MACHINE LEARNING-BASED CATEGORIZATION OF RACISM IN AMHARIC TEXT USING MULTI-CLASS CLASSIFICATION

Abstract—The study addresses the prevalence of racist content on Facebook and Telegram, utilizing deep learning and supervised machine learning methods to classify Amharic racist comments and posts. Data collection involved FacePager, Beautifulsoup, and data exporting techniques, with annotation into non-racist, individual racist, regional racist, and country racist categories. LSTM, BI-LSTM, DT, SVM, NB, and KNN models were employed for data classification, using word2vec for feature extraction. The dataset was split into training, testing, and validation sets, and the BI-LSTM model exhibited superior accuracy. Future recommendations include preparing an Amharic language racist text dataset and applying pre-trained models for multi-class classification.

Keywords— Racism; Amharic Text Posts and Comments; Multi-Class Classification; Deep Learning; Supervised Machine Learning; Social Media

I. INTRODUCTION

In today's digital era, the widespread use of mobile phones and social media has reached unprecedented levels, with over 5.15 billion mobile phone users and 3.96 billion social media users globally [1]. In Ethiopia alone, investigated that [2] there are approximately 21.14 million users, actively engaging on popular platforms such as Facebook and Telegram [3]. These platforms have become valuable sources of data, including content that promotes hate speech, making them crucial for academic research, especially in the context of racism.

Social media has emerged as a vital communication tool, serving as a platform for news dissemination and raising social awareness [4]. Bowling [5] investigated Amharic one of the widely spoken languages in Ethiopia, plays a significant role in this digital landscape. However, this flexibility also poses a challenge as it allows for the propagation of racism. The objective of this research is to employ machine learning and deep learning algorithms to classify racist comments and posts in Amharic on platforms like Facebook and Telegram. By automatically detecting and categorizing racist content, the study aims to comprehend the individual, regional, and country-specific contexts surrounding this issue.

While acknowledging the significant role of social media as a communication tool in Ethiopia, this research also sheds light on the concerns regarding the spread of hate speech, particularly racism, on these platforms. To address this pressing issue, the researchers aim to harness the power of widely used platforms like Facebook and Telegram to develop a text classification system capable of automatically categorizing racist comments and posts [6]. Through content analysis and the utilization of supervised machine learning and deep learning techniques, the system aspires to detect and comprehend the prevalence of racism in online spaces, specifically in the Amharic language [7]. By contributing to the field of text classification, the research endeavors to mitigate racism on social media platforms.

The primary goal of this study is to combat the detrimental effects of racism on social media, particularly within the Amharic language context. Inadequate language resources have allowed racist ideas to propagate, resulting in adverse mental health outcomes and even inciting violence. While previous research has focused on hate speech, there is a notable lack of attention dedicated explicitly to the impact of racism on social media posts and comments like Facebook and Telegram.

To address this gap, the research employs supervised machine learning and deep learning algorithms to construct a classification model. Racist text is categorized into four distinct types: individual, regional, country, and non-racist. Through this categorization, the study aims to gain insights into the prevalence and impact of racism on social media platforms.

The findings of this research hold significant potential in assisting policymakers, social media companies, and legal experts in tackling the issue of racism and promoting inclusivity within society. By addressing the challenges faced in collecting evidence of racism, identifying suitable machine learning and deep learning algorithms for building a robust classification model, and assessing the effectiveness of the proposed model in classifying racist content in Amharic texts, this study seeks to contribute to the fight against racism on social media platforms.

This research aims to combat racism's harmful effects on social media, focusing on the Amharic language. It addresses the inadequacy of language tools in curbing the spread of racist ideas and emphasizes the significance of understanding racism's specific impact. By employing supervised machine learning and deep learning algorithms, the study aims to create a classification model that can categorize racist text effectively. The research's outcomes are expected to benefit policymakers, social media companies, and law experts in their efforts to tackle racism and create a more inclusive society.

This thesis is structured as follows: Section one: Introduction - Provides an overview of the research topic, its significance, and the study objectives. Section two: Related Work - Explores existing literature, studies, and theories relevant to the research topic, identifying knowledge gaps and areas for further investigation. Section three: Research Methodology - Describes the approach, techniques, and tools used to gather and analyze data, ensuring the validity and reliability of the research findings. Section four: Research Evaluation - Presents the results obtained through the research methodology. Finally, the Conclusion - Summarizes the key finding.

II. RELATED WORKS

In their study, Dereje & Abraham [8] utilized a machine learning approach to detect instances of racism in Amharic texts. The researchers categorized racist issues in both online and offline documents as either racist or non-racist. They labeled the data using supervised, semi-supervised, and deep learning algorithms. The BiLSTM model achieved the highest accuracy at 84.44%. However, the study did not address racism in social media posts and comments, and it only explored instances of racism after they had been posted, rather than before.

Tesfaye & Tune [9] developed a recurrent neural network (RNN) model for automatic hate speech classification in Amharic text on social media. The research used manually labeled datasets and applied data preprocessing techniques. The LSTM model achieved an accuracy of 97.9%. However, the research did not classify specific reasons for hate speech or racist issues, highlighting a research gap in the field.

Mossie & Wang [10] researched hate speech detection in Amharic text on Facebook using machine learning algorithms. They used RF and NB models with Word-2-vec and TF-IDF for feature selection. The highest accuracy of 79.83% was achieved using Word2vector. The research did not include categories such as offensiveness, sexism, and racism, indicating the need for future research on racism classification in Amharic text.

Gambäck & Sikdar [11] addressed hate speech classification on Twitter using deep learning CNN. The research classified sexism, racism, both, and non-hate speech categories. Word2vector achieved the highest F-score of 78.3%. However, the research lacked annotated data, multilingual classification, contextual understanding, interpretability, and protection against adversarial attacks. Future work should focus on the multiclass classification of racism in social media comments and posts.

Greevy & Smeaton [6] researched racist text classification on the web using support vector machines (SVM). The research focused on categorizing web documents into racist and racist-free using SVM with bag-of-words, bigrams, and part-of-speech tagging. The research did not specialize in the language of racist news classification, suggesting a research gap in this area.

Tulkens et al. [12] examined racist speech classification in Dutch social media using SVM. The research labeled datasets and achieved a 73% RUC-AUC value for identifying racist discourse. However, challenges related to sarcasm, sardonicism, and humorous posts were not addressed.

Williams et al. [13] explored the impact of racism on health disparities and outcomes. The research highlighted the link between racism and physical and mental health issues, emphasizing the role of institutional racism. Future research should focus on researching racism at the individual and group levels and its effects on access to resources.

Pei & Mehta [14] developed a multidimensional model for racism detection on social media, categorizing racist acts into stigmatization, offensiveness, blame, and exclusion. The research used BERT with cross-validation and achieved good performance. The research aimed to address the gap in understanding various dimensions of racism identification.

Preot & Ungar [15] researched user-level race and ethnicity prediction using Twitter data. Logistic regression with Elastic Net regularization was used for classification. The research collected data through online polls and achieved prediction accuracy, considering demographic information.

Based on related research, a gap exists in addressing racism through text classification in the Amharic language. While there are studies on detecting hate speech, none specifically analyze racist text in Amharic. Given the growing prevalence of racism in Amharic text on social media, further research is needed to bridge these gaps and gain a comprehensive understanding of racism, its expressions, and effective classification techniques in the Amharic context.

III. MATERIALS AND METHODS

This research proposes a machine learning-based approach to classify racist Amharic text on Facebook and Telegram.

The approach involves steps such as scraping posts and comments, preprocessing the data by removing punctuation, ASCII characters, numbers, links, HTML, and emojis, and normalizing the text. The labeled data is then split into training and test sets, followed by building a classification model to identify racist text. Finally, the proposed model is used to detect racism in Amharic text comments and posts depicted in Fig. 1.

A. Data Collection

This research aimed to collect posts and comments from Facebook and Telegram, chosen for their large user bases and group creation capabilities. The data collection process followed specific criteria, including the selection of sources, types of data collected, locations of data collection, methods employed, and data collection techniques.

Moreover to obtain representative data, popular Facebook and Telegram pages related to activism, news, and broadcasting channels were monitored. Various methods, such as scraping and exporting, were used to extract relevant text-based content and metadata. This research ensured that the collected data was comprehensive and applicable to the research questions, allowing for meaningful conclusions and insights to be drawn.

1) Sources of Data

This research focuses on the multi-class classification of Amharic text to detect racism in social media posts and comments. The choice of data sources plays a crucial role in determining research outcomes. Facebook and Telegram are selected as valuable platforms for collecting Amharic text data related to posting and commenting [16].

Researchers must carefully select relevant and appropriate data sources aligned with their research objectives. Specific criteria were established to ensure data relevancy, including seeking active public links and pages that frequently use the Amharic language. Pages addressing toxic issues and those

with a high frequency of Amharic language text were prioritized to capture the sentiment, tone, and comprehensive representation of the language.

2) Scraping of Data

Collecting data is a critical step in researching multi-class classification of racist issues on social media, and it entails gathering comments and posts from publicly accessible Facebook and Telegram pages. To achieve this, various data scraping and exporting techniques are used, including Face Pager, Beautiful Soup, and data exporting. Face Pager or Beautiful Soup are utilized for scraping specifically from Facebook while exporting is necessary for collecting data from Telegram.

B. Data Preprocessing

After preparing the dataset, the next step in the research involves preprocessing the data to construct a multi-class classification model for identifying racist text. Preprocessing is crucial in converting text into machine-readable formats, simplifying the complexity of supervised machine learning and deep learning algorithms.

Previous research by Gebeyehu & Sreenivasa [17] focused on text categorization and proposed various preprocessing techniques to enhance data quality. These techniques aim to eliminate irrelevant words and improve the effectiveness of the classification model for social media scraped data.

The preprocessing activities in this research include stop word removal, padding, normalization, and tokenization. By applying these techniques, the research aims to optimize the dataset for analysis and classification. This process allows supervised machine learning and deep learning algorithms to better understand and learn patterns from the preprocessed text data, thereby improving the accuracy and performance of the classification model.

1) Data Preparation

After data collection, the dataset contained diverse elements such as Amharic text, English text, punctuation, numbers, emails, links/URLs, HTML tags, and emojis. However, for the research, only the Amharic text is relevant. Preparing the Amharic text dataset involves removing irrelevant characters, applying transliteration techniques, and refining the data. Dataset annotation for detecting racist content in Amharic text requires using tools like Google Sheets and MS Excel. Cleaning and transliteration are essential in this process. The subsequent step is designing a training model to detect racist content.

2) Data Annotation

Data annotation in multi-class text classification research is challenging and time-consuming. This research focuses on classifying different types of racism, including non-racist, individual racist, regional racist, and country racist. A team of six law experts serves as annotators. To streamline the annotation process, a web-based tool with a user-friendly interface and comprehensive annotation guidelines has been developed.

The annotators are divided into two groups. Five annotators directly annotate the data using the web-based tool depicted in TABLE I, following the guidelines to assign class labels. One annotator is responsible for re-annotating any unclear annotations to ensure accuracy and consistency. By leveraging the web-based tool and guidelines, the research aims to improve efficiency and accuracy in classifying racist texts and extracting valuable insights from the annotated dataset.

TABLE II DATA SET ANNOTATORS

User Name	Phone	Address	Satus
AMM/Nati	+251930707411	Bahir Dar	Active
Bewuketu	+251922936537	Kimrdngay	Active
Addis	+251943486861	Ebnat	Active
BDUabebe	+251947018103	Bahir Dar	Active
Sisay	+251947018103	Seattle	Active

3) Annotation Guide Lines

A comprehensive annotation guideline has been established to ensure consistency and standardization during the annotation process. It provides detailed instructions and rules for annotators to accurately assign class labels related to racism, with input from a law expert. Each racist class label is precisely defined within the guidelines, including specific criteria for labeling text instances. Adhering to these guidelines improves the reliability and validity of the annotated dataset, enabling consistent and effective categorization.

The annotation guidelines play a crucial role in guiding annotators, promoting uniformity, and facilitating analysis and interpretation of the multi-class classification of racist text.

Annotators use web tools and guidelines to annotate data, with assigned subsets to prevent bias and promote collaboration. This ensures thorough annotations and enables cross-validation. Web tools facilitate efficient and organized annotation, resulting in reliable annotations for multi-class classification of racist texts into individual racist, regional racist, country racist, and non-racist as depicted in Fig.2.

Fig. 3.Interface of Web Annotator

4) Feature Extraction

The literature review focuses on classifying racist text and the importance of features like word2vec, word embedding, and padding. Advanced techniques such as word2vec and word embedding are crucial for enhancing Amharic text understanding and representation by capturing semantic relationships and contextual information about racism.

Padding is highlighted as a fundamental technique for handling varying Amharic text lengths and ensuring consistent input dimensions for effective training and classification by creating matrices form. This research highlights the significance and application of these features in the multi-class classification of racist text, contributing to the development of more accurate text classification models.

5) Data Set Splitting

Data splitting involves dividing 4545 datasets into subsets for training, validation, and testing. In this research, the Amharic racist text dataset was divided into training (80%), testing (10%), and validation (10%) sets. The training set was used to train the model, the validation set aided in hyperparameter tuning, and the testing set assessed the final model's performance, preventing overfitting and enhancing generalization to unseen data.

C. Model Building

This research focuses on utilizing long sequence text data to build models for multi-class classification of Amharic language racist text into four both racist and non-racist issues. The data's observable relationship between the words used in comments and the assigned labels was found to be highly intricate, resulting in linear inseparability by both supervised and deep learning algorithms.

Nave Bayan's (NBs) rely on the Structural Risk Minimization principle, a concept derived from computational learning theory from supervised machine learning algorithms. NBs are highly versatile learners, initially grasping linear threshold functions in their basic form. However, they can be easily adapted to learn more complex classifiers such as polynomial classifiers, radial basic function (RBF) networks, and three-layer sigmoid neural nets by simply incorporating an appropriate kernel function.

An outstanding characteristic of NBs lies in their capacity to learn independently of the feature space's dimensionality. Instead of considering the number of features, NBs evaluate the complexity of hypotheses based on the margin that separates the data. In other words, they focus on the space between data points rather than the features themselves. This advantageous property allows us to achieve generalization even when dealing with a large number of features, as long as the data is separable with a sufficiently wide margin using functions from the hypothesis space.

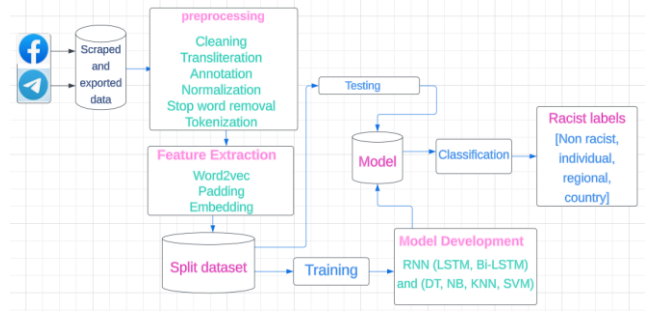Similar to the above activity bi-directional long short memory (Bi-LSTM) is better than others of RNN algorism.

Fig. 4.Proposed Model of Racist Text Classification

IV. RESULT AND DISCUSSION

A. Experimenting with Supervised Machine Learning Algorithm

Various experiments were conducted in this study using different classification algorithms to accurately classify racist Amharic text. The performance of each algorithm was evaluated through model accuracy with feature selection and hyper-parameter optimization.

In addition to accuracy, the models were also assessed using precision, F1 score, and recall metrics of hyper-parameter. One of the main objectives of this research was to develop a supervised machine-learning model for Amharic racist text classification. To achieve this, several classification algorithms such as Naive Bayes, Support Vector Machines, k-Nearest Neighbors, and Decision Trees were investigated to identify the most effective algorithm for optimal classification performance in the given task.

In general, when evaluating the results obtained through hyper-parameter experimentation in supervised machine learning algorithms, Naive Bayes (NBs) tends to achieve better performance compared to other algorithms depicted in TABLE III.

The measure of hyper-parameters plays a crucial role in fine-tuning the algorithm's behavior and optimizing its performance. Through a systematic exploration of various hyper-parameter settings, the NBs algorithm can be tailored to better fit the characteristics of the given dataset, leading to improved results.

B. Experimenting with Deep Learning Algorithm

LSTM and Bi-LSTM RNN algorithms are used for racist text classification in Amharic. They use neural networks to identify patterns and overcome the vanishing gradient problem. LSTM networks retain information from previous time steps to handle long-term dependencies. The implementation involves importing the tf.keras.layers.

LSTM module and using the softmax function to generate a probability distribution for multi-class classification. Bi-LSTM uses two LSTM layers to capture the past and future context in the classification of racist texts. The models are compiled using the sparse_categorical_crossentropy loss function and the Adam optimizer, and their performance is evaluated using accuracy....

The models is influenced by factors such as network architecture, dataset quality, and hyperparameter selection. Evaluation is based on the accuracy of the test dataset, with generalizability being important to avoid overfitting. TABLE IV illustrates how accuracy varies during training.

TABLE V EVALUATION RESULTS OF RNN MODEL

Models	Value s	Valu e size	Valid ation split	Accu racy	Loss value
BLST M:1	Batch	50	0.1	0.9948	0.0269
	Epoch	8			
BLST M:2	Batch	128	0.1	0.9954	0.0180
	Epoch	10			
BLST M:3	Batch	128	0.2	0.9968	0.0129
	Epoch	10			
BLST M:4	Batch	5	0.2	0.964	0.0114
	Epoch	128			
BLST M:5	Batch	128	0.15	0.9954	0.0231
	Epoch	50			
LSTM: 1	Batch	32	0.15	0.9861	0.0171
	Epoch	30			
LSTM: 2	Batch	128	0.15	0.9710	0.1332
	Epoch	50			
LSTM: 3	Batch	128	0.2	0.9710	0.1332
	Epoch	10			

TABLE VI EVALUATION RESULTS OF SUPERVISED MACHINE LEARNING MODEL

	Precis ion	recall	F1-score	Accur acy	Model
NB	78	96	88	78	Selecte d
DT	80	77	78	63	
KNN	78	97	87	76	
SVM	43	96	60	42	

Descusion of the result

The study results indicate that the deep learning model outperformed the supervised machine learning models. The evaluation was based on metrics such as accuracy, with the deep learning model achieving an impressive accuracy rate of over 99.46% TABLE VII while the supervised machine learning models achieved accuracy values of 78% and below TABLE VIII.

The experiments involved developing supervised machine-learning algorithms for classifying racist Amharic text using labeled instances. However, sequential Amharic text data requires consideration of data point order, which is not optimal for traditional supervised machine learning methods. In contrast, RNNs (Recurrent Neural Networks) are designed to handle sequential text data, allowing them to learn from each point in the sequence and outperform conventional supervised machine learning methods in multi-class classification of racist content in Amharic.

V. CONCLUSION

In the present era of social media, an overwhelming amount of data in the form of posts and comments is exchanged among users daily, significantly impacting their lives either positively or negatively. Among this vast volume of data, racist texts on platforms like Facebook play a dominant role in negatively affecting users' lives by promoting conflicts based on race-related issues of individuals, regions, and countries.

In this research, a model has been developed to categorize racist texts using recurrent neural networks and supervised machine learning. The successful implementation of the model began with defining racist texts, which encompass individual racism, regional racism, country racism, and non-racist language. These definitions were based on formal annotation guidelines prepared through the agreement of legal experts and Amhara Mass Media Company. The dataset was collected through various tools and websites and then labeled using data annotation tools, resulting in a dataset of 4545 samples that were prepared for multi-class classification.

To train and test the model, the dataset was split into training, validation, and test sets in an 80:10:10 ratio. The experiment employed a recurrent neural network model along with supervised machine learning, utilizing word2vec for feature representation. Through various evaluation metrics, the model achieved an impressive accuracy of 99.68% using Bi-LSTM (Bidirectional Long Short-Term Memory).

The findings of this study demonstrate that deep learning neural network models can effectively categorize racist posts and comments on platforms like Facebook and Telegram for Amharic text data analysis. Among the different hyper-parameters and evaluation metrics, the Bi-LSTM model outperformed other RNN and supervised machine learning models in terms of efficiency and accuracy. Furthermore, altering the neural network's hyper-parameters directly affected the accuracy of the deep learning model, reinforcing the significance of careful parameter tuning in such analyses.

Abbreviations

TABLE IX ABBREVIATIONS AND ACRONYMS

Abbreviations and Acronyms

ANN	Artificial Neural Network
Bi-LSTM	Bi-directional Long Short-Term Memory
DT	Diction Tree
KNN	K Nearest Neighbor
LSTM	Long Short-Term Memory
NB	Nave Bayan's
NLP	Natural Language Processing
RNN	Recurrent Neural Network
SVM	Support vector machine

Ethics approval and consent to participate

The written text has been confirmed and recognized as accurate, representing the actual findings derived from an experiment.

Authors' contributions

This study makes the following contributions to knowledge:

- The creation of a data set tailored explicitly for Amharic racist text categorization.

- The development of classification models in Amharic to identify and categorize racist text.

REFERENCE

[1] S. G. Tesfaye, S. G. Tesfaye, and K. K. Tune, "Automated Amharic Hate Speech Posts and Comments Detection Model using Recurrent Neural Network," pp. 1–2, 2020, [Online]. Available: https://doi.org/10.21203/rs.3.rs-114533/v1%0ALicense:

[2] W. Akram and R. Kumar, "A Study on Positive and Negative Effects of Social Media on Society," *Int. J. Comput. Sci. Eng.*, vol. 5, no. 10, p. 347, 2017, doi: 10.26438/ijcse/v5i10.351354.

[3] Z. Mossie and J. Wang, "Vulnerable community identification using hate speech detection on social media," *ScienceDirect*, no. November 2018, pp. 1–16, 2019, [Online]. Available: www.elsevier.com/locate/infoproman

[4] Wasim Ahmed, "Using Twitter as a data source: An overview of social media research tools (2015) | Impact of Social Sciences," 2021, [Online]. Available: https://blogs.lse.ac.uk/impactofsocialsciences/2015/07/10/social-media-research-tools-overview/

[5] Bowling, "RACIAL HARASSMENT AND THE PROCESS OF VICTIMIZATION : Conceptual and Methodological Implications for the Local Crime Survey Author (s): Benjamin Bowling Source : The British Journal of Criminology, Spring 1993, Vol . 33, No . 2 (Spring 1993), Stable," *Oxford*, vol. 33, no. 2, pp. 230–250, 1993, [Online]. Available: %0AAll use subject to the

[6] E. Greevy and A. F. Smeaton, "Classifying Racist Texts Using A Support Vector Machine," *SIGIR*, vol. 04, pp. 468–469, 2004.

[7] D. Endalie and G. Haile, "Automated Amharic News Categorization Using Deep Learning Models," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–9, 2021, doi: 10.1155/2021/3774607.

[8] S. Aweke, "Amharic Racism Text Detection from The Writer Using Machine learning approach," 2022. [Online]. Available: http://ir.bdu.edu.et/handle/123456789/15402

[9] S. G. Tesfaye and K. K. Tune, "Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network," *Res. Sq.*, pp. 1–20, 2020, [Online]. Available: https://orcid.org/0000-0001-6933-3442

[10] Z. Mossie and J. Wang, "S OCIAL N ETWORK H ATE S SPEECH DETECTION FOR AMHARIC LANGUAGE," *Proc. Comput. Sci. Inf. Technol*, pp. 41–55, 2018.

[11] B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate-Speech," *Proc. of the First Work. Abus. Lang. Online*, no. 7491, pp. 85–90, 2017.

[12] S. Tulkens, L. Hilte, E. Lodewyckx, B. Verhoeven, and W. Daelemans, "The Automated Detection of Racist Discourse in Dutch Social Media," *Comput. Linguist. Netherlands J.*, vol. 6, no. March, pp. 3–20, 2016.

[13] D. R. Williams, J. A. Lawrence, and B. A. Davis, "Racism and Health : Evidence and Needed Research," vol. 40, pp. 105–125, 2019.

[14] X. Pei and D. Mehta, "Multidimensional racism classification during COVID-19: stigmatization, offensiveness, blame, and exclusion," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, pp. 1–13, 2022, doi: 10.1007/s13278-022-00967-9.

[15] D. Preot and L. Ungar, "User-Level Race and Ethnicity Predictors from Twitter Text," *Int. Conf. Comput. Linguist.*, vol. 27, pp. 1534–1545, 2018.

[16] R. Chugh and U. Ruhi, "Social media in higher education : A literature review of Facebook," *Springer*, no. 2017, pp. 605–616, 2018, doi: 10.1007/s10639-017-9621-2.

[17] S. Gebeyehu and V. Sreenivasa, "A Two-Step Data Mining Approach for Amharic Text Classification," *AJER*, vol. 03, no. 04, p. 254, 2014.