Amharic Language Hate Speech Detection on Social Media

Ermias Tadesse * , Beyene Kassa , Tarekegn Walle

Posted Date: 11 March 2025

doi: 10.20944/preprints202503.0820.v1

Keywords: Ge'ez; Fidel; LSTM; BiLSTM; GRU; and BiGRU

 doi:10.20944/preprints202503.0820.v1

Article

Amharic Language Hate Speech Detection on Social Media

Ermias Melku Tadesse [1,*], Beyene Kassa Wondie [2] and Tarekegn Walle Yirdaw [1]

[1] Department of Information System, Kombolcha Institute of Technology, Wollo University, Ethiopia

[2] Information Technology Department, Kombolcha Institute of Technology, Wollo University, Ethiopia

* Correspondence: ermiasmelku3400@gmail.com

Abstract: Social media platforms enable rapid communication, information sharing, and opinion expression. However, their misuse of hate speech targeting race, religion and political differences has become a growing concern. This issue is particularly sensitive for underrepresented languages like Amharic, a Semitic language with the second-largest number of speakers after Arabic and the working language of Ethiopia. This study addresses the challenge of detecting hate speech in Amharic text by analyzing posts and comments from Facebook, YouTube, and Twitter. A dataset of 7,590 labelled entries was collected using the Face Pager tool, focusing on hate speech related to race, religion, politics, and neutral content. The dataset was annotated with the guidance of researchers, legal experts, and language specialists. Preprocessing techniques, including data cleaning, tokenization, and normalization, were applied, and feature extraction was performed using embedding layers. The dataset was split into training (80%), validation (10%), and testing (10%) sets. Several deep learning models LSTM, BiLSTM, GRU, BiGRU, and RoBERTa were developed and evaluated using precision, recall, F1-score, and accuracy metrics. The RoBERTa model outperformed others, achieving an accuracy of 91%. This research highlights the effectiveness of advanced deep learning techniques in detecting Amharic hate speech, offering a valuable tool for mitigating this critical issue in Ethiopian social media contexts.

Keywords: Ge'ez; Fidel; LSTM; BiLSTM; GRU; and BiGRU

1. Introduction

Social media is transforming lifestyles, communication, and culture globally. In Ethiopia, despite poor internet services and occasional outages or blockages, social media usage has grown significantly. Platforms enable communication, opinion-sharing, and interaction through text, audio, video, and images. However, the misuse of social media, particularly for hate speech, is a growing concern. Hate speech based on race, gender, religion, politics, or ethnicity fuels offensive behavior and violence. The anonymity of social media allows individuals to spread harmful content without consequences [1].

To address this, the Ethiopian government has introduced regulations to combat hate speech and misinformation. Penalties include imprisonment and fines, especially for offenders with large followings or those using broadcast media. Hate speech manifests in various forms, such as fake news, defamation, offensive language, and targeted hate. Distinguishing between these types is crucial for effective detection and mitigation [2,3].

This research focuses on detecting hate speech in Amharic, Ethiopia's widely spoken language. We proposed a model using a newly labelled dataset categorized into racial hate, religious hate, political hate, positive, and neutral speech. The study leverages posts and comments from Facebook, YouTube, and Twitter, employing embedding feature extraction and machine learning techniques, including RNN and transformer-based deep learning algorithms. We aim to compare these models and identify the most effective approach for Amharic hate speech detection.

 doi:10.20944/preprints202503.0820.v1

2. Related Work

Even though this study focused on Amharic text hate speech detection, For clearly understand, we have seen different literature on local languages (Ethiopian language) and foreign languages social media hate speech text detection research. In this study, the local language means that the language is mainly spoken by Ethiopian and many Ethiopian mother tongue languages.

The studied [4] Are They, Our Brothers? Analysis and Detect of Religious Hate Speech in the Arabic Twitter sphere and creation of the first publicly available Arabic dataset annotated for the task of spiritual hate speech come upon and the number one Arabic lexicon which includes terms normally positioned in spiritual discussions collectively with ratings representing their polarity and strength. Developed various detection models using lexicon-based, n-gram-based, and deep learning-based approaches. The study focuses only religious class of hate and the models were confined to the Arabic language.

One of the attempts [5] also used convolutional Neural Networks to Detect Speech, to classifier assign each tweet to four redefined categories: racism, politics, both (racism and politics), and impartial speech. Character 4-grams, phrase vectors primarily based completely on semantic records constructed the usage of word2vector, randomly generated phrase vectors, and phrase vectors blended with individual n-grams. The feature set emerges as downsized inside the networks through the manner of way of max pooling and SoftMax function used to come across tweets. Tested with the useful resource of the use of 10-fold circulate validation, the model based mostly on word2vector completed best, with higher precision than recall, and a 78.3% F-score.

The other study [6] also English tweets annotated with three labels hate speech, offensive language but no hate speech, and no offensive content (OK) classes by using a linear SVM to perform multi-class detection for experiments the researcher used machine learning and statistical method of feature extraction such as character n-grams, word n-grams, and word skip-grams and obtained results of 78 % accuracy in identifying posts across three classes but we focus deep learning algorithms and word embedding feature.

Amharic is one of the sub-Saharan countries' Ethiopian working languages which is written left to right in its unique script which lacks capitalization and contains characters mainly consonant vowel pairs. It is the second largest Semitic language in the world after Arabic and is spoken by about 40% of the population as a first or second language [7]. Amharic is the legitimate language of Ethiopia. It is a Semitic language family that has the largest number of speakers after Arabic [1]. It is spoken, as per the 1998 census, by 17.4 million people as a mother tongue and 5.1 million people as a second language [2]. Amharic has five dialectical variations spoken in different Amharic regions, Addis Ababa, Gojjam, Gonder, Wollo, and Menz [3].

A set of 38 phones, seven vowels, and thirty-one consonants, makes up the entire stock of sounds for the Amharic language. Amharic consonants are generally classified as stops, fricatives, nasals, liquids, and semi-vowels. The vowels (ኧ, , ኡ , ኢ , ኣ , ኤ ,ኤ and ኦ) are categorized as rounded (ኡ and ኦ) and unrounded (ኧ , ኢ , ኣ , ኤ and ኤ) [8,9,10].

Amharic makes use of a script that originated from the Ge'ez alphabet It has 33 smooth characters with every having 7 one in every of a type for every consonant-vowel combination. In Amharic, seven vowels are used, each in seven distinct forms that reflect the seven vowel sounds they are ኣ ፣ ኡ ፣ ኢ ፣ ኣ ፣ ኤ ፣ ኣ ፣ ኦ cross pond to (ä, u, I, a, e, ï and o). There are 33 simple characters, every one of which has seven different relying on which vowel is to be stated inside the syllable. Unlike the North Semitic languages which include Arabic, Hebrew, or Syrian, the language is written from left to right. Amharic is written in a barely changed shape of the alphabet used for writing the Geʿez language. The alphabets are closely similar and only differ with the addition or omission of a few Ge'ez letters. [10] The version, that the Amharic language uses is known as Fidel or most commonly, the Amharic Alphabet. It includes extra letters such as ፕ (Ce), ሽ (She), and ቼ (Che) which are not in the Ge'ez script [11].

 doi:10.20944/preprints202503.0820.v1

3. Methodology of the Study

The goal of this research is to develop a hate speech detection model for the Amharic language using deep learning. Amharic is Ethiopia's national language, widely spoken across the country. However, political instability and ethnic tensions have led to an increase in hate speech on social media platforms. Social media is often used to spread political propaganda and malicious messages in Amharic, making it crucial to address this issue. This study focuses on detecting and mitigating hate speech in Amharic language social media content.to accomplish the study we used the subsequent methods and techniques.

3.1. Research Design

The study employed an experimental research methodology, a scientific approach where independent variables are manipulated to observe their effect on dependent variables. This method helps establish relationships between variables and draw meaningful conclusions. The research involves three main tasks: Preparation of hate speech data, Selection of implementation tools and Evaluation of the proposed hate speech detection model to measure its performance [12].

3.2. Data Collection and Preprocessing

Data was collected from social media platforms such as Facebook, YouTube, and Twitter using tools like Face pager. The collected Amharic text data was labelled by legal experts into five categories: racism, politics, religion, and positive and neutral speech.

3.3. Design and Development

The study utilized **Python**, an open-source programming language, for implementation and prototype development. Python was chosen its because of Free and open-source nature, large community support, and Compatibility with free cloud services like Google's Lightweight and user-friendly features [13].

3.4. Testing and Evaluation

The proposed hate speech detection model was evaluated using standard metrics such as precision, recall and F1-Score. These metrics were used to assess the model's accuracy and effectiveness in detecting hate speech.

3.5. System Design and Architecture

In this part of the study, we proposed a model for detecting hate speech on social media. For this proposal, we focused on the main components, the interactions between them describe tools and techniques. Discussed the proposed architecture of the recognition model, how the data preprocessing is performed, model recognizes the sentence-level components of hate speech and their subcomponents from a workflow perspective.

As shown in **Figure 1,** The architecture detects Amharic language hate speech posts and comments on social media. The proposed solution is primarily based on the architecture shown in Figure 1 It takes Amharic datasets as input from different social media and then preprocessed based on the language nature, which puts off punctuation, normalization, tokenize, and another basic necessary preprocess. After all the preprocessing, then feature extraction takes vicinity to extract by use of the Keras embedding layer. The output of this task is an essential feature vector of the dataset for training the model. After feature extraction, models were developed model with the use of machine learning algorithms (RNN deep learning) LSTM, BiLSTM, GRU, BiGRU, transformer deep learning (Roberta), and training set with feature vectors data frame of the whole dataset. The models evaluated the use of precision, recall, and f1- score. The evaluation result is used to select the best detection model. Finally, the detection model is evaluated and chosen primarily based on the

 doi:10.20944/preprints202503.0820.v1

consequences obtained from the usage of the model evaluation method. The final selected detection model is used to develop a prototype that can take new Amharic texts as input and it classifies the enter whether or not it contains political hate, religious hate, racist hate, or positive and neutral speech.

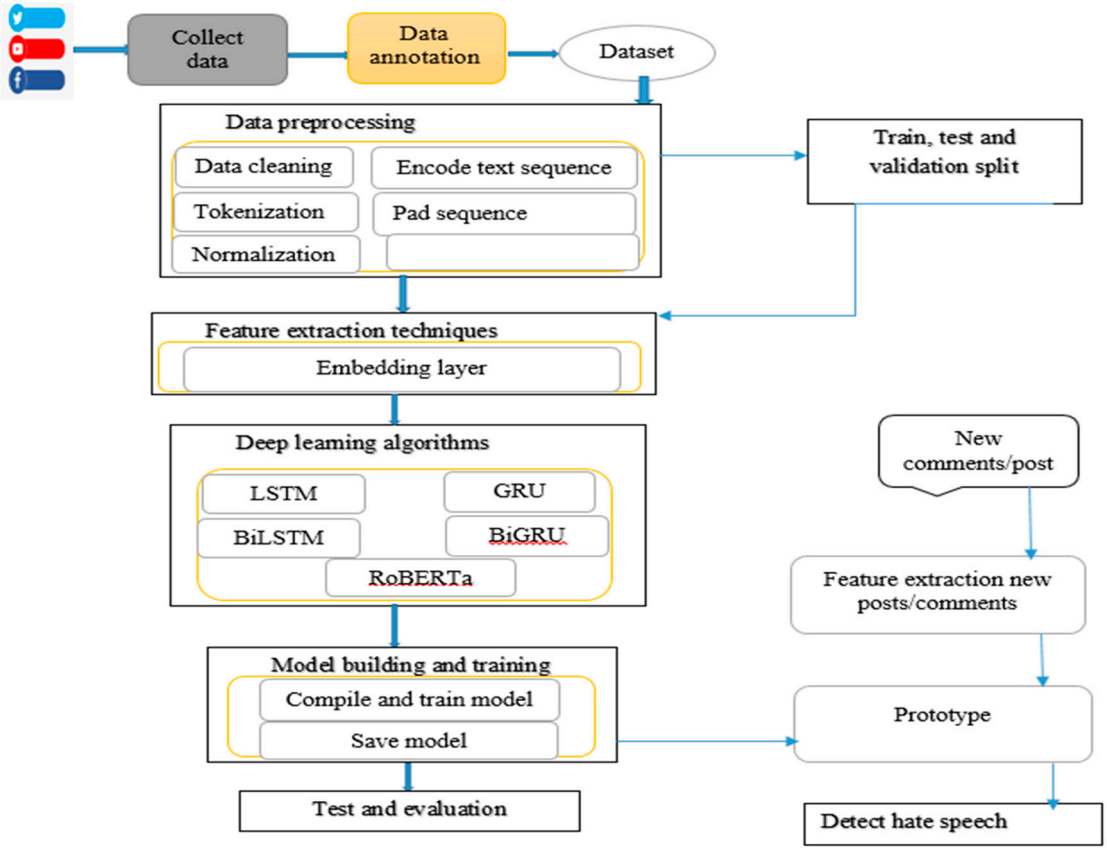

Figure 1. Architectural Design proposed Amharic hate speech detection.

4. Experiment Result and Discussion

In this study we used different experimental setups and conducted five experiments with deep learning such as Long Short-Term Memory (LSTM), Bidirectional Long-Short Term Memory (BILSTM), Gated Recurrent Unit (GRU), Bidirectional Gated Recurrent Unit (BiGRU), and A Robustly Optimized Bidirectional Encoding Representation Transformer (Roberta) with embedding layer.

To accomplish the experiment, we used different types of tools like Microsoft Excel 2013 to store the collected comments and posts before and after labelling. Microsoft Word was also used to prepare the text documents for writing and editing. The other one is the Google Collaboratory environment using a GPU usage of Python which assists one of the excessive levels, open source, and powerful deep learning library Keras constructed on top of Tensor Flow, which is used to create and share documents that contain live code for the data cleaning, statistical modeling, for the deep learning technique. To create the dataset, a Facepager has been used to gather comments and posts from different Facebook, Youtube, and Twitter accounts.

The research is divided into a variety of phases, setting out with the gathering of posts and comments, labelling of these posts and comments; preprocessing, model creation, splitting of the dataset training, and assessment of the hate speech detection. We have used datasets of 7590 with that 80:10:10 train, validate, and test splitting ratio, which means, 80% of the dataset used for training the model,10% of the dataset used for validating the model and the remaining 10 % of the dataset used for testing the model. in this experiment also observed that calculating the precision, recall, f1-score, and accuracy of each class by the formula described in chapter three and seen below figure

 doi:10.20944/preprints202503.0820.v1

here, the column support represents the number of samples that were present in each class of the test set. The macro-average was calculated as the arithmetic mean of individual classes' precision and recall.

	precision	recall	f1-score	support
positive speech	0.83	0.79	0.81	156
political hate	0.95	0.96	0.95	154
religious hate	0.95	0.99	0.97	170
racist hate	0.92	0.88	0.90	145
neutral	0.90	0.92	0.91	134
accuracy			0.91	759
macro avg	0.91	0.91	0.91	759
weighted avg	0.91	0.91	0.91	759

Figure 2. Performance evaluation of the Roberta model

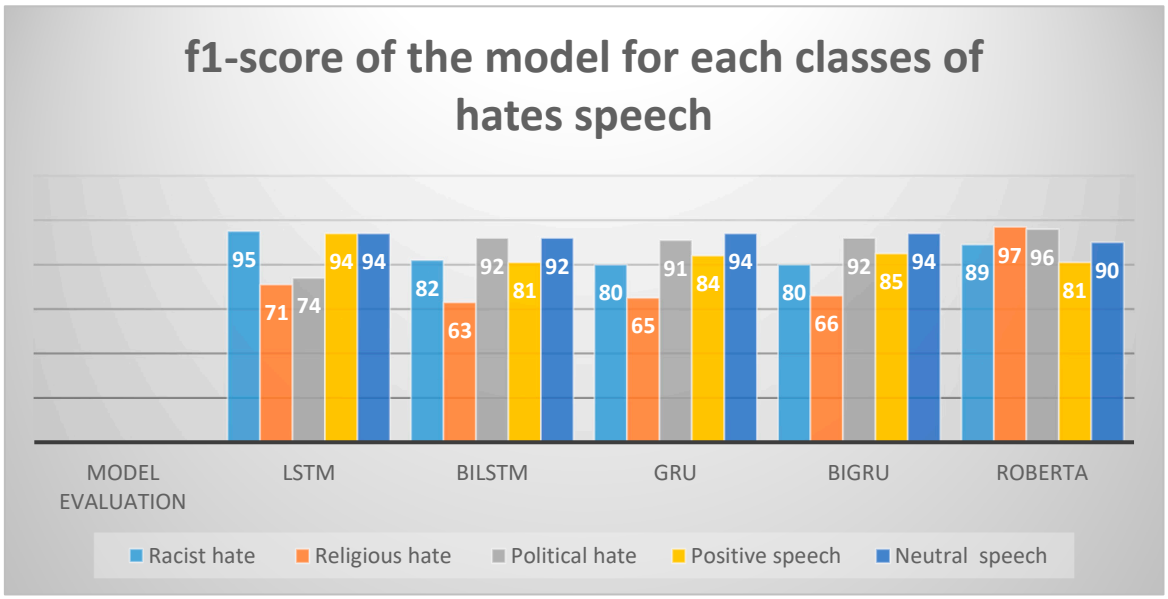

Figure 3. f1-score of each class of hate speech.

As shown in **Table 1,** in the five experiments, the performance result in Roberta shows higher performance than compared to others, the Performance of Roberta is 13% higher than GRU, 11% higher than LSTM, same 9% more than BiLSTM and 8% higher than BiGRU. Even though LSTM has three gates (input, output, and forget gate) in our experiment GRU better performance than LSTM and BiGRU better than BiLSTM because in terms of model training speed, GRU is faster than LSTM for processing the same dataset; and in terms of performance, GRU performance will surpass LSTM in the scenario of long text and small dataset. In this experiment, we observed that when compared the result of the accuracy of RNN deep learning with the transformer deep learning (Roberta) model using the embedding layer feature extraction technique was 91% which is the highest score of the other models.

 doi:10.20944/preprints202503.0820.v1

Table 1. summarizes the performance (in %) of the deep learning with each class.

Feature	Classes	Positive speech			Political hate			Religious hate			Racist hate			Neutral speech			Acc
	Model	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	
Keras embedding layer	LSTM	85	93	94	94	65	74	69	86	71	78	91	95	93	84	94	**79**
	BiLSTM	75	88	81	95	90	92	68	59	63	84	83	82	90	94	92	**82**
	GRU	86	81	84	92	91	91	71	59	65	72	89	80	93	94	94	**83**
	BiGRU	88	82	85	93	92	92	64	70	66	81	79	80	93	94	94	**84**
	Roberta	85	78	81	94	97	96	95	99	97	91	87	89	88	93	90	**91**

Transformer deep learning (Roberta) is Non-sequential, Self-Attention, and Positional embedding. Non-sequential that the sentences are processed as a whole in the desire to work with the resource of the usage of words. Self-Attention is used to compute similarity scores between words in a sentence. Positional embedding characteristic of Transformer deep learning the idea is to apply regular or positioned weights which encode facts associated with a selected function of a token in a sentence. Although the ultimate goal of our study has been detecting Amharic text hate speech as the class of racist hate, religious hate, political hate, and positive and neutral speech the Roberta model performance

Therefore, using contextual embedding in transformer-based deep learning works better for the Amharic hate speech task. We used Roberta tokenizer for using Byte Pair Encoding subword segmentation. This tokenizer class will tokenize raw strings into integer sequences and is based on keras nlp and also contains a contextual embedding layer. This study also experimented with embedding layer feature extraction techniques and RNN and transformer deep learning algorithms LSTM, BiLSTM, GRU, BiGRU, and Roberta model hate speech detection using Amharic text datasets evaluated by precision, recall, f1-score and accuracy.

4.1. Prototype

The raw textual content prototype has developed the usage of the Python programming language and the usage of Google Collaboratory. The developed raw textual content detection offers functionalities for inspecting user-generated entered texts. Users can enter texts into the raw text area by way of imparting the sentences one with the aid of one on the furnished space textual content, that is accessed by way of the user Then, the system generates the detected value of the sentences on the other raw text detect the class of hate speech.

 doi:10.20944/preprints202503.0820.v1

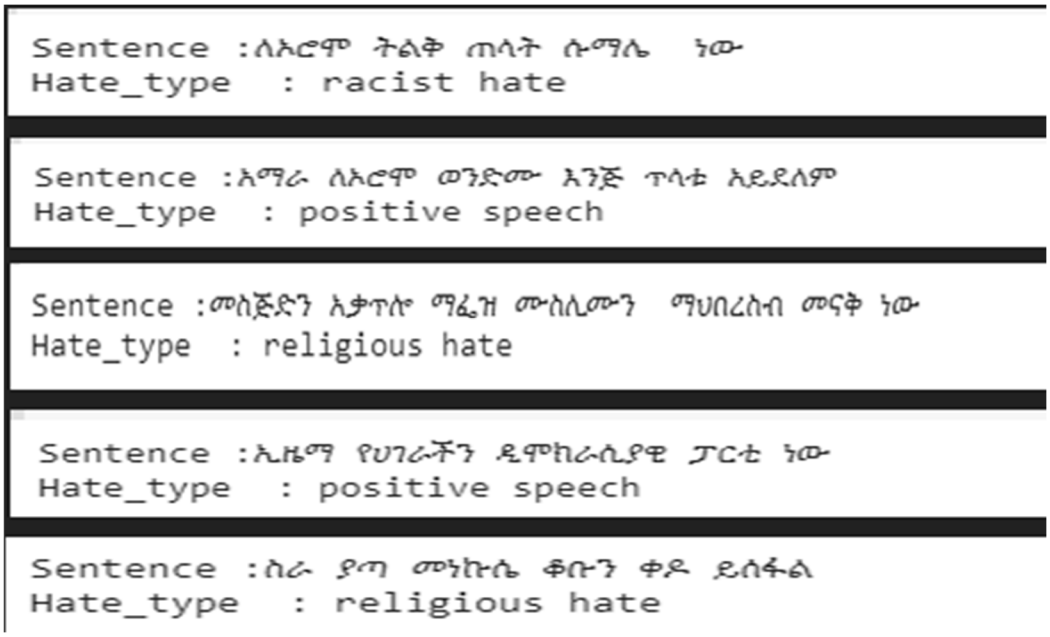

Figure 4. Shows Detect of the raw text or Amharic sentences.

5. Conclusion and Recommendation

Currently, on social media, there is a huge amount of data exchanged among users daily and poses a lot of influence on users' lives positively or negatively, among those huge amounts of data hate speech texts on Facebook, Youtube, and Twitter take the major role to affect users' life negatively by imitating them for conflict based on their race, religion, political which could differentiate individuals. In this research, we developed a model for hate speech detection for Amharic texts on Facebook, youtube, and Twitter. We collected 7590 cleaned Amharic posts and comments from suspicious social media public pages of organizations and individuals. The vital preprocessing steps like data cleaning, tokenization, and normalization were performed based on the requirement of the language to get a cleaned corpus. The dataset has five classes named hate speech such as racist, religious, political, positive, and neutral speech classes by experts based on the prepared Amharic annotation guidelines. In this thesis, we have used embedding layers feature selector techniques to generate word Vectors that can able to capture syntactic and semantic relations of words and randomly generated vectors with the use of the embedding layer. We have experimented using state-of-the-art deep learning algorithms such as LSTM, GRU, BiLSTM, BiGRU, and Roberta. Roberta with embedding layers and Adam's optimization algorithm achieved better performance with an accuracy of 91%. Finally, we developed a prototype for our model to detect and classify Amharic text hate speech within input raw text and detected text space.

The proposed solution is confined only to the Amharic language; however, there are more than eighty unique languages used in the country. For an extra complete detection model, the future researcher can focus on growing datasets and models for the dominant language spoken in Ethiopia such as Somali, afar, wolayta, Sidama, and other languages that are used on social media platforms. Additionally, posts and comments frequently comprise non-textual content images, emoji, and sarcasm or figurative speech that influence hate speech and may also include hateful expressions. Future research may want to centre attention on such kind of content.

References

1. Y. Kenenisa and T. Melak, "Adama, Ethiopia, September 2019," *Hate Speech Detect. Amharic Lang. Soc. Media Using Mach. Learn. Tech. By*, vol. Unpublishe, pp. 1–103, 2019.
2. Z. Mossie and J. Wang, "SOCIAL NETWORK HATE SPEECH," pp. 41–55, 2018.

 doi:10.20944/preprints202503.0820.v1

3. B. Emuye, "Amharic Text Hate Speech Detection in Social Media Using Deep Learning Approach," no. july, 2020.
4. N. Albadi, M. Kurdi, and S. Mishra, "Are They Our Brothers ? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere," 2018.
5. B. Gambäck and U. K. Sikdar, "Using Convolutional Neural Networks to Classify Hate Speech," no. 7491, pp. 85–90, 2017.
6. M. Zampieri, "Detecting Hate Speech in Social Media," pp. 467–472, 2017.
7. Z. Mossie and J. Wang, "SOCIAL NETWORK HATE SPEECH," no. April, 2018, doi: 10.5121/csit.2018.80604.
8. S. Teferra and W. Menzel, "Automatic Speech Recognition for an Under-Resourced Language-Amharic."
9. F. A. Melat, "Hate Speech Detection for Amharic Language on Facebook Using Deep Learning," pp. 1–23, 2022.
10. A. G. Debele and M. M. Woldeyohannis, "Multimodal Amharic Hate Speech Detection Using Deep Learning," *2022 Int. Conf. Inf. Commun. Technol. Dev. Africa, ICT4DA 2022*, no. December, pp. 102–107, 2022, doi: 10.1109/ICT4DA56482.2022.9971436.
11. S. G. Tesfaye and K. Kakeba, "Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network," 2020, doi: 10.21203/rs.3.rs-114533/v1.
12. M. Bhardwaj, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Hostility Detection Dataset in Hindi," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.03588
13. C. Ozgur, T. Colliau, G. Rogers, Z. Hughes, E. " Bennie, and " Myer-Tyson, "The Selection of Independent Variables for A Multiple Regression Problem Using LASSO methods," 2017. [Online]. Available:https://www.researchgate.net/publication/328175547