaper	71-1131	384-6079
6	Dereje Tegenaw	106	Journalist	Newspaper	90-1656	475-8943
7	Fissaha Getnet	100	Journalist	Newspaper	134-1363	713-7386
8	Henok Yared	120	Journalist	Newspaper	36-2242	2917-11881
9	Kaleyesus Bekele	108	Journalist	Newspaper	60-1602	309-9076
10	Mihret Moges	108	Journalist	Newspaper	42-2033	1768-11299
11	Nathan Dawit	111	Journalist	Newspaper	477-1046	2680-5710
12	NeamnAshenafi	110	Journalist	Newspaper	56-1627	311-9014
13	Samson Abdela	66	Journalist	Newspaper	330-1719	1816-9560
14	Shahida Hussen	115	Journalist	Newspaper	207-2038	1142-11045
15	Tadesse Gebremariam	100	Journalist	Newspaper	114-984	626-5497
16	Tamiru Tsige	106	Journalist	Newspaper	53-2537	259-13895
17	Tamrat Getachew	70	Journalist	Newspaper	99-2926	541-15833
18	Widneh Zenebe	102	Journalist	Newspaper	129-1656	706-8698

19	Yohannes Enberbr	101	Journalist	Newspaper	16-3193	841-17597
20	Zemenu Tenagne	100	Journalist	Newspaper	126-2974	700-15517

Source: Modified from [**12**].

Proper description of the dataset for author identification using classification models is very important because the nature and characteristics of the articles in the input dataset determines the features and directly affects the classification models. For example, the difference in the range of the number of words and characters among authors might affect the accuracy of classification algorithms and will be discussed in the result section.

4.1.2. Experimental setup

The factors investigated in this research, the number of authors and number of articles per author. The input files are organized separately and run with different codes to get the results of the relationship among all other factors such as features and algorithm models used for author identification. The first experiment is based on the number of authors. The 31 authors were divided into ten categories where the first group contains three authors and the subsequent groups increase by three till the last group of 31 authors. The second experiment uses inputs of datasets of ten where the first group is rearranged to constitute ten articles for each author, the second group to be 20 articles per author, and so on until the last group where all authors' articles are around 100 articles per author.

The two experiments are performed by using HP EliteBook 840 G1 core i7 2.70 GHz and 8GB RAM 4th Generation laptop. The Integrated development environment was the Linux Debian version operating system of Ubuntu 18.04.LTS. Sklearn python3 version of 0.24.2 was broadly used for these experiments. Due to grid search hyper parameter tuning and performance intensity after the fifth interactions of both experiments on two of the algorithms such as SVM and Neural Network Multilayer Perceptron, an Ubuntu linux server of 56 core and 80GB RAM was used. This reduces run time from around four hours to a time which is less than an hour.

Features and models are the two factors common to the two experiments. Six classification algorithms are included out of the three major models. The selected classification algorithm setup for the two experiments are Naive Bayes Multinomial (NB-Multinomial), Neural Network Multilayer Perceptron Classifier logistic(NN-MLP-Logistic), NN-MLP Rectified Linear Unit (NN-MLP-relu), NN-MLP-hyperbolic tangent (NN-MLP-tanh), Support Vector Machine Linear (SVM-Linear) and SVM-Radial Basis Function (SVM-RBF).

MultinomialNB function of the NB-Multinomial algorithm contains a smoothing parameter alpha is equal to one (1.0), Prior probabilities of the classes set to none and learned class prior probabilities (fit_prior) used the default, True.

NN-MLP uses an MLPClassifier function of Multi-layer perceptron backward propagation that was formulated to allow 1000 iterations or aka epochs. The GridSearchCV function runs these iterations with hyper parameter tuning of five cross-validations. The same function also uses three activation functions 'tanh', 'logistic' and 'ReLU'with a hidden layer

size setup of 10, 20 and 30 the two learning rate is selected ‘constant’ and ‘adaptive’ two alpha 0.0001 and 0.05 and also three solver parameters are used ‘Stochastic Gradient Descent(sgd)’, 'adam' and ‘lbfgs’ for hyper parameter tuning of GridSearchCV.This function used by combining every parameter and ranked based on their accuracy finally predict the unknown authors using the first rank of parameters.

SVM algorithm as a parameter used a “linear” and “rbf” kernel parameter and 10 and 20 C parameter of optimization inside svm. SVC (support vector classifier) functions, respectively. The gridsearchcv is another function used in the SVM algorithm. This function delivers four ranked results based on the accuracy but linear and rbf optimization selects one optimum value.

There are three features in both experiments. These include stylometry, n-gram and the combination of the two (Merged). In both experiments, n-gram feature extraction used 1-garm and 2-gram character based extraction with scikit-learn library methods of text feature extraction called TfidfVectorizer (frequency and inverse document frequency) and CountVectorizer (term frequency). As a feature selection, SVM and Neural Network Multilayer Perceptron (MLP) classification algorithms use TfidfVectorizer in n-gram features. However, Naive Bayes MultinomialNB uses CountVectorizer for feature selection. SVM and NN needs scaled data and also have good discrimination and accuracy capability with term frequency and inverse document frequency rather than term frequency. Similarly, stylometry feature selection uses sklearn MinMaxScaler for SVM and NN algorithms, but Naive Bayes Multinomial uses term frequency function for stylometry feature selection. For better performance of SVM and NN, normalizing the data reduces bias of extreme high and to extreme low value.

In addition to feature pre-processing, feature extraction, feature selection, model estimation, and prediction, the sklearn function also gives model performance metrics using score and cv_results_ to measure accuracy of the model and also classification_report method to measure performance measure (precision, recall and f-score). Only accuracy is not sufficient to measure performance in classification models because it cannot correctly measure an unbalanced class dataset. However, performance measure can appropriately measure any proportion of class dataset. For example, precision represents the relevance of a result, i.e., how many times the classifier is correct in identifying a particular text. The classifier's recall

indicates how many texts from a collection of predictions are properly identified to the appropriate author. The F1-score is calculated by taking the harmonic mean of precision and recall. Table 4 and 7 record the macro average of all authors' of precision, recall and f1-score with respect to the two experiments. In general, both experiments run for three features, six models and ten number of authors (first experiment) and ten number of articles per author (second experiment). Therefore, there are three by ten and by six (3*10*6=180) iterations for each experiment and hence expected outputs from each experiment.

CHAPTER FIVE
RESULTS AND DISCUSSION

5.1. **Introduction**

The output terminal shows four performance measurement metrics, one accuracy value in percent, precision, recall and f-score for both experiments. The results are, therefore, recorded in four tables. The first two tables are from the first experiment and the last two tables from the second experiment. The first table records accuracy with respect to ten categorical groups of the number of authors, three features and six types of classification models. This delivers 180 accuracy values as shown in Table 5 below. The second table records 180 values of the three performance metrics that make up a total of 540 outputs for each parameter pair of the number of authors, features and types of the models, as shown in Appendix F performance measure table. The results of the last two tables compose the same type of parameters except the number of authors being replaced by the number of articles per author. Appendix D and E are screen shots of the results from experiment one and two as displayed on the Linux terminal.

These two experiments are intended to meet the objectives of the research. Experiment one analyzed the effect of the number of authors on the accuracy, precision, recall and f1-score by increasing the number of authors by three up to 31 starting from three authors. The numbers of authors in each group dataset are selected randomly to generate results from ten cumulative iterations. Experiment two constitutes 27 authors and 100 articles per author. The other four authors were excluded because the number of articles for each author is less than hundred as shown in Table 5 above. As mentioned before, experiment two analyzes the effect of the number of articles per author on the accuracy, precision, recall and f1-score of the models calculated using different Amharic features. The following sections discussed the results of the effect of increasing the number of authors as well as the number of articles per author based on accuracy, precision, recall, f1-score and discrimination capability (prediction).

5.1.1.The Effect of the number of Authors

In the first experiment, results are explained in a tabular and graph form to visualize the effect of the increasing number of authors from 3 to 31 on the confusion matrix measures. The results are broken down by a more popular scientific evaluation of classification algorithms, accuracy, precision, recall and F1-score. Based on these model performance metrics, the six models were compared for each feature and the features were also compared and ranked for each model in detail in the following subsections. Table 5 displays the accuracy of the three features and for six types of classification models. The table also displays discrimination capability results by italics those that are not predicted correctly and bolding those that were predicted correct. However, Table 7 shows separately the number of correctly (TRUE) predicted values. Table 7 shows the values of the three performance metrics (precision, recall and f1-score) accompanied with Figures 6 & 7 to illustrate the trends and variation among features and classification models for the number of authors and similar composition of tables and figures illustrated the results and finding of the research for the number of articles per author as well.

5.1.2.Model Selection

As shown in table 6, this research obtained promising results as compared to previous research. For example, Amharic author identification study by BahirHussen[**12**].used 20 authors and found 87% accuracy with a combination of character 3-gram and word pos 4-garm features. This research used 11 more authors and 1-gram with 2-garm character based features. NN-MLP-rule and merged features 90.91% accuracy at a similar number of authors of the above study [**12**]. Therefore, this research brought absolute improvement in accuracy by more than 3.91% of the previous study.

As shown in Figure 6 below, Neural Network classification models perform best and have the highest accuracy in almost all numbers of authors. More specifically, NN-MLP Logistic and relu are the best models that could be used for author identification.

Table 6 The 20 Authors accuracy of previous researcher dataset used **[**12**]**.

Features	NB-Multinomial[1]	SVM-Linear[2]	SVM-RBF[3]	NN-MLP-tanh[4]	NN-MLP-rule 5	NN-MLP-logisitic[6]
Stylometry	**68**	**71.47**	**77.43**	**78.68**	**76.60**	**81.19**
N-gram-char	**75.75**	**86.21**	**83.7**	**89.03**	**84.32**	**86.52**
Merged	*75.75*	**77.74**	**79.01**	**89.03**	**90.91**	**89.03**

Key 1. NB-Multinomial Naïve Bayesian Multinomial algorithm

2. SVM-Linear support vector machine with linear kernel function

3. SVM-rbf support vector machine with rbf kernel function

4. NN-MLP-tanh neural network multilayer perceptron with tanh activation function

5. NN-MLP-rule neural network multilayer perceptron with Rectified Linear Unit activation function

6. NN-MLP-logistic neural network multilayer perceptron with logistic activation function

Table 7.The number authors on accuracy by feature and model in hyper parameter tuning, %

Features	No-of Author's	NB-Multinomial[1]	SVM-Linear[2]	SVM-RBF[3]	NN-MLP-tanh[4]	NN-MLP-rule 5	NN-MLP-logisitic[6]
	3	**90**	**85**	**70.36**	**91.93**	**89.89**	**90.24**
	6	**89.74**	**78.63**	**76.06**	**83.35**	**81.36**	**86.83**
	9	**78.29**	**74.77**	**69.14**	**80.05**	**76.83**	**84.17**
	12	**80.09**	**78.97**	**69.03**	**81.72**	**72.56**	**85.09**
	15	**77.08**	**79.51**	**72.51**	**81.07**	**76.5**	**85.27**
	18	**81.69**	**79.51**	**75.77**	**77.99**	**72.67**	**85.15**
	21	**83.33**	*78.47*	*62.90*	**72.8**	**77.43**	**84.42**
	24	**80.43**	**77.06**	**69.36**	**76.08**	**72.42**	**81.19**
Stylometry	27	**76.40**	**76.77**	*66.29*	**74.54**	*65.54*	**79.64**
	31	**73.51**	**74.34**	*62.75*	**72.8**	**69.42**	**79.64**
	3	*85*	**95.3**	**94.63**	**96.64**	**96.65**	**95.97**
	6	*86.32*	**91.45**	**88.37**	**91.28**	**91.79**	**90.77**
	9	**83.42**	**88.64**	**86.93**	**89.45**	**88.88**	**88.76**
	12	**82.74**	**90.24**	**86.28**	**91.48**	**90.95**	**90.41**
	15	*77.8*	**89.3**	**88.74**	**89.9**	**86.11**	**90.13**
	18	*78.59*	**88.65**	**88.43**	**90.01**	**83.94**	**89.56**
	21	*69.28*	**87.92**	**83.81**	**89.28**	**81.90**	**89.57**
	24	*81.28*	**88.85**	**88.04**	**90**	**89.06**	**89.36**
N-gram-char	27	*79.21*	**87.81**	**86.72**	**83.52**	**88.22**	**88.31**
	31	*63.24*	**87.06**	**86.27**	**86.63**	**86.33**	**86.87**
	3	*80*	**91.67**	**91.25**	**95.63**	**96**	**95**
	6	*87*	**93.16**	**85.29**	**91.62**	**92.13**	**91.79**
	9	*88.57*	**88.57**	**84.86**	**84.57**	**91.17**	**90.83**
	12	*82.30*	**90.06**	**90.26**	**92.37**	**92.10**	**92.72**
	15	*77.43*	**89.36**	**88.39**	**92.35**	**92.77**	**92.56**
	18	*79.71*	**89.44**	**88.73**	**92.1**	**92.21**	**92.04**
	21	*81.19*	**90.71**	**86.81**	**91.67**	**85.48**	**92.2**
	24	*82.77*	**89.23**	**87.87**	**91.82**	**91.53**	**91.53**
	27	*80.15*	**87.44**	**87.40**	**90.62**	**90.25**	**90.43**
Merged	31	*76.99*	**86.47**	**84.05**	**89.25**	**89.05**	**88.85**

Key
1. NB-Multinomial Naïve Bayesian Multinomial algorithm
2. SVM-Linear support vector machine with linear kernel function
3. SVM-rbf support vector machine with rbf kernel function
4. NN-MLP-tanh neural network multilayer perceptron with tanh activation function
5. NN-MLP-rule neural network multilayer perceptron with Rectified Linear Unit activation function
6. NN-MLP-logistic neural network multilayer perceptron with logistic activation function

At the maximum number of authors, NN-MLP-tanh has the maximum accuracy (89.25%).

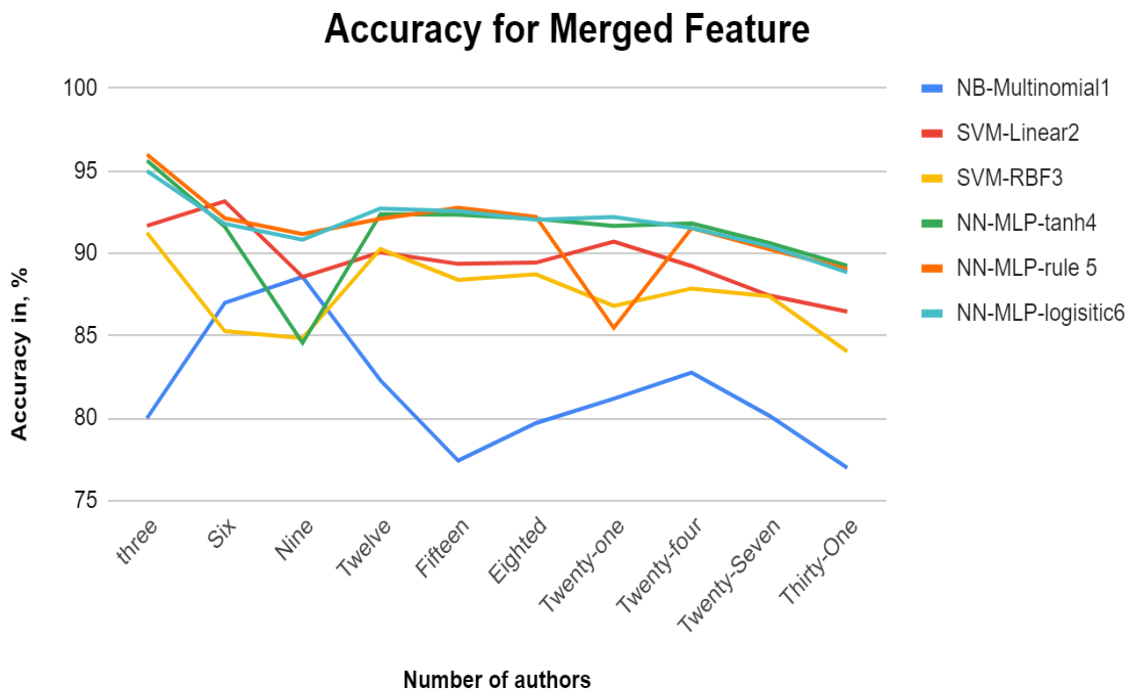

Figure 6. Effect of number of authors on accuracy for merged features

Figure 6 clearly depicted the six models' accuracy with merged features. The graph shows that NN-MLP outperformed throughout the number of authors.

Figure 7 shows the values of performance metrics such as precision, recall and f1-score by the features, number of authors and classification models. All models except NB-Multinomial show higher values in all the three metrics. For example, SVM-Linear, NN-MLP-relu and NN-MLP logistic achieve a maximum f1-score 87%, 86% and 88% respectively at 31 maximum numbers of authors from the performance matrix Appendix F table.

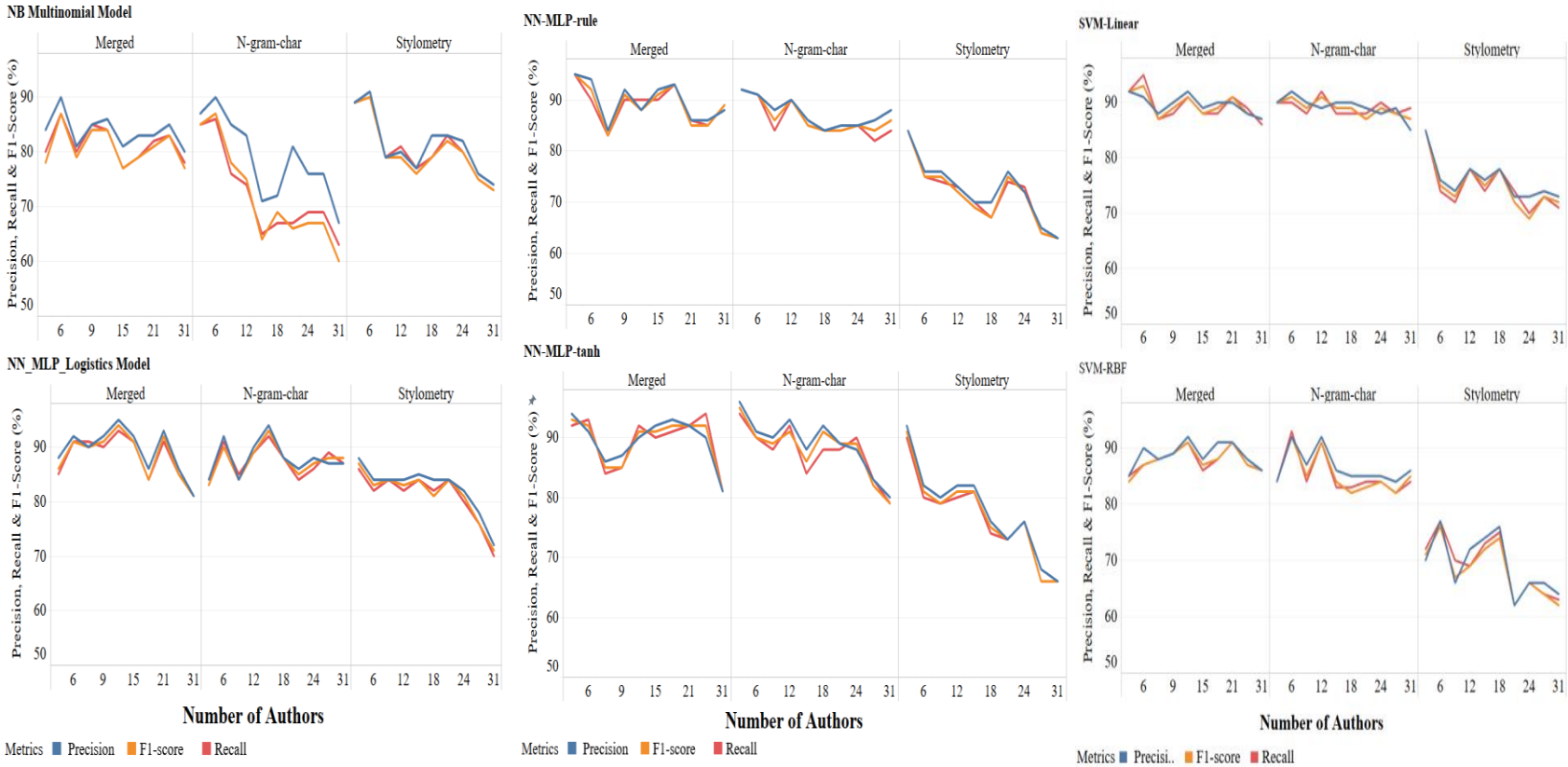

Figure 7.The trend of the performance metrics over increasing the number of authors and broken down by features for each model.

Table 8 shows the number of correctly predicted (TRUE) and incorrectly predicted (FALSE) pairs of features and classification models. Prediction capability of unknown authors from suspected list those 3 models with merged and n-gram correctly predicts the unknown author which means the models have high discrimination capability. For example, six models NN-MLP-Logistic and SVM linear can correctly predict the unknown author with all features and all iterations of the number of authors.

Table 8 The Number of correct (TRUE) and wrong (FALSE) predictions of unknown author

		Models					
Features	**Prediction**	**NB-Multinomial**	**NN-MLP-Logistic**	**NN-MLP-rule**	**NN-MLP-tanh**	**SVM Linear**	**SVM-RBF**
Merged	FALSE	10					
	TRUE		10	10	10	10	10
N-gram-char	FALSE	8					
	TRUE	2 (9&12)	10	10	10	10	10
Stylometry	FALSE			1(27)	1(27)		2(21&27)
	TRUE	10	10	9	9	10	8

Naïve Bayes models do not predict the unknown author using the merged features across all numbers of authors. However, the other two features predict the unknown author correctly using the same model across all numbers of authors except for the n-gram feature where the unknown author was not predicted when nine and twelve authors were used in the experiment. Moreover, except for four experiments, all 56 possible experiments of the stylometry feature predict the unknown author correctly using all the six models.

5.1.3.Features Comparison

Results presented in Table 7 show that merged features registered maximum accuracy at maximum number of authors corresponding to four models except the two SVM models. Even though discrimination capability table 8 shows that stylometry has best, but based on figure 8below and table 7 merged features outperformed. Therefore, the use of merged features is better than using the two features separately. The combination of merged features with Neural Network outperformed accuracy throughout the increasing number of authors. The maximum accuracy at the maximum 31 number of authors merged features with NN-MLP-tanh is 89.25%.

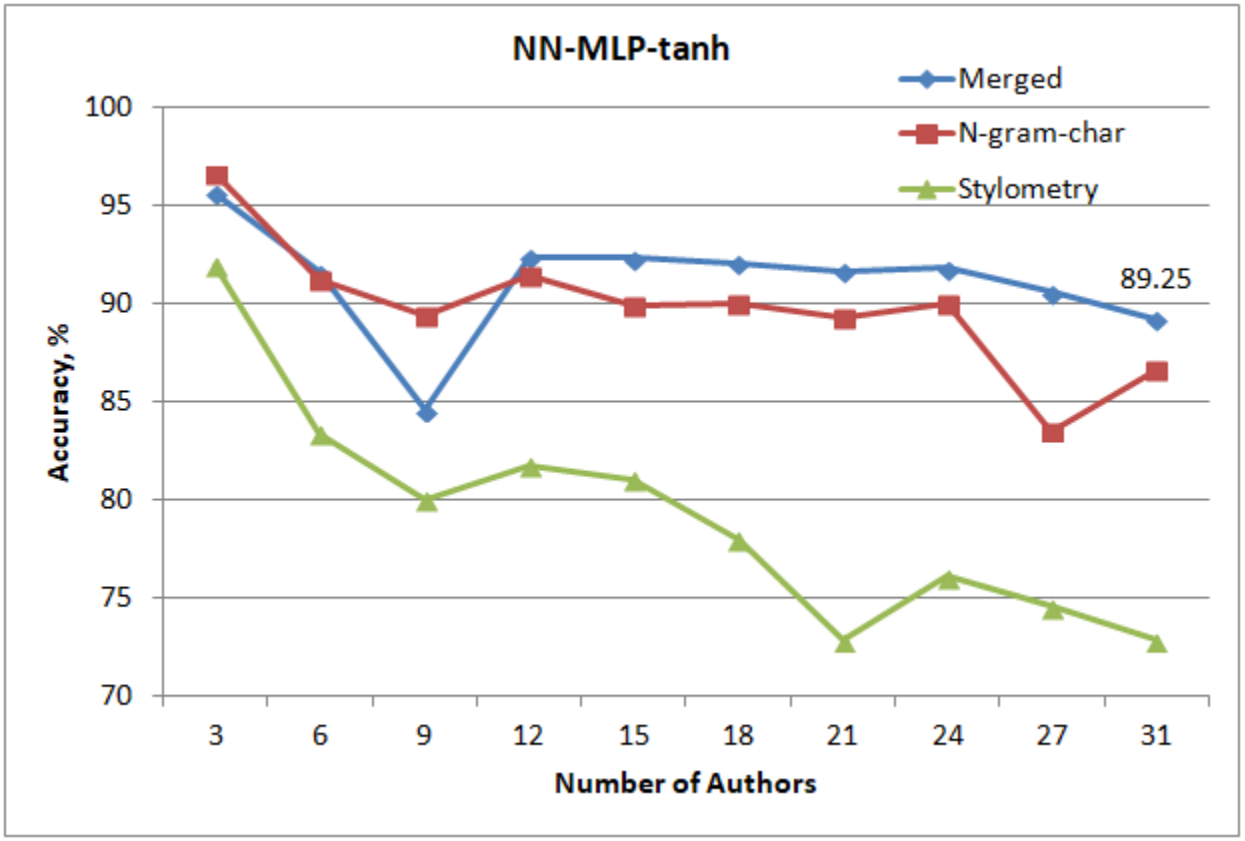

Figure 8. Trend of accuracy over the number of authors by features and models

Table 7 also shows n-gram(1-gram and 2-gram character based) features have relatively better accuracy values than stylometry features across all numbers of authors and classification models except NB-Multinomial. Figure 7 shows the best model (NN-MLP-tanh). This result is consistent with the previous studies by Nicol Mariah et al[**6**]and Smita et al [**16**]where they have

found that n-gram character based features give the best accuracy and discrimination capability in English language.

NN-MLP-logistic and SVM-Linear has better discrimination capability. In all iterations the models developed by stylometry features could make 56 correct predictions, n-gram features 52 and Merged features 50. However, n-gram increases its discrimination capability as we increase the number of iterations except with NB-Multinomial algorithm. Figure 7 shows the trend of performance metrics over increasing the number of authors across the three features. The three scores in stylometry feature looks highly fluctuating and decreasing for all models

5.2.The Effect of the number of articles per author

Table 9 shows the accuracy of classification models by features and the number of articles per author. Based on the results on the table 9, merged features with Neural Network, NN-MLP-tanh, outperformed the most relative to other models. For example, NN-MLP-tanh has the highest accuracy (82.7%) with only 20 articles per author as compared to other models. Similarly, the same model performs the highest (84.75%) accuracy with 40 articles per author. This indicates that Neural Network models give a promising result for an author identification system with a small number of articles per authors, because the second experiment performed with balanced class which means each author contain the same number of articles or texts in each 10 iteration. The previous articles confirmed that rather than imbalanced class balanced class dataset outperformed [**83**] [**84**].

Moreover, referring to the values at the bottom of the table, SVM linear performs the maximum accuracy value (97.52%) at 100 articles per author using merged features for 27 authors. This is relatively higher than any other research findings that used either stylometry or n-gram. Comparing the accuracy value of the same number of authors, the same model and features in table 7, but with an unbalanced number of articles per authors, the accuracy drops by large and becomes 87.44%. This shows that balanced number of articles per author for analysis gives better accuracy results than balanced number of articles per author in the author identification process.

Table 9 The effects the number of articles per author on accuracy by features and models, %

Features	No-of Author's	NB-Multinomial[1]	SVM-Linear[2]	SVM-RBF[3]	NN-MLP-tanh[4]	NN-MLP-rule[5]	NN-MLP-r logistic[6]
	10	*71.93*	*52.63*	*43.47*	*50.01*	*42.10*	*50.04*
	20	**61.26**	*63.96*	*56.75*	*51.71*	**54.95**	*61.08*
	30	**67.27**	**65.45**	**67.12**	*68.93*	*54.54*	*70.6*
	40	**73.97**	*52.33*	*50.01*	*60.2*	*62.23*	*64.51*
	50	**79.12**	*53.47*	*53.39*	*53.48*	*65.93*	*65.65*
	60	**75.23**	*58.71*	*53.75*	**68.04**	*51.68*	**67.97**
	70	**74.28**	**69.39**	*59.06*	**65.45**	**60.36**	**73.96**
Stylometry	80	**76.55**	**71.3**	*60.23*	**69.33**	*62.29*	**74.85**
	90	**76.55**	**70.42**	*60.23*	**71.67**	*59.77*	**75.97**
	100	**80.89**	**88.01**	**83.27**	**89.78**	**88.66**	**88.92**
	10	*28.07*	*59.2*	*42.11*	*63.6*	*56.14*	*62.1*
	20	*71.17*	**75.86**	*60.36*	**79.27**	69.36	**77.12**
	30	69.69	**78.78**	*63.03*	**81.81**	**64.84**	**80.6**
	40	*74.42*	**79.18**	*72.15*	**80.55**	**78.08**	**80.27**
	50	*79.85*	81.31	*77.47*	**72.89**	**81.78**	**81.59**
	60	*77.68*	**81.96**	**74.92**	**83.79**	**76.15**	**82.87**
	70	*79*	**83.30**	**77.69**	**84.46**	**78.22**	**84.36**
N-gram-char	80	*78.62*	**82.2**	**78.62**	**85.89**	**77.47**	**86.2**
	90	*78.62*	**84.76**	**78.62**	**86.34**	**77.47**	**85.69**
	100	*81.66*	**95.30**	**94.04**	**95.92**	**94.22**	**94.33**
	10	*70.18*	*61.4*	*47.41*	*64.51*	*40.35*	*62.71*
	20	*67.58*	**75.68**	**70.27**	**82.70**	**81.80**	**62.16**
	30	*70.30*	*78.18*	*70.91*	*62.42*	**84.85**	**85.91**
	40	*74.42*	**83.11**	**77.99**	**84.75**	*60.27*	**84.11**
	50	*80.25*	**80.21**	**75.69**	**85.34**	**71.43**	**85.08**
	60	*77.06*	**81.5**	**81.04**	**85.78**	**72.78**	**85.7**
	70	*77.95*	**83.04**	**83.99**	**87.03**	**77.42**	**87.24**
	80	*79.08*	**84.87**	**83.67**	**88.32**	**76.00**	**88.41**
Merged	90	*81.98*	**86.63**	**83.67**	**89.3**	**82.38**	**88.65**
	100	*83.06*	**97.52**	**93.92**	**95.8**	**95.68**	**95.41**

key1. NB-Multinomial Naïve Bayesian Multinomial algorithm

2.SVM-Linear support vector machine with linear kernel function

3.SVM-rbf support vector machine with rbf kernel function

4.NN-MLP-tanh neural network multilayer perceptron with tanh activation function

5. NN-MLP-rule neural network multilayer perceptron with Rectified Linear Unit activation function

6.NN-MLP-logistic neural network multilayer perceptron with logistic activation function

5.2.1. Model Selection

Figure 10 shows the trend of the three performance scores (precision, recall and F1 score) over the number of articles per author broken by features and types of classification models. The graphs are directly drawn from Appendix G table. All graphs show that the values of all the three scores increase as the number of articles per authors increases. This shows that the performance of the classification models improves as more online text data per author is collected. Since it is difficult to collect more data, the best classification model is the one that gives the highest values of the score with the lowest number of articles collected per author. To explain this statement with an example and compare the six models, the maximum F1-score and the corresponding number of articles per author are filtered. These pairs of information are shown in Figure 9 below.

For example, for n-gram features, SVM-RBF classification model has the maximum F1-score (94%) but at 100 articles per author. NN-MLP-tanh performs 81% F1-score with the lowest number of articles per author (10 articles only). Stylometry looks to perform the maximum (95%) of all other models by NN-MLP-tanh but it exhausted all articles to get this result. However, changing the feature from stylometry to n-gram or to the combination of the two improves the performance of the same model drastically and makes its score the highest with the lowest number of articles per author. A model that gives a higher performance at a smaller number of articles per author serves to analyze short and dangerously illegal online texts that are usually posted on social media. Thus, the Neural Network models are preferred to other classification models for short online text author identification.

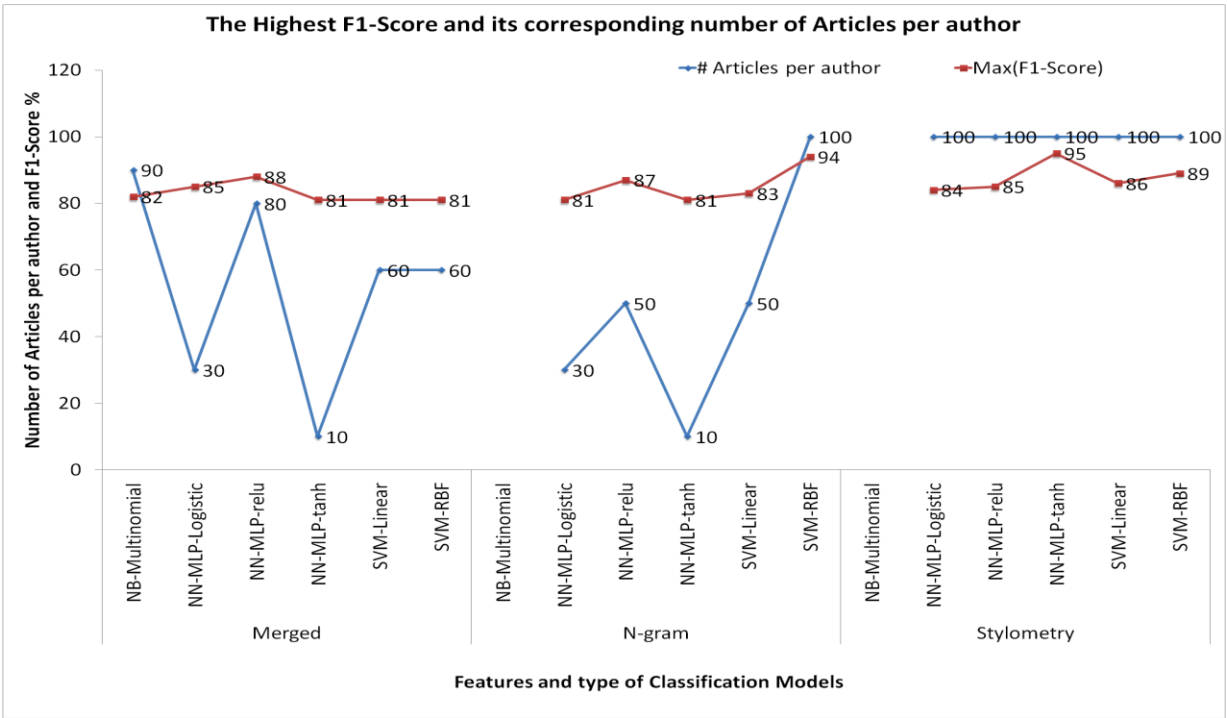

Figure 9. The maximum F1-score and its corresponding number of articles

This figure 9 depicted that outperformed model with small number of articles per author for example NN-MLP-tanh 10 number of articles 81% performance with merged and n-gram features.

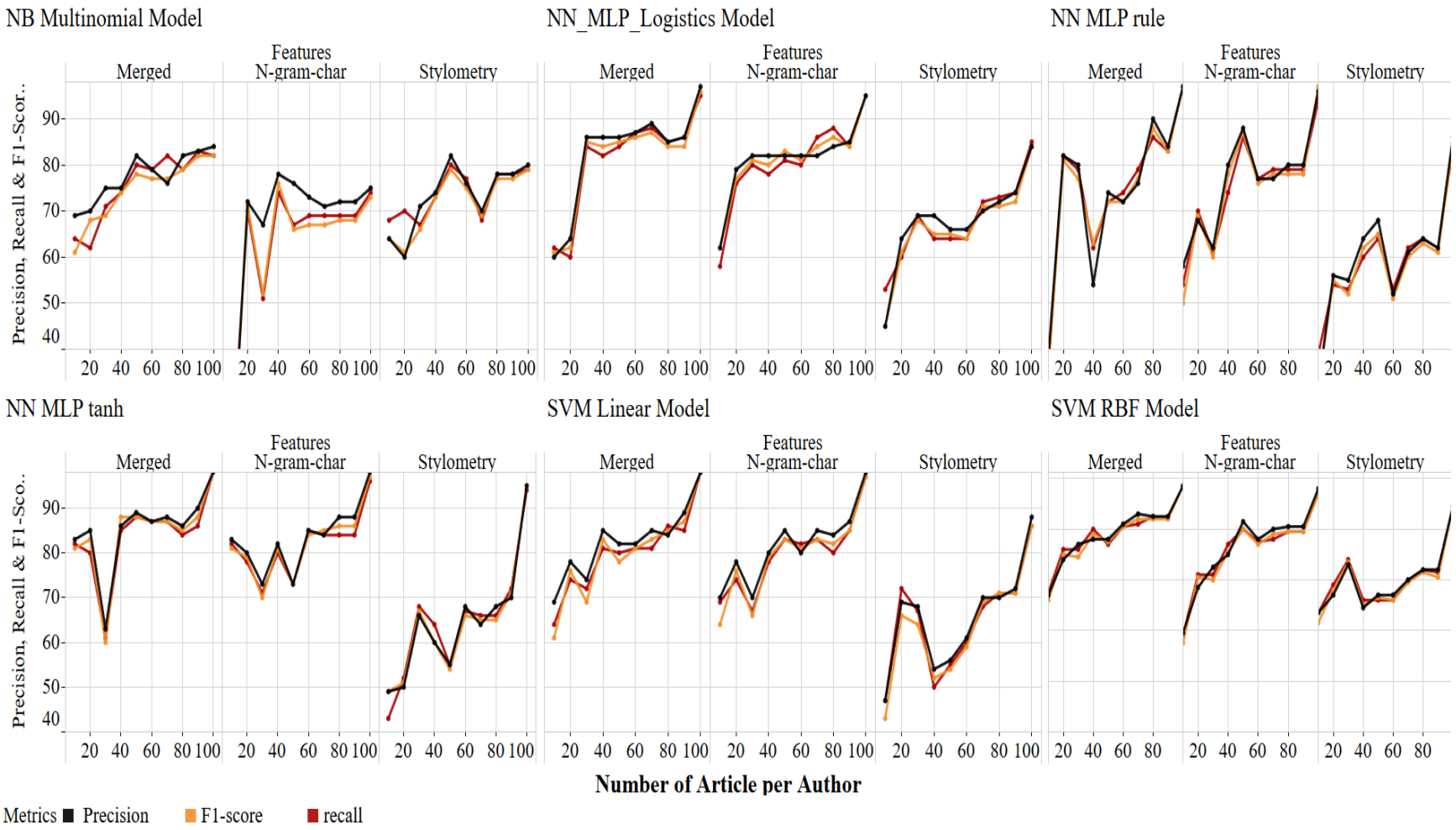

Figure 10. Effects of precision, recall f1-score in increasing number of articles per author

This figure shows clearly that affects in performance of number of articles or texts per author additionally the tradeoff between precision, recall and the f-score (harmonic mean) within each of the classification models.

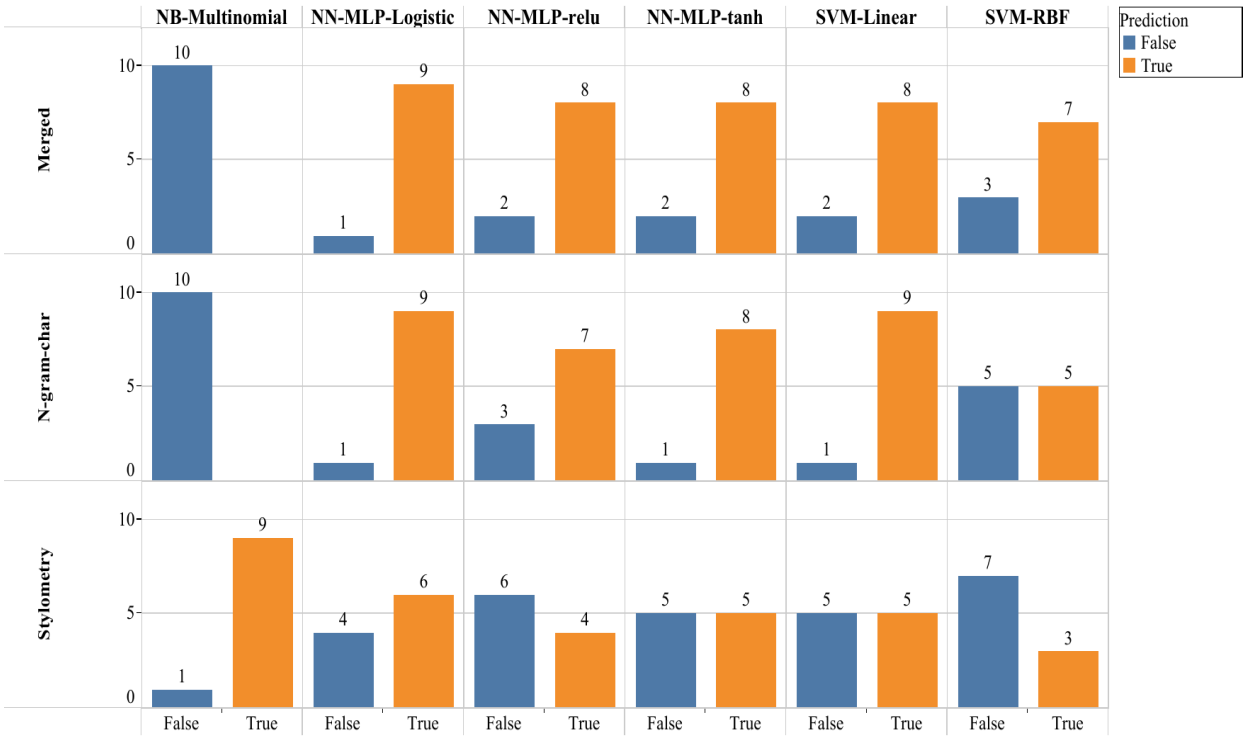

Figure 11. The number of correct and incorrectly predicted combination of features and models

Figure 11 shows the number of correctly predicted experiments through the combination of features, model and number of articles per author as shown in bold in the original table of the results. For example, the Naïve Bayes model does not predict the unknown author using the n-gram and merged features across all numbers of articles per author, because of in the two experiment alpha parameter used the default 1.0that affects the discrimination capability and performance of this algorithm. It predicts well under the stylometry feature. Neural Network MLP logistic predicts the unknown author using most of the number of articles per author in all three features. Merged features predict the unknown author better than others in all models except NB Multinomial.

5.2.2. Comparison of features

Figure 12 shows the trend of F1-Score over the number of articles per author broken by the three features, classification models for both correctly and incorrectly predicted unknown author combinations. Merged feature gives the highest performance metric and predicts correctly for the unknown author using the NN-MLP-Logistic, NN-MLP-relu model, NN-MLP-tanh model. N-gram features give the highest performance metric and predict correctly for the unknown author using the SVM-Linear model. Merged feature seems to give the highest performance metric and predict correct for the unknown author using SVM-Linear model

As shown in Figure 12, the stylometry features show the highest discrimination capability and the highest confusion matrix scores or F1-score in NB- Multinomial (a) classification model. Therefore, stylometry is the most preferred feature in Naive Bayes multinomial Model. However, for all other five classification models, merged features have higher values of F1-Score with 40 articles per author and above. This result is more prevalent in the three classification models: NN-MLP Logistic (b), NN-MLP-tanh (d) and SVM linear (e).

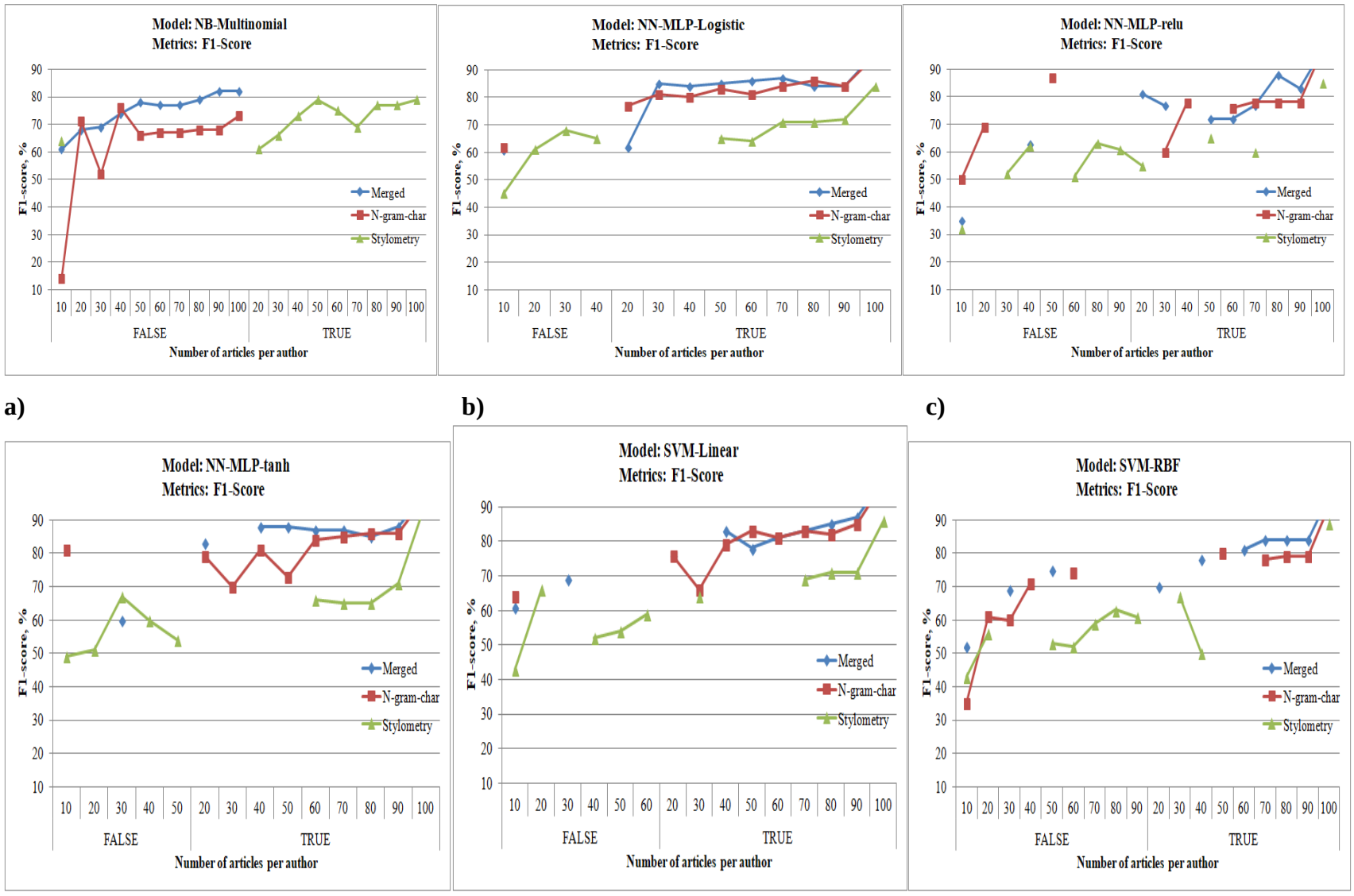

Figure 12. The values of F1-score for each feature and classification model broken down by correct and incorrect predictions

5.3.**Summary**

In this section the results of the two experiments are described and interpreted with respect to our research questions. Both experiments contribute many findings which are amazing for Amharic text author identification models and features whether we have a small number of articles per author or with a maximum number of authors. Those findings of the two experiments give a lot of ideas to develop Amharic author identification system. For example, when we see the accuracy presented in Table 5 of experiment one, the merged features deliver a new best result of 89.25% accuracy from the Neural Network multi-layer perceptron algorithm using the three activation functions such as tanh ,relu and logistics. Similarly SVM linear gives 86.47% accuracy with merged features. As shown in Figure 4, the use of the merged features achieved consistently higher accuracy and the other performance metrics (confusion matrix results) for all the number of authors.

This research obtained better results than the previous research that has been conducted by previous researcher [**12**] who achieved 87% accuracy with 20 authors and combination of character 3-gram and word pos 4-garm features. Because of the neural network resource intensive we can't perform experiment more than 1-gram and 2-gram character-based with stylometric features. In this research we used 31 authors and stylometric(lexical and syntactic),1-gram combined with 2-garm character based features and NN-MLP-tanh(Neural Network Multilayer perceptron with tanh activation function) to achieve a better result that goes upto 89.25%. The findings of this research are, therefore, organized in two subsections as we describe it in the experiment set up section above. The first discussion is about the effect of the increasing number of authors in each model and features discrimination capability with their performance matrix. The second subsection describes the findings related to the effects of increasing the number of articles per author on the accuracy and the confusion matrix of the models with respect to the three features.

CHAPTER SIX
CONCLUSION AND RECOMMENDATION

6.1.**Conclusion**

Distributing information throughout the world using the internet is one of the great opportunities and the internet also makes the world one society. All the distributed text information through the internet is not legal. The distribution of misinformation anonymously brings a challenge to identify these illegal authors. In Ethiopia, nowadays, there are many anonymous writers that distribute false narrative texts through different languages.

In this research we studied the Amharic text author identification using stylometry and character based n-gram features with classification algorithms of support vector machine, Naive Bayes and Neural Network multilayer perceptron. The Amharic language problems are studied through literature review, experimentations, and the solution of the research was investigated and compared to each other. The experiments are systematically and scientifically arranged which is directly related to research problems. The effect of the number of authors over the features and the six classification models. Likewise, the effect of the number of articles per author is also studied. The output values are shown by using accuracy, precision, recall and f1-scores through all possible combinations of three features and six algorithms.

The combined or merged features perform better than the individual stylometry and n-gram features in all models except NB-Multinomial. Moreover, for merged features, support vector machine linear kernel function, Neural Network multilayer perceptron with both logistic and tanh activation function outperformed other models. Based on the findings, we concluded that when the number of authors increases, the performance of the model goes down, because increasing the number of authors means increasing the number of classes in classification

problems. So in the classification problem when the number of classes increases the accuracy decreases. On the other hand, the number of articles per author and model performance metrics is directly proportional to each other. This is because in classification problems the increasing number tuples in a dataset the classification algorithms accuracy increase. The results of the three NN and the two SVM with linear kernel parameter models have shown promising results with merged features. Especially Neural Network, the results are stable and show the best identification capability from the given suspect. Amharic author identification system can be developed using merged features with those three models.

Changing the feature from stylometry to n-gram or to the combination of the two improves the performance of the same model drastically and makes its score higher with the lowest number of articles per author. A model that gives a higher performance at a smaller number of articles per author serves to analyze short and dangerously illegal online texts that are usually posted on social media. Thus, the Neural Network models are preferred to other classification models for short online text author identification.

6.2. **Contribution of the research**

The contributions of this research are listed below:

- Collecting, preparing and adding eleven (11) authors of Amharic text corpus on the 20 authors corpus collected from previous, a total 31 authors of Amharic corpus is organized and ready for the next related studies.
- Prototype Amharic Author Identification is developed
- New stylometric Amharic language features of lexical and syntactic are extracted and can be used as initial data for future studies
- Three numbers of algorithms and features (three) were applied to increase the possibility of getting the best method that delivers the most effective model for author identification analysis.
- Identified the best features and classification algorithms for Amharic Author identification.
- Studied the effects of increasing the number authors and the number of articles per author on Amharic author identification using the performance metrics by collecting and organizing Amharic online text.

6.3.**Recommendation**

Despite considerable and promising results, this research's performance can still be enhanced further. The findings of this study open up other research questions and the author identified numerous potential research areas. The following tasks are recommendations and future research works:

- Based on this research result the discrimination capability and accuracy of stylometry lexical character-based, syntactic and n-gram features are promising result.So these imply that this research will be extended to other Ethiopian language.
- This research uses only lexical character-based and syntactic features of stylometry with n-garm.From previous related work four stylometric features to other languages like English and Chinese Grigori et al. [48] and Zeng et al. [16] perform best using four categories of stylometry.So additionally using other two stylometry (structural and content-specific) features might give promising results.
- This research result shows promising performance of models at a smaller number of texts per author. However each article does not have equal number words33 and characters based on corpus analysis table this implies an indication of need future studies effects of authorship identification specifically for short messages that have very small words and characters.

REFERENCES

[1] Meseret Assefa Adamu, "Role of social media in Ethiopia's recent political transition ," *Journal of Media and Communication Studies*, vol. 12(2), no. 2141-2545, pp. 13-22, june 2020.

[2] Imran Sarwar Bajwa and Shabana Ramzan Waheed Anwar, "Design and Implementation of a Machine Learning-Based Authorship Identification Model," *Hindawi*, vol. 2019, no. 9431073, p. 14, october 2018.

[3] Vijayshri Khedkar Sileshi Girmaw Miretie, "Automatic Generation of Stopwords in the Amharic Text," *International Journal of Computer Applications*, vol. 180, pp. 0975 – 8887, 2018.

[4] Yohannes Eneyew Ayalew. (2020, may 1) ethiopia-insight.com. [Online]. https://www.ethiopia-insight.com/2020/05/01/is-ethiopias-first-fake-news-case-in-line-with-human-rights-norms/

[5] Halefom H. Abraha and Marie Curie. (2019, june 4th 2019) blogs.lse.ac.uk. [Online]. https://blogs.lse.ac.uk/medialse/2019/06/04/the-problems-with-ethiopias-proposed-hate-speech-and-misinformation-law/

[6] Naveed Muhammad and Fernando Alonso-Fernandez Nicole Mariah Sharon Belvisi, "Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features," *8th International Workshop on Biometrics and Forensics (IWBF)*, no. 10.1109/IWBF49977.2020.9107953., pp. 1-6, march 2020.

[7] Zewdie & Wang, Jenq-Haur Mossie, "SOCIAL NETWORK HATE SPEECH DETECTION FOR AMHARIC LANGUAGE," , 2018, pp. 41-55.

[8] David Recine. (2021, june) https://magoosh.com. [Online]. https://magoosh.com/toefl/english-writing-structure-compared-to-other-languages/

[9] Rong Zeng, "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, vol. 57(3), pp. 378–393, 2006.

[10] Hongjun Zhu, Ariel Apostoli and Pei Cao Stewart Yang, "N-gram Statistics in English and Chinese: Similarities and Differences," in *International Conference on Semantic Computing Google, Inc.*, California Mountain View, 2007, pp. 7695-2997.

[11] Elena Boychuk, Inna Vorontsova and Ilya Paramonov Ksenia Lagutina, "A Survey on Stylometric Text Features," in *PROCEEDING OF THE 25TH CONFERENCE OF FRUCT ASSOCIATION*, Yaroslavl State , 2019, pp. 2305-7254.

[12] Baher Hussen Geletu, "Authorship Attribution Model for Amharic Documents using Machine Learning," Addis Ababa University, Addis Ababa, Master Thesis 2020.

[13] Rong & Li, Jiexun & Chen, Hsiu-chin & Huang, Zan Zheng, "A framework for authorship identification of Online messages: Writing-style features and classification techniques," *Writing-style features and classification techniques*, no. 57, pp. 378-393, 2006.

[14] Daniel Pavelec, Edson Justino, and Luiz S. Oliveira, "Author Identification using Stylometric Features," *Inteligencia Artificial*, vol. 11, p. 5965, 2007.

[15] A. Anderson, M. Corney, O. de Vel, and G. Mohay, "Identifying the Authors of Suspect E-mail," *Communications of the ACM*, 2001. (Submitted).

[16] Smita & Dharaskar, Rajiv & Thakare, V. M. Nirkhi, "Authorship Identification using Generalized Features and Analysis of Computational Method.," *Transactions on Machine Learning and Artificial Intelligence*, vol. 3, no. 2, pp. 2054-7390, 2018.

[17] Mrs. Sujata Khedkar,Shashank Agnihotri,Anshul Agarwal,Mahak Pancholi,Pooja Hande, "Stylometry Based Authorship Identification," *International Journal for Research in Applied Science & Engineering Technology*, vol. 6, no. IV, p. 2321, 2018.

[18] Smita Nirkhi, Dr. R. V. Dharaskar,Dr. V. M. Thakare, "An Experimental Study on Authorship Identification for Cyber Forensics," *International Journal of Computer Science and Network*, vol. 4, pp. 2277-5420, 2015.

[19] Mubin Shauka, "Authorship Analysis and Identification Techniques: A ReviewAuthorship Analysis and Identification Techniques: A," *International Journal of Computer Applications* , vol. 77, p. 2013, 0975-8887.

[20] Sara El Manar El Bouanani,Ismail Kassou, "Authorship Analysis Studies: A Survey," *International Journal of Computer Applications*, vol. 86, pp. 0975 – 8887, 2014.

[21] H.Chen,Z Huang,J Li,R Zheng, "Authorship Analysis Studies:A Survey," *International Journal of Computer Applications*, vol. 86, pp. 0975-8887, 2014.

[22] Gray, Sallis, and MacDonell, "Software forensics: Extending authorship analysis techniques to computer programs," in *Paper presented at the 3rd biannual conference of the International Association of Forensic (IAFL '97)*, 1997.

[23] Farkhund Iqbal, "Messaging Forensic Framework for Cybercrime Investigation," CONCORDIA UNIVERSITY, Montréal, Québec, Canada, Thesis 2011.

[24] D., & Loader, B.D. Thomas, "Cybercrime: Law enforcement, security and surveillance in the information age. Routledge.," 2000.

[25] M. L. Jockers and D. M. Witten, "A comparative study of machine learning methods for authorship," *Literary and Linguistic Computing*, vol. 25, p. 2, 2010.

[26] Haifa Alharthi, Diana Inkpen, and Stan Szpakowicz, "A survey of book recommender systems," *Journal of Intelligent Information Systems*, vol. 9, pp. 1-22, 2017.

[27] Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods," *J. Am. Soc. Inf. Sci. Technol 60(3):*, pp. 538–556, 2009.

[28] E. Stamatatos, "Authorship attribution using text distortion," 2017.

[29] W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola, M. A. Sanchez-Perez and A. Barrón-Cedeño E. Stamatatos, "Overview of the Author Identification Task at PAN 2014," *CLEF 2013 Evaluation Labs and Workshop Working Notes Papers*, pp. 877-897, September 2014.

[30] L. Sarkute and A. Utka J. Kapociute-Dzeikiene, "The Effect of Author Set Size in Authorship Attribution for Lithuanian," *Proceedings of the 20th Nordic Conference of Computational Linguistics*, 2015.

[31] A., and Chen, H. Abbasi, "A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Transactions on Information Systems (TOIS)* , vol. 26(2), p. 7, 2008.

[32] Saud Alotaibi and Abdulrahman Alruban Abdulaziz Altamimi, "Surveying the Development of Authorship Identification of Text Messages," *International Journal of Intelligent Computing Research (IJICR)*, vol. 10, no. 1, pp. 953-966, March 2018.

[33] Anderson A, Corney M, Mohay G de Vel O, "Mining e-mail content for author identification forensics," pp. 55-64, Apr. 2001.

[34] Eugene Liang Kenji Takahashi, "Analysis and design of Web-based information systems," *Computer Networks and ISDN Systems*, vol. 29, no. 8–13, pp. 1167-1180, 1997.

[35] Google developers. (2021, march) developers.google.com. [Online]. https://developers.google.com/machine-learning/guides/text-classification/step-3

[36] scikit-learn. (2021, march) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html

[37] Micheline Jawie haline, *Data mining concept and techniques Kamberand jian pei*. 225Wyman Street,Waltham, MA 02451, USA: Elsevier Inc., 2012.

[38] (2021) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/svm.html#

[39] Steven Busuttil, "Support Vector Machines," *Department of Computer Science and AI*, 2014.

[40] scikit-learn. (2021, February) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/naive_bayes.html

[41] Soner Yıldırım. (2020, feb) towardsdatascience.com. [Online]. https://towardsdatascience.com/naive-bayes-classifier-explained-50f9723571ed

[42] IBM Cloud Education. (2020, august) www.ibm.com. [Online]. https://www.ibm.com/cloud/learn/neural-networks

[43] Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao Shervin Minaee, "Deep Learning Based Text Classification: A Comprehensive Review," *Microsoft Research, Redmond*, vol. 1, p. 1, January 2020.

[44] Ujjawal Verma. (2021, Feb) medium.com. [Online]. https://medium.com/analytics-vidhya/text-preprocessing-for-nlp-natural-language-processing-beginners-to-master-fd82dfecf95

[45] Richard Sproat and Steven Bedrick. (2021, Feb) en.wikipedia.org. [Online]. https://en.wikipedia.org/wiki/Text_normalization#cite_note-cs506-1

[46] bitext. (2021, feb) /blog.bitext.com. [Online]. https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/

[47] Walle Engedayehu, "The Ethiopian Orthodox Tewahedo Church in the Diaspora: Expansion in the Midst of Division," *African Social Science Review*, vol. 6, no. 1, may 2013.

[48] Gebremeskel Hagos Gebremedhin and Abera Asefa Mebrahtu, "Linguistic Evolution of Ethiopic Language: A Comparative Discussion ," *International Journal of Interdisciplinary Research and Innovations*, vol. 8, no. 1, pp. 1-9, march 2020.

[49] Björn Gambäck and Lars Asker, "Experiences with Developing Language Processing Tools and Corpora for Amharic," *ResearchGet*, pp. 1-9, January 2010.

[50] wokru Kelemu, "Automatic Amharic text news classification: Aneural networks approach," *Ethiop. J. Sci. & Technol*, vol. 6(2), pp. 127-137, 2013.

[51] Wolf Leslau, "Concise Amharic dictionary. Amharic-English, English-Amharic," *Wiesbaden: Otto Harrassowitz*, pp. 668-669, 1976.

[52] Beletu Reda, "A Graphemic Analysis of the Writing System of Amharic," Addis Ababa University, Addis Ababa, Paper for the Requirement of the Degree of Bachelor of Art in Linguistics. 1982.

[53] M., and C. Ferguson. eds Bender, "The Ethiopian Writing System. In Languages in Ethiopia," *Oxford University Press*, 1976.

[54] Y. Mohammed, "Amharic Grammar and Literature," *International Leadership Printing Press*, 2017.

[55] T. H. GEBERMARIAM, "AMHARIC TEXT RETRIEVAL: AN EXPERIMENT USING LATENT SEMANTIC INDEXING (LSI) WITH SINGULAR VALUE DECOMPOSITION (SVD)," Addis Ababa Univerisity, Addis Ababa, Unpublished Masters Thesis 2003.

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

[57] A. Tefera and Y. Assabie, "Automatic construction of Amharic semantic networks from unstructured text using Amharic wordNet," in *Proceedings of the 7th Global Wordnet Conference*, GWC, 2014.

[58] Lars & Argaw, Atelach & Gambäck, Björn & Asfeha, Samuel & Habte, Lemma Asker, "Classifying AmharicWebnews," *Information Retrieval*, vol. 12, no. 10.1007/s10791-008-9080-x, pp. 416-435, 2009.

[59] Addis Ababa : Eleni plc.

[60] Getachew Haile, "The Problems of Amharic Writing System. Unpublished," , 1967.

[61] BEZZA TESFAW AYALEW, "THE SUBMORPHEMIC STRUCTURE OF AMHARIC:TOWARD A PHONOSEMANTIC ANALYSIS," University of Illinois at Urbana-Champaign, Urbana, Illinois, Doctor of Philosophy in Linguistics 2013.

[62] Taye, and Shiferaw Bekele Assefa, "The Study of Amharic Literature: An Overview.," *Journal of Ethiopian Studies*, vol. 33, no. 2, pp. 27–73, 2000.

[63] Aynadis Temesgen Gebru, "Design and Development of Amharic Grammar Checker," Addis Ababa univerisity , Addis Ababa, Degree of Master in Computer Science 2013.

[64] MARKOS KASSA GOBENA, "Implementing An Open Source Amharic Resource Grammar," Chalmers University of Technology , Göteborg, Sweden, Master of Science Thesis in Intelligent Systems Design 2010.

[65] Afsaneh & Hirst, Graeme Fazly, "Testing the Efficacy of Part-of-Speech Information in Word Completion," no. 10.3115/1628195.1628197, 2003.

[66] Tsegaye Andargie, "INCORPORATING LINGUISTIC FEATURES IN BI-DIRECTIONAL AMHARIC - ENGLISH STATISTICAL MACHINE TRANSLATION," Addis Ababa univerisity, Addis Ababa, Unpublished Masters Thesis 2018.

[67] Ruth Kramer, "The Amharic Definite Marker and the Syntax–Morphology Interface," *A jornal of theoretical, expermental and interdesiplinary study* , vol. 13, no. 3, pp. 196-240, september 2010.

[68] Taye, and Shiferaw Bekele Assefa, "the Study-of-Amharic-literature: An Overview ," *Journal of Ethiopian Studies*, vol. 33, no. 2, pp. 27–73, 2000.

[69] Vlado Ke seljy Fuchun Pengz Nick Cerconey Calvin Thomasy, "N-GRAM-BASED AUTHOR PROFILES FOR AUTHORSHIP ATTRIBUTION," august 2003.

[70] John Houvardas and Efstathios Stamatatos, "N-Gram Feature Selection for Authorship Identification," *Dept. of Information and Communication Systems Eng.*, 2006.

[71] Grigori Sidorov, "Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts," *computational system*, vol. 22, pp. 47-53, 2018.

[72] R., Herath, P., and Senanayake, U Ragel, "Authorship detection of SMS messages using unigrams," in *2013 8th IEEE International Conference*, Sri Lanka, 2013, p. In Industrial and Information Systems (ICIIS).

[73] John Houvardas and Efstathios Stamatatos, "N-Gram Feature Selection for Authorship Identification ," *Dept. of Information and Communication Systems Eng.*, 2006.

[74] Teshome Kassie. (2009, April) WORD SENSE DISAMBIGUATION FOR AMHARIC TEXT RETRIEVAL: A CASE STUDY FOR LEGAL DOCUMENTS.

[75] scikit-learn. (2021, april) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[76] creately. (2021, aprill) www.creately.com. [Online]. https://app.creately.com

[77] sklearn. (2021, april) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/naive_bayes.html

[78] sklearn. (2021, april) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[79] machinelearningmastery. (2021, april) machinelearningmastery.com. [Online]. https://machinelearningmastery.com/feature-selection-machine-learning-python/

[80] sklearn2. (2021, april) https://scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[81] sklearn4. (2021, april) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[82] scikit-learn. (2021, april) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[83] ANDREWK. C. WONG and MOHAMED S. KAMEL YANMIN SUN, "CLASSIFICATION OF IMBALANCED DATA: A REVIEW," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687–719, March 2009.

[84] Justin M. Johnson* and Taghi M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Jornal of Big Data*, p. 6, March 2019.

[85] sklearn. (2021, april) scikit-learn.org. [Online]. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

APPENDIX

Appendix A.

Appendix I. List of Function words [74]

አንቀጽ	ሀያ	የሆኑትን	ስድስት	እስከ	ይኸኛው	-	205/	-4	8/ሀ/	/እንቀጽ.164/	(ሰ)	336)	2-55/	36//
ወይም	ሆነ	በሆኑ	ምንም	ብሎ	ከርሱ	/1/	90/	248-250)	1እ5	62/	481//	107/	/pubic/	356-34
እንደሆነ	በኋላ	ጀምሮ	በሆነው	ከሰላሳ	መሆኑን	1/	84/	248-22/	784/	613/3//	134/	335/3/	2)%//1/	355- 34/
ላይ	በአንድ	እንደዚህ	ስለ	የሚሆኑ	ለዚያው	/	78/	248-22	784-791/	599/	479//	332)	187//ሰ/)	327/ሀ
ማንም	የሆኑ	በመሆን	ይሁን	ላይም	ለዚሁ	//	200/	247/ሀ/	784-790/	///ሀ/	478/	327/ሀ	/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/	2እ5	///ሀ/	284- 337/
እና	መሆኑ	እንዲቆይ	በአንዳንድ	እስከ	ለ/	66/	684/	/ሀ//	19-200/	587/	46(1)/	103/	136//	27/
ይሆናል	ሌላውን	ሌላው	በእያንዳንዱ	ለሆኑት	ሐ/	179/	218/	237/	770/	583/	459/	298/	(እንቀጽ18 7//ሀ/)	269- 24)
ሆኖ	ከሰባት	የሚሆነው	ጊዜም	አለው	መ/	5/	671/	238-260	754/	58/3/	455//	297/	13/	26/1// 2/
ነው	ለሌላ	በአንዱ	እስከ	እነዚሁ	ሠ/	/ከእንቀጽ	161/	231/	750/	58-582/	129-15)	ሆን½	129-15)	(1)እስ ክ
በዚህ	አለበት	አይሆንም	የሌሎች	እንደሆኑ	ረ/	(አንቀጽ	6/	229/	187/1/	(መ)	451/	29/	123/ሀ	
እስከ	ሲል	ማለት	የሚሆኑት	ስለማናቸውም	ሸ/	/3/	16/	228/	/1//ሰ/	(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	ሳለ	የነበረውን	ከአንዳንድ	-ለ/	12/	156/	224/	187//ሰ/)	57/	449/	289-283/	11-154/	
በማናቸውም	በሙሉ	የሆነው	ያሉ	በእነዚሁ	-ሐ/	9/	/መ//	(እንቀጽ187//ሀ/)	ማናቸWM	57-59/	435/	284-337	1/3/	
ወር	አስራ	መሆናቸው	ከሌሎች	በአምስት	-መ/	83/	15//	224-228	728-30/	560/	127/	284-325/	9-	
ከአምስት	አስር	በዋና	እንዲት	ወይም	-ሠ/	188/	145/	22)	72/	150/	426/	284-317/	84እ	
በላይ	እንደ	በማቀድ	ለሌሎች	በሆኑ	-ረ/	140/	145-150/	217/	187//ሀ/	56-599	423/	1/ሰ/	8/ሰ/	
ሲሆን	ቢሆንም	ጊዜና	እንኳን	የሆኑበታል	ሀ.	-1	48/	213/	71/	559/	123/ሀ	28/1/	728-30/	
በሆነ	እንዱ	ለዚህ	ለሆነው	ለነዚህ	ለ.	4/	49/	211-213/	70/	55/3/	(ሀ)/	28(3)/	72/	
ከዚህ	የሌላውን	ለስተኛ	አንዳንድ	ለማንኛውም	ሐ.	208/	13/	21//	187/	55/	422-424	1/3/	ን/እ	
የሆነ	ከሁለት	የነገሩ	ለአት	አንደኛ	መ.	20/	407/	/5/	ኘ/	15/1/	420/	273/	690//	

አምስት	ሰስት	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		9ሕ	37/3/	94-95/	186/	533/	42-47/	ሀ	676/1//	
አንድ	ቢሆን	ከዘምሳ	ይህንንም	እነደሆነ		183/	110(2)/	937	ረ	146-150/	123/ሀ/	አንቀጽ	66//	
ልዩ	እነዚህን	ይኸው	የአንድን	እነደሆነ		/ፋርማሊስት/	34/	207/	69/	532-534/	41/	12/	656/	
ከሰስት	ናቸው	ለአንድ	በመ·ሉም	ይህችው		131/	106/	93/	184/	53/	4010	5/	622-31	
በተለይም	አንዱን	የሚችሰውን	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)	ነ/	(ሰ))	121-128/	9ሕ	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//ለ/	51(ሰ)/	397)	14/	511///5/	
ከአስር	ላይሆን	ከዚያ	ማንኛውንም	የለውም		81/	27//	84/1/መ//	670/	144-149/	39/	15//	510//	
የማይበልጥ	እንደሆነና	እንዲሆኑ·	ሁለቱ	በሚችሉ		19/	27/	84/1/(መ)/	665/	506/3/	39-641/	16/	51(ሰ)/	
ብቻ	ብሎ	ከሌላ	እንጂ	የሌላቸውን		/1//ሀ/	269-322	84/1/	180)	505-513/	117/	10/	4ሕ	
እንዲሁም	ከብር	ለሆነ	ከስምንት	በሶስተኛ		68/	269-24)	2/	ች/	143/	379/	766/	4ሕ7/	
ሌሎች	ሆኖም	በሌሎች	ሁለቱንም	በቀር		18/	266/	/36//	66/1/ሀ/	4ሕ	376/	59/	493//	
ይህ	በታች	እንደሆነ	በሁለት	በነሱ		90/3/	261/1	835/	66//	4ሕ9	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)%//1/	654/	4ሕ7	37)	26/1/	"ሀ"	
የዚህ	ይህንኑ·	በእንደዚህ	ቁጥር	ው		154/	26/1/	829/	179/·	ዚህ	367/	254/4/	479//	
ማናቸውም	ሆነው	ስምንት	ባሉ·	በዚህ		54/	257/ለ/	821/	/አንቀጽ 101/	497/	36//	25//	46(1)/	
ከስድስት	በስተቀር	ሲሆንና	ከመቶ	በዘህም		14/	257)	2(ለ)	(ሰ)	14/1	11-419	248-22/	455//	
መቶ	መሆን	ምንጊዜም	እነዚህም	በዚህና		101/	256/	809/	65-687/	494-500	36/	248-22	45(2)/	
ያለ	ስም	ለማናቸውም	ሲኖር	ከዚህም		/4/	255)	2(ሀ)	640/	493//	በ/	243//3/	41/	
መሆኑን	እንደገና	የአንድ	ሰላሳ	በሁኔታው		50/	ሀ/	/3	168-162	491/	356-34	22)	42/	
አንድን	የማያንስ	እነዚህኑ·	ማንም	ከነዚሁ		232-237/	254/	80/	627/	488/	11-154/	21//	(Manual)	
ያላቸውን	እጅግ	ሲሆኑ·	ለሆኑ·	ሌሎች		21/	25//	8/ለ/	622-31	486/	355-34/	20//)	39/	
ሊሆን	ሚን	በሁለቱም	አለ	ይህን		98/3/	25/	2(5)	168(2)/ማዘዝ	136//	35/	2/	39-641/	

Appendix B

	ā/ä [a]	u [u]	ī/i [i]	a [a]	ē/e [e/ɛ]	(i)/(ə) [ə]	o [o/ɔ]
h [h]	ሀ ha	ሁ hu	ሂ hi	ሃ ha	ሄ he	ህ h(ə)	ሆ ho
l [l]	ለ le	ሉ lu	ሊ li	ላ la	ሌ le	ል l(ə)	ሎ lo
h/ḥ [h]	ሐ ha	ሑ hu	ሒ hi	ሓ ha	ሔ he	ሕ h(ə)	ሖ ho
m [m]	መ me	ሙ mu	ሚ mi	ማ ma	ሜ me	ም m(ə)	ሞ mo
s/ś [s]	ሠ se	ሡ su	ሢ si	ሣ sa	ሤ se	ሥ s(ə)	ሦ so
r [r]	ረ re	ሩ ru	ሪ ri	ራ ra	ሬ re	ር r(ə)	ሮ ro
s [s]	ሰ se	ሱ su	ሲ si	ሳ sa	ሴ se	ስ s(ə)	ሶ so
sh/š [ʃ]	ሸ ʃe	ሹ ʃu	ሺ ʃi	ሻ ʃa	ሼ ʃe	ሽ ʃ(ə)	ሾ ʃo
k'/q [k']	ቀ k'e	ቁ k'u	ቂ k'i	ቃ k'a	ቄ k'e	ቅ k'(ə)	ቆ k'o
b [b]	በ be	ቡ bu	ቢ bi	ባ ba	ቤ be	ብ b(ə)	ቦ bo
t [t]	ተ te	ቱ tu	ቲ ti	ታ ta	ቴ te	ት t(ə)	ቶ to
ch/č [ʧ]	ቸ ʧe	ቹ ʧu	ቺ ʧi	ቻ ʧa	ቼ ʧe	ች ʧ(ə)	ቾ ʧo
h/ḫ [h]	ኀ ha	ኁ hu	ኂ hi	ኃ ha	ኄ he	ኅ h(ə)	ኆ ho
n [n]	ነ ne	ኑ nu	ኒ ni	ና na	ኔ ne	ን n(ə)	ኖ no
ny/ñ [ɲ]	ኘ ɲe	ኙ ɲu	ኚ ɲi	ኛ ɲa	ኜ ɲe	ኝ ɲ(ə)	ኞ ɲo
ʾ/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
ʿ/ʿ [ʔ]	ዐ ʔa	ዑ ʔu	ዒ ʔi	ዓ ʔa	ዔ ʔe	ዕ ʔ(ə)	ዖ ʔo
z [z]	ዘ ze	ዙ zu	ዚ zi	ዛ za	ዜ ze	ዝ z(ə)	ዞ zo
zh/ž [ʒ]	ዠ ʒe	ዡ ʒu	ዢ ʒi	ዣ ʒa	ዤ ʒe	ዥ ʒ(ə)	ዦ ʒo
y [j]	የ je	ዩ ju	ዪ ji	ያ ja	ዬ je	ይ j(ə)	ዮ jo
d [d]	ደ de	ዱ du	ዲ di	ዳ da	ዴ de	ድ d(ə)	ዶ do
j/ǧ [ʤ]	ጀ ʤe	ጁ ʤu	ጂ ʤi	ጃ ʤa	ጄ ʤe	ጅ ʤ(ə)	ጆ ʤo
g [g]	ገ ge	ጉ gu	ጊ gi	ጋ ga	ጌ ge	ግ g(ə)	ጎ go
t'/ṭ [t']	ጠ t'e	ጡ t'u	ጢ t'i	ጣ t'a	ጤ t'e	ጥ t'(ə)	ጦ t'o
ch'/č̣ [ʧ']	ጨ ʧ'e	ጩ ʧ'u	ጪ ʧ'i	ጫ ʧ'a	ጬ ʧ'e	ጭ ʧ'(ə)	ጮ ʧ'o
p'/p̣ [p']	ጰ p'e	ጱ p'u	ጲ p'i	ጳ p'a	ጴ p'e	ጵ p'(ə)	ጶ p'o
ts'/ṣ [ts]	ጸ ts'e	ጹ ts'u	ጺ ts'i	ጻ ts'a	ጼ ts'e	ጽ ts'(ə)	ጾ ts'o
ts'/ṣ́ [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

Appendix C

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

Appendix D

First output on terminal

@@@

@@@

@@@@ **ADDIS ABABA UNIVERSITY 2021** @@@@

@@@@ SisayZinabuGesit @@@@

@@@@ Advisor:-Solomon Teferra Abate (PhD) @@@@

@@@@ **Author identification of Amharic online text using** @@@@

@@@@ **stylometric and n-gram features by classification techniques** @@@@

@@@

@@@

===

27 Number of Authors with 100 articles per author accuracy and confusion matrix

===

Stylometry Accuracy of this model SVM ==> 0.8801042571676803

Stylometry Accuracy of this model NB ==> 0.8088618592528236

Stylometry Accuracy of this model NN ==> 0.8731537793223284

Stylometry predict using NB = ['YohannesEnberbr']

Stylometry predict using SVM= ['YohannesEnberbr']

Stylometry predict using NN===== ['YohannesEnberbr']

==

confusion_matrix_NB_27_authors==>

	precision	recall	f1-score	support
BerhanuFekade	0.54	0.90	0.68	41
Bewketu_seyoum	0.94	0.98	0.96	45
Biruk Abdu	0.78	0.86	0.82	36
DawitEndeshaw	0.38	0.35	0.36	17
DawitTaye	0.63	0.70	0.67	47
DawitTolossa	0.75	0.71	0.73	38
DerejeTegenaw	0.83	0.85	0.84	40

Diyakon_Daniel_kibret	0.95	0.79	0.86	47
Dn_Yaregal_Abegaz	0.94	0.97	0.95	32
FissahaGetnet	0.87	0.94	0.91	36
HenokYared	0.84	0.82	0.83	56
KaleyesusBekele	0.82	0.84	0.83	49
MihretMoges	0.94	0.67	0.78	43
Nahusenay_Tsedalu_psycho	0.92	1.00	0.96	47
Nathan Dawit	0.94	0.90	0.92	49
Natneal_Mekonin	0.84	1.00	0.91	41
NeamnAshenafi	0.89	0.66	0.76	47
Rodas_Tadesse	0.97	0.85	0.90	39
Samson Abdela	0.67	0.80	0.73	5
Seyoum_Teshome	1.00	0.54	0.70	35
ShahidaHussen	0.62	0.70	0.66	43
TadesseGebremariam	0.79	0.75	0.77	36
TamiruTsigea	0.88	0.61	0.72	38
TamratGetachew	0.62	0.42	0.50	12
WidnehZenebe	0.77	0.72	0.75	57
YohannesEnberbr	0.66	0.75	0.70	28
ZemenuTenagne	0.59	0.67	0.63	24
dr_Zebene_Lema	0.77	0.92	0.84	25
prof_mesfin_w_maryam	0.97	0.97	0.97	37
sisay_sahlu	0.86	0.95	0.90	19
zemedkun	0.89	1.00	0.94	42
accuracy	0.81	1151		
macroavg	0.80	0.79	0.79	1151
weightedavg	0.82	0.81	0.81	1151

Appendix E

Secondoutput on csv file sample cv_results_ GridSearchCV of SVM with merged features

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_C	param_kernel	params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	mean_test_score	std_test_score	rank_test_score
0	14.0753	0.10201	6.01737	0.0709	10	linear	{'C': 10, 'kernel': 'linear'}	0.955	0.9403	0.95765	0.95	0.9457	0.95	0.00635	2
1	22.9966	0.38168	11.4091	0.1087	10	rbf	{'C': 10, 'kernel': 'rbf'}	0.945	0.9327	0.95114	0.93	0.937	0.939	0.0077	3
2	14.2041	0.10859	6.19631	0.0696	20	linear	{'C': 20, 'kernel': 'linear'}	0.955	0.9403	0.95765	0.95	0.9467	0.95	0.00622	1
3	23.1088	0.32686	11.5021	0.0648	20	rbf	{'C': 20, 'kernel': 'rbf'}	0.945	0.9327	0.95114	0.93	0.937	0.939	0.0077	3

Appendix F :-The effect of the number of authors on precision, recall and f1-score measures by features and models, percent

Features	No of Author's	NB-Multinomial[1]			SVM-Linear[2]			SVM-RBF[3]			NN-MLP-tanh[4]			NN-MLP-rule 5			NN-MLP-Logistic[6]		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
	3	89	89	89	85	85	85	70	72	71	92	90	91	84	84	84	88	86	87
	6	91	90	90	76	74	75	77	77	76	82	80	81	76	75	75	84	82	83
	9	79	79	79	74	72	73	66	70	67	80	79	79	76	74	75	84	84	84
	12	80	81	79	78	78	78	72	69	69	82	80	81	73	73	72	84	82	83
	15	77	77	76	76	74	75	74	73	72	82	81	81	70	70	69	85	84	84
	18	83	79	79	78	78	78	76	75	74	76	74	75	70	67	67	84	82	81
	21	83	83	82	73	74	72	62	62	62	73	73	73	76	74	75	84	84	84
	24	82	80	80	73	70	69	66	66	66	76	76	76	72	73	72	82	80	81
	27	76	75	75	74	73	73	66	64	64	68	66	66	65	64	64	78	76	76
Stylometry	31	74	73	73	73	71	72	64	63	62	66	66	66	63	63	63	72	70	71
	3	87	85	85	90	90	90	84	84	84	96	94	95	92	92	92	84	84	83
	6	90	86	87	92	90	91	92	93	92	91	90	90	91	91	91	92	91	90
	9	85	76	78	90	88	89	87	84	85	90	88	89	88	84	86	84	85	84
	12	83	74	75	89	92	91	92	91	91	93	92	91	90	90	90	90	89	89
	15	71	65	64	90	88	89	86	83	84	88	84	86	86	86	85	94	92	93
	18	72	67	69	90	88	89	85	83	82	92	88	91	84	84	84	88	88	88
	21	81	67	66	89	88	87	85	84	83	89	88	89	85	85	84	86	84	85
	24	76	69	67	88	90	89	85	84	84	88	90	89	85	85	85	88	86	87
N-gram-char	27	76	69	67	89	88	88	84	82	82	83	83	82	86	82	84	87	89	88
	31	67	63	60	85	89	87	86	84	85	80	79	79	88	84	86	87	87	88
Merged	3	84	80	78	92	92	92	85	85	84	94	92	93	95	95	95	88	85	86

	6	*90*	*87*	*87*	**91**	**95**	**93**	**90**	**87**	**87**	**91**	**93**	**92**	**94**	**90**	**92**	**92**	**91**	**91**
	9	*85*	*85*	*84*	**90**	**88**	**89**	**89**	**89**	**89**	**87**	**85**	**85**	**92**	**90**	**91**	**92**	**90**	**91**
	12	*86*	*84*	*84*	**92**	**91**	**91**	**92**	**91**	**91**	**90**	**92**	**91**	**88**	**90**	**88**	**95**	**93**	**94**
	15	*81*	*77*	*77*	**89**	**88**	**88**	**88**	**86**	**87**	**92**	**90**	**91**	**92**	**90**	**91**	**92**	**91**	**91**
	18	*83*	*79*	*79*	**90**	**88**	**89**	**91**	**88**	**88**	**93**	**91**	**92**	**93**	**93**	93	**86**	**84**	**84**
	21	*83*	*82*	*81*	**90**	**91**	**91**	**91**	**91**	**91**	**92**	**92**	**92**	**86**	**86**	**85**	**93**	**91**	**92**
	24	*85*	*83*	*83*	**88**	**89**	**88**	**88**	**87**	**87**	**90**	**94**	**92**	**86**	**85**	**85**	**86**	**85**	**85**
	7	*81*	*80*	*79*	**88**	**87**	**87**	**88**	**88**	**88**	**86**	**84**	**85**	**84**	**83**	**83**	**90**	**91**	**90**
	31	*80*	*78*	*77*	**87**	**86**	**87**	**86**	**86**	**86**	**81**	**81**	**81**	**88**	**89**	**89**	**81**	**81**	**81**

Key: 1. P=Precision, R=Recall and F=F1-score

Appendix G:- The effect of the number of articles per author on the three metrics by features and by models

Features	No_of Articles 's	NB-Multinomial1			SVM-Linear4			SVM-RBF5			NN-MLP-tanh2			NN-MLP-rule 3			NN-MLP-Logistic3		
	Articles	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
	10	64	68	64	47	47	43	47	47	43	49	43	49	32	39	32	45	53	45
	20	60	70	61	69	72	66	54	58	56	50	52	51	56	54	55	64	60	61
	30	71	67	66	68	67	64	66	68	67	66	68	67	55	53	52	69	69	68
	40	74	73	73	54	50	52	49	52	50	60	64	60	64	60	62	69	64	65
	50	82	80	79	56	55	54	54	52	53	55	55	54	68	64	65	66	64	65
	60	76	77	75	61	60	59	54	52	52	68	67	66	52	53	51	66	64	64
	70	70	68	69	70	68	69	60	60	59	64	66	65	61	62	60	70	72	71
	80	78	78	77	70	71	71	64	64	63	68	66	65	64	64	63	72	73	71
Stylometry	90	78	78	77	72	71	71	64	63	61	70	72	71	62	62	61	74	74	72
	100	80	79	79	88	86	86	89	89	89	95	94	95	86	84	85	84	85	84
	10	14	17	14	70	69	64	39	38	35	83	82	81	58	54	50	62	58	62
	20	72	70	71	78	74	76	57	62	61	80	78	79	68	70	69	79	76	77
	30	67	51	52	70	67	66	65	62	60	73	71	70	62	61	60	82	80	81
	40	78	74	76	80	78	79	70	74	71	82	80	81	80	74	78	82	78	80
	50	76	67	66	85	83	83	83	80	80	73	73	73	88	86	87	82	81	83
	60	73	69	67	80	82	81	76	75	74	85	85	84	77	77	76	82	80	81
	70	71	69	67	85	83	83	80	76	78	84	84	85	77	79	78	82	86	84
N-gram-char	80	72	69	68	84	80	82	81	79	79	88	84	86	80	79	78	84	88	86
	90	72	69	68	87	85	85	81	79	79	88	84	86	80	79	78	85	84	84
	100	75	74	73	98	97	97	96	94	94	98	96	97	96	94	97	95	95	95
	10	69	64	61	69	64	61	54	55	52	83	82	81	38	37	35	60	62	61

Merged	20	70	62	68	78	74	76	68	72	70	85	80	83	82	82	81	64	60	62
	30	75	71	69	74	72	69	74	72	69	63	61	60	80	79	77	86	84	85
	40	75	74	74	85	81	83	76	80	78	86	85	88	54	62	63	86	82	84
	50	82	80	78	82	80	78	76	74	75	89	88	88	74	72	72	86	84	85
	60	79	79	77	82	81	81	82	81	81	87	87	87	72	74	72	87	87	86
	70	76	82	77	85	81	83	86	82	84	88	87	87	76	79	77	89	88	87
	80	82	79	79	84	86	85	85	85	84	86	84	85	90	86	88	85	85	84
	90	83	83	82	89	85	87	85	85	84	90	86	88	84	83	83	86	86	84
	100	84	82	82	98	98	98	97	97	97	98	98	98	97	97	97	97	95	96