Sentiment analysis based- Hate speech detection of Amharic social media post and comments

A Thesis Presented

by

Tsion Shita Bedassa

to

The Faculty of Informatics

of

St. Mary's University

In Partial Fulfillment of the Requirements for the Degree of Master of Science

in

Computer Science

January, 2025

ACCEPTANCE

Sentiment analysis based- Hate speech detection of Amharic social media post and comments

By

TSION SHITA BEDASSA

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

Mulugeta Adbaru
Internal Examiner
{Full Name, Signature and Date}

Mesfin Abebe
External Examiner
{Full Name, Signature and Date}

_____

Dean, Faculty of Informatics
{Full Name, Signature and Date}

January 2025

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Tsion Shita Bedassa
Full Name of Student

_____
Signature

Addis Ababa

Ethiopia

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

Alembante mulu(PHD)
Full Name of Advisor

_____
Signature

Addis Ababa

Ethiopia

January 2025

ACKNOWLEDGEMENT

I want to start by giving grateful to God for providing me with the patience necessary to complete this research. I also want to thank Dr. Alembante M., my advisor, for his kind assistance, ongoing counsel, helpful criticism, and recommendations during the preparation on this research report.

lastly, I would want to express my gratitude to my colleagues who helped with this research's annotation.

TABLE OF CONTENTS

List of tables

List of figures

List of Abbreviations

BILSTM	Bidirectional long short-term memory
BOW	Bag of Words
FP	False Positive
FN	False Negative
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
NLTK	Natural language toolkit
RNN	Recurrent Neural Network
TF-IDF	Term frequency-inverse document frequency
TP	True Positive
TN	True Negative
CNN	Convolutional neural network

Abstract

Social media allows the user to post, comment and communicate freely. And this led to an increasing amount of online hate speech. Online hate speech has different offline repercussions, according to studies. In recent years, hate speech have led to internal violence, relocation, and human rights violations against specific social groups around the world. And Ethiopian societies are among the victims. To lessen the spread of hate speech, this study develops Amharic hate speech detection. The study's main goal is to create a model for detecting hate speech by taking into account sentiment analysis of the relevant datasets and proving a link between hate speech and sentiment analysis. Peacemakers can take action when hate speech comments are being circulated online by using an Amharic-language hate speech detection system. Additionally, it will assist owners of social media platforms by automatically reporting hate speech remarks before they are seen by a wider audience

Comments were gathered from Facebook, TikTok, and YouTube channels in order to create a labeled large Amharic dataset. Following data cleaning, 79991 hate and hate-free annotated datasets along with their sentiment were obtained. To label the dataset as hate and hate-free, new annotation guidelines were created. Despite previous related work, recent and large dataset were collected and their sentiment were also considered. To construct the model, CNN and GRU deep learnings were used in conjunction with Word embedding features.

Negative sentiment was revealed to be the source of hate speech content. And most of the hate free dataset were found to be having positive sentiment. Using datasets that have been annotated by humans as a hate and hate free, the GRU and CNN models demonstrated respective accuracies of 0.90 and 0.72. And when both hate and non-hate annotated datasets along with their sentiment were used in the hate speech detection model, the models' respective accuracies become 0.75 and 0.74. As a result, in both model GRU outperform CNN model, and the CNN approach shows good performance for the hate speech detection model that was developed by integrating sentiment analysis.

Key words: Deep Learning, Hate speech detection, Amharic post and comment dataset, sentiment analysis, Gated Recurrent Unit; Convolutional Neural Networks

CHAPTER ONE

INTRODUCTION

1.1 Background

The definition of hate speech is not universally agreed upon. According to a recent study [1] speech which incites violence against or fosters an atmosphere of prejudice that could lead to actual violent acts against that group. According to some definitions, using language that disparages groups on the basis of particular traits constitutes it.

Due to the quick development of blogs, forums, and several other social networks, peoples now a days acquire helpful information, exchange ideas, thoughts, and experiences, and even influence one another's opinion-expressing. Accessing important information has become simpler for people due to the quick growth of these digital platforms. Users can stay informed, research various topics, and share their findings with a larger community in these online places. People utilize these platforms to exchange ideas, share personal stories, and have deep conversations in addition to acquiring information. Because people influence one another through conversation and the sharing of different points of view, these encounters frequently have an impact on public opinion.

Any content posted, shared, accessed, or received on social media include anything might be construed as discriminating, to individuals in general. This suggests the necessity for an automated system to filter material from online communication. Social media content, whether posted, shared, accessed, or received, can sometimes be interpreted as discriminatory toward certain individuals or groups. With the vast amount of information being exchanged online, there is an increasing concern that harmful or biased material may influence public opinion and contribute to social inequality. Effectively managing and monitoring such content is essential to fostering a more inclusive and respectful digital space.

This issue emphasizes the importance of implementing an automated system to filter inappropriate or offensive material in online interactions. Leveraging technologies like artificial intelligence and machine learning, such a system could help identify and limit the

spread of discriminatory content. By adopting these solutions, online platforms can promote a safer and more ethical environment while maintaining a balance between free speech and responsible communication.

With the help of attorneys, a speech that is hate and disseminated through books, magazines, or other multimedia platforms can be quickly located and tracked down in a matter of hours. Nonetheless, the vast bulk of speech that are hate on the platform of social media can originate from anonymous people hiding behind a screen, making it challenging to police with physical force. Speech that are Hate and spread on the online communication in Ethiopia has recently come under fire, particularly for being the impetus for ethnic violence. In Ethiopia, hate speech circulating on online platforms has recently drawn significant criticism, particularly for its role in provoking ethnic violence. These harmful messages, often shared on social media, have intensified ethnic tensions and contributed to outbreaks of violence in various regions.

The growing presence of online hate speech has raised alarms about its effect on societal unity and public security. As these divisive messages continue to spread, there is an increasing need for strategies to prevent their dissemination and address the root causes of the hostility they promote. Strengthening online regulations and encouraging responsible communication are crucial steps toward reducing conflict and safeguarding vulnerable populations. Accordingly , government of Ethiopia primarily blocks social media sites when rebellion breaks out in order to lessen the impact of posts made by various users that, whether intentional or not, incite animosity and lead to disputes[2].

Amharic, is one of Semitic language, serves as Ethiopia's official working language. It is, nevertheless, one of the languages with the fewest computational linguistic works and the least amount of study and resources. Amharic is a common language used by both public and commercial broadcast media in Ethiopia to communicate with viewers.
Although Amharic is essential for everyday communication and media, the lack of progress in computational linguistics for the language creates challenges in adapting it to modern technological advancements. This gap in resources and research limits the ability of

Amharic to be effectively integrated into digital platforms, highlighting the need for further investment in its linguistic and technological development.

Moreover, the absence of adequate tools for Amharic in areas like natural language processing and machine translation restricts its broader application in technology. As Ethiopia continues to evolve and connect with global digital trends, prioritizing the development of computational resources for Amharic will enable speakers to fully access digital tools and services. This investment is key to reducing the technological divide and ensuring Amharic's continued relevance in the digital era.

Based on the material that is currently available, investigation into identifying speech that are hate and written in Amharic is remaining in its early stages. Sentiment analysis tasks are important for many applications; hence sentiment analysis will be applied in the development of the hate detection mode. With the aid of APIs, browser extensions, Python scripts, and scrapers, new datasets will be gathered from a range of social media sites. The data undergoes preprocessing and labeled as hate and hate-free datasets using a newly developed annotation guideline, followed by annotating the corresponding sentiment, using existing annotation guideline[3].

A hate speech detection model will next be developed using the labeled dataset. Deep learning methods and a variety of feature extraction techniques will be used to build the model. Additionally, this work will attempt to take into account the sentiment of Amharic social media texts By including sentiment analysis, the model will be better equipped to assess the tone and underlying intent of messages, improving its ability to detect hate speech. This added dimension will offer more detailed insights into the content shared on social media and enhance the effectiveness of detecting harmful language in Amharic online interactions.

1.2 Motivation

Hate speech are constantly posted and shared among Ethiopians social media platforms. This might result in violent crimes or hate crimes, which would ruin people's lives as well as those of their families, communities, and the entire nation. The increase of online hate speech served as the impetus for this topic's investigation. Hate speech is pervasive on social media platforms because they give people a free and unrestricted platform to share their thoughts, opinions, and ideas. Using social media become common among Ethiopians youth and the more these young generation read the hateful comments, the more this inappropriate speech be normalized among the them. This intern is affecting the peaceful existence of people.

Hate speech that incites violence and distracts from property must be monitored in order to provide a safe environment for society. Ethiopian government's response to hate speech consists of shutting down the Internet countrywide and temporarily blocking social media accounts. Blocking the internet connection can have a negative consequence since most peoples of the county day to day life start to depend on the internet connection. Hence, these works propose more accurate Amharic hateful comment detection model incorporating sentiment analysis to tackle proliferation of hateful comment. The methodology used is based on the notion that social media comments and posts with negative emotion are more likely to contain hate content. These comments and replies can also mislead data mining activities and lead to incorrect classification.it worth to develop hate speech detection taking this fact in to consideration.

1.3 Statement of the problem

Hateful content that spreads on social media has the power to exacerbate societal unrest, spark conflicts, and bring about crises. Hate speech spread via social media has also horrifying effects on the mental health of each victim because it instills dread, damages one's sense of self-worth, and inflicts psychological injury. Information filtering solutions that are based on sentiment analysis of social media texts can be used to block access to inappropriate or illegal social media content as well as politically and socially sensitive

content. Social media filtering proved costly in a number of ways. In the Ethiopian internet community, social media hate speech is becoming common and problematic because people disseminate them while hiding behind their computers and it gives them an unrestricted liberty to voice their opinions

The Ethiopian government has frequently blocked parts or all of the Internet out of concern that hate speech spreading on social media could exacerbate the security issues. Thus, it is undeniable that hate speech spreads via social media. As in study conducted in[4],the quantity of tweets in the nation that contain abusive terms is rising over time. Only Throughout the past years has the field of identifying hateful speech research grown, and the majority of these studies were done in foreign languages.

Because Amharic differs from other languages in its semantic, lexical, and syntactic structure, methodologies and procedures designed for other languages do not apply directly to Amharic. As a result, it is necessary to research the morphology that suit the nature of the Amharic language and prepare a well annotated dataset. Sentiment analysis is incredibly helpful for social media monitoring since it gives us a general idea of how the public feels about a given subject. It is helpful for rapidly deriving conclusions from vast amounts of textual data. And this work proposes a classification model by utilizing the sentiment

1.4 Research Question

RQ1. What are the steps to develop Amharic language sentiment analysis-based hate speech detection model for social media be developed?

RQ2. Which feature extraction methods should be used to extract relevant features from Amharic hate speech data?

RQ3. Which deep learning algorithms detect hateful comments more accurately?

RQ4. To what extent the proposed approach identifies hate speech in Amharic online hateful comments?

RQ5. To what extend the accuracy of hate speech detection of Amharic social media texts improved by considering the sentiment of the dataset?

1.5 Objective of the Study

1.5.1 General Objective

This study's primary goal is to create and develop framework for detecting hate speech by taking account the sentiment of social media posts and comments that are written in Amharic text using deep learning techniques.

1.5.2 Specific Objective

Specific Objectives

The following particular goals are noted in order to accomplish the overall goal:

- Conduct a literature review
- Collect the dataset and get the necessary corpus ready.
- Prepare standards to label the dataset
- Annotate the dataset
- Embrace a deep learning approach appropriate for the Amharic language's morphological characteristics.
- Design and develop Hate Speech detection model.
- Explore the relationship between sentiment analysis and hate speech
- Assess the detection model performance

1.6 Methodology

1.6.1 Literature Review

To improve understanding in this area, a thorough assessment of literature will be conducted.

1.6.2 Research Design

The design science research technique will be applied

1.6.3 Data Preparation

Comments will be gathered from Tik Tok, Facebook and YouTube. comment collector extensions, different python scripts and Api will be utilized for this dataset acquisition. The collected data will be labeled with the annotated guideline as a hate and hate free along with their sentiment polarity. Techniques such as removing the unnecessary data, normalization, tokenization, stop word removal, and expanding abbreviated texts followed by feature extraction will be implemented to prepare the relevant corpus.

1.6.4 Tools and Techniques

- To collect the data from the social media platforms you tube API, comment collector extensions, and different python codes will be used.
- The tools and packages to be employed using the Python along with different library including Scikit Learn, NumPy, Pandas, NLTK, TensorFlow and Jupyter note book on google collab.

1.6.5 Evaluation

Accuracy, recall, f-measure and precision, are the assessment measures that will be utilized in the experiment to assess performance.

1.7 Significance of the study

This work will introduce new hate and non-hate dataset along with their sentiment and create a model for detecting hate speech by taking into account the sentiment of social media comments and responses, which might skew data mining efforts and lead to incorrect categorization. Owing to TikTok’s increasing popularity, the majority of the data was collected from its interns, this work provides a recent dataset and a variety of hate speech content.

And explore to what extent the consideration of sentiment help the effectiveness of identifying hateful comments on social media Additionally, it backs the government of Ethiopia's ongoing efforts to prevent hate speech from spreading on social media.

1.8 Scope and limitation

The extent of this research is restricted to identifying hate comments that are written in the language of Amharic. The study excludes social media data components, such as images, audio, videos, and other emotive signals. comments and responses could be conveyed in a number of forms, Comprising audio and video. This work focuses on only texts.

1.9 Organization of the Thesis

There are five chapters in this work. The study is introduced in the first chapter, which also covers the background, problem identification, goal, main contribution, and study scope. The second chapter then reviews similar works. Model construction techniques and tools are covered in the third chapter. The experiments, analysis, and findings are the main topics of chapter four. The study's conclusions and recommendations were discussed in the last chapter.

CHAPTER TWO

LITRATURE REVIEW

2.1 Overview

This section covers the research done on social media hate speech detection by various researchers. The morphology of Amharic language has also been thoroughly discussed, and the procedures in place to track hateful comments on social platforms have been examined. Furthermore, a thorough investigation of pertinent literature has been conducted to enhance comprehension of the concept and explore the research subject.

2.2 Hate speech

Hate speech can be viewed as inherently wrong, harmful directly, or harmful indirectly, according to[5]. It does not uphold the dignity of hate speech speakers and instead damages the self-respect of those who are targeted by it. Furthermore, self-respect at least in a great many cases is adequate to explain the harm caused by hate speech due to its tremendous significance and relevance in our psychic existence. A "trigger" event is all that is needed for hate crimes to occur, according to many research, and negative sentiments towards minorities and stereotypes tend to grow over time.

2.3 Hate Speech on Social Media

The fact that almost all of the Content of social media is created by the users, makes the platform unique and can be contributed by anybody who wants to express their opinions in public, not just a select set of journalists or persons. Therefore, editorial or any other type of previous oversight by competent agencies, nor the position of an individual who can share information, limit communication through social media.

Hate speech has long been a problem that disproportionately affects minorities across a broad spectrum of society. Social media platforms are web-based tools that allow anybody to share anything, including ideas, images, audio and video files, and more[6]. Hate speech has been a persistent issue that primarily impacts minority groups, leading to discrimination and social division across various areas of society. It can take many forms, such as offensive language or violent threats, and has serious emotional, psychological, and

societal repercussions for those affected. Minorities, in particular, face a disproportionate amount of this harmful communication, which can lead to marginalization. With the rise of digital connectivity, the reach of hate speech has expanded, causing harm to more individuals and communities.

Social media platforms are a key factor in this problem, as they allow users to freely share ideas, images, videos, and other content. While these platforms have made communication more accessible, they have also contributed to the spread of harmful speech. The open nature of social media enables rapid distribution of such content, often without enough oversight or regulation. As a result, hate speech can spread unchecked, requiring a balance between protecting free speech and taking action against dangerous content

Today's online social media platforms provide free communication between users for almost nothing. Users are using these platforms more frequently to exchange news as well as communicate with one another. Although these platforms offer an open forum for individuals to express themselves, they also have a negative aspect. Based on empirical evidence, it has been observed that refugees and immigrants, together with religious and ethnic minorities, are the most frequently targeted groups for such speech. These groups are marginalized, and acts of terrorism and transgressions by members of specific minority groups are sometimes used as an excuse to damage the reputation of those who are weaker than others.

This has caused social media to grow quickly and brought about certain unfavorable effects, such as the lack of specific laws pertaining to free speech and an increase in the use of derogatory and cruel words on the platform. Hate speech spread via social media has also horrifying effects on the mental health of each victim because it instills dread, damages one's sense of self-worth, and inflicts psychological injury[7].

Social media define hate speech as follows,

Hate speech, according to Facebook [31], is when someone directly attacks someone because of their identity, or a severe sickness .Assaults are depicted as acts of viciousness or dehumanization, mediocrity complex comments, or requests for segregation. Twitter characterizes despise discourse as ‘Scornful conduct: that specifically assault or debilitate others on the premise of a race, national origin, and ethnicity, age, incapacity’[32]

In addition to the lack of control mechanism particularly in Ethiopia, nothing shared on social media can be relied upon. since hate speech and disinformation, there has been a lot of unrest and instability in Ethiopia lately. Hate speech is spreading quickly across the globe on various social media sites. It is the obligation of the platform owners to regulate the spread of hate speech in order to combat its spread on social media.

Even though Ethiopia's government issued a decree to outlaw hate speech, the country has seen a sharp rise in the spread of hate speech. Ethiopians are currently suffering from the root causes of hate speech as well as associated issues like political, economic, religious, and racial conflict.

Although the Ethiopian government has enacted a law to prohibit hate speech, the country has witnessed a sharp increase in its proliferation. This rise in harmful speech is a symptom of deeper societal issues, including ongoing political, economic, religious, and ethnic conflicts that continue to divide the nation. Despite the government's attempts to combat hate speech, these underlying problems remain unresolved, making it difficult to control the spread of such rhetoric.

The people of Ethiopia are bearing the brunt of these underlying causes, as hate speech exacerbates existing tensions and deepens divisions. Political unrest, economic disparities, and religious or ethnic discrimination have become more entangled with the spread of hate speech, making it harder to promote national unity and harmony. In this context, the challenge of addressing hate speech extends beyond legal measures, requiring a broader approach to address the root causes of conflict within society.

2.4 Sentiment Analysis

Sentiment analysis involves the identification and extraction of subjective information from language, such as opinions, emotions, and attitudes. It is a crucial part of natural language processing (NLP) that helps in understanding how individuals feel or think about certain topics, products, or issues. By analyzing text or speech, sentiment analysis can classify sentiments as positive, negative, or neutral, and even detect more specific emotions like anger or joy.

According to [8], sentiment analysis is typically categorized into three levels: document-level, sentence-level, and aspect-level. Document-level analysis looks at the overall sentiment of a full text, such as an article or review, while sentence-level analysis examines the sentiment of each individual sentence within the text. Aspect-level analysis focuses on particular aspects or features mentioned in the text, such as specific products or services, and determines the sentiment tied to those elements. This layered approach enables more detailed and accurate insights into the sentiments expressed, making sentiment analysis a valuable tool in areas like market research and social media analysis.

To establish a peaceful living environment in Ethiopia, it is crucial to analyze hate speech identification in Amharic. In this work, in addition to other natural language processing techniques, sentiment analysis will be considered as additional feature on Amharic social media texts to detect and identify hate speech.

2.5 Amharic Language

Amharic comes in second after Arabic in terms of being semitic language spoken worldwide. It uses distinct "Fidel" scripts, where a syllable-based writing system in which each graphic character contains both vowels and consonants. According to a research in [9] Amharic, there are twenty numerals, eight punctuation marks, and 51 labeled characters. Amharic is a highly inflected, morphologically complicated language with around 310 distinct characters. In terms of how syntactic and grammatical relationships are expressed at the word level, it is morphologically rich.

In texts written in Amharic, homophonic letters are frequently noted to be used interchangeably, and the unusual labiovelars are not consistently used. This brings the total number of Amharic characters to 231 without counting the additional letters that are not part of the main row and have separate sounds[10].The language of Amharic exhibits a morphological phenomenon known as the root-pattern. A root is a group of consonants with a fundamental lexical meaning. They are also known as radicals. A pattern is a group of vowels that are positioned in between a root's consonants to create a stem. Since the majority of Ethiopians speak Amharic as their mother tongue or a secondary language and use it as a medium of internet communication, it makes sense to consider building the model.

Table2.1: Summery of morphology of Amharic language

Category	**Description**	**Examples**
Structure of words	Roots are primarily triconsonantal, whereas patterns add vowels or affixes for derivation and inflection.	**Root**: "□□□" (to break) **Verb**: "□□□□" (was broken)
Morphology of Nouns	- **Gender**: Masculine (default) and feminine (marked by -it or -wa). - **Number**: Singular (default) and plural (suffix -och). - **Definiteness**: Suffix -u (m.), -wa (f.).	"□□□" (*wend* - "man"), "□□" (*sēt* - "woman"), "□□□□" (*məṣḥaf* - "book") "□□□□□" (*məṣḥaf-och* - "books").
Morphology of verb	Using templates, prefixes, and suffixes, verbs inflect for person, gender, number, tense/aspect, and mood.	"□□□□" ("he spoke"), "□□□□□" ("he speaks").
Tense/Aspect	- Perfective: Actions finished. Imperfective: Consistent or routine behavior. Imperative/Jussive: Orders or demands.	Perfective: "□□" "he worked." Imperative/Jussive: "□□" "work!"

Category	Description	Examples
Subject Agreement	When it comes to person, gender, and number, verbs agree with the subject.	"□□□□□□□□□" ("I did not speak to you all").

2.6 Existing Hate Speech Detection approaches

These include deep learning, machine learning and hybrid approach. An attempt has been made to determine which of these the researchers use the most.

2.6.1Machine learning algorithms

Supervised learning

When the dataset is pre-labeled, supervised learning can be applied in this method. Although labeling tasks requires a great deal of work and time, it works better for events that are domain-dependent.

Unsupervised learning

Instead of employing human labor to label a large training set, it dynamically extracts important terms linked to the domain. Semantic Hate and theme-based elements were added to their model, which produced the best result[11].

Reinforcement Learning

Semi-supervised learning which combines supervised and unsupervised learning is a technique used by several well-known algorithms. In this case, an ML model is trained using both labeled and unlabeled examples.

In addition to machine learning algorithms, several manual feature engineering and rules are used in machine learning methodologies. Hate speech can be best classified using the bag of words, word, and character n-grams features, according to research. Other popular classifiers include Naïve Bayes, SVM, Random Forests, Decision Trees, and Logistic Regression.

2.6.2 Deep Learning Models

Multiple layers of interconnected nodes make up deep neural networks, and each layer builds on the one before it to optimize and improve classification or prediction. Visible layers are the input and output layers of a deep neural network[12].

A deep learning network consists of the below components. Neurons computes the weighted average of the input data after processing it through a nonlinear function. Two propagation functions are involved in deep learning: forward propagation, which provides the output value based on the input value, and backward propagation, which provides the "error value." Each connection, which connects a neuron from one layer to another, either from the same layer or from a different layer, has a value for weight. To reduce the chance of inaccuracy, the weight value should be decreased. The learning rate determines how fast or slowly we choose to update the model's weight (parameter) values.

Recurrent Neural Network (RNN)

Instead of producing a single result for a single input, RNNs can plan out several inputs and products to yield one-to-many, many-to-one, or many-to-many outputs[12].

CNN

Three layers: input, output, and one or more hidden levels are all included in CNNs, a particular kind of neural network made up of node layers. Convolutional Neural Networks (CNNs) use word embeddings to extract important elements that are used to classify hate speech. The first step in the process is text preprocessing, which involves turning words into dense vector representations, normalizing sequence lengths through padding, and tokenizing words. Convolutional layers receive these embeddings after which filters detect important local patterns, including dangerous phrases or particular word combinations. Pooling layers are used to further refine the generated feature maps, which facilitate calculations and enhance the model's capacity for generalization.

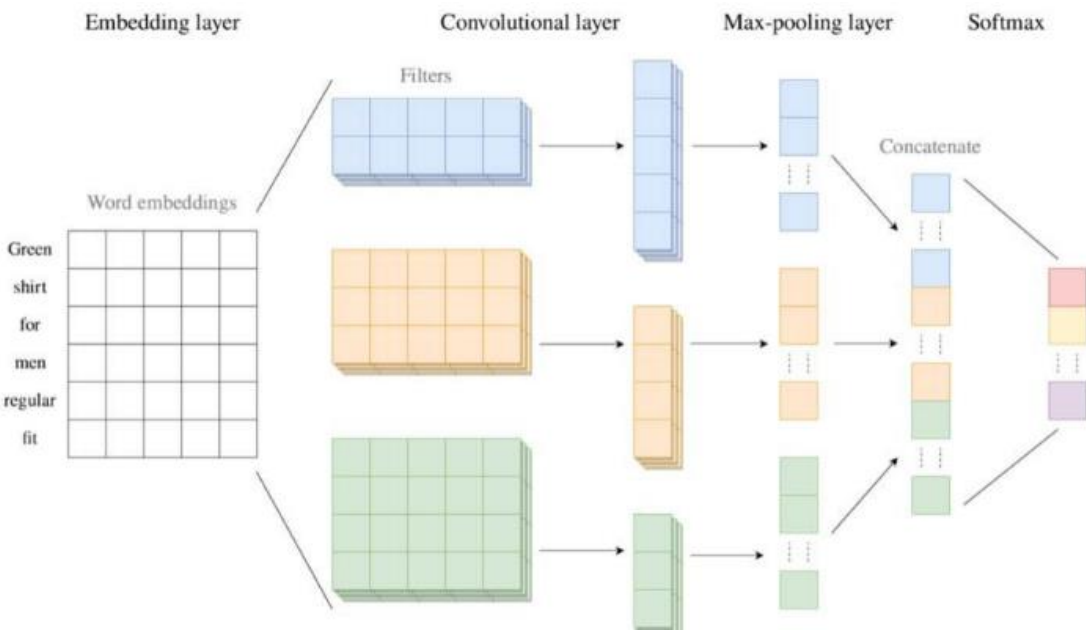


Figure 2.1 CNN architecture

Refined characteristics are sent to fully linked layers, which use the patterns found to create predictions. The final output layer determines whether hate speech is included in the input text by using activation functions such as sigmoid for binary classification or SoftMax for multi-class assignments. Because of their capacity to effectively recognize pertinent local patterns, manage text of various lengths, and maintain their resilience to slight input alterations, CNNs are especially well-suited for this application.

Including contextualized models like BERT or pre-trained embeddings like Word2Vec can help increase accuracy. The model's capacity to capture contextual and local characteristics is improved when CNNs are combined with other architectures, such as RNNs or attention mechanisms. CNNs are very useful tools for identifying hate speech in text when they are trained on properly labeled data.

GRU

By using unique "gates" to control whether information it retains or discards, it is able to spot patterns in lengthy sequences. A particular kind of RNN called Gated Recurrent Units (GRUs) is used to identify hate speech by examining textual context and sequential patterns. Prior to processing, text is transformed into dense vector representations, tokenized, and padded to consistent lengths. Using reset and update gates to regulate the information flow, GRUs process these sequences one step at a time. By capturing both short-term and long-term relationships, these techniques enable the model to successfully detect detrimental circumstances or offensive language.

The final hidden state or an attention mechanism highlights important textual elements in the hidden states generated by GRUs, which serve to summarize the sequence. To find out if the input contains hate speech, this condensed representation is fed into a classification layer and fully linked layers. Because they retain pertinent information from previous segments of the sequence, GRUs are very good at comprehending contextual nuances and are hence very useful for text categorization tasks.

Pre-trained embeddings such as Word2Vec or BERT for richer word representations can be used by GRUs to increase accuracy. The model can capture both local text patterns and a larger sequential context when GRUs and CNNs are combined, and attention methods help the model concentrate on key elements of the input. With these tactics and top-notch training data, GRUs are a reliable method for identifying both overt and covert hate speech in text.

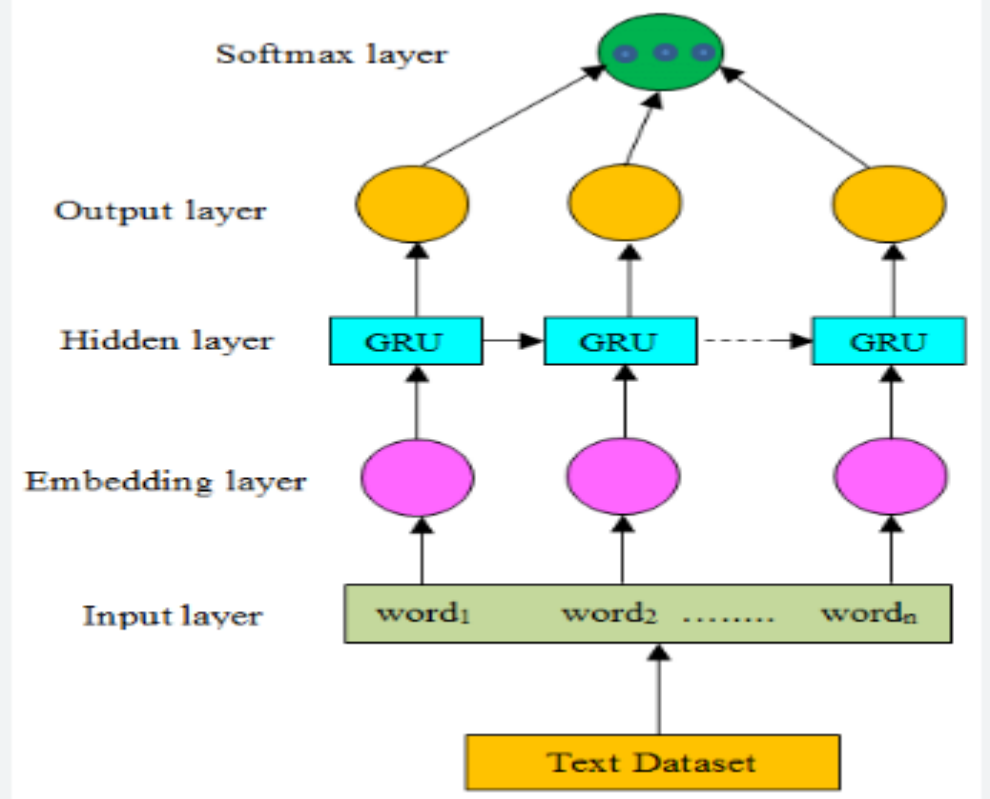


Figure 2.2 GRU architecture

Long Short-Term Memory (LSTM)

The purpose of this kind of recurrent neural network (RNN) is to model time-series data and sequences. Long Short-Term Memory (LSTM) networks are effective for detecting hate speech because of their ability to process and retain sequential information over long text spans. In this process, the text is transformed into numerical formats like word embeddings, allowing the LSTM to understand the relationships between words in a sentence. The network analyzes the sequence while maintaining crucial contextual information through its memory system, enabling it to recognize subtle expressions of hate that may appear in different parts of the text.

LSTMs excel at identifying complex linguistic patterns, such as sarcasm or indirect forms of hate speech, by focusing on relevant parts of the text and disregarding less important details. Through specialized gates that regulate the flow of information, LSTMs can

prioritize critical context, improving the accuracy of their predictions. This capability helps the model differentiate between hate speech and non-hate speech, even in cases where the offensive content is implied rather than directly stated.

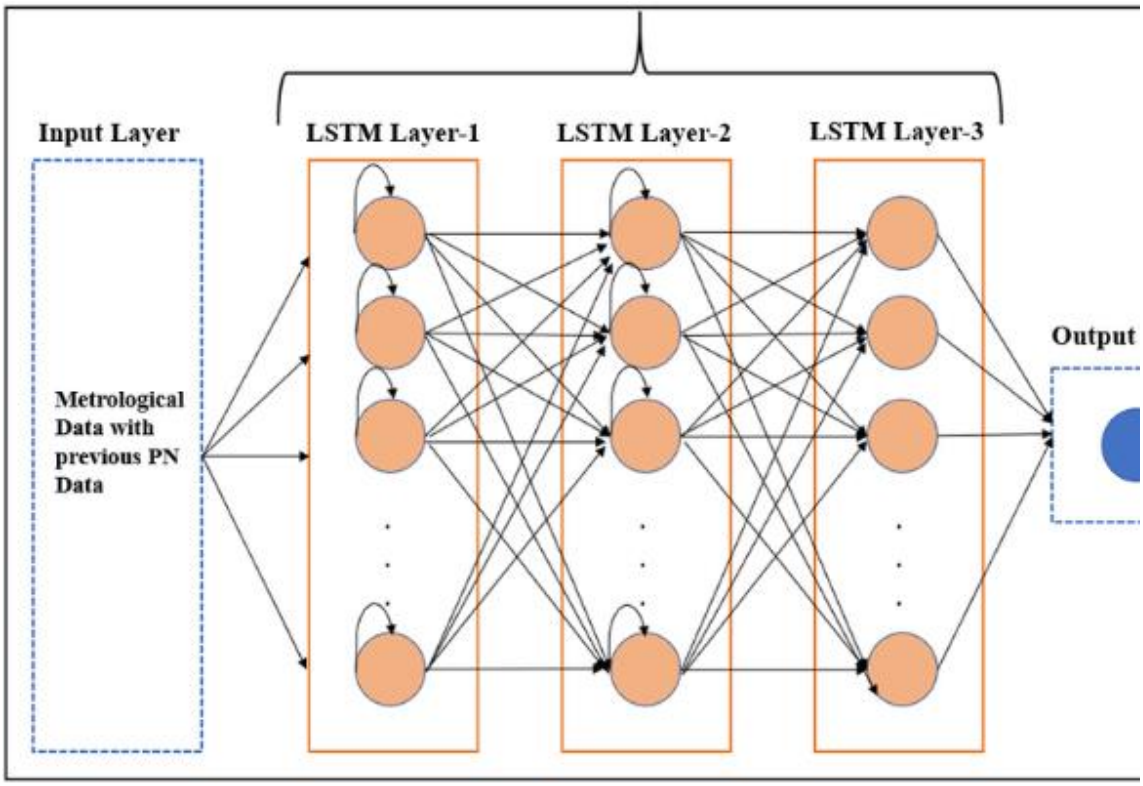


Figure 2.3 Architecture of LSTM

Bidirectional RNN/LSTM Bidirectional RNNs

Two LSTMs make up this sequence processing model; one forwards the input, and the other backwards. For problems involving natural language processing, bidirectional LSTM is a particularly popular option.

BILSTM are effective for detecting hate speech because they process text in both directions, forward and backward, enabling them to capture a more comprehensive context. By considering both previous and subsequent words, BILSTMs can identify subtle or complex instances of hate speech that may rely on surrounding context. This bidirectional

approach helps the model recognize nuanced forms of hate, such as sarcasm or indirect offensive language, where the meaning can change depending on the words before or after. As a result, BILSTMs offer improved accuracy in classifying text as hate speech or non-hate speech, even when the harmful content is implied rather than explicitly stated.

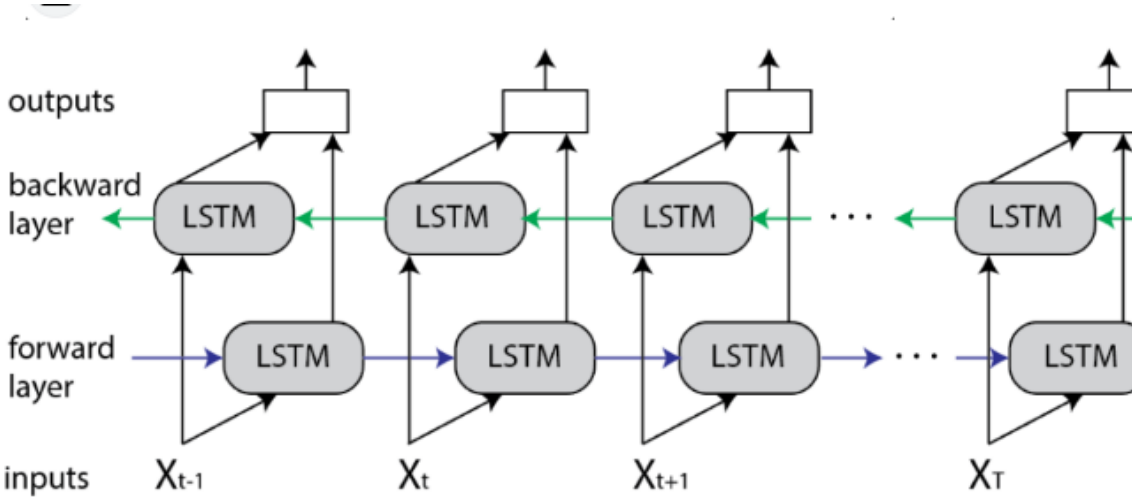


Figure 2.4 BILSTM architecture

Multilayer Perceptron (MLPs)

The fundamental idea underlying a multilayer perceptron's operation is backpropagation, a crucial network training procedure. Multilayer Perceptron (MLPs) work for hate speech detection by converting text into numerical features, typically through word embeddings, and processing these features through multiple layers of neurons. During training, the network learns to identify patterns and relationships in the data by adjusting the weights and biases. Each layer applies an activation function to help capture relevant features that indicate hate speech. Once trained, the MLP produces a classification, such as determining whether the text is hating speech or not. While MLPs might not capture sequential dependencies as well as models like LSTMs, they can still effectively classify hate speech when paired with well-constructed input features.

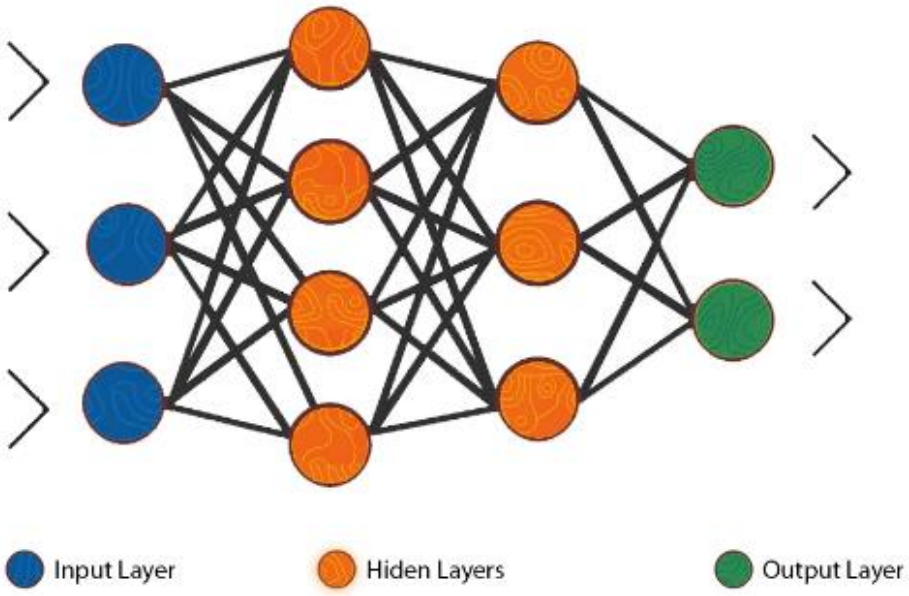


Figure 2.5 MLP architecture

2.6.3Hybrid Technique

It is a strategy for getting around an approach's drawbacks. since every solution has a unique set of constraints. Additionally, it seems like a good idea to combine two or more ways into a hybrid strategy where they complement one another.

2.7 Feature Extraction for Hate speech Detection

The act of depicting and changing unprocessed data into numerical features while keeping the original data intact is known as feature extraction.

Hate speech detection methods employ a variety of feature extraction approaches. Below is a discussion of a few of them.

Bag-of-words (BOW)

The bag of words technique breaks down each text into its constituent words without retaining any knowledge about the text's grammar or syntax. It then executes tokenization and vector construction operations.

Term frequency-inverse document frequency (TF-IDF)

The word "importance of inverse document frequency" (IDF), which appears infrequently in the corpus but might have significant information, is significant. The most effective method for handling basic machine-learning tasks is TF-IDF.

Word Embedding

Words connected to the closer vector can be represented vector-based using word embedding. The word embedding method makes it possible to identify a word's syntactic and semantic relationships. Mathematical representations of words within the corpus are created using the word embedding technique[13].

Word2vec:

A feature extraction method called Word2vec is used to extract words from documents and learn their associations, including their synonyms. Using word2vec algorithms, which scan the entire corpus, frequently occurring words are used to construct the vector. The corpus size and the location of each vector in the vector space can be used to calculate the vector size in the word2vec technique. Stochastic gradient descent is the training method. The mapping of a word within an n-gram context into continuous vectors is known as the projection layer.

The skip-gram technique is applied as an output context word and as an input word center. Skip-gram works well with tiny datasets and words that appear seldom. The center of the word is the output and context words are the input in the continuous bag of words approach. When dealing with large datasets and the most frequently occurring terms, a continuous bag of words works well[13].

GloVe:

As a feature extraction method, global vector offers a vector representation of the entire content. The global word-to-word co-occurrence matrix is utilized in the gloVe approach to represent global contextual information in the corpus[13].

FastText:

FastText is a modification of Word2Vec that improves word representations by using n-grams, which are sequences of adjacent characters, in the training process. Rather than assigning a single vector to each word, FastText breaks down words into smaller n-grams, enabling the model to capture subword information. This approach enhances the representation of rare or out-of-vocabulary words. FastText then combines the vectors of these n-grams to form a word's overall representation, resulting in more accurate and reliable word embeddings, especially for languages with complex structures or when dealing with new or misspelled words.

The Facebook Artificial Intelligence Research Lab unveiled a brand-new word embedding technology called Fast Text as a solution to this problem. A character n-gram, or bag of characters, represents each word. Fast Text, on the other hand, generates a vector for every letter n-gram[13]

2.8 Related work

Extensive research has been conducted on earlier works to determine what is lacking and to explore the gap. In[2], automatic hate speech identification was developed using 30,000 Amharic Facebook comments. Pré trained BERT were used in conjunction with n-grams for feature extraction. The hyper parameter is adjusted using the random search technique. The tagging of the dataset as hate and hate-free was done by hand. Word2vec and n-gram were employed to extract features. Models using BERT were able to obtain a 91% accuracy rate. Unoptimized hyperparameter tweaking could be viewed as a gap. A work in [4] examine the patterns and distribution of hate speech over time, and contrast the hate speech on Twitter with that of general reference. The Amharic corpus was the paper's intended audience. Five native speakers provided about 144 abusive speech keywords, which were then divided into hate and offensive speech categories.

In[14], an initial attempt was made to identify and categorize multilingual hate speech (Amharic and Afaan Oromo) using 3 approaches of Extracting of features and four distinct deep learning classifiers Data minor tools along with a Face pager are used to collect the data. Every piece of data has three distinct domain experts annotating it. With the gathered 30,000 annotated bilingual datasets in Amharic and Afaan Oromo. an accuracy of 78.05% was obtained by BILSTM. With more data, the model's performance could be enhanced.

Datasets in Amharic were prepared in [15]from YouTube, Twitter, and Facebook for the purpose of hateful writing detection model building. One way to extract features is by word embedding. One of the issues encountered during the experiment was classifying the scraped comments and deleting the dataset. The two groups that performed the best in the experiments were BILSTM and GRU (91%), followed by LSTM and MLP (90%).

The Amharic posts and comments on Facebook and Twitter were utilized as a dataset in[16]. Techniques for wdata augmentation were applied to equalize the labeled training

set. The BILSTM-based word2vec model performs somewhat better, achieving an accuracy of 88.89% for both the enhanced and original datasets. In order to create binary class datasets, hate speech published in the Tigrigna language was gathered in [17] Facebook API and Face Pager were utilized to get data. According to an experimental result, hate speech identification with 79% accuracy was obtained with a slightly better performance by the NB. A tiny dataset with few features was employed.

Using a separate dataset, researchers in[18]suggest fuzzy multi-task learning for Hate Speech recognition. Four categories of hate speech were identified through an experimental study. Fuzzy approach works better than cutting-edge probabilistic techniques like SVM and DNNs. For the Arabic Twittersphere, religious hate speech identification has been studied by[19]. Six thousand Arabic tweets were gathered using Twitter's search API2. The purpose of crowd sourcing platforms is to get annotations for training and testing datasets. Additionally, it was discovered that religious hate speech may be accurately detected with 0.79 accuracy.

In[20], hate speech detection system for foreign languages was created. Using the Bidirectional-LSTM, the researchers developed a detection model on Vietnamese social media text. The social media text classifications that the model predicts are Hate, Offensive, and Clean (not-hate). The word embedding methods that were employed were Word2Vec and FastText. On the Vietnamese Language and Speech Processing (VLSP) 2019 public standard test set, they obtained an F1 Score of 71.43% using this model.
Two classifiers were constructed in[21]. The Part-Of-Speech tagger automatically morpho-syntactically identified the hate speech corpus because the approach depends on morpho-syntactically marked texts. When a binary classification was used, the classifier produced outcomes that were similar to those of sentiment analysis tasks for Italian that were primarily studied.
A study in [22] suggested a multitasking technique according to a popular Transformer-oriented
model to identify hate speech in Spanish tweets. It has been observed that in a multitask learning environment, the quality of corpora is crucial, and that locating such materials is

not always feasible, particularly in languages with limited resources. The model's shortcomings stem from the fact that multitasking increases the computational cost of categorization by utilizing additional corpora.

A study published in [23] suggests a clever deep learning technique for automatically identifying vile comments on Twitter in the Arabic language. The NLTK library was used to process the data collection. a collection of characteristics that were taken from the dataset using a word embedding technique. The. The deep learning strategy that has been put into practice is a cross between an LSTM network and a convolutional neural network (CNN). The suggested method performed well in categorizing tweets as Normal or Hateful.

The other related work by[24],for sarcasm detection ,separated the features in their investigation into features that were connected to syntax, sentiment, punctuation, and patterns. In machine learning classification models, this thorough feature selection worked well. Models based on deep learning offer automated feature extraction. They could be able to recognize context and pick up on minute semantic patterns. They are also quite configurable and demand a lot of data. Even just figuring out which model to use proved to be a challenging process.

Table2.2: Summary of related work

Ref	Feature Extraction Method	Algorithms Used	Selected Model	Gap
[2]	Word n-grams, Word2Vec	LSTM, GRU	RNN (97.9% & 88% accuracy)	Polarity of sentiment was not considered
[25]	n-gram	Naïve Bayes	Naïve Bayes	Limited training dataset
[14]	CNN, BiLSTM, CNN-BiLSTM, BiGRU	Keras word embedding, Word2Vec, FastText	BiLSTM with FastText	High computational cost
[26]	SVM, MNB, LR, DT, RF	TF-IDF, n-gram blend	Linear SVC (highest performance)	Special characters not considered
[17]	SVM, NB, RF	N-gram, TF-IDF	NB with TF-IDF	Lack of detailed annotation guidelines
[27]	Word2Vec, TF-IDF	NB, RF	Naïve Bayes	Limited dataset (6,120)
[19]	Pre-trained word embedding	LR, SVM, RNN-GRU	RNN-GRU	Focused only on religious text speech

2.9 Research gap and Summary

Literature on the morphology of the Amharic language, sentiment analysis and hate speech detection mechanisms were reviewed. The reviews show that the majority of the study was conducted on languages with more resources; in contrast, the Amharic social media hate speech domain has received very less research.

Hate speech has a significant negative impact on the targeted person's right to freedom and equality. Although this provides sufficient incentive to combat it, history has shown that if action is not taken, there will be long-term effects. Hate speech widens the gulf between social groupings, perhaps causing severe disruptions to social cohesiveness. It would be ineffective to use a straightforward blacklist technique, hence automated hate speech detection models must be created.

The majority of earlier research used regular machine learning models, whereas this study uses deep learning; previous works have poorer accuracy because of limited datasets and a lack of sentiment consideration. Sentiment-aware models provide better context comprehension than datasets that are only tagged with hate speech. Through the collection of new datasets from Facebook, Tik Tok, and YouTube, this work prepares a large labeled dataset and attempts to assess how sentiment analysis affects the classifier's functionality.

CHAPTER THREE

DESIGN AND METHODOLOGY

3.1 Overview

This chapter goes into great detail on the overall design and methodology of the proposed hate speech detection model for Amharic written comment on social media. The model architecture, data preparation and collecting process, annotation approaches, and guidelines will also be presented. The proposed model is built using newly collected datasets from social media, which are the primary source of data for this study.

The final portion covered word vectorization and categorization using various algorithm to build models that predict whether they are hate speech or hate free. Furthermore, the system's implementation tools, the algorithms to be used, and assessment methods are all provided in this chapter.

3.2 The proposed system architecture

The system builds hate speech detection models for Amharic languages. The model is made up of various parts. The Amharic social media dataset is the first component. It is compiled from various social media page posts, comments, and replies. From the gathered Amharic corpus's relevant data set has then been filtered.

The second component is annotation, in which a text is classified by human annotators into hate and non-hate classes using a newly developed annotation guideline. The hate and hate free class will further be annotated as having positive, negative and neutral sentiment by existing sentiment guideline of Amharic sentiment analysis work. Followed by a preprocessing step that clean and format the input dataset so that it will be ready for the subsequent components. During this stage, unnecessary character removal, tokenization,

normalization, and other fundamental preprocessing are completed. Then, feature extraction that performs vectorization, will be put into practice. And GRU and CNN will be used in the model's construction. The detection and classification component are in charge of classifying test data into the appropriate categories as a hate and free. The last task will be evaluating the model using several evaluation matrices to determine the model's performance level.

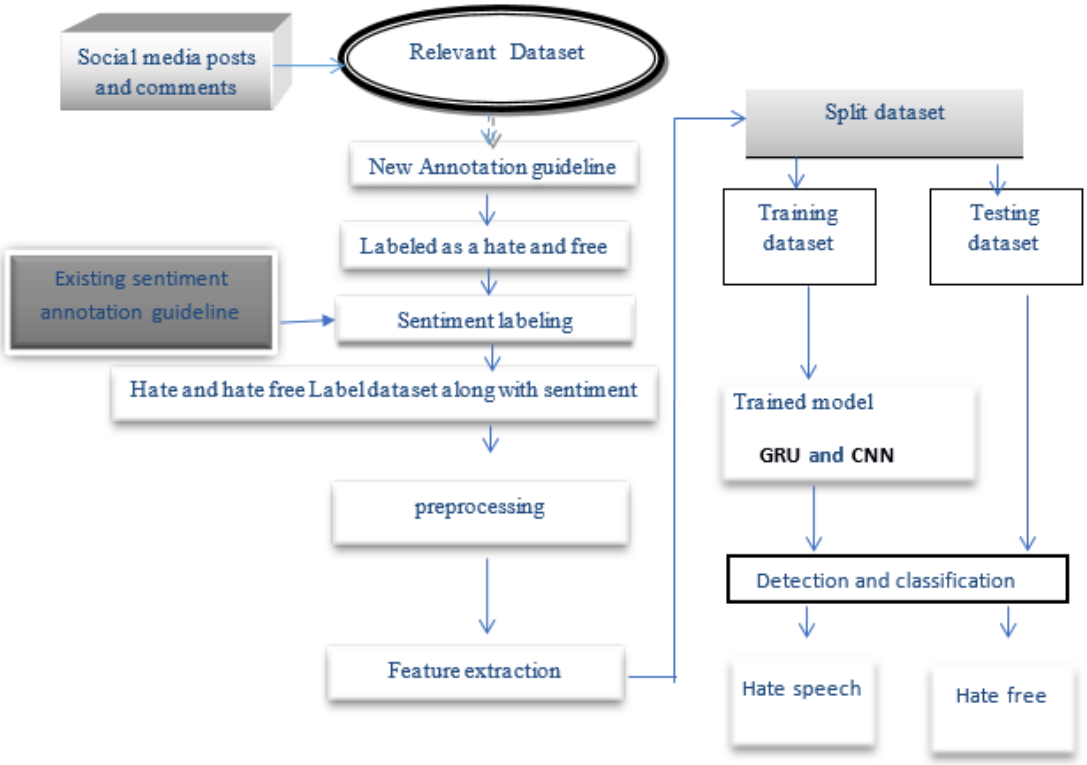


Figure 3.1: The proposed architecture for detecting hate speech for Amharic language

3.3Data collection and preparation

A new dataset is collected for this study. The following steps are completed in order to meet the dataset's construction purpose. Based on their popularity and the fact that these

pages usually post discussions covering a wide range of topics, social media pages of different influentials including journalists, politicians, local celebrities' pages, and channels that are most frequently watched were selected.

The chosen platforms are TikTok, Facebook and you tube. Comment extractor extensions, you tube API, and different python codes were utilized to gather the data. Since this comment collector's extension limit the amount of comment to be collected python codes along with the social media API were mainly implement to collect dataset. A total of 160,000 data set is collected.

The collected dataset is in the form of excel. And comments that are not written in Amharic words were eliminated, along with emoji, images, blank spaces, null and other language sentences, as only Amharic-language data will be considered in this instance.

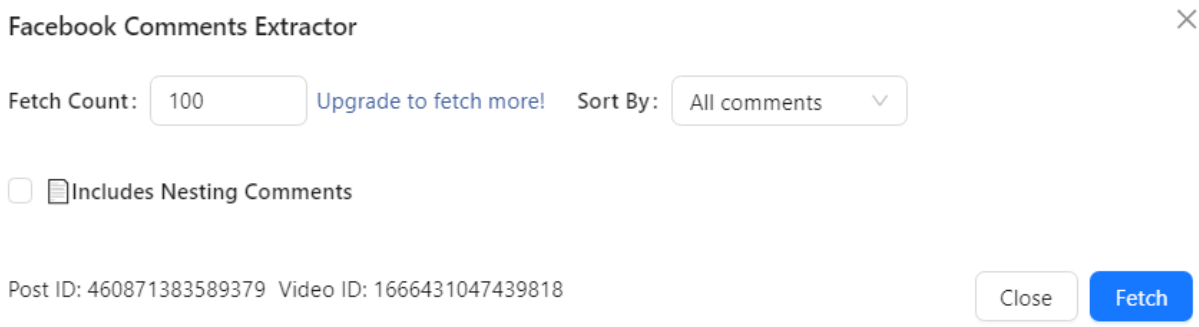


Figure 3.2: Face book comment extractor

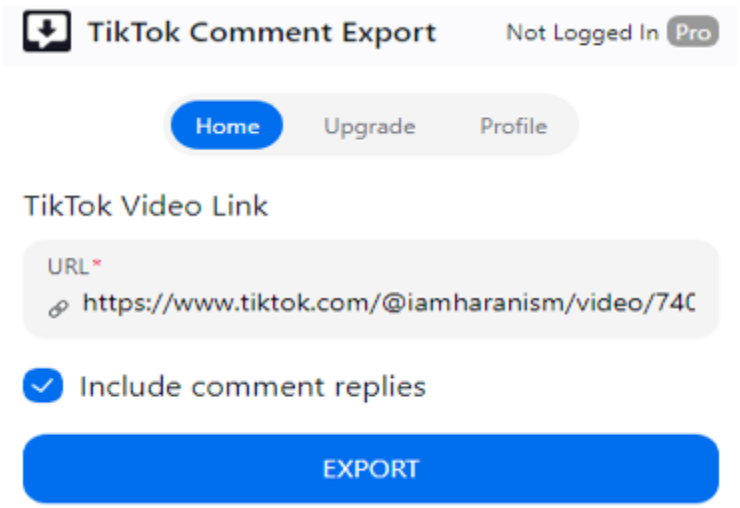


Figure 3.3: TikTok comment extractor

```
Run  Debug  myFunction  Execution log

function scrapeCommentsWithoutReplies(){
  var ss = SpreadsheetApp.getActiveSpreadsheet();
  var result=[['Name','Comment','Time','Likes','Reply Count']];
  var vid = ss.getSheets()[0].getRange(1,1).getValue();
  var nextPageToken=undefined;

  while(1){
    var data = YouTube.CommentThreads.list('snippet', {videoId: vid, maxResults: 100, pageToken: nextPageToken})
    nextPageToken=data.nextPageToken
    //console.log(nextPageToken);
    for (var row=0; row<data.items.length; row++) {
      result.push([data.items[row].snippet.topLevelComment.snippet.authorDisplayName,
                   data.items[row].snippet.topLevelComment.snippet.textDisplay,
```

Figure 3.4: screenshot of python code to collect dataset from you tube

The code in the above figure was modified from GitHub, which extracted the data in Excel format using the YouTube API and post ID.

The following table shows a list of some of the chosen social media pages along with the amount of filtered dataset collected from them.

Table 3.1: Some of the selected social media pages

NO	Social media page	Number of filtered comments
1	Gursha tube	2015
2	Getu temesgen	1100
3	Zehabesha	2000
4	Yonimagna _scabogna01	8900
5	Motta keraniyo	9700
6	ZemedkunBekele	8500
7	Liya show	7500
8	Iamharanism	7800
9	Amhara media corporation	2100
10	ESATtv Ethiopia	1100
11	FDRE Ministry of justice	1000
12	ESAT	1100
13	Dire Tube	2200
14	ebc_tiktok	1150
15	Amharafanofirst	2150
16	Tamagn.beyene	800

some unofficial accounts that went viral and had a lot of comments and views were the source of the rest collected datasets.

3.4Annotation guidelines

Following data cleansing, 79,991 datasets remained. Emojis, empty comments, Amharic comments written in English letters, sentences in non-Amharic, and comments with just special characters are all eliminated as part of the data cleaning process.

Table 3.2:sample social media comments before preprocessing

Name	Comment	Time	Likes	Reply Count
@Add****	□□□□□□□□□□□□	2024-09-16T06:51:00Z	0	1
@eli***	□□□□□□□□□□□□□□□□□□□□	2024-09-15T19:12:35Z	0	0
@□□***	□□□□	2024-09-15T19:28:41Z	2	0
@tase***	□□□□□□□□□□	2024-09-15T17:39:26Z	1	0
@Redi***	□□□□□□□□□□□□□□□□□□	2024-09-15T17:24:20Z	7	0
@av***	Left political wing of Prosperity party	2024-09-15T18:52:11Z	0	0
@An****	🎉🎉🎉	2024-09-15T18:23:50Z	1	0
@OneL***	□□□□□□□□□□□□□□□□□□□□	2024-09-15T17:35:07Z	1	1

The next step is to classify this 79,991 Amharic dataset as hate or non-hate after gathering the relevant data set. 30 Human annotators annotated manually the dataset as hate and hate free based on a newly developed annotation guideline that is presented in the Annex A.

The annotators are staff members of Ethio Telecom's Digital Customer Care Division, with experience in assigning sentiment to customer comments on the company's posts as part of their job. They are also proficient in the Amharic language.

The annotators were given a brief description of the annotation guideline in order to classify posts and comments into the binary classifier. Before starting the annotation, the relevant rules and standards were established to make sure the annotator understood the assignment well.

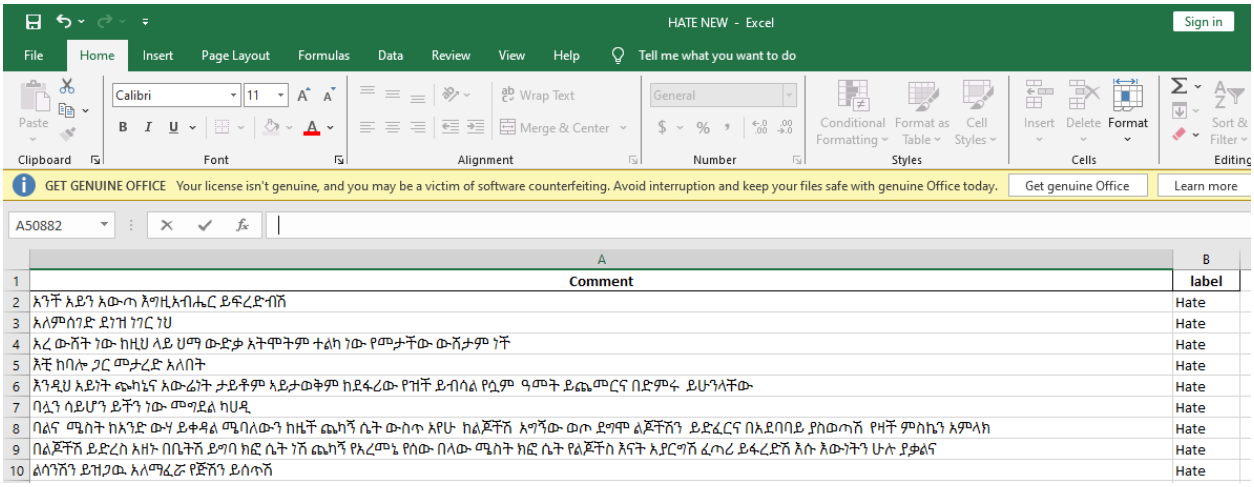


Figure 3.5: screenshot of some of the labeled dataset

Since this work consider to include the sentiment of each dataset, an existing Amharic sentiment analysis guideline were adopted from the work of [3] ,and the same annotators label each dataset as having positive, negative and neutral sentiment.

	A	B	C	D
1	Comment	label	sentiment	
2	አንች አይን አውጣ እግዚአብሔር ይፍረድብሽ	Hate	negative	
3	አለምሰገድ ደነዝ ነገር ነህ	Hate	negative	
4	አረ ውሸት ነው ከዚህ ላይ ህማ ውድቃ አትሞትም ተልካ ነው የመታችው ውሸታም ነች	Hate	negative	
5	እቺ ከባሎ ጋር መታረድ አለበት	Hate	negative	
6	እንዲህ አይነት ጭካኔና አውሬነት ታይቶም አይታወቅም ከደፋረው የዝች ይብሳል የሷም ዓመት ይጨመርና በድምሩ ይሁንላቸው	Hate	negative	
7	ባሏን ሳይሆን ይችን ነው መግደል ካሀዲ	Hate	negative	
8	ባልና ሚስት ከአንድ ውሃ ይቀዳል ሚባለውን ከዚች ጨካኝ ሴት ውስጥ አየሁ ከልጆቻሽ አግኝው ወጦ ደማሞ ልጆቻሽን ይድፈርና በአደባባይ ያስወጣሽ የዛች ምስኪን አምላክ	Hate	negative	
9	በልጆችሽ ይድረስ አዘኑ በቤትሽ ይግባ ክፎ ሴት ነሽ ጨካኝ የአረመኔ የሰው በላው ሚስት ክፎ ሴት የልጆችስ እናት አያርግሽ ፈጣሪ ይፋረድሽ እሱ እውነትን ሁሉ ያቃልና	Hate	negative	
10	ልሳንሽን ይዝጋዉ አለማፈሯ የጅሽን ይሰጥሽ	Hate	negative	

Figure 3.6: screenshot of some of the labeled dataset along with their sentiment

3.5 Preprocessing

Preparing the input texts into appropriate format for additional analysis is known as preprocessing. The preparation steps involved removing all non-Amharic characters from the provided comments and posts, including alphabets, punctuations, white space, emoji and unnecessary columns. The dataset cleaning, normalization, dataset balancing, tokenization, stop-word removal, and morphological analysis are the preprocessing phases in the architecture.

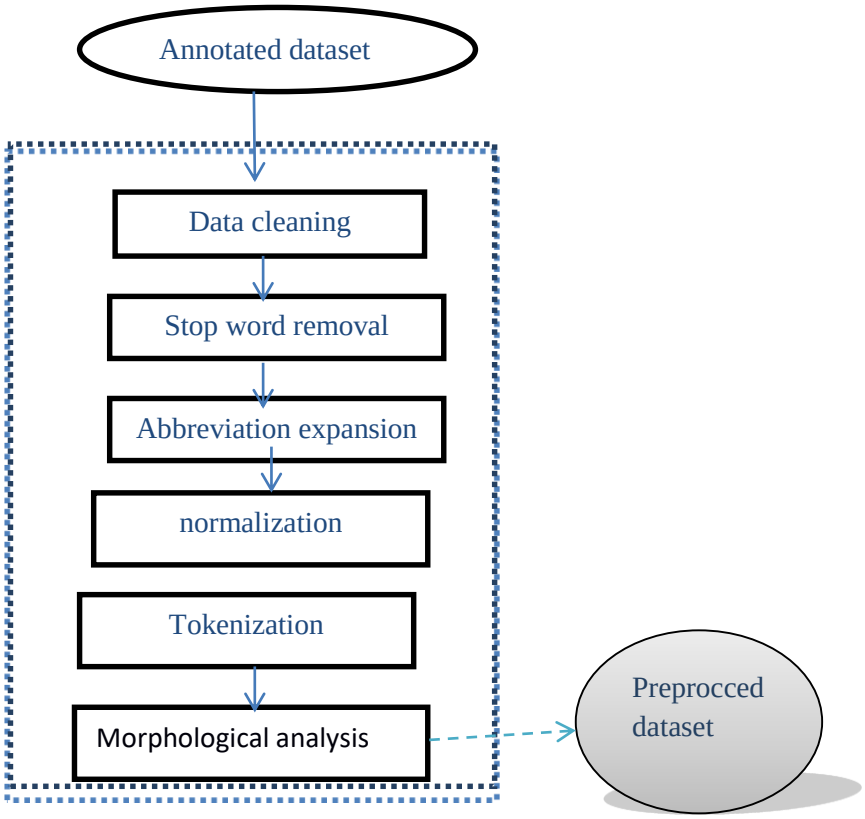


Figure 3.7: preprocessing Component

3.5.1 Data Cleaning

Only Amharic texts are included in the dataset; texts in any other language have been eliminated, and text has been cleaned according to the property of the language. Non-standard terms like null values, empty spaces, numbers, non-word characters, HTML, URLs, numbers, emojis, and comments that are written in non-Amharic words were eliminated.

```
def removing_url(string_list):
    result_list = []
    for string in string_list:
        if isinstance(string, str):
            string = re.sub(r'http\S+', '', string)
            string = re.sub('https?://\S+|www\.\S+','',string)
            string= re.sub("@[A-Za-z0-9_]+","", string)
            string=re.sub(r"(?:\@|https?\://)\S+", "", string)
            string = re.sub('([\w\.-]+)@([\w\.-]+)','',string)
            result_list.append(string)
    return result_list
```

Figure 3.8: python code for Data set cleaning

3.5.2 Stop word Removal

It is one of the preprocessing techniques that identifies the words that appear in the corpus most or least frequently. This step allows a more focus to be given to relevant content in the dataset, by eliminating these words. Stop word lists were developed for this study from the gathered corpus.

```python
import re
path = '/content/stopwords-am.txt'
words_to_remove = set()
with open(path, 'r', encoding='utf-8') as file:
    words_to_remove = {word.strip() for word in file.readlines()}

def remove_words_from_sentence(sentence):
    tokens = re.findall(r'\w+|[^\w\s]', sentence, flags=re.UNICODE)
    filtered_tokens = [token for token in tokens if token not in words_to_remove]
    cleaned_sentence = ' '.join(filtered_tokens)
    return cleaned_sentence
```

Figure 3.9: python code for stop word removal

3.5.3Abbreviation Expansion

The user may utilize the abbreviation word in Amharic. These abbreviations' ways of writing style must be expanded in the dataset such as □/□□ (fe/bet) □□□ □□ (court).

3.5.4 Data normalization

The process of normalization reduces duplication in several languages. It shows the consistency of characters. These Fidels are □and□, □and□, □and□, □, □, and □. Character inconsistencies like these could lead to an unneeded rise in the number of document representative words, which would require processing of enormous amounts of data. Consequently, certain kinds of characters were made into a single character in order to standardize them.

```python
import re

def normalization(input):
    rep1=re.sub('[ሃኅኃሐኻ]','ሀ',input)
    rep2=re.sub('[ሑኁዅ]','ሁ',rep1)
    rep3=re.sub('[ኂሒኺ]','ሂ',rep2)
    rep4=re.sub('[ኌሔኼ]','ሄ',rep3)
    rep5=re.sub('[ሕኅ]','ህ',rep4)
    rep6=re.sub('[ኆሖኾ]','ሆ',rep5)
    rep7=re.sub('[ሠ]','ሰ',rep6)
    rep8=re.sub('[ሡ]','ሱ',rep7)
    rep9=re.sub('[ሢ]','ሲ',rep8)
    rep10=re.sub('[ሣ]','ሳ',rep9)
    rep11=re.sub('[ሤ]','ሴ',rep10)
    rep12=re.sub('[ሥ]','ስ',rep11)
```

Figure 3.10: python code for Normalization

3.5.5 Tokenization

Using punctuation or space to separate the collected text into individual words or tokens is a technique known as tokenization. The method determines a written text's limits.

```
[ ]  def tokenize(sentence):
       return sentence.split()
```

```
[ ]  df['Comment'] = df_new['Comment'].apply(tokenize)
```

Figure 3.11: python code for Tokenization

3.5.6 Sequence Padding

The goal of this preprocessing step is to equalize the input length so that the embedding matrix may be created. This will be done with Keras function pad sequences ().

3.6 Morphological Analysis

For morphologically rich languages like Amharic, morphological analysis is required. In order to extract the necessary characteristics and root or stem word to construct a classifier, and HornMorpho[30] a tool were adapted.

3.7 Sentiment Analysis

The 30 annotators first annotate the dataset as a hate and hate free based on the newly developed guideline as described in Annex A. Since considering the sentiment analysis of each dataset is part of the model building ,existing Amharic sentiment annotation guidelines[3] were implemented to assign each dataset sentiments

In the paper[3] the authors used the following criteria when analyzing the sentiment of social media texts:

Sentiment Categories:

Each social media post's sentiment was divided into three primary classes:

- Positive: Posts that convey a feeling of approval or favorability.
- Negative: Posts that express criticism, discontent, or a negative attitude.
- Posts that are vague in their sentiment or lack a strong opinion or sentiment are considered neutral.

The annotators were directed to give considerable thought to the social media messages' context.

The analysis included distinguishing between sentiment, even though the main focus was on sentiment as a whole. The sentiment analysis was intended to identify the general positive or negative tone, independent of the particular emotions stated.
The annotators were also instructed to take into consideration linguistic subtleties that could influence sentiment classification, given the nature of social media language. This included slang, informal language, acronyms, and other unconventional Amharic usages that may make it difficult to distinguish between different feeling groups. Posts containing complicated or conflicting emotions should carefully examine the tone of the whole.

3.8 Feature Extraction

In this step, a subset of relevant features from the labeled dataset that were helpful in recognizing hate and hate-free content are chosen so they can be employed in the model building. At this point, the dataset is converted into a numerical vector. The experiment will utilize pre trained embedding feature extraction methods.

This step constructs a word embedding matrix using pre-trained GloVe embeddings [44] to map words in a dataset's vocabulary to their corresponding vector representations. It begins by loading the GloVe file and storing each word and its associated embedding vector in a dictionary. Then, an embedding matrix is initialized with zeros.

By iterating over the vocabulary produced by a tokenizer, the code checks if each word has a pre-trained embedding in the dictionary. If found, the embedding vector is ascribed to the respective row regarding on index of words. This embedding matrix aligns pre-trained GloVe vectors with the dataset's vocabulary, enabling the model to incorporate rich semantic information for tasks like text analysis and classification.

The selected deep learning algorithms

GRU and **CNN** are the selected classifier algorithms, to be utilized to categorize the gathered comments as hate and hate free. Several related works suggest to obtain good performance utilizing them, therefore testing with more than one classification algorithms gave comparison indications for identifying top performing algorithms for the proposed model.

3.9 Classification

A sum of 79,991relevant dataset is filtered from collected160k dataset from social media Amharic textual comments and posts, with the help of comment extractor extension, python code, API and scrappers. And this relevant dataset is annotated manually with 30 annotators, that are expert on the language based on the newly developed annotation guideline, that is described in Annex A.

Beside the hate and hate free class, the annotators also manually annotate the sentiment of each dataset as a positive, negative and neutral based on already existed a sentiment analysis guideline of the work in [3].Then the annotated dataset along with their sentiments

will be preprocessed and the feature extracted data and their polarity labels will be fed into the deep learning model for training. Eighty percent of the corpus was utilized to train the model, and twenty percent will be used to test its performance. Using quality metrices, the model execution ability in categorizing the test sets will be assessed in the final stage. (RQ1)

3.10 Development Tools and Techniques

Comment extractor extensions, Python codes along with social media API and scrape storm software: were utilized to retrieve comments and posts from the social media pages.

Microsoft Excel

used to do data preparation operations such as cleaning, filtering, sorting, and deleting redundant data from the collected data. utilized to oversee the annotating task as well.

Google Collab: since the dataset is large google Collab will be Used to execute python code from Google Drive.

Python programming language: The data was preprocessed and the model will be developed using the Python programming language

The Python programming language is an interpreted, dynamically typed, object-oriented language that is ideal for natural language processing (NLP) due to its ease of use, debugging capabilities (exceptions and interpreted language), ease of structuring (modules and object-oriented language), and robust string manipulation capabilities.

Among the most widely utilized libraries for the research are the ones listed below.

•**NumPy: provides** support for matrices, huge arrays, and high-level mathematical functions. NumPy is particularly helpful when utilizing its random number and linear algebra features.

Scikit-learn NumPy and SciPy, two math packages, serve as its foundation. It has many more features, like different clustering, regression, and classification methods.

NLTK: It's a Python package that comes in handy when handling data in human languages. The collection includes over fifty distinct lexical resources and datasets, including WordNet, parsing, and tagging. Assignments of different complexity and size can be

supported by NLTK. Additionally, by combining and expanding existing NLTK components with completely new ones, students can use NLTK to create a flexible framework for more complex projects like creating a multicomponent system.

3.11 Performance Evaluation

To achieve that, this study will make use of accuracy, recall, F-measure, and precision to assess the model.

Precision

When the precession result is high, the system is substantially more likely to return the right value than the wrong one. Mathematically

$$Precision\ (P)\ =\ \ TP/(TP\ +\ FP)(1)$$

Recall: is about how many related instances are projected out of all the relevant instances; a high recall value suggests the model forecasts the most relevant outcome expressed as

$$Recall\ (R)\ =\ \ \ \ TP/\ (TP\ +\ FN)\ \ \ \ \ (2)$$

F-measure is an accuracy metric calculated as a balanced average of a test's precision and recall.

$$\text{F\_Measure (F)}\ =\ 2\ *\ (\text{P}\ *\ \text{R})/\ (\text{P}\ +\ \text{R})\ (3)$$

Accuracy

$$Classification\ Accuracy\ =\ (TP + TN)\ /\ (TP\ +\ FN\ +\ FP\ +\ TN\ \ \ (4)$$

True Positive: These are the posts and comments that the machine learning algorithm has identified as hateful and classed as such.

False positive: These are postings and comments that are labeled as hateful even if machine learning predicts them to be non-hateful.

True Negative: These are the posts or comments that the machine learning system classifies as not hateful and predicts as such.

False Negative: These are postings and comments that are labeled as hateful but that the machine learning algorithm predicts are not hateful.

Confusion Matrix

The algorithm's effectiveness is measured using a confusion matrix. The predicted values for the proper and incorrect classes are displayed in the confusion matrix.

CHAPTER FOUR

EXPERIMENTATION AND DISCUSSION

4.1 Overview

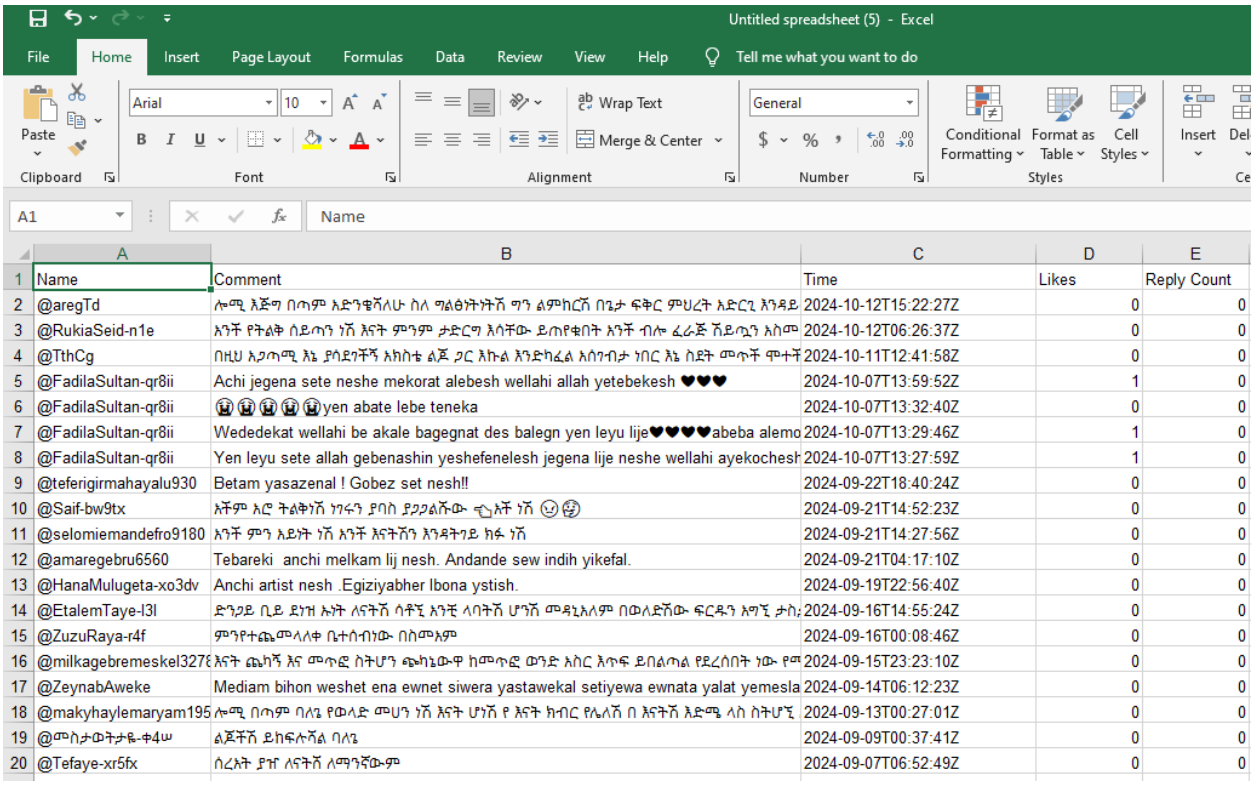
The dataset to be used, the experiment's execution, and its outcome are all covered in this chapter. The model evaluation will come after a detailed presentation of the preprocessing stages, feature extractions and the construction of the Amharic hate speech detection model. This study will also examine the impact of sentiment analysis on hate speech detection.

4.2 Data Collection

Amharic sentences were collected from comments of social media platform mainly from TikTok, YouTube and Facebook posts. Facebook comment extractor extensions and python scripts were employed to extract Facebook, YouTube and TikTok comments. A total of 160k dataset were retrieved from public pages, political activist pages, famous artists private pages and most followed YouTube channels. Eyoha media, Dallot entertainment, Egreghaw media, yonimagha_scabogha01, Motta keraniyo ,Zemedkune Bekele, Gursha ,Zehabesha and Getu temesegen were some of the social media pages used for the dataset source.

4.3 Building the Corpus

Numerous forms of information are included in the gathered dataset, such as the commenter's name, the comment, time, and number of likes and replies. Since the comments is the only dataset to be used in building the model, the rest are manually removed. After performing data cleaning on the crawled data to remove unnecessary information a total of 79,991datasets were gathered.



Untitled spreadsheet (5) - Excel

File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do

Paste Clipboard Arial 10 Font Wrap Text Merge & Center Alignment General Number Conditional Formatting Format as Table Cell Styles Styles Insert

A1 Name

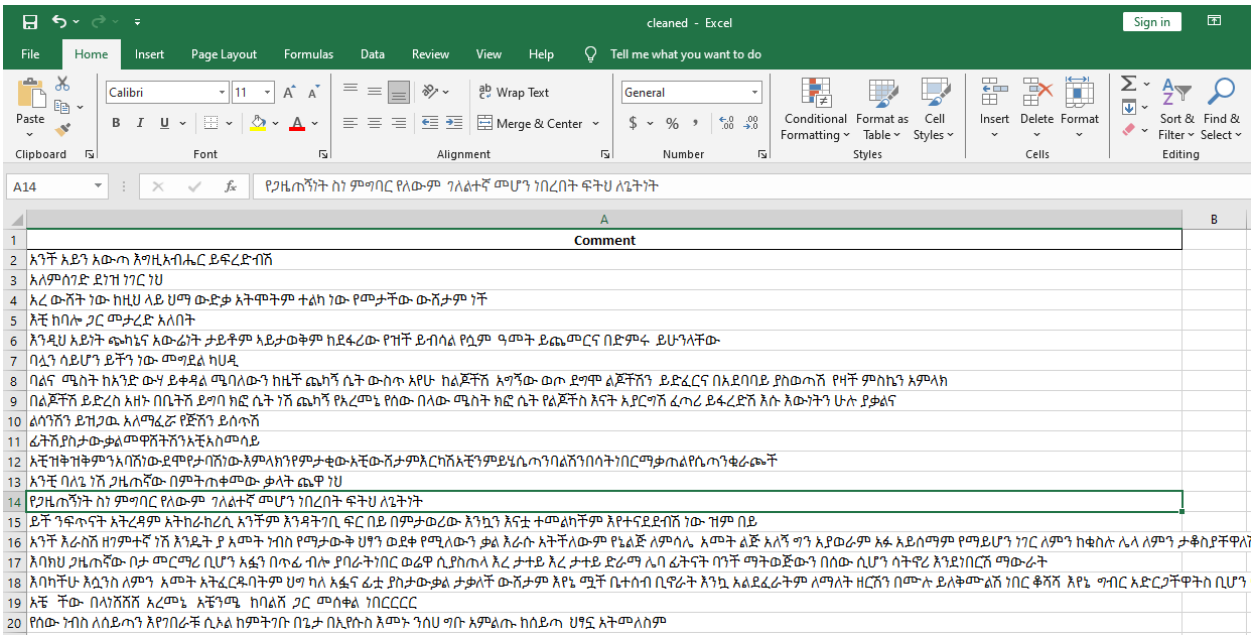
	A	B	C	D	E
1	Name	Comment	Time	Likes	Reply Count
2	@aregTd	ሎሚ እጅግ በጣም አድጎቄሻለሁ ስለ ግልፅነትነሽ ግን ልምከርሽ በጌታ ፍቅር ምሀረት አድርጊ እንዳይ	2024-10-12T15:22:27Z	0	0
3	@RukiaSeid-n1e	አንች የትልቅ ሰይጣን ነሽ እናት ምንም ታድርግ እሳቸው ይጠየቁበት አንች ብሎ ፈራጅ ሽይጪን አስመ	2024-10-12T06:26:37Z	0	0
4	@TthCg	በዚህ አጋጣሚ እኔ ያሳደገችኝ አክስቴ ልጅ ጋር እኩል እንድካፈል አሰባብታ ነበር እኔ ስደት መጥቶ ሞተች	2024-10-11T12:41:58Z	0	0
5	@FadilaSultan-qr8ii	Achi jegena sete neshe mekorat alebesh wellahi allah yetebekesh ❤❤❤	2024-10-07T13:59:52Z	1	0
6	@FadilaSultan-qr8ii	😭😭😭😭😭yen abate lebe teneka	2024-10-07T13:32:40Z	0	0
7	@FadilaSultan-qr8ii	Wededekat wellahi be akale bagegnat des balegn yen leyu lije❤❤❤❤abeba alemo	2024-10-07T13:29:46Z	1	0
8	@FadilaSultan-qr8ii	Yen leyu sete allah gebenashin yeshefenelesh jegena lije neshe wellahi ayekochesh	2024-10-07T13:27:59Z	1	0
9	@teferigirmahayalu930	Betam yasazenal ! Gobez set nesh!!	2024-09-22T18:40:24Z	0	0
10	@Saif-bw9tx	አችም አሮ ትልቅነሽ ነገሩን ያባስ ያጋጋልሹው 👈አች ነሽ 😠🤮	2024-09-21T14:52:23Z	0	0
11	@selomiemandefro9180	አንች ምን አይነት ነሽ አንች እናትሽን እንዳትገይ ክፉ ነሽ	2024-09-21T14:27:56Z	0	0
12	@amaregebru6560	Tebareki anchi melkam lij nesh. Andande sew indih yikefal.	2024-09-21T04:17:10Z	0	0
13	@HanaMulugeta-xo3dv	Anchi artist nesh .Egiziyabher Ibona ystish.	2024-09-19T22:56:40Z	0	0
14	@EtalemTaye-l3l	ድንጋይ ቢይ ደነዝ ኡነት ለናትሽ ሳቶኚ አንቺ ላባትሽ ሆንሽ መዳኒአለም በወለድሽው ፍርዱን አማኚ ታስ	2024-09-16T14:55:24Z	0	0
15	@ZuzuRaya-r4f	ምንየተጨመላለቀ ቤተሰብነው በስመአም	2024-09-16T00:08:46Z	0	0
16	@milkagebremeskel3278	እናት ጨካኝ እና መጥፎ ስትሆን ጭካኔውዋ ከመጥፎ ውንድ አስር እጥፍ ይበልጣል የደረሰበት ነው የማ	2024-09-15T23:23:10Z	0	0
17	@ZeynabAweke	Mediam bihon weshet ena ewnet siwera yastawekal setiyewa ewnata yalat yemesla	2024-09-14T06:12:23Z	0	0
18	@makyhaylemaryam195	ሎሚ በጣም ባለጌ የወላድ መሀን ነሽ እናት ሆነሽ የ እናት ክብር የሌለሽ በ እናትሽ እድሜ ላስ ስትሆኚ	2024-09-13T00:27:01Z	0	0
19	@መስታውትታቴ-ቀ4ሠ	ልጆችሽ ይከፍሉሻል ባለጌ	2024-09-09T00:37:41Z	0	0
20	@Tefaye-xr5fx	ሰረአት ያዝ ለናትሽ ለማንሺውም	2024-09-07T06:52:49Z	0	0

Figure 4.1 sample dataset before preprocessing

4.4 Execution of preprocessing

4.4.1 Eliminating extraneous symbols and punctuation

Python programming using NLTK modules and regular expression (Regex) was utilized to apply preprocessing to the collected data. Prior to saving the preprocessed data in the Excel format, the Python program read the Excel file and remove unnecessary information such as symbols, emojis, Amharic comments written in English letters, URLs, HTML, numbers and other language scripts



cleaned - Excel

	Comment
2	አንች አይን አውጣ እግዚአብሔር ይፍረድብሽ
3	አለምሰገድ ደነዝ ነገር ነህ
4	አረ ውሸት ነው ከዚህ ላይ ሀማ ውድቃ አትሞትም ተልካ ነው የመታችው ውሸታም ነች
5	እቺ ከባሎ ጋር መታረድ አለበት
6	እንዲህ አይነት ጭካኔና አውሬነት ታይቶም አይታወቅም ከደፋሪው የዝቸ ይብሳል የሷም ዓመት ይጨመርና በድምሩ ይሁንላቸው
7	ባሏን ሳይሆን ይቸን ነው መግደል ካሀዲ
8	ባልና ሚስት ከአንድ ውሃ ይቀዳል ሚባለውን ከዚች ጨካኝ ሴት ውስጥ አየሁ ከልጆቸሽ አማኝው ወጥ ደግሞ ልጆቸሽን ይድፈርና በአደባባይ ያስወጣሽ የዛች ምስኪን አምላክ
9	በልጆቸሽ ይድረስ አዘኑ በቤትሽ ይግባ ክፎ ሴት ነሽ ጨካኝ የእረመኔ የሰው በላው ሚስት ክፎ ሴት የልጆቸስ እናት አያርግሽ ፈጣሪ ይፋረድሽ እሱ እውነትን ሁሉ ያቃልና
10	ልሳንሽን ይዝጋዉ አለማፈሯ የጅሽን ይሰጥሽ
11	ፊትሽያስታውቃልመዋሸትሽንአቺአስመሳይ
12	አቺዝቅዝቅምንአባሽነውደሞየታባሽነውእምላክንየምታቂውአቺውሸታምእርካሽአቺንምይሄሴጣንባልሽንበሳትነበርማቃጠልየሴጣንቁራጮች
13	አንቺ ባለጌ ነሽ ጋዜጠኛው በምትጠቀመው ቃላት ጨዋ ነህ
14	የጋዜጠኝነት ስነ ምግባር የለውም ገለልተኛ መሆን ነበረበት ፍትህ ለጌትነት
15	ይች ንፍጥናት አትረዳም አትከራከረሊ አንችም እንዳትገቢ ፍር በይ በምታወሪው እንኳን እናቷ ተመልከቸም እየተናደደብሽ ነው ዝም በይ
16	አንች እራስሽ ዘገምተኛ ነሽ እንዴት ያ እመት ነብስ የማታውቅ ህፃን ወደቀ የሚለውን ቃል እራሱ አትቸለውም የኔልጅ ለምሳሌ አመት ልጅ አለኝ ግን አያወራም አፉ አይሰማም የማይሆን ነገር ለምን ከቁስሉ ሌላ ለምን ታቆስያቸዋለ
17	እባክህ ጋዜጠኛው ቦታ መርማሪ ቢሆን አፏን በጥፊ ብሎ ያባራትነበር ወሬዋ ሲያስጠላ እረ ታተይ እረ ታተይ ድራማ ሌባ ፊትናት በጎች ማትወጅውን በሰው ሲሆን ሳትኖሪ እንደነበርሽ ማውራት
18	እባክቸሁ እሷንስ ለምን አመት አትፈርዱባትም ህግ ካለ አፏና ፊቷ ያስታውቃል ታቃለች ውሸታም እየኔ ሚቸ ቤተሰብ ቢኖራት እንኳ አልደፈራትም ለማለት ዘርሽን በሙሉ ይለቅሙልሽ ነበር ቆሻሻ እየኔ ማብር አድርጋቸዋትስ ቢሆን
19	አቼ ቸው በላነሽሽሽ አረመኔ አቼንሜ ከባልሽ ጋር መሰቀል ነበርርርር
20	የሰው ነብስ ለሰይጣን እየገበራቹ ሲኦል ከምትገቡ በጌታ በኢየሱስ እመኑ ንሰሀ ግቡ አምልጡ ከሰይጣ ህፃኗ አትመለስም

Figure 4.2 screenshot of the dataset after data cleaning stage

4.4.2 Normalization

Normalization is a technique that eliminates redundancy across different languages by standardizing text formats. It ensures consistency in spelling, punctuation, and structure, enhancing linguistic analysis and improving the effectiveness of natural language processing (NLP) applications.

```
df['Comment'] = df['Comment'].apply(lambda x: normalization(x))
```

Figure 4.3 screenshot of implementation of normalization python program

4.4.3 Implementation of Tokenization

Following the cleaning and normalization processes, the tokenization approach separates the sentences into discrete words.

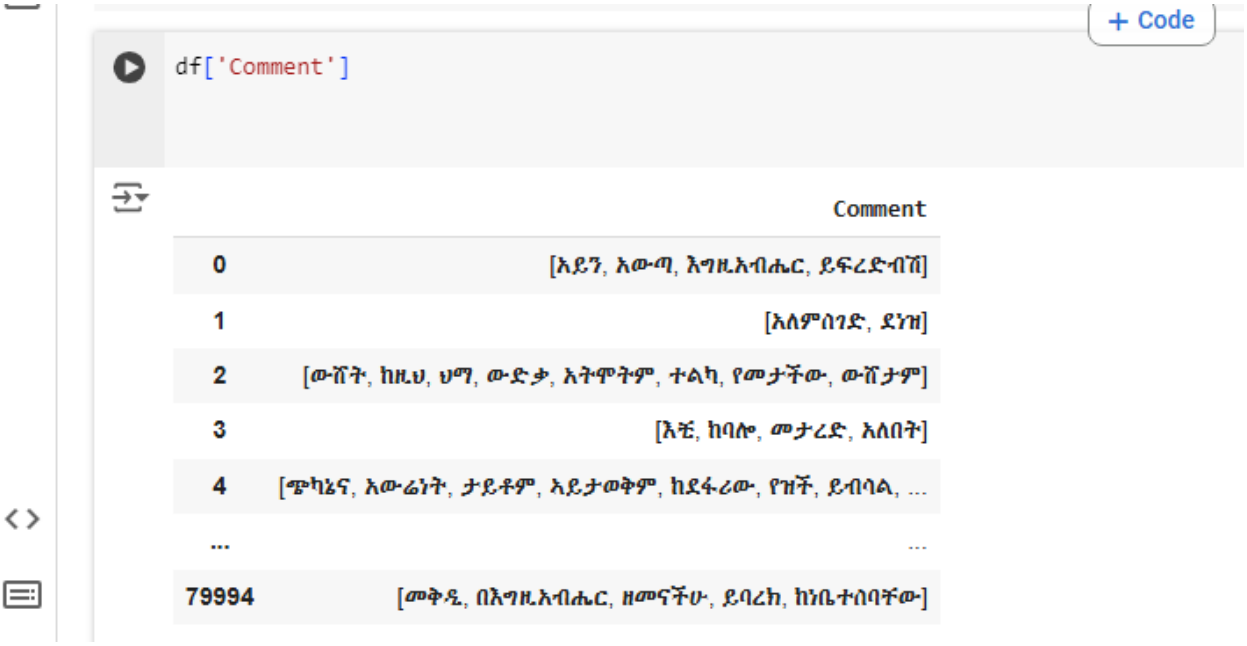


Figure 4.4 screenshot of tokenized data set

4.5 Feature extraction

Words were represented in vector numerical format using the tokenized dataset as input for the extraction of features procedure. This work used Word2vec to implement word vectorization by Python Scikit-learn module.

4.5.2 Implementation of Word2vec

For identification of hate speech, these embeddings are crucial since they capture the semantic links between words. In deep learning, Word2Vec embeddings are used as input features in conjunction with networks. In this work pretrained Amharic word embedding from [44] was implemented. First pretrained Word2Vec embeddings were loaded into a dictionary and prepare an embedding matrix for use in a neural network model. Followed

by retrieving the pretrained embedding for the word. If the dictionary does not contain the word, it skips assigning an embedding.

If a word's embedding exists, it is stored in the corresponding row of embedding matrix, where the row index corresponds to the word's integer index from the tokenizer. The embedding matrix is typically passed to an embedding layer in a deep learning model. (RQ2)

```
embedding_matrix = np.zeros((vocab_size, embedding_dim))
for word, i in tokenizer.word_index.items():
    if i < vocab_size:
        embedding_vector = embeddings_dictionary.get(word)
        if embedding_vector is not None:
            embedding_matrix[i] = embedding_vector
```

Figure 4.5 screenshot of python program to prepare embedding matrix

4.6 Exploring data

Using pie chart the below figure visualize the distribution of hate and hate free destitution

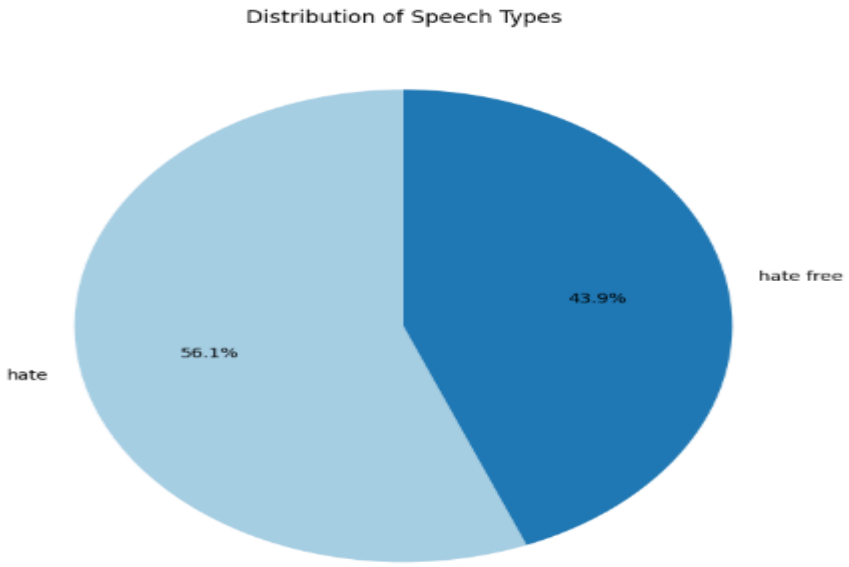


Figure 4.6: Dataset label distribution

Using a word cloud, the most commonly used terms for both hate and non-hate classed comments are visually displayed. In the diagram, terms that have been used more frequently than others are shown by bolding and enlarging the font size.

Figure 4.7 Word cloud of hate free Classified comments

Figure 4.8 Word cloud of hate Classified comments

4.7 Sentiment Analysis

The sentiment of each dataset was annotated manually based on the guideline in available on in the work of [3]. Each class of the dataset is categorized by its polarity such as positive, neutral, and positive.

The dataset was presented in two ways The dataset that was labeled as hate and hate free and a dataset that was labeled as hate and hate free along with their corresponding distribution of sentiment polarity.

	Comment	label	sentiment
0	['አይን', 'አውጣ', 'እግዚአብሔር', 'ይፍረድብሽ']	Hate	negative
1	['አለምበገድ', 'ደነዝ']	Hate	negative
2	['ውሸት', 'ከዚህ', 'ህማ', 'ውድቃ', 'አትሞትም', 'ተልካ', 'የ...	Hate	negative
3	['እቺ', 'ከባሎ', 'መታረድ', 'አለበት']	Hate	negative
4	['ሞካኔና', 'አውሬነት', 'ታይቶም', 'አይታወቅም', 'ከደፋሪው', '...	Hate	negative
5	['ባሏን', 'ይችን', 'መግደል', 'ካሀዲ']	Hate	negative
6	['ባልና', 'ሜስት', 'ከአንድ', 'ውሃ', 'ይቀዳል', 'ሜባለውን', ...	Hate	negative
7	['በልጆችሽ', 'ይድረስ', 'አዙኑ', 'በቤትሽ', 'ይግባ', 'ክፎ', ...	Hate	negative
8	['ልባንሽን', 'ይዝጋዉ', 'አለማፈሯ', 'የጅሽን', 'ይስጥሽ']	Hate	negative
9	['ፊትሽያስታውቃልመዋሸትሽንአቺአስመሳይ']	Hate	negative

Figure 4.9 screenshot of the pre procced dataset along with their label

Table 4.1 the sentiment analysis comparison of both hate and non-hate class

Sentiment polarity	Hate free		Hate		Total	
Neutral	4419	12.5%	4814	10.72%	9233	11.5%
Positive	22998	65.5%	7655	17.05%	30,653	38.32%
negative	7692	21.9%	32413	72.21%	40,105	50.13%
Total	35109		44882		79,991	

Table 4.1 shows the sentiment analysis comparison of both hate and hate-free class datasets. There are 35109 comments in the datasets that have been annotated by human annotators as hate-free class. Sentiment analysis of the Hate-free class dataset reveals that 65.5% have positive polarity, 12.5% have neutral polarity, and 21.9% have negative polarity. Positive polarity is found in most datasets, but negative and neutral polarity are uncommon.

The 44882 comments in the datasets that have been annotated by human annotators and classified as a hateful speech. And the Sentiment analysis of the Hateful class dataset reveals that 72.21% of the hate dataset have negative polarity, 10.72% have neutral polarity, and 17.05% have positive polarity. Neutral polarity is uncommon and most of the sample is classified as having negative polarity

The hate speech detection model uses 44,882 hate datasets in total. The sentiment analysis shows that 32413 comments are negative polarity, 4814 comments are neutral polarity, and 7655 comments are positive polarity from the hated class. While the total hate free dataset used for the hate speech detection model is 35109 dataset and the sentiment analysis shows the hate free class dataset as 22998 comments are positive polarity, 4419 comments are neutral polarity, and 7692 comments are negative polarity.

Table 4.1 show the relationship and correlation between sentiment analysis and hate speech for the Amharic language (RQ2).

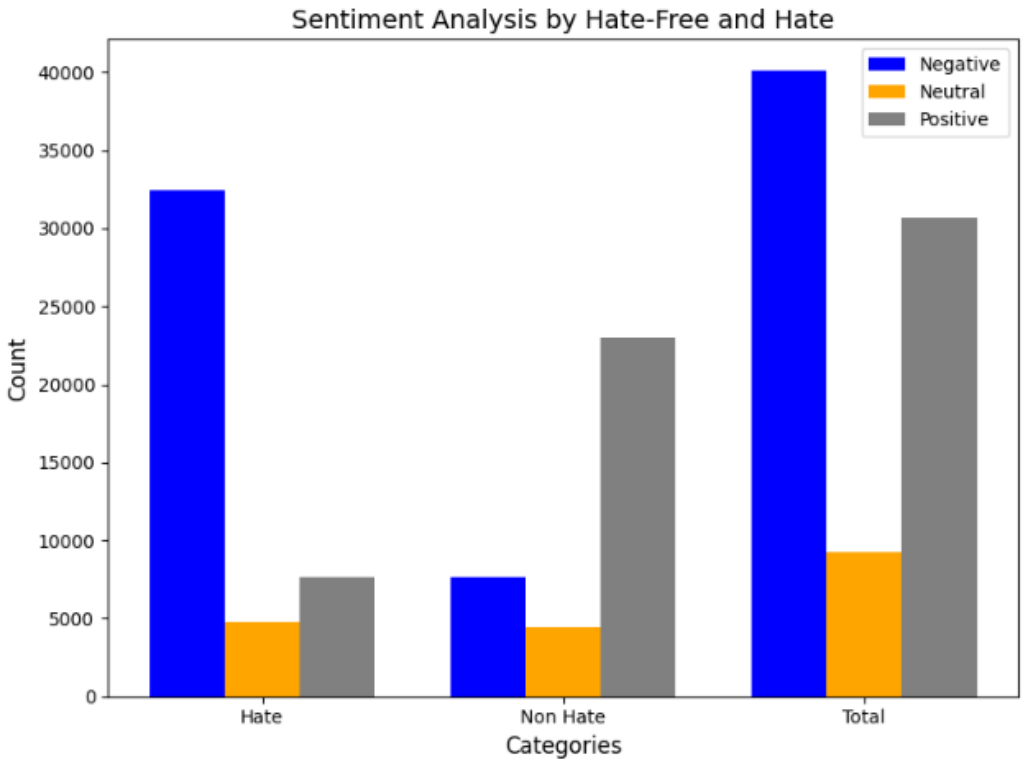


Figure 4.10 the comparison results of hate speech and hate-free datasets along with the sentiment.

The sentiment analysis generally demonstrates that negative polarity is the source of most of the hate speech. Positively oriented comments, however, hardly ever lead to hate speech. A positive polarity with marginally more data than other polarities contribute for the non-hate speech (RQ2).

4.8 Model building

Two experiments were carried out using the gathered dataset with a total amount of 79,991. The first experiment was conducted only with a hate and hate-free labeled dataset. The second experiment uses the dataset containing hate and hate-free speech labels along with

their sentiment polarity. Python programing language were used on Google collab. The experiment was also conducted using the deep learning models CNN and GRU.

Table 4.2: GRU hyperparameters along with their values

Hyperparameter	**Value**
Batch Size	128
Epochs	5
Dense Layer (Activation)	Sigmoid

Table 4.3: CNN hyperparameters along with their values

Hyperparameter	**Epochs**	**Dense Layer (Units)**	**Dense Layer (Activation)**	**Dense Layer (Output Activation)**	**Loss Function**	**Optimizer**
Value	9	64	'relu'	'sigmoid'	'binary_crossentropy'	'adam'

Deep learning model performance without consideration of sentiment analysis

Result for GRU (Gate Recurrent Unit)

Classification Report:

	precision	recall	f1-score	support
Non-Hate	0.90	0.86	0.88	3547
Hate	0.89	0.92	0.91	4453
accuracy			0.90	8000
macro avg	0.90	0.89	0.90	8000
weighted avg	0.90	0.90	0.90	8000

Figure 4.11: GRU categorization report

Evaluation metrics demonstrate how well the GRU model performs in two categories: "Non-hate" and "Hate." 89% of forecasts classified as "Hate" are accurate, whilst 90% of predictions labeled as "Non-hate" are accurate for Precision. The algorithm accurately detects 92% of real "Hate" samples and 86% of real "Non-hate" samples in terms of recall.

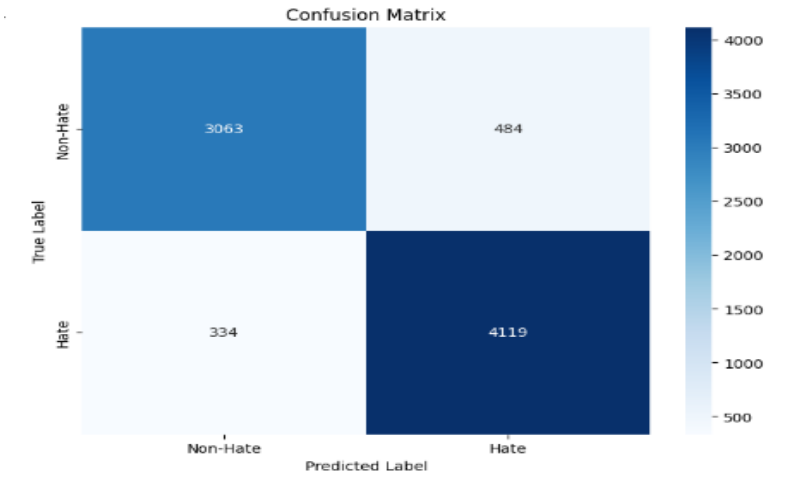


Figure 4.12 GRU model confusion matrix

Result for CNN

Classification Report:

	precision	recall	f1-score	support
Non-Hate	0.68	0.67	0.68	7028
Hate	0.74	0.76	0.75	8972
accuracy			0.72	16000
macro avg	0.71	0.71	0.71	16000
weighted avg	0.72	0.72	0.72	16000

Figure 4.15: categorization report for the CNN model

With a comparatively low rate of false positives and a greater capacity to detect real positives, this implies that the model is more successful at identifying "Hate" content.

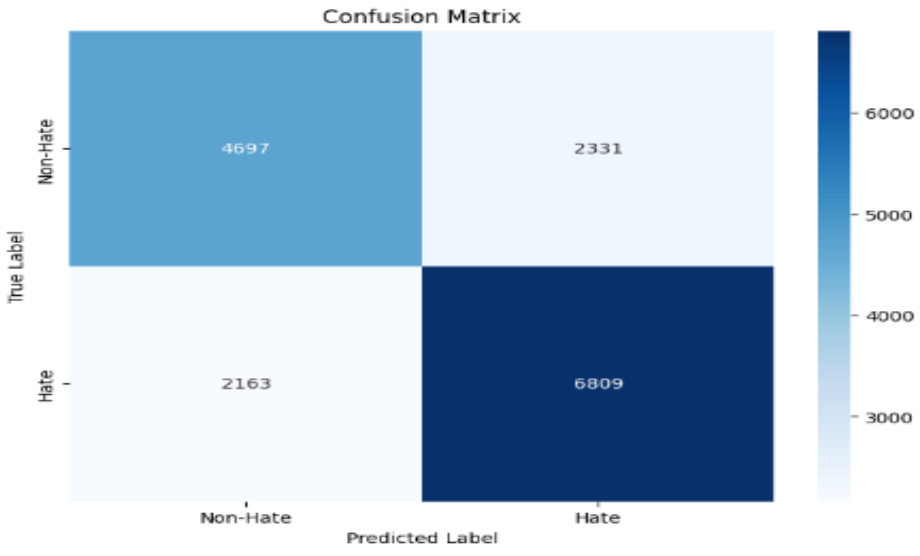


Figure 4.16 : matrix of confusion for the CNN model

Deep learning model performance with consideration of sentiment analysis

Sentiment information whether positive, negative, or neutral is added as an extra input feature. Since some sentiments (such "negative") may be more strongly correlated with hate speech, adding sentiment to the model could help it better differentiate between hateful and non-hateful content.

Result for GRU (Gate Recurrent Unit) considering the sentiment

Prior to final classification, the GRU-based text embeddings are concatenated with the sentiment input. Sentiment is an additional feature that enhances the GRU layer's text-based representation.

Classification Report:

	precision	recall	f1-score	support
Non-Hate	0.74	0.66	0.70	7028
Hate	0.75	0.82	0.79	8972
accuracy			0.75	16000
macro avg	0.75	0.74	0.74	16000
weighted avg	0.75	0.75	0.75	16000

Figure 4.19: classification report of GRU

The above result indicates that it is 74% accurate in predicting a comment to be non-hate. The Non-Hate class's recall, on the other hand, is 0.66, implying that while the model accurately detects 66% of all genuine non-hate remarks, it does miss some of them. The model correctly detects a significant percentage (82%) of real hate comments while significantly reducing false positives (only 25% of the projected hate comments are wrong)

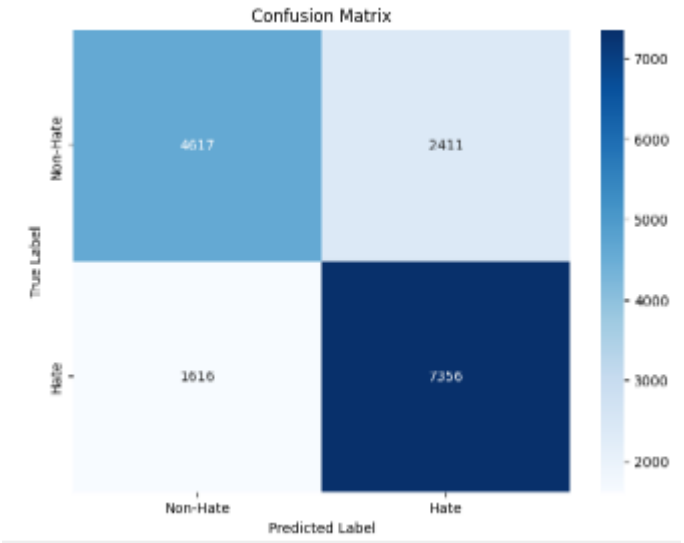


Figure 4.20: matrix of confusion for the GRU model

Result for CNN considering the sentiment

Classification Report:

	precision	recall	f1-score	support
Non-Hate	0.68	0.79	0.73	7028
Hate	0.81	0.71	0.76	8972
accuracy			0.74	16000
macro avg	0.75	0.75	0.74	16000
weighted avg	0.75	0.74	0.75	16000

Figure 4.21: classification report regarding the CNN model

For the Non-Hate class, the model gets a Precision of 0.68, implying that when it predicts a comment as "Non-Hate," 68% of the time it is true. The algorithm correctly detects 79% of the real "Non-Hate" remarks, according to the comparatively high recall of 0.79. This implies that the model is typically good at recognizing non-hate comments, even when it occasionally misclassifies hate speech as non -hate (lower precision).

when the model predicts a comment as "Hate," 81% of the time it is correct. This is a high precision, meaning the model is cautious and confident when predicting hate speech. However, the Recall for the hate class is lower at 0.71, meaning the model only identifies 71% of the actual hate speech in the dataset. While it is relatively good at identifying hate speech, it misses about 29% of the hate comments, which affects the recall.

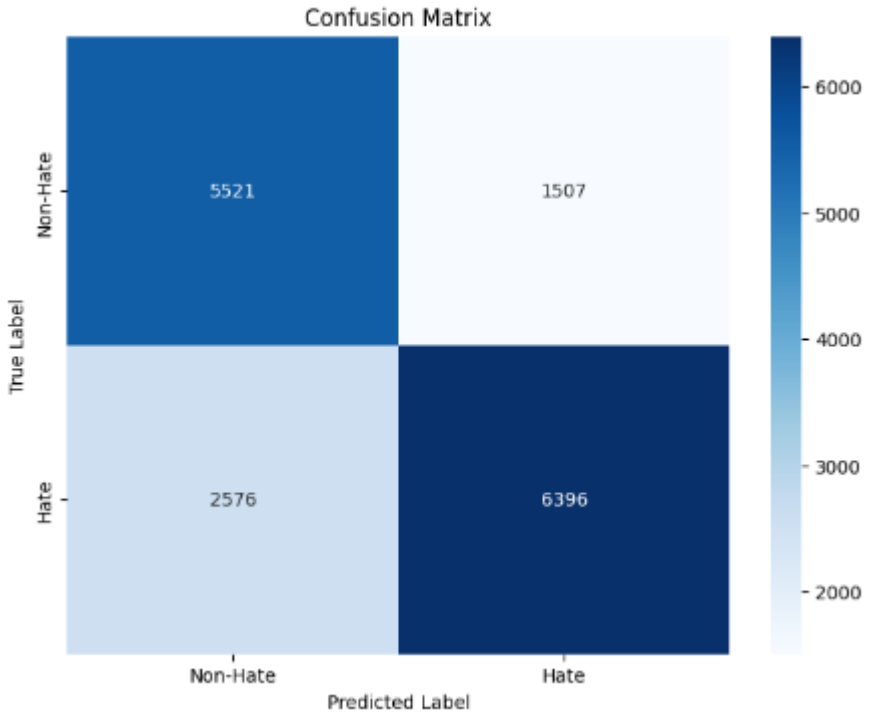


Figure 4.22: CNN model's confusion matrix

Actual and anticipated hate and hate-free comment values are represented by the confusion matrices. Considering sentiment as a feature in the dataset did impact the accuracy of the hate speech detection models. The GRU model, which initially achieved an accuracy of 90% without sentiment analysis, saw a drop in accuracy to 75% when sentiment features were included. However, the CNN model performed better with sentiment features, achieving 74%accuracy. This shows that sentiment analysis has a significant impact on improving the CNN model's performance but reduces the GRU model's ability to detect hate speech accurately. This highlights that the effectiveness of sentiment analysis varies across models, and should be considered carefully. (RQ5)

Since the consideration of sentiment of the dataset shows a model improvement on the CNN model, Here’s the comparison of the result between the two CNN models

Table 4.4 CNN model comparison

Metric	CNN model trained with out sentiment consideration	CNN model trained with sentiment consideration	Conclusion
Non-hate Precision	0.68	0.68	Precision is the same in both results. The model equally identifies "Non-hate" examples correctly.
Non-hate Recall	0.67	0.79	Recall improved significantly, meaning better identification of actual "Non-hate" cases.
Non-hate F1-score	0.68	0.73	F1-score increased due to improved recall, reflecting better overall performance for "Non-hate."
Hate Precision	0.74	0.81	Precision improved, meaning fewer false positives in predicting the "Hate" class.
Hate Recall	0.76	0.71	Recall decreased, meaning the model now misses more "Hate" cases (higher false negatives).
Hate F1-score	0.75	0.76	F1-score slightly improved due to better precision, despite the drop in recall.
Overall Accuracy	0.72	0.74	Overall accuracy improved, reflecting better balanced performance across both classes.

While the recall for the "Hate" class is marginally compromised, the CNN model that consider sentiment of the dataset, shows a superior balance. Overall accuracy improved from 0.72 to 0.74, indicating the model as a whole performed better in compared to the one that does not take sentiment is to consideration.

GRU outperforms CNN in both experimental setups. Without sentiment analysis, GRU achieves 90% accuracy, while CNN achieves 72%. With sentiment analysis, CNN slightly improves (74%) while GRU's accuracy drops to 75%, indicating sentiment's effect on detection. (RQ3)

The proposed approach identifies hate speech in Amharic social media with a high degree of accuracy, as demonstrated by the GRU model, which achieved an accuracy of 90% in detecting hate speech on human-annotated datasets. The inclusion of sentiment analysis had a negative impact on the GRU model's accuracy, reducing it to 75%, but the model still remained effective in detecting hate speech. Therefore, the proposed approach can identify hate speech in Amharic on social media with a significant degree of accuracy, though sentiment features can sometimes reduce the model's effectiveness. (RQ4)

4.9 Error Analysis of the Experiment

The following factors contribute to our suggested model's misclassification issue with posts and comments. There is ambiguity in correctly labeling the dataset; even after a brief explanation of the annotation guidelines, some annotators interpret certain sentences differently. Some of the written comments in the gathered dataset are misspelled.

4.10 Comparison with related works

The accuracy, feature extraction, and algorithm utilized in a related hate speech detection model were compared.

Table 4.5: Performance comparison of related works

Ref	**Used Model(s)**	**Feature Extraction**	**Result**
[27]	Naive Bayes, Random Forest	TF-IDF, Word2Vec	79.83%
[23]	CNN and LSTM hybrid	Word2Vec	66.56%
[26]	RF, LR, DT, LSVC, MNB, SVC	N-gram, TF-IDF	LSVC (F1-score: 64%)
[17]	SVM, NB, RF	N-gram, TF-IDF	NB (Accuracy: 79%)

CHAPTER FIVE

CONCLUSION, RECOMMENDATION AND FUTURE WORK

5.1 Overview

The topics covered in this chapter include fundamental implementation procedure as well as a summary of the analysis and findings in building the Amharic hate speech detection model. Followed by contribution, recommendation and future works for Amharic hate speech detection model.

5.2 Conclusion

With the speed at which technology is developing, using social media platforms has become common habit. These platforms allow users to express their feelings through posting, sharing, commenting, and replying, which makes it crucial to keep an eye out for hate speech and create measures to monitor it from spreading. This study created a hate speech detection model by taking into account the sentiment of each Amharic dataset that was gathered.

A total of 79,991 dataset have been used which were represented in an excel format. These collected datasets were extracted from a total of 160k comments and posts which are found on different kinds of personal and public Channels of TikTok, you tube and Facebook.

The annex section provides a detailed description of the new annotation guidelines that were created based on the dataset that was gathered. The annotation process of the hate speech dataset has a number of challenges, including word interpretation and contextual meaning, as well as time consumption.

Important pre-processing procedures were then carried out in accordance with the language's requirements to clean the corpus. Additionally, pretrained word embedding were implemented for to extract features. In the development of the system GRU and CNN were used. The datasets used for this thesis have been allocated into training dataset and testing datasets, which each contains 80% and 20% of the content, respectively.

The experiments were performed in two ways. The first experiment was performed using human-annotated datasets that were hateful and hate-free. A dataset with human annotations of hate and hate-free data along with their respective sentiments. The hate and hate free label were annotated by new annotation guideline that is presented in detail in the annex section. Whereas each dataset's sentiment analysis was examined utilizing an already existing sentiment analysis annotation guideline.

There are 35109 comments in the datasets that have been annotated by human annotators as hate-free class. Sentiment analysis of the Hate-free class dataset reveals that 65.5% have positive polarity, 12.5% have neutral polarity, and 21.9% have negative polarity. Positive polarity is found in most datasets, but negative and neutral polarity are uncommon.

The 44882 comments in the datasets that have been annotated by human annotators and classified as a hateful comment. And the Sentiment analysis of the Hateful class dataset reveals that 72.21% of the hate dataset have negative polarity, 10.72% have neutral polarity, and 17.05% have positive polarity. Neutral polarity is uncommon and most of the sample is classified as having negative polarity.

From the first experiment, the GRU model scored and 0.90, 0.895, 0.895, and 0.89 While CNN model scored, and 0.72 , 0.715 ,0.71 and 0.715, and in accuracy, F1-score, precision, and recall, respectively with only human-annotated hate speech datasets. This implies GRU outperform in every evaluation parameter. The second experiment was done using annotated dataset along with their respective sentiment. And GRU model scored and 0.75 , 0.74, 0.745 and 0.745 and While the CNN scored, and 0.74, 0.75, 0.745, and 0.745 in accuracy, recall, F1-score and precision respectively

Although, the second experiment's outcomes shows that sentiment analysis has an impact on the performance of GRU, reducing its accuracy from 90% to 75%, CNN model achieved better performance under sentiment analysis. The CNN model that takes into account the sentiment of the dataset performs better, overall, in terms of recall for the Hate class and precision for the non-hate class, which results in higher F1-scores for both categories, when compared to the one without sentiment consideration

Although both models do rather well, the CNN model trained with sentiment consideration performs better at distinguishing occurrences of hate and non-hate.

5.3 Contribution

- A new dataset that includes the sentiment polarity associated with hatred and free hate labels.
- Due to TikTok's growing popularity, the majority of the data was gathered from its interns, that supply a range of hate speech content and a recent dataset.
- A new set of annotation requirements for datasets containing hate speech
- The dataset's sentiment polarity class assigns labels of negative, neutral, and positive polarity, while the hate class assigns labels of hate and non-hate.
- The relationship between sentiment analysis and hate speech in Amharic were presented.
- Assists in reducing the dissemination of divisive and hateful content on news websites, social media, and other online forums.
- Enhances the tranquility and inclusivity of the digital environment by detecting and restricting offensive content that may provoke violence or prejudice.

5.4 Recommendation and future work

- It is advised to apply technical annotation techniques to enhance dataset quality, prevent bias, and reduce time consumed for annotating.
- In order to improve accuracy and further the research, certain ethics and religions be identified in order to determine who is more exposed.
- Future studies can incorporate Amharic language character that impact hate speech detection models, such as metaphorical discourse and sarcasm.

Reference

[1] D. M. E. D. M. Hussein, "A survey on sentiment analysis challenges," *J. King Saud Univ. - Eng. Sci.*, vol. 30, no. 4, pp. 330–338, 2018, doi: 10.1016/j.jksues.2016.04.002.

[2] S. G. Tesfaye and K. K. Tune, "Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network," *Res. Sq.*, pp. 1–14, 2020, [Online]. Available: https://www.researchsquare.com/article/rs-114533/latest?utm_source=researcher_app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound

[3] S. M. Yimam, H. M. Alemayehu, and U. Hamburg, "Exploring Amharic Sentiment Analysis from Social Media Texts : Building Annotation Tools and Classification Models," pp. 1048–1060, 2020.

[4] S. M. Yimam, A. A. Ayele, and C. Biemann, "Analysis of the Ethiopic Twitter Dataset for Abusive Speech in Amharic," pp. 1–5, 2018.

[5] J. Seglow, "Hate Speech , Dignity and Self-Respect," *Ethical Theory Moral Pract.*, no. July, 2016, doi: 10.1007/s10677-016-9744-3.

[6] T. Mikhael, "Hate Speech on Social Media Highlights of 2018 in Lebanon," *Maharat Found. Credit.*, p. 13, 2018.

[7] L. A. Silva, "A Measurement Study of Hate Speech in Social Media," pp. 85–94, 2017.

[8] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, p. 107134, 2021, doi: 10.1016/j.knosys.2021.107134.

[9] A. M. Gezmu, A. Nürnberger, and B. E. Seyoum, "Portable Spelling Corrector for a Less-Resourced Language : Amharic," pp. 4127–4132, 2014.

[10] M. Abate and Y. Assabie, “Development of Amharic Morphological Analyzer Using Memory-Based Learning,” pp. 1–13, 2014.

[11] I. H. Sarker, “Machine Learning : Algorithms , Real - World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[12] L. Alzubaidi *et al.*, *Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*, vol. 8, no. 1. Springer International Publishing, 2021. doi: 10.1186/s40537-021-00444-8.

[13] E. M. Dharma, F. L. Gaol, H. L. H. S. Warnars, and B. Soewito, “the Accuracy Comparison Among Word2Vec, Glove, and Fasttext Towards Convolution Neural Network (Cnn) Text Classification,” *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 2, pp. 349–359, 2022.

[14] T. M. Ababu and M. M. Woldeyohannis, “Bilingual Hate Speech Detection on Social Media : Amharic and Afaan Oromo,” *2022 Lang. Resour. Eval. Conf. Lr. 2022*, pp. 6612–6619, 2022, [Online]. Available: https://www.researchsquare.com/article/rs-3355274/v1

[15] F. A. Melat, “Hate Speech Detection for Amharic Language on Facebook Using Deep Learning,” *Bahir Dar Inst. Technol.*, pp. 1–23, 2022, [Online]. Available: http://ir.bdu.edu.et/handle/123456789/14487

[16] E. Bawoke, “Amharic text hate speech detection in social media using deep learning approach,” 2020, [Online]. Available: http://ir.bdu.edu.et/handle/123456789/11266%0Ahttp://ir.bdu.edu.et/bitstream/handle/123456789/11266/Emuye Bawoke Final Thesis 2020 G.C.pdf?sequence=1&isAllowed=y

[17] Weldemariam B., “Hate Speech Detection from Facebook Social Media Posts and Comments in Tigrigna language,” *St. Mary’s Univ.*, pp. 1–3, 2022.

[18] H. Liu, W. Alorainy, P. Burnap, and M. L. Williams, “Fuzzy multi-task learning

for hate speech type identification," *Web Conf. 2019 - Proc. World Wide Web Conf. WWW 2019*, pp. 3006–3012, 2019, doi: 10.1145/3308558.3313546.

[19] N. Albadi, M. Kurdi, and S. Mishra, "Are they our brothers? analysis and detection of religious hate speech in the Arabic Twittersphere," *Proc. 2018 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2018*, pp. 69–76, 2018, doi: 10.1109/ASONAM.2018.8508247.

[20] H. T. Do, H. D. Huynh, K. Van Nguyen, N. L. Nguyen, and A. G. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text using the Bidirectional-LSTM Model," pp. 4–7.

[21] F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," *CEUR Workshop Proc.*, vol. 1816, no. January, pp. 86–95, 2017.

[22] F. M. Plaza-del-arco, M. D. Molina-gonzález, L. A. Ureña-lópez, and M. T. Martín-valdivia, "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis," vol. 9, 2021, doi: 10.1109/ACCESS.2021.3103697.

[23] H. Faris, I. Aljarah, M. Habib, and P. A. Castillo, "Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context," no. Icpram 2020, pp. 453–460, 2022, doi: 10.5220/0008954004530460.

[24] M. Bouazizi, T. O. Ohtsuki, and S. Member, "A Pattern-Based Approach for Sarcasm Detection on Twitter," *IEEE Access*, vol. 4, pp. 5477–5488, 2016, doi: 10.1109/ACCESS.2016.2594194.

[25] W. Philemon and W. Mulugeta, "a Machine Learning Approach To Multi- Scale Sentiment Analysis of Tigrigna Online Posts," *HiLCoE J. Comput. Sci. Technol.*, vol. 2, no. September, pp. 80–87, 2021.

[26] N. B. Defersha and K. K. Tune, "Detection of Hate Speech Text in Afan Oromo Social Media using Machine Learning Approach," *Indian J. Sci. Technol.*, vol. 14,

no. 31, pp. 2567–2578, 2021, doi: 10.17485/ijst/v14i31.1019.

[27] Z. Mossie and J.-H. Wang, “Social Network Hate Speech Detection for Amharic Language,” pp. 41–55, 2018, doi: 10.5121/csit.2018.80604.

[28] F. Del Vigna, A. Cimino, and F. D. Orletta, “Hate me , hate me not : Hate speech detection on Facebook Hate me , hate me not : Hate speech detection on Facebook,” no. January, 2017.

[29] S. Ro, “Greek a d roman mytholoGy,” pp. 473–481, 1999, [Online]. Available: https://support.twitter.eom/articles/l

[30] M. Gasser, “HornMorpho: a system for morphological processing of Amharic,” *Oromo, Tigrinya, Indiana Univ.*, 2012.

[31] “Facebook Community Standards | Transparency Center.” Accessed: Sep. 27, 2024. [Online]. Available: https://transparency.meta.com/policies/community-standards/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2F%2520objectionable_content.

[32] L. W. Levy, K. L. . Karst, and A. Winkler, “Encyclopedia of the American Constitution. Vol. 2,” 2004.

[33] A. A. Ayele, S. M. Yimam, T. D. Belay, T. T. Asfaw, and C. Biemann, “Exploring Amharic Hate Speech Data Collection and Classification Approaches,” pp. 49–59, 2023.

[34] N. S. Mullah, W. Mohd, and N. Wan, “Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media : A Review,” *IEEE Access*, vol. 9, pp. 88364–88376, 2021, doi: 10.1109/ACCESS.2021.3089515.

[35] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[36] K. Shah, H. Patel, D. Sanghvi, and M. Shah, “A Comparative Analysis of Logistic

Regression , Random Forest and KNN Models for the Text Classification," *Augment. Hum. Res.*, vol. 0, 2020, doi: 10.1007/s41133-020-00032-0.

[37] B. M. Jadav and M. E. Scholar, "Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis," *Int. J. Comput. Appl.*, vol. 146, no. 13, pp. 975–8887, 2016.

[38] S. Zhang and S. Member, "Challenges in KNN Classification," pp. 1–13, 2021, doi: 10.1109/TKDE.2021.3049250.

[39] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019, doi: 10.1109/ACCESS.2019.2912200.

[40] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[41] S. Hunegegnaw, "Amharic Sentiment Mining Model for Opinionted Text of Social Media Using Machine Learning," *Master Sci. Comput. Sci.*, 2023.

[42] N. D. T. Ruwandika and A. R. Weerasinghe, "Identification of Hate Speech in Social Media," *2018 18th Int. Conf. Adv. ICT Emerg. Reg.*, pp. 273–278, 2018.

[43] A. M. Iddrisu, S. Mensah, F. Boafo, G. R. Yeluripati, and P. Kudjo, "A sentiment analysis framework to classify instances of sarcastic sentiments within the aviation sector," *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 2, p. 100180, 2023, doi: 10.1016/j.jjimei.2023.100180.

[44] G. Demlew, "Amharic-Word-Embedding-Word2vec," GitHub, 2020. [Online]. Available: https://github.com/gashawdemlew/Amharic-Word-Embedding-Word2vec. [Accessed: Sep. 12, 2024].

Annexes

Annex A: Dataset Annotation Guideline

The guidelines for labeling the post or the comments as hate speech are as follows:
The annotator's task is to label the sentence as a hate and hate free. The annotators are staff members of Ethio Telecom's Digital Customer Care Division, with experience in assigning sentiment to customer comments on the company's posts as part of their job. They are also proficient in the Amharic language.

List of the key queries to be considered before labeling

1.Are the commenters use words to call a person, group member a liars, cheaters or other terms to provoke a reaction

2.Does the text equate a certain group or individual to culturally despised non-human entities, such as germs or dirt?

3.Does the statement characterize a person or groups of people in extremely negative ways, by comparing them with a type of animals that are culturally perceived as inferior?

4.Does the sentence refer person or a group as a Certain Inanimate Objects and Non-Human States?

5.Does the text Mock people on the basis of their Characteristics?

6.Does the sentence use harsh words leads to divisions in society

7.Do the statements promote malicious stereotypes against a group?

8.Does the sentence use words that undermines or causes psychological harm to its victims

9.Does the sentence consist of statements that attack or insult a person or group?

10. Does the sentence contain language that disparages someone based on who they are?

11. Are the comments makes any allusions to the purported superiority or inferiority of particular target groups.

12.Does the sentence influence a person's various traits and inspires viewers to act or break the law.

13. Does the sentence include stereotypes, which are oversimplified beliefs about a certain subject

14. Does the sentence criticize or accuses someone based on the target groups they belong to

15. Does the sentence phrase is intended towards persons that have similar culture and way of life

16. Does the sentence use violent language in its communications.

17. Does the sentence links to and endorses other hostile information, hateful tweets, and organizations.

18 Does the sentence it expresses violent messages.

19. Does the sentence aim to suppress a minority or other race or religion.

20. Does the sentence criticize a minority without providing a solid justification

Since Amharic understanding level is different among the annotators, some comments that are ambiguous to be annotated as a hate and non-hate, were labeled by discussion among the annotators.

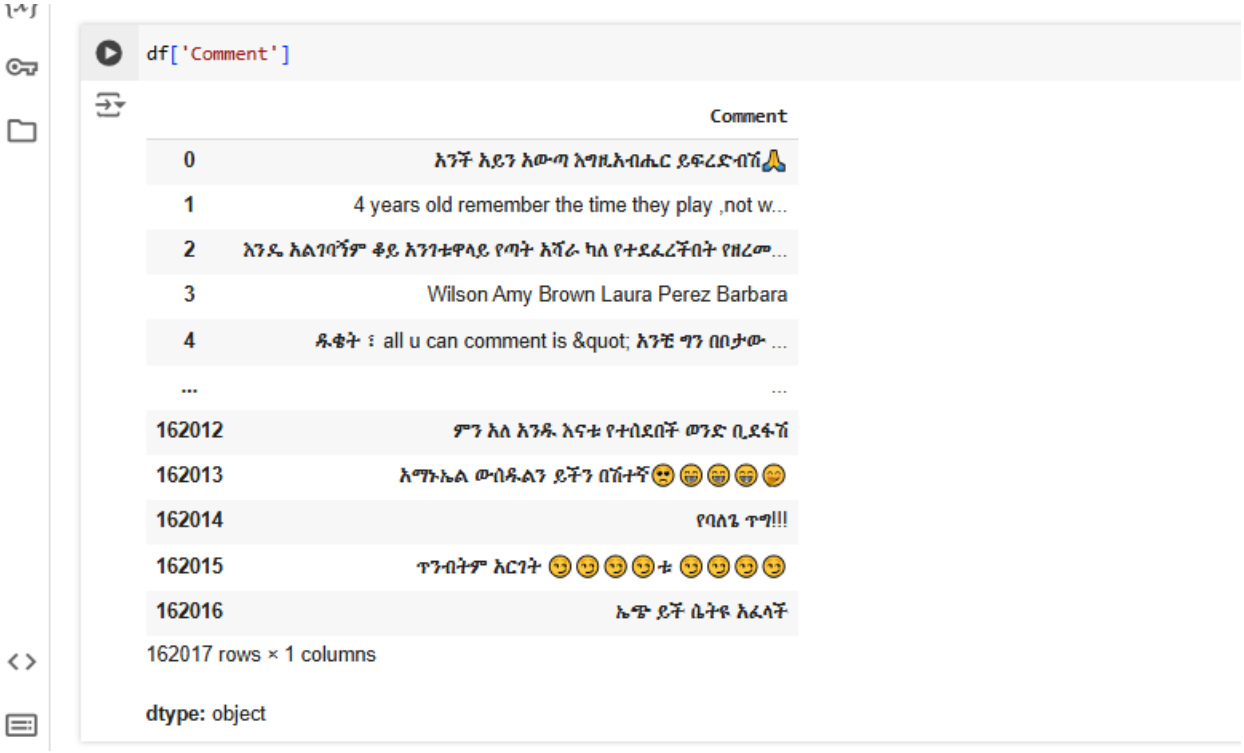
The guidelines for labeling the post or the comments as non-hate speech are as follows:

●If a post or remark does not contain hate speech or harsh language

● If a post or comment is made by the government or a duly authorized government agency with the aim of promoting and enlightening the public.

●Facts and typical viewpoints fall under the scope of normal communication.

Annex B: Portion of the acquired dataset



```
df['Comment']
```

	Comment
0	አንች አይን አውጣ አግዚአብሔር ይፍረድብሽ🙏
1	4 years old remember the time they play ,not w...
2	እንዴ አልገባኝም ቆይ አንገቱዋላይ የጣት አሻራ ካለ የተደፈረችበት የዘረመ...
3	Wilson Amy Brown Laura Perez Barbara
4	ዱቄት ፣ all u can comment is " አንቺ ግን በቦታው ...
...	...
162012	ምን አለ አንዱ እናቱ የተበደበች መንድ ቢደፋሽ
162013	አማኑኤል ውስዱልን ይችን በሽተኛ😖😁😁😁😊
162014	የባለጌ ጥግ!!!
162015	ጥንብትም አርገት 😡😡😡😡ቱ 😡😡😡😡
162016	ኤጭ ይች ቤትዩ አፈላች

162017 rows × 1 columns

dtype: object

Annex C: Portion of the acquired dataset during preprocessing

```
dtype: object

[ ] for i in range(200):
        print(df['Comment'][i])

አይን አውጣ አግዚአብሄር ይፍረድብሽ
አለምስባድ ደነዝ
ውሸት ከዚህ ሀግ ውድቃ አትሞትም ተልካ የመታችው ውሸታም
እቺ ከባሎ መታረድ አለበት
ጭካኔና አውሬነት ታይቶም አይታወቅም ከደፋረው የዝች ይብሳል የሷም አመት ይጨመርና በድምሩ ይሁንላቸው
ባሷን ይችን መግደል ካህዲ
ባልና ሚስት ከአንድ ውሀ ይቀዳል ሚባለውን ከዚች ጨካኝ አየሁ ከልጅቾሽ አሳኝው ወጣ ልጅቾሽን ይድፈርና በአደባባይ ያስወጣሽ የዛች ምስኬን አምላክ
በልጅቾሽ ይድረስ አዙ በቤትሽ ይግባ ክፎ ጨካኝ የአረመኔ በላው ሚስት ክፎ የልጅችስ እናት አያርግሽ ይፋረድሽ አውነትን ያቃልና
ልባንሽን ይዝጋዉ አለማፈሯ የጅሽን ይሰጥሽ
ፊትሽያስታውቃልመዋሸትሽንአቺአስመሳይ
አቺዝቅዝቅምንአባሽነውደሞየታባሽነውአምላክንየምታቂውአቺውሽታምአርካሽአቺንምይሄቤጣንባልሽንበባትነበርማታጠልየቤጣንቁራጮች
ባለጌ ጋዜጠኛው በምትጠቀመው ቃላት ጨዋ
የጋዜጠኝነት ስነ ምግባር ገለልተኛ ነበረበት ፍትህ ለጌትነት
ይች ንፍጥናት አትረዳም አትከራከረሊ አንችም እንዳትገቢ ፍር በምታወረው እንኳን እናቷ ተመልካችም እየተናደደብሽ
እራስሽ ዘገምተኛ ያ አመት ነብስ የማታውቅ ህፃን ወደቀ የሚለውን ቃል እራሱ አትችለውም የኔልጅ ለምባሉ አመት አለኝ እያወራም አፉ አይለማም የማይሆን ከቁስሉ ታቆስያቾዋለሽ
ጋዜጠኛው መርማሪ አቷን በጥፊ ያባራትነበር ወሬዋ ቢያስጠላ ታተይ ታተይ ድራማ ሌባ ፊትናት ባንች ማትወጅውን በነው ባትኖሪ እንደነበርሽ ማውራት
አባካችሁ አሏንስ አመት አትፈርዱባትም ህግ ካስ አቷና ፊታ ያስታውቃል ታቃለች ውሸታም እየኔ ሚች ቤተሰብ ቢኖራት አልደፈራትም ዘርሽን በሙሉ ይለቅሙልሽ ቆሻሻ እየኔ ግብር አድርጋቾዋትስ ያቃል
እቺ ችው በላነሽሽሽ አረመኔ አቺንሚ ከባልሽ መበቀል ነበርርር
ነብስ ለበይጣን አየገበራቹ ሊአል ከምትገቡ በጌታ በኢየሱስ አመኑ ንስሀ ግቡ አምልጡ ከበይጣ ህፃኗ አትመለስም
```

Annex D: sample of the collected dataset after tokenization

```
[ ] for i in range(200):
        print(df['Comment'][i])
```

```
['አይገ', 'እውጣ', 'እግዚአብሄር', 'ይፍረድብሽ']
['አለምስባድ', 'ደነዝ']
['ውሽት', 'ከዚህ', 'ህማ', 'ውድቃ', 'አትሞትም', 'ተልካ', 'የመታቸው', 'ውሸታም']
['እቺ', 'ከባሎ', 'መታረድ', 'አለበት']
['ጭካኔና', 'እውሬነት', 'ታይቶም', 'አይታወቅም', 'ከደፋረው', 'የዝች', 'ይብላል', 'የሷም', 'አመት', 'ይጨመርና', 'በድምሩ', 'ይሁንላቸው']
['ባሷን', 'ይችን', 'መግደል', 'ካህዲ']
['ባልና', 'ሚስት', 'ከአንድ', 'ውሀ', 'ይቀዳል', 'ሚባለውን', 'ከዜች', 'ጨካኝ', 'አየሁ', 'ከልጅችሽ', 'አማኝው', 'ወጦ', 'ልጆችሽን', 'ይድፈርና', 'በአደባባይ', 'ያስወጣሽ', 'የዛች'
['በልጆችሽ', 'ይድረስ', 'እዙ', 'በቤትሽ', 'ይግባ', 'ክፎ', 'ጨካኝ', 'የአረመኔ', 'በለው', 'ሚስት', 'ክፎ', 'የልጆችስ', 'እናት', 'አያርግሽ', 'ይፋረድሽ', 'እውነትን', 'ያቃልና']
['ልባንሽን', 'ይዝጋዉ', 'አለማፈሯ', 'የጅሽን', 'ይሰጥሽ']
['ፌትሽያስታውቃልመዋሸትሽንአቺአለመሳይ']
['አቺህቅዝምንአባሽነውደሞየታባሽነውአምላክንየምታቂውአቺውሸታም እርካሽአቺንምይሄሌጣንብልሽንበባትክበርግቃጠልየቤጣንቂራጮች']
['ባልጌ', 'ጋዜጠኛው', 'በምትጠቀመው', 'ቃላት', 'ጨዋ']
['የጋዜጠኝነት', 'ስነ', 'ምግባር', 'ገለልተኛ', 'ነበረበት', 'ፍትህ', 'ለጌትነት']
['ይች', 'ንፍጥናት', 'አትረፈም', 'አትከራከረሊ', 'አንችም', 'እንዳትገቢ', 'ፍር', 'በምታወሪው', 'እንኳን', 'እናቷ', 'ተመልካችም', 'አየተናደደብሽ']
['እራስሽ', 'ዘገምተኛ', 'ያ', 'አመት', 'ኑብስ', 'የማታውቅ', 'ህፃን', 'ወደቀ', 'የሚለውን', 'ቃል', 'አራሁ', 'አትችለውም', 'የኔልጅ', 'ስምባሉ', 'አመት', 'አለኝ', 'አያወራም', '
['ጋዜጠኛው', 'መርማሪ', 'አፏን', 'በጥፊ', 'ያባራትነበር', 'ወሬዋ', 'ሲያስጠላ', 'ታተይ', 'ታተይ', 'ድራማ', 'ሌባ', 'ፌትናት', 'ባንች', 'ማትወጅውን', 'በስው', 'ባትኖሪ', 'እንደነ
['እባክችሁ', 'አሏእስ', 'አመት', 'አትፈርጹባትም', 'ህግ', 'ካለ', 'እፏና', 'ፌቷ', 'ያስታውቃል', 'ታቃለች', 'ውሸታም', 'እቦኔ', 'ሚች', 'ቤተሰብ', 'ቢኖራት', 'አልደፈራትም', '
['እቼ', 'ችው', 'በላነሽሽሽ', 'አረመኔ', 'እቼንሚ', 'ከባልሽ', 'መሰቀል', 'ነበርርርር']
```

Annex E: sample of the labeled and tokenized dataset

```
[ ]  dataset_path= "/content/cleaned2.xlsx"
     df = pd.read_excel(dataset_path)
     df.head(10)
```

	Comment	label	sentiment
0	['አይን', 'አውጣ', 'እግዚአብሄር', 'ይፍረድብሽ']	Hate	negative
1	['አለምበገድ', 'ደነዝ']	Hate	negative
2	['ውሸት', 'ከዚህ', 'ህማ', 'ውድቃ', 'እትሞትም', 'ተልካ', 'የ...	Hate	negative
3	['እቺ', 'ከባሎ', 'መታረድ', 'አለበት']	Hate	negative
4	['ጭካኔና', 'አውሬነት', 'ታይቶም', 'አይታወቅም', 'ከደፋሪው', '...	Hate	negative
5	['ባሏን', 'ይችን', 'መግደል', 'ካሀዲ']	Hate	negative
6	['ባልና', 'ሚስት', 'ከአንድ', 'ውሀ', 'ይቀዳል', 'ሚባለውን', ...	Hate	negative
7	['በልጅቸሽ', 'ይድረስ', 'እዙኑ', 'በቤትሽ', 'ይግባ', 'ክፎ', ...	Hate	negative
8	['ልባንሽን', 'ይዝጋዉ', 'አለማፈሯ', 'የጅሽን', 'ይስጥሽ']	Hate	negative
9	['ፌትሽያስታውቃልመዋሸትሽንእቺአስመላይ']	Hate	negative

Annex F: sample of python codes

```
[ ]  # special symbol removal
     df['Comment']=df['Comment'].str.replace('[^\w\s]','',regex=True)
     # removing digits from the dataset
     df['Comment']=df['Comment'].str.replace('\d+','',regex=True)
```

```
#character level normalization

import re

def normalize_char_level_missmatch(input_token):
    rep1=re.sub('[ሃኅኃሐኻ]','ሀ',input_token)
    rep2=re.sub('[ሑኁዅ]','ሁ',rep1)
    rep3=re.sub('[ኂሒኺ]','ሂ',rep2)
    rep4=re.sub('[ኌሔኼ]','ሄ',rep3)
    rep5=re.sub('[ሕኅ]','ህ',rep4)
    rep6=re.sub('[ኆሖኾ]','ሆ',rep5)
    rep7=re.sub('[ሠ]','ሰ',rep6)
    rep8=re.sub('[ሡ]','ሱ',rep7)
    rep9=re.sub('[ሢ]','ሲ',rep8)
    rep10=re.sub('[ሣ]','ሳ',rep9)
    rep11=re.sub('[ሤ]','ሴ',rep10)
    rep12=re.sub('[ሥ]','ስ',rep11)
```

```
# Build the CNN model
model = Sequential()
model.add(Embedding(vocab_size, 50, input_length=maxlen, trainable=False))  # Reduced embedding size
model.add(Conv1D(16, kernel_size=3, padding='same', activation='relu', strides=1))  # Fewer filters
model.add(Dropout(0.7))  # Increased dropout rate
model.add(GlobalMaxPooling1D())
model.add(Dense(16, activation='relu'))  # Reduced units in Dense layer
model.add(Dropout(0.7))  # Increased dropout rate
model.add(Dense(1, activation='sigmoid'))
```

```
model2_GRU=Sequential()
#model2_GRU.add(Embedding(vocab_size, 100, input_length=maxlen))
model2_GRU.add(Embedding(max_features,100,mask_zero=True))
model2_GRU.add(GRU(64,dropout=0.5,return_sequences=True))
#model2_GRU.add(Dense(32))
model2_GRU.add(GRU(32,dropout=0.5,return_sequences=False))
model2_GRU.add(Dense(num_classes,activation='sigmoid'))
model2_GRU.compile(loss='binary_crossentropy',optimizer=Adam(learning_rate = 0.001),metrics=['accuracy', Precision(), Recall()])
history =model2_GRU.fit(X_train, y_train, batch_size=batch_size, epochs=epochs, verbose=1, validation_split=0.2)
```

```
# Load your dataset
dataset_path = "/content/preproccessed_hate and non hate_sentiment_labeled.xlsx"
df = pd.read_excel(dataset_path)

# Drop rows with NaN in Hate or Sentiment columns
df_new = df.dropna(subset=['label', 'sentiment'])

# Extract text data, hate labels, and sentiment labels
x = df_new['Comment']

# Encode hate speech labels (Hate = 1, Non-Hate = 0)
y = np.array(list(map(lambda x: 1 if x == "Hate" else 0, df_new['label'])))

# Encode sentiment labels (positive = 0, negative = 1, neutral = 2)
sentiment_mapping = {"positive": 0, "negative": 1, "neutral": 2}
z = df_new['sentiment'].map(sentiment_mapping).values
z = to_categorical(z, num_classes=3)  # One-hot encoding for sentiment

# Train-test split
X_train, X_test, y_train, y_test, z_train, z_test = train_test_split(x, y, z, test_size=0.20, random_state=1)
```