ACADEMY JOURNAL
OF
SCIENCE AND ENGINEERING
Available online @
www.academyjsekad.edu.ng
AJSE 17 (1) 2023
Original Research

HATE SPEECH DETECTION USING MACHINE LEARNING: A SURVEY

**Seble, H., Muluken, S., Edemealem, D., Kafte, T., Terefe, F.,
Mekashaw, G., Abiyot, B. and Senait, T.**

Cyber Security Research Division, Data Science, Information Network Security Administration (INSA),
Addis Ababa, Ethiopia

Abstract

Social media platforms provide an opportunity to create and grow anonymous online friends and followers, as well as an online forum for discussion about community life, culture, politics, and other topics. Therefore, hate speech is a growing challenge for society, individuals, policymakers, and researchers. This is the problem we are noticing in our continent and even in our world. Therefore, studies to identify, and detect hate speech are needed in terms of quality and performance. This paper provides a systematic review of literature in this field, with a focus on techniques like word embedding, machine learning, deep learning techniques, hate speech terminology, and other state-of-the-art technologies with their gaps and challenges.In this paper, we have made a systematic review of the last 6 years of literature.Furthermore, limitations, along with algorithm selection and use challenges, data collection, and cleaning challenges, and future research directions are discussed in detail.

Keywords: Hate Speech, Machine Learning, Deep Learning, social media

1.0 INTRODUCTION

The extensive use of social media not only allows for unlimited communication between people but also greatly increases the level of information exchange. Social media in Ethiopia is trusted by many as an important source of information and they tend to believe everything from these sources(Assefa, 2020). Social media in Ethiopia have become a new paradigm for the widespread dissemination of hate speech that threatens the safety of citizens(Getahun, 2023). Recently, we have seen a lot of chaos in Ethiopia due to misinformation and abusive thought spreading on social media. The spread of hate speech and fake news has affected the lives of many people(freedomhouse, 2021). In addition to the damage to education, schools, and universities have been destroyed, trade between cities has been severely disrupted by

Corresponding author's e-mail: desuking1717@gmail.com *Website: www.academyjsekad.edu.ng*

road closures, civil mobility has been severely disrupted, millions have been displaced, and Hundreds died(Morris Kiruga, 2019). Following the continuation of the Tigray conflict in Ethiopia, calls have been made to attack ethnic minorities on social media(MUNA SHIFA, 2022). It is remembered that the dispute between the Federal Government of Ethiopia and the Tigray People's Liberation Front (TPLF) resulted in the closure of media organizations and the arrest of journalists and bloggers(Fred Harter, 2022).

Hate speech and fake news are receiving international attention because they cause great damage. Some countries have created laws that respect freedom of speech, but preventing hate speech does not mean restricting or denying freedom of speech, but rather preventing hate speech from escalating into something dangerous(Brüggemann et al., 2022). United Nations Educational, Scientific and Cultural Organization (UNESCO) listed five ways to counter hate speech in the Media through Ethics and Self-regulation. The five ways are Education on media ethics, encouraging conflict-sensitive reporting and multicultural awareness campaigns, regulating social media, Encouraging victims and witnesses to report hate speech-related crimes, and Ending impunity against hate crimes(Poni Alice JameKolok, 2022).

However, since this document is a scientific survey that explores the state-of-the-art of hate speech and its challenges, it will also help to draw out the challenges in Ethiopian languages and help future researchers to find direction easily. Hate speech has become a problem in the world that contributes to the loss of lives for many people, the disintegration of the countries, and various historical fragments(Zachary Laub, 2019). This kind of disease can be cured not only by the awareness of the public about it but by a scientific solution supported by many studies. Identifying challenges and developing something that can be a direction for further research is one step toward a scientific solution. In this sense, the contribution of this survey is significant. Because the main purpose of this survey is not only to identify the existing challenges, but also to put the existing state of art knowledge in the field and to show the direction of the future.

1.1 Hate Speech

Any form of communication in speech, writing, or behavior that attacks or uses highly offensive or discriminatory language in reference to a person or group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender, or other identity factor(Lumenwaymaker, 2019). Hate speech is an offensive language that targets a person or group based on inborn traits like race, religion, or gender and may disturb the peace in a community. Alternatively, hate speech is any kind of communication that attacks or uses pejorative or discriminatory language concerning a person or a group based on who they are (Nazmine et al., 2021). In response to the rising levels of racism, xenophobia, and hate speech throughout the world in 2019, the UN released the UN Strategy and Plan of Action on Hate Speech. The UN acknowledged that hatred is becoming more prevalent. The proposal suggests a two-pronged strategy to combat hate speech: to address the underlying causes and to make it possible for the UN to effectively respond to the effects on society. Thirteen commitments were included United Nations Strategy and Plan of Action on Hate Speech, such as helping victims to report hate speech crimes, using social media to create awareness, and using education to prevent hate speech (UN, 2019).

Therefore, it is true that hate speech is having a great impact on social media nowadays. This is because social media is a fundamental part of our daily lives and we use it as an important source of communication, information, and entertainment. Studies have been done that can help our society protect itself from the hateful speech on social media. In addition to this, big media such as Facebook, Twitter, YouTube, and others are doing the job of distinguishing hate information from real information by developing models (Burnap & Williams, 2015). However, the increase in the number of languages and the morphological differences in interpretations continue to be a challenge to detect hate speech. Hate speech is often not based on just one identity. It can target based on gender, religion, race, and disability(Seglow, 2016).

1.1.1Gender-based Hate Speech

Expressions that spread, incite, promote, or justify sexist hatred are considered gender-based hate speech(Sękowska-Kozłowska et al., 2022). The victims of this type of hate speech are generally women and girls. In our world, there is much hate speech toward women because of their gender. This is known as sexist hate speech and is a form of social shaming that disrespects women and aims to promote fear and distrust of women in society. Easy access to the Internet, the rapid growth of technology, communication technologies, and social networking sites have exacerbated violence against women and girls. These advancements are being utilized to abuse women and girls(Violence, 2023). Online violence against women and girls is a worldwide issue. Social networks are the primary means of gender-based online harassment. Such harassment of women affects the personal lives and professional careers of women (Nova et al., 2019). Studies show that violence and harassment against women and girls in the society can be one of the reasons for a woman to join terrorist organizations (Edwards, 2017). (Rahman et al., 2018)Revealed that social media are at a high level of cyber harassment against women. Online bullying is a common aspect of digital life, particularly for young adults, who are more exposed to more serious types of harassing conducts. So, we can understand that something should be done to solve this problem. That is why this research paper synthesis work has been done to narrow down the research gap.

1.1.2 Religious Hate Speech

The hatred against religions like Islam, Hinduism, and Christianity(Kiper, 2023). As religion involves a group of people, hate speech against it is more harmful than an individual. Extremist individuals are subjected to negative stereotypes, discrimination, physical abuse, and violence online. Research shows that anti-Muslim abuse is increasing online, so it is necessary to address the issue of Islam on social networks(Ghasiya &Sasahara, 2022). The internet serves as an amplifier, reflecting and amplifying existing discourses into networks, resulting in a powerful polarizing influence.

Individuals use social media as a cover for their illegal activities to create misunderstandings, intolerance, hatred, and extreme debates between religions. When such an event occurs, it causes great tension among the followers of the religion(Asians et al., n.d.). This illegal activity is a common phenomenon in Europe, Asia, and Africa as a whole. One of the things that lead to the extreme push and hatred of religions is religious hate speech that is posted online(Strategic Communications, 2022). History has shown that if religion cannot be protected from individuals who want to use it as a tunnel for politics, it can become dangerous. As Ethiopia is a country of many religions, religious conflicts or hatred are directly or indirectly spread, so important things are being done through scientific research to prevent such things from happening again.

1.1.3 Racist Hate Speech

An expression directed towards the look of a person or group is referred to as racist hate speech(Association, 2017). These differences render certain racial groups inferior to others. Most of the time such speech is done at the international level. The frequency and impact of this speech depend on the needs and attitudes of a nation's government and vary from one leadership to another.

The number of people who do not hate racism is increasing in our world(Hate, 2020). They claim that they exercise their right to freedom by spreading and sowing racial hatred on social media. A scientific method is needed to single out people who have chosen to live in moral depravity by rejecting human equality. This synthesis of studies has something to contribute as it contains studies that can respond to this problem.

1.1.4 Hate speech on disability

Disability hate speech refers to hate speech directed at people having physical or mental disabilities to make them feel less than human. In the medical field, disability, like race and gender, is considered a social category instead of an independent reality. Disability can occur in humans due to various conditions(Runswick-Cole, 2014). Disability can happen to humans due to medical errors, accidents, and natural and other causes. However, it is seen that online social media users are harming people's living conditions by spreading hate speech that can harm people with disabilities(K. Saha et al., 2019). This kind of hate speech keeps people with disabilities from being seen as equals and from mixing with people. This can cause great tension in the community. Because the greatest wealth for human beings is to think that they are equal to other people. Therefore, this type of thing should be presented to human beings in the form of education, people should improve themselves, take care of their brothers and sisters, and forget their disabilities.

1.2 Stages of Hate Speech

Hate speech is divided into four stages(Chetty, Naganna,Alathur, 2018). The first is called the influence stage, and this is because there is a heavy flow of traffic on every social media as soon as the event occurs. This will aggravate the hate speech. After a few days, the impact of the event will decrease and this level is called the intervention stage. After some more days and after a long time once again, the impact will reduce to zero level which is called the response stage. A dashed line on the figure shows an optional stage in the rebirth stage. Hate speech may or may not reappear over a lengthy period, depending on the nature and impact of the occurrence. Figure 1 shows different stages of hate speech.

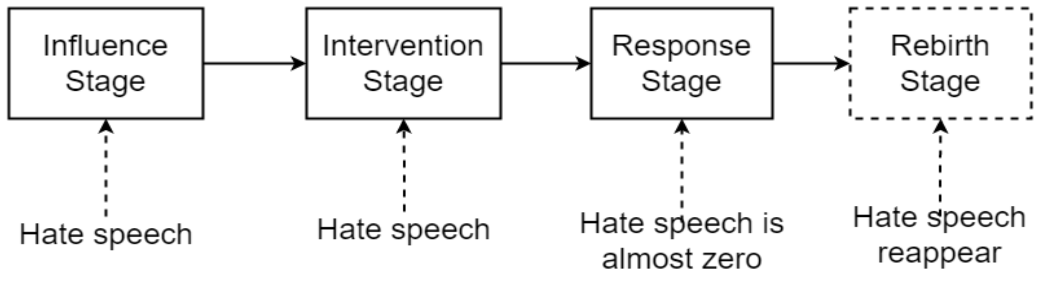

Figure 1:Stages of hate speech (Chetty, 2018)

1.3 Hate Speech Techniques

Mostly Studies have used three types of techniques to identify hate speech. They are Keyword-Based Techniques, Machine Learning Techniques, Deep learning Techniques, and hybrid. It has been tried to see which of these are the most used by the researchers. The overall composition is given below:-

1.3.1 Keyword-Based Technique

The keyword-based approach is a basic technique for identifying hate speech(Njagi et al., 2015). By using a dictionary, the text contains potentially hateful keywords(Njagi et al., 2015). For instance, collecting hate keywords from various social media platforms such as Facebook, Twitter, blogs, forums, and YouTube is one ways to organize hateful keyword data. Although the keywords collected might change their meaning depending on the time and situation, it should be clear that they are mostly used for disgusting and hateful actions. However, it is not possible to detect hate speech by just using hateful slur words to identify hate speech. The keyword-based hate speech system has severe limitations. The content contains hate speech but not the keyword might not be marked as hate speech. This is one of the challenges of the keyword-based approach. For example: "*መንጋውእንደተለመደውተነሳ*" This literal interpretation shows a herd of animals rose together, From the point of view of the politicians, they interpret it as a blindly traveling society as class or religion, so they change the meaning according to the situation of the time.

Furthermore, keyword-based approaches cannot identify hate speech that does not contain any hate keywords (eg, metaphors or slang)(MacAvaney et al., 2019). Slang such as "*አህያለአህያቢራገጥጥርስአይራገፍም*" literally means it shows donkeys are not careful when playing. However, with the political context, some interpret this as a catalyst for conflict and war between religion and society.

1.3.2 Machine Learning Technique

The scientific approach of algorithms and statistical models that computer systems use to perform a specific task effectively without being programmed, instead relying on patterns and data(Mahesh, 2019). It is often seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of trained data to make predictions or decisions without being explicitly programmed to perform the task. Machine learning aims to construct a classifier or regression model by learning a training dataset and evaluating the performance of the classifier or regression model using test data. Machine learning can be categorized as Supervised, unsupervised, or semi-supervised learning.

1.3.3 Deep Learning Technique

It is a machine learning technique that teaches computers to do things that come naturally to humans: learn by example. A deep learning approach uses neural networks to solve more complex problems innovatively(Alzubaidi et al., 2021). Using deep learning a computer model can learn to carry out classification tasks directly from images, text, or sound. State-of-the-art accuracy can be attained by deep learning models. Sizable labeled data and multi-layered neural network architectures are used to train models(Sarker, 2021).

1.3.4 Hybrid Technique

It is a method used to overcome the limitations of an approach. Because each solution has its own set of limitations. And combining two or more approaches into a hybrid approach where they complement each other seems like a good solution.

2.0 RELATED WORKS

In this section, studies conducted in various techniques were reviewed and discussed.Those techniques are keyword-based, machine learning, deep learning, and hybrid techniques. It is presented as follows.

2.1 Keyword-Based Technique

It's a technique that works by collecting and organizing keywords in each context. When analyzing keywords, it counts how many of those keywords are found in a document and gives an idea about the document. It is presented below:

Yimam et al., (2019)analyzed the Ethiopic Twitter Dataset for Abusive Speech in Amharic. Primarily in this study, the data was collected to train linguistic models for language identification tasks and to analyze the distribution of selected keywords for abusive language. The textual data were obtained only for Amharic, Tigrinya, and Ge'ez languages. This study collected around three million tweets from 154,477 users from mid-August 2014 till 2019. To analyze the distribution of keywords in the General reference corps and Ethiopia tweets dataset 99 hate speech and 48 offensive speech keywords for the Amharic language were collected from five native speakers. The native speakers collected the keywords from Facebook posts and comments, Twitter tweets and re-tweets, and YouTube comments from popular pages. In this study the year 2015 data was not analyzed due to an encoding error. The five native speakers collected 147 keywords and categorized them into hate and offensive speech. According to this, the research indicated that the number of Amharic tweets as well as the number of tweets containing offensive keywords is increasing from time to time.

2.2 Machine Learning Techniques

The study in Thomas et al., (2017) was done with a focus on the quality of data. The study began with a hate speech lexicon containing words and phrases identified as hate speech by Internet users, compiled by Hatebase.org. Twitter API was also used to search tweets containing terms from the lexicon resulting in a sample of tweets from 33,458 Twitter users. The timeline for each user was extracted and got around 84.4 million tweets. From this data extracted, the author took 25k tweets containing terms from the lexicon and manually labeled them by CrowdFlower workers. Workers were asked to label each tweet as hate speech, offensive but not hate speech, or neither offensive nor hate speech. The workers were asked to think not about the words existing in a given tweet, but also about the context in which they were used. In this approach, all offensive words did not necessarily indicate hate speech. Each tweet was coded by three or more people. Each tweet was encoded by three or more people by using an inter-coder agreement by using the majority decision for each tweet to assign a label. Only 24,802 coded samples were taken from the total dataset as the majority of tweets were not coded by the agreement of all people. From this 5% of tweets were coded as hate speech by the majority of coders and 1.3 were coded without opposition from encoders. Because the study used stricter criteria to classify hate speech, most tweets were classified as offensive language and the rest were deemed not offensive. The Porter stemming algorithm was used for stemming and TF-IDF (Term Frequency - Inverse Document Frequency). To capture the quality of each tweet the study use modified Flesch-Kincaid Grade Level and Flesch Reading Ease scores were used. Then after the study used five different classical algorithms which are logistic regression, naive Bayes, decision trees, random forests, and linear SVMs. rather than using default parameters, the Grid Search() function was used to iterate over the data to select the best-tuned parameters. 5-fold cross-validation, the technique was also used to split the data to prevent overfitting. The Logistic Regression and Linear SVM tended to perform significantly better than other models. But based on the previous works Logistic regression was selected. The best-performing model has an overall precision of 0.91, a recall of 0.90, and an F1 score of 0.90. This study focused entirely on quality data. This is to understand the context in which offensive language and hate speech inflame race, religion, and identity in society.

Authors in Mossie & Wang, (2018) aimed to build a hate speech detection model on the Amharic language. The authors have built a corpus of comments retrieved from Facebook public pages of Ethiopian newspapers, individual politicians, activists, TV and radio broadcasts, and groups. Authors employ Facepager, a versatile Facebook crawler that uses the Graph API to extract the content of comments from Facebook posts. To preprocess the extracted information some rules were followed: - such as only Amharic comments were kept, all null values were removed, HTML tages and any other symbols are removed, and checked to assure that no repetitions, and all elongations were removed to the same fixed size character. After preprocessing authors considered three bases for future annotation like Discourse analysis, Content analysis, and Automated techniques. After the first cleaning authors got 25,850 posts and comments but due to limitations in resources for the annotation task authors sampled 10,000 posts and comments. Three Ph.D, 2 MSC students, and 1 assistant professor from Amharic Language studies were selected as annotators. To annotate the comments or opinions the author had identified the categories of the opinion or comments if categorized as politics, ethnicity, religion, and socioeconomic. In addition to the annotation rule, kappa agreement was also used. For Feature selection, both Word2vec and TF-IDF were used. The Naive Bayes and Random Forest algorithms were trained on 4,882 posts and evaluated on 1238 raw data. For model evaluation 10-fold cross-validation was used. The model based on word2vec embedding performed best with 79.83%accuracy. The proposed method achieved a promising result with the unique feature of spark for big data.

Yimam, Ayele and Biemann, (2019)(Yimam et al., 2019)the study defined the toxic language and divided toxic language into three categories: hate speech, offensive language, and clean. Based on the previous researches the author used n-gram for feature extraction and weight them according to their TF-IDF values. The dataset was obtained publically from Crowdflower and Github which contains tweets that have been manually classified into Hateful, Offensive, and Clean. The data obtained from different sources were integrated by using Twitter API. Unnecessary content from the tweets like Space Patterns, URLs, Twitter Mentions, Retweet Symbols, and Stopwords were removed by converting the tweets to lowercase. Porter Stemmer algorithm was used to reduce word inflection. For the model building, the author selected three prominent machine learning algorithms like Logistic Regression, Naive Bayes, and Support Vector Machines. The grid search algorithm was used for hyperparameter tuning and performing 10-fold cross-validation. The results showed that Logistic Regression performed better with the optimal n-gram range of 1 to 3 for the L2 normalization of TF-IDF. Upon evaluating the model on test data, the promising Logistic regression model achieved 95.6% accuracy.

The study in Raufi & Xhaferri, (2018) aimed to build a lightweight machine-learning model for detecting hate speech in the Albanian Language for mobile applications. 10,268 raw data was collected from the local Albanian forum jeta osh qef and Xing me Ermalin. From whole collected data 6648 data is offensive and 3620 data is Normal. The dataset collected from forums was small so the author used simple resample and SMOTE (Synthetic Minority Oversampling Technique) to resample and balance the data. The training process used both 10-fold cross-validation and a 30/70 percent split. For the model building, the author selected an artificial neural network Classifier. Here different experiments were conducted in this study. The first is training the model with SMOTE with a 10-fold CV, Resample with a 10-fold CV, SMOTE with a 30/60 percentage split, and Resample with a 30/70 percentage split. SMOTE with a 30/70 percentage split shows good accuracy. Therefore, the model learned in 30/70 produced good results, and the researcher built a mobile application as a prototype.

Arabic Offensive and Hate Speech Detection Aldjanabi et al., (2021) using a Cross-Corpora Multi-Task Learning Model, aimed to develop a Classification system for determining offensive and hate speech using a multi-task learning (MTL) Model built on top of a pre-trained Arabic language model using three different available Arabic offensive and hate speech datasets, such as OSACT, L-HSAB and –HSAB. According to the result of the experiment the developed MTL model showed better classification performance and outperformed existing selected models.

Authors in Lata Guta kanessaa, (2021) focused on Hate Speech Detection Framework from Social Media Content: The Case of Afaan Oromoo Language. The data collected from the number of likes and followers of the page must be greater than 10,000, which allowed more active public pages. The data was collected by using Facepager and ScrapeStorm software. After the data was collected, pre-processing of texts was held such as stop words removal, removal of unnecessary characters, removing all non Afaan Oromoo and non-textual posts and comments, short word expansion and stemming. Annotation development, mainly achieved by the researcher, also performed annotation with two additional annotators. The researcher gave brief insights into the annotation guidelines provided for labelling posts and comments into the binary classifier. After finalizing the annotation, the dataset was given to the respective model as input in csv format. During training time this data was split into two with 80:20 ratio for training and testing purposes. Used python programming language for implementing and experimenting with each proposed solution from the data pre-processing to the model building steps. In the end, the Flask framework was used to develop a web application since it provided tools, libraries, and technologies that are useful to build a web application. The author used four machine learning classification algorithms (Support Vector Machine, Logistic Regression, Naïve Bayes and Random Forest). Support Vector Machine, Logistic Regression, Naïve Bayes, and Random Forest used the same TF-IDF feature and the result shows 96.0%, 94.0%, 94.0%, and 94.0% average accuracy were achieved respectively. From the classification machine learning algorithm, SVM outperformed the TF-IDF feature extraction techniques which achieved 0.96 percentage of accuracy for two classes of classification.

The authors in admin et al., (2022) used two types of datasets English and Malayalam. The Malayalam data set is created.3400 English data sets and 1700 for Malayalam. After the data was collected, filter out corrupted data, Change the given data into a lower case. Also, removed all the URLs, usernames, white areas, hashtags, punctuations and stop-words using sample matching strategies from the collected speech. Stemming and tokenization were also used because the machine learning algorithms cannot understand the classification rules from the raw text. These algorithms need numerical features to understand classification rules. Authors used TF-IDF and bag of words techniques. During training time this data was split into two with 70:30 ratios for training and testing purposes. The authors first, built machine learning algorithms model for English language with TF-IDF features. SVM, logistic regression and random forest were different machine learning algorithms used for experiment. For SVM, logistic regression and random forest the result showed 90.0%, 81.0%, and 86.0% average accuracy. Secondly, machine learning algorithms model for Malayalam language with TF-IDF features was also built. And SVM, logistic regression and random forest; the authors recorded 94.0%, 90.0%, and 92.0% average accuracy. The performance evaluation of the selected machine learning models was done with the help of confusion matrix. Results showed that SVM gave the best performance with 90% Accuracy Score for English dataset and 94% accuracy for Malayalam dataset.

Automatic Hate Speech Detection using Machine Learning: A Comparative Study was the focus in Abro et al., (2020). This paper aimed to compare the performance of three feature engineering techniques and eight machine learning algorithms to evaluate their performance on a publicly available dataset having three distinct classes. Authors collected publicly available hate speech tweets of 14509 datasets and applied different pre-processing techniques to filter noisy and non-informative features from the tweets. The tweets were changed into lowercase and then removed all the URLs, usernames, white spaces, hashtags, punctuations, and stop-words using pattern-matching techniques from the collected tweets. Tokenization and stemming were used to to find the root word. The tweets were labelled by CrowdFlower into three distinct classes, namely, hate speech, not offensive, and offensive but not hate speech. In all 24 analyses, the lowest precision (0.58), recall (0.57), accuracy (57%) and F-measure (0.47) found in MLP and KNN classifier using TFIDF features representation with bigram features. Moreover, the highest recall (0.79), precision (0.77), accuracy (79%) and F-measure (0.77) were obtained by SVM using TFIDF features representation with bigram features. In feature representation, bigram features with TFIDF obtained the best performance as compared to Word2vec and Doc2vec. However, there was a fringe difference between the result observed in bigram, and Doc2vec. In text-classification models, the SVM classifier best performed among all the eight classifiers. However, the AdaBoost and RF classifiers results were lesser than SVM results and were better than

LR, DT, NB, KNN, and MLP results. The experimental results showed that the SVM algorithm with the combination of bigram with TFIDF FE techniques gave the best result.

Authors in Defersha, Naol Bakala Tune, (2021) studied the Detection of Hate Speech Text in Afan Oromo Social Media using Machine Learning Technique. Social media is used as a source of data for research. The authors collected 13600 comments and posts between September 2019 and 2020 on the respective public page using Face pager in which 7000 and 6600 data were collected from Twitter and Facebook. After the data was collected, pre-processing steps such as Spell correction, cleaning punctuation marks, special symbols, emoji, numbers, URL, and stop words and converting the upper case to lower case was done. The authors used five experts to annotate data depending on the annotation procedure prepared. The number of experts is limited to five due to resource scarcity. Experts recruited for data annotation were MA and above MA holders. Using N-Gram and TF-IDF for feature extraction. The results of the study indicated that Linear support vector Classifiers achieved Performance Precision of 66%, recall, of 66%, and F-score of 64%. The Multinomial NB achieved a performance Precision of 60%, recall of 65%, and F-score of 62%. The Random Forest classifier achieved a performance Precision of 64%, recall of 64%, and F-score of 63%. The Logistic Regression classifier achieved a Performance Precision of 65%, recall of 64%, and F-score of 61%. The Decision Tree Classifier achieved the Performance Precision of 59%, recall of 59%, and F-score of 59%. The result of the experiment showed that the performance of the Linear Support Vector Classifier scored an f1-score value was+ 64%. The authors confirmed that Linear Support Vector Classifier scored the highest performance compared with others. Therefore, the researchers agreed to use linear support vector classifiers to deploy Afan Oromo hate speech detection mode.

2.3 Deep Learning Techniques

Authors in Badjatiya et al., (2017) applied a deep learning approach for hate speech detection in tweets. The tweet was classified as racist, sexist, or neither. For experimentation 16K annotated dataset was made available(Waseem & Hovy, 2016). The 10-Fold Cross Validation and calculated weighted macro precision, recall, and F1-scores were used. In this paper, different experiments were conducted with multiple classifiers such as Logistic Regression, Random Forest, SVMs, Gradient Boosted Decision Trees (GBDTs), and Deep Neural Networks (DNNs). Three deep learning architectures were also used such as FastText, CNNs, and LSTMs to find the best algorithm with different embedding approaches. For each of the three methods, the author initialized the word embeddings with either random embeddings or Global Vectors for word representation (GloVe) embeddings with embedding size of 200 for GloVe as well as for Random Embedding. For comparative analysis, state-of-the-art methods were utilized like char n-grams, TF-IDF, and Bag of Words Vector approach (BoWV) as a baseline. Balanced SVM and GBDT with TF-IDF, SVM and GBDT with BoWV, and logistic regression with Char n-gramwere used together. In addition to the baseline, the author conducted two different experiments. First, the author experimented with three selected deep learning models (LSTM, FastText, and CNN) with GloVe and Random Embedding. Secondly, the author compared three selected deep learning models, ensembled with Global Vectors for Word Representation (GloVR), Random Embedding, and Gradient Boosted Decision Trees (GBDT). Finally, LSTM scored best with an accuracy of 93% when combined with Random Embedding and GBDT.

Comparative analysis of deep learning based on Afaan Oromo hate speech detection was carried out in Ganfure, (2022). The main focus of the study was to presents an empirical evaluation of five deep learning models (i.e., CNN, LSTM, GRU, BiLSTM, and CNN-LSTM) for detecting Afaan Oromo hate speech by conducting experiments and to prepare Text dataset for Afaan Oromo hate speech detection. The author of this paper retrieved 35,200 comments and posts published on Facebook and Twitter public pages from January 2019 to June 2019. To remove the noise from the data set, rigorous preprocessing was carried out, which resulted in the removal of HTML, URLs, tags, emoticons, and other language scripts. And this dataset was annotated by the language experts into four classes (neutral, hate, offensive, and both). To present the results investigated, three series of experiments were conducted using the five different deep learning models. The First one involved the case where the word embedding is pre-trained and is used for feature extraction; whereas the second one is the case where word embedding is trained together with the model itself and the third experiments were conducted to assess the impact of data augmentation on classification performance. In the first experiment, the BiLSTM and CNN accomplished the best performance (with a weighted average F1-score of 87%). The average F1-score of CNN-LSTM, GRU, and LSTM were 85%, 86%, and 82%, respectively. By comparing the experimental results of the neural network, first, a model trained with embedding representation captured syntactic and semantic relations of Afaan Oromo words. Secondly, the data augmentation mechanism improved the performance of the hate detection models. Finally, BILSTM achieved the highest F-score of all classifiers used in the experiments. In conclusion, considering the size of the data set examined in this paper, the performance of the deep learning model in detecting Afan Oromo hate speech is promising.

Automated Amharic Hate Speech Posts and Comments Detection Model Using a Recurrent Neural Network was the work done in Tesfaye & Tune, (2020). Aiming to develop hate speech posts and comments detection models using a deep learning approach and to prepare a labeled dataset for Amharic hate speech detection. The research began with the literature review covering the traditions that approached online hate speeches from complementary perspectives, including the legal literature that studied how hate speech is addressed in different continents and countries on social media.30,000 data were collected manually from mostly followed pages of activists and news pages and annotated to the binary class of hate and free speech based on the guidelines given by the researcher and pre-processed by applying data cleaning and normalization techniques. A Recurrent Neural Network was developed by using LSTM and Gated Recurrent Unit (GRU) with word n-grams for feature extraction and word2vec to represent each unique word by vector representation. Based on the dataset the LSTM and GRU model were trained and tested by splitting the dataset into a training, validation, and test sets using the split ratio of 80:10:10. The experiment performed with different parameters on GRU and LSTM based RNN model by feature representation of word2vec resulted in better test accuracy of 97.9% by RNN-LSTM.

The study conducted on Das et al., (2021) , Bangla hate speech detection on social media using attention-based recurrent neural network. For this study 7,425 comments were collected from Facebook with 80% and 20% training and testing set respectively. For this study encoder–decoder-based machine learning models such as the attention mechanism, LSTM, and GRU-based decoders were used for predicting hate speech categories. Among the three encoder–decoder algorithms, the attention-based decoder obtained the best accuracy of 77%.

Conducted a study on Samuel, (2012) Hate Speech Detection and Classification System in Amharic Text with Deep Learning. The author collected a total of more than 1 million data from social media: Facebook, Twitter, and YouTube using both manual and automatic ways. For the automatic collection method, the researcher used the FacePager tool and Twitter API. The author consolidated every data and filtered racial, religious, and gender hate speeches using their list of hate speech keywords. The keywords were collected by analyzing some sample hate speeches. These identified keywords include 14 gender keywords, 30 religious keywords, 168 hate-related keywords, 70 offensive keywords which can be a head start for hate speeches, and 56 known Ethiopian popular ethnic group names. From 1 million data collected after filtering 162,179 data were remaining.The collected data was named in two rounds by people from different areas so that the data was finally reduced to 5000 speeches. The annotation was done by 100 annotators who have different demographic and sociocultural backgrounds, besides giving the developed guideline. During training time this data was split into three with an 80:10:10 ratio for training, validation, and testing purposes. The algorithms for pre-processing and model training were developed using Python and the hate speech detection model was developed using Google Collab. The author applied the word embedding technique by using FastText for vector representation. For feature extraction, TF-IDF and N-gram were used. To measure the model performance; accuracy, precision, recall, and f1-score were used. The researcher first, builds a dummy classifier model with different classifier strategies: stratified, most frequent, prior, and uniform with TFIDF features. The lower average accuracy was 26.94% using the “Stratified” classifier strategy by applying TFIDF features. The higher accuracy achieved was 40.19% using the “Most Frequent” classifier strategy by applying TF-IDF feature engineering techniques. Secondly, author built a classical machine learning model: Linear SVC, Logistic Regression, Multinomial NB, and Random Forest classifier and the result showed 80.3%, 72.1%, 70.4%, and 41.2% average accuracies. Finally, Deep Learning: Stacked Bidirectional LSTM-based RNN model was applied with different hyper-parameter value combinations, and the highest accuracy of 94.8% was achieved. Lastly, the author introduced an approach for classifying the hate speech of Amharic Twitter, posts, and comments. The classical Machine learning model, deep learning model, and dummy classifier model. From these three models, the deep learning model has shown the highest accuracy result in comparison with the two baseline approaches by having a 94.8% accuracy result while the dummy classifier scores 40.1% accuracy and the classical machine learning scores 80.3% accuracy.

The authours Getachew & Kakeba, (2020), began their research with preparing labeled Amharic dataset and they used recurrent neural network algorithm with LSTM and Gated Recurrent Unit (GRU) with word n-grams for feature extraction and word2vec to represent each unique word by vector representation. They achieved accuracy of 97.9% with the LSTM based RNN model which has better performance.

HASOC provides a forum and a data challenge for multilingual research on the identification of problematic content. Based on this team(Mohtaj et al., 2022)was one of

HASOC (Hate Speech and Offensive Content Identification) forum competitors in 2021. There were two tasks raised by the forum Subtask 1A is a coarse-grained binary classification task where tweets should be classified into two classes: (NOT) Non Hate-Offensive: These posts do not contain any hate speech, profane or offensive content, (HOF) Hate and Offensive: These posts contain hate, offensive and profane content. Subtask 1B is a three-class classification task offered for English and Hindi, where hate speech, profane and offensive posts from subtask 1A were further classified into the following categories (HATE) Hate speech: this class contains posts which hate-speech content, (OFFN) Offensive: posts in this class contain offensive content, (PRFN) Profane: posts in this class contain profane content. They tested different NLP models like recurrent neural networks in word and character levels and transfer learning approaches based on Bert for the two sub tasks and achieved the best result with transfer learning approaches based on Bert.

This team(P. Saha et al., 2019)was one of HASOC (Hate Speech and Offensive Content Identification) forum competitors in 2019. There were three tasks raised by the forum for three languages: Hindi, English, and German. Dataset in Hindi and English had three subtasks each, while German had only two subtasks. Sub-task A predict if a given piece of text is hateful and offensive (HOF) or not (NOT). All the three languages have this sub-task. Sub-task B predicts the three different classes in the data points annotated as HOF: Hate speech (HATE), Offensive language (OFFN), and Profane (PRFN). Again all the three languages have this sub-task. Sub-task C predicted the type of offense: Targeted (TIN) and Untargeted (UNT). Sub-task C was not conducted for the German dataset. They generate two types of feature vector: the BERT and LASER Embedding, which were then concatenated and fed as input to the final classifier. They use Light Gradient Boosting Machine (LGBM) for classification because the amount of data in each category was insufficient to train a deep learning model. The working team got the first position in the German sub-task with a macro F1 score of 62%. This means that the accuracy of the model is about 62% when compared to other algorithms used in the study.

2.4 Hybrid Techniques

Multiple deep learning and machine learning algorithms work together to complement and augment each other this refers to hybrid techniques. This type of method is often used by researchers to develop effective models.

A study entitled (Ababu & Woldeyohannis, 2022) ‘Afaan Oromo Hate Speech Detection and Classification on Social Media’ was conducted at the School of Computing, Dire Dawa University Institute of Technology, School of Information Science, Addis Ababa University in 2022. To develop and test a model used to detect and classify Afaan Oromo hate speech on social media, total of 12,812 data was collected from the Facebook account, which has most frequent posts and comments in Afaan Oromo languages (which has a minimum of 500 followers on Facebook).after this data was preprocessed, they use different machine learning algorithm from classical (SVM and NB), ensemble (RF and XGBoost) and deep learning (CNN and BiLSTM) with different feature extraction techniques such as BOW, TF-IDF, Word2vec and embedding layer. To test the result, two experiments were performed with eight and two classes. From classical and ensemble machine learning algorithm SVM is outperformed machine learning algorithm with word2vec feature extraction techniques which achieved 82 % of accuracy for eight class classification and from the deep learning algorithm, BiLSTM algorithm achieved better accuracy of 84% with pre-trained word2vec feature extraction techniques for eight class classification. BiLSTM achieved better performance result of 0.88 percent accuracy with pre trained word2vec. Finally, authors recommended that further research is required for hate speech detection for audio, video, emoji, memes models that detects and classifies hate speech from social media with multilingual language.

Astudy entitled as (Mnassri et al., 2022) ‘BERT-based Ensemble Approaches for Hate Speech Detection’. The study mainly focused on classifying hate speech in social media using multiple deep learning models implemented by integrating recent transformer-based language models such as BERT with several NNs such as MLP, CNN and LSTM to enhance hate speech detection performance via ensemble learning. The analysis was based on three publicly available Twitter datasets, such as Davidson, hateval2019, OLID that was generated to identify offensive languages. And the authors, fused all these datasets to generate a single dataset (DHO dataset), which is more balanced across different labels, to perform multi-label classification. First, the study assessed the contextual information derived from BERT, next fine-tune them using the datasets to get its contextual representations and then, ensemble models (combining BERT+MLP, BERT+CNN and BERT+LSTM) with several ensemble learning techniques: aggregation and stacking with several voting techniques such as Soft Voting or averaging, Maximum voting, Hard voting and Stacked Generalization ensemble, aiming to improve performance and robustness, and to get better classification. As for the aggregation ensembles, all the approaches outperformed single models, it shows obviously better results, especially the Soft Voting of BERT+LSTM with BERT+CNN, as well as Hard Voting ensemble that outperformed both of the other aggregation ensembles. Moreover, aggregation ensembles outperformed each of these single models, getting the best result when ensemble the 2 most performed models: BERT+MLP and BERT+LSTM. Unlike BERT-CNN, BERT-MLP and BERT-LSTM gave the best performance on DHO and Davidson datasets respectively.

The authors (Ababu & Woldeyohannis, 2022) Developed and test a model used to detect and classify Afaan Oromo hate speech on social media. In the data collection process firstly, authors searched the data on Facebook by using the respective thematic areas keywords. The keywords are selected by the domain expert based on the four thematic areas such as gender class (related), religion class (related), race class (related), and offensive class (related). If the data is related to their thematic area then check the account has a minimum of five hundred followers or members. If the data is achieved by both criteria they are scraping the posts or comments by using Facepager and/or data miner open-source tools. The researcher has collected a total of 12,812 posts and comments from Facebook. After collecting data remove special characters, emojis, punctuation marks, HTML tags, and stop words. Used Tokenization and normalization. Then every four categories of the dataset are annotated individually by three persons who are voluntary to do the task then apply mode to the annotation of three annotators and select the annotation upon which two persons agreed upon it. In the first experiment, the authors used eight classes of classification that is implemented with different machine learning algorithm (classical, ensemble, and deep) and with different feature extraction techniques such as BOW, TF-IDF, Word2vec, and embedding layer. Firstly, from the classical classifier, the highest accuracy which is 0.82 is recorded by SVM with word2vec feature extraction. This is because word2vec feature extraction is capture more semantic and syntactic text data than BOW and TFIDF feature extraction techniques. However, the low accuracy recorded a score of 0.74 is recorded by the Naïve Bayes algorithm with word2vec feature extraction. Secondly, from the ensemble classifier, the highest accuracy which is 0.81 is recorded by both RF and XGboost with word2vec feature extraction. Thirdly, from the deep learning classifier, the highest accuracy which is 0.84 is achieved by BiLSTM with pre-trained word2vec feature extraction techniques. However, the low accuracy is recorded and a score of 0.81 is recorded by CNN with word2vec feature extraction. In the second experiment, the authors used only two classes of classification by consolidating all hate classes and offensive classes as hate speech and all free speech (FS) classes into another class. Firstly, from classical ML classifiers like experiment one the highest accuracy which is 0.88 is recorded by SVM with TF-IDF and word2vec feature extraction. However, with Naıve Bayes algorithm and word2vec feature extraction technique achieved low accuracy likewise eight classes of classification. Secondly, from the ensemble ML classifier, the highest accuracy which is 0.87 is recorded by RF with TFIDF and word2vec feature extraction. However, a low accuracy is recorded which scores the XGboost algorithm with TF-IDF and word2vec feature extraction. Thirdly, from the deep learning classifier, the highest accuracy which is 0.88 is recorded by BiLSTM with direct embedding feature extraction of text. However, the lowest accuracy 0.82 is recorded by the CNN algorithm in combination with word2vec feature extraction techniques. From classical and ensemble machine learning algorithms SVM outperformed machine learning algorithms with word2vec feature extraction techniques which achieved 0.82 percentage of accuracy for eight classes of classification. From the deep learning algorithm, the BiLSTM algorithm achieved better accuracy which is 0.84 with pre-trained word2vec feature extraction techniques for eight classes of classification.

The authors built(Del Vigna et al., 2017)a corpus of comments retrieved from the Facebook public pages and groups of Italian newspapers, politicians, and artists by using web crawler. Then ascribe human annotators to assign one class to each post then they computed the Fleiss’ kappa κ inter-annotator agreement metric, which measured the level of agreement of different annotators on a task.In this study two different classification experiments were conducted the first considering the three different category of hate (Strong hate, Weak hate and No hate) the second considering only two categories, No hate and Hate, where the last category was obtained by merging the Strong hate and Weak hate classes by using two different classifiers Support Vector Machines (SVM) and the second one on a Recurrent Neural Network named Long Short Term Memory (LSTM). Lastly developed the first hate speech classifier for Italian texts. They suggested considering distinction among hate levels.

Shankar Biradar, (Biradar et al., 2022), The data collection contained 4575 code-mixed tweets, of which 1661 were Hate speech, and the remaining 2914 code-mixed tweets in the data set consist of Non-Hate speech. They used two transformer models for feature selection mBERT and IndicBERT pre trained on different Indian languages. For classification they tested conventional machine learning classifiers like Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RM), Naïve Bayes (NB), and K Nearest Neighbors (KNN) on translated and transliterated Devanagari script using mBERT embeddings. Then they experimented with the Deep Neural Network (DNN) model. Their experimental results found that their model outperformed existing state-of-art methods for Hate speech identification in Hinglish language with an accuracy of 73%. The experimental trials found that the mBERT model and traditional machine learning classifiers have performed better than IndicBERT for hate content detection in Hind and English datasets were used for this study.

3.0 COMPARATIVE ANALYSIS

There are studies done in different approaches and languages. Below an attempt has been made to show the full statistics of the approaches, and the language in which the research was conducted. Figure 2 shows different languages that hate speech detection study were conducted.

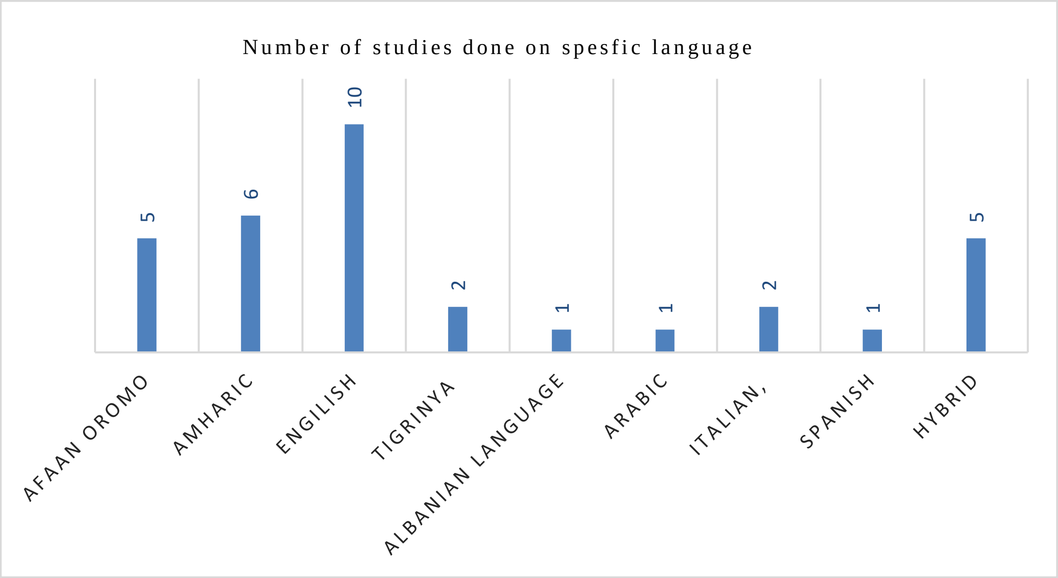

Figure 2: Number of Studies on different languages

Studies used different techniques and languages in order to detect hate speech content. Figure 3 shows techniques and languages used in different studies.The **Figure 3** shows that, firstly, most of the research was done with classical machine learning algorithms, and secondly, deep learning algorithms are being used by researchers and used deep learning as a hybrid with other machine learning techniques.Using deep learning algorithms with other machine learning algorithms, it is possible to develop a model that can identify good hate speech. Although deep learning algorithms require a large amount of data, their accuracy is high according to the reviewed studies, so it can be said that they are preferred.

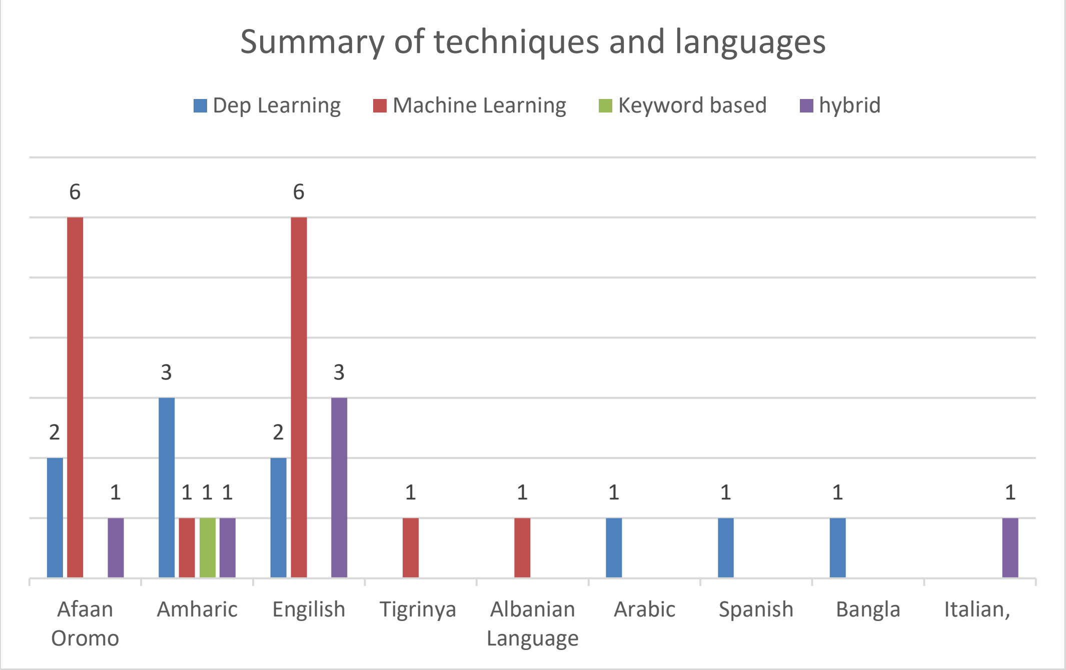

Figure 3: Summary of Techniques and Language

4.0 FINDINGS

NLP tools are needed to identify and predict hate speech. In addition, by adding a model complex, it was possible to confirm that the effort alone is not enough to achieve good results. Languages have their own dialects. As a result, a machine learning algorithm that is good for one language may not be good for another language at all. However, if the gaps in languages are carefully observed and enough research is done, it is possible to organize information for such studies.

Overall, the focus of this survey is to review the research done to identify and predict hate speech in the Amharic language, looking at their gaps and overall results. Based on this, it has been confirmed that Ethiopian languages need serious research. In addition, the lack of benchmarks and the lack of various NLP tools have created a huge gap in the language. Also, there is no data annotation standard used to label the data, and this has led to the release of research that seems careless and insufficient knowledge to be created.The problems we noticed in this survey look like this. Many studies and research have been done, but they have not been able to create enough knowledge on Amharic and other Ethiopian languages.

5.0 CHALLENGES

Significant research is being done on hate speech detection in English-language data. However, hate speech has become so widespread that it is seriously affecting non-English speaking communities. In particular, online posts containing hate speech to the extent that it looks like a campaign against African countries can be seen circulating on social media(Beltrami, 2021). However, in order to overcome the limitations of addressing the hate speech that is happening in African countries and other areas, the scientific community is required to work together. Ethiopia as an African country has gone through very difficult times due to hate speech.In addition to international pressure, there are elements that hide in society and spread hate speech in order to achieve their personal political needs and desires to disrupt people's coexistence.

Meanwhile, Ethiopian languages have not yet developed, resulting in many gaps in the field.However, there are a number of challenges in developing a hate speech identification and detection model using Ethiopian languages. The challenges are a lack of benchmark datasets for identifying hate speech, a lack of guidelines for data annotation, a lack of NLP tools for Amharic and other official languages, the diversity of language used by people on social media and its contextual meaning, accuracy, and choice of algorithms, way to preprocess data (Emoji, Figurative speech), etc(Aldjanabi et al., 2021). Researchers must continue to experiment with different methods to ensure that these experiments do not continue to be a problem. This is one of the aims of this document. All concerned parties should make an effort not to increase this hateful speech, especially the growing violence against girls. If not, it is safe to imagine that our world may be destroyed due to hate speech and fake information.

6.0 CONCLUSION

In this document, we presented a survey on the automatic detection of hate speech. This task is usually framed as a terminologies of hate speech, Stage of Hate Speech, approaches used to detect hate speech, features, such as a word2vec, bag of words or embedding’s, and the performance of the selected models. It has been tried to examine the methods used by many studies and evaluate the process by which they have preprocessed the data. In addition, efforts have been made to identify challenges. In general, the studies did not include figurative language and images that could convey specific hate messages. We have also confirmed that the existence of such studies has little implications for reducing the impact of hate speech. One of the challenges we've identified in this document has to do with data collection. This collection of data is aggregated and organized at the individual level, and the algorithms are evaluated only on the data collected and aggregated at the individual level. In this case, the data collected will only be labeled as hate speech when it comes to speech that appears to be bullying, offensive to ethnic minorities, etc. We propose for a benchmark dataset for hate speech identification to improve the comparison of different features and methods.

Acknowledgements

The authors would like to thank Cyber Security Research Division in Information Network Security Administration, for supporting this survey through the initiative of the Data Science Research Team.

REFERENCES

Ababu, T. M., & Woldeyohannis, M. M. (2022). Afaan Oromo Hate Speech Detection and Classification on Social Media. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, *June*, 6612–6619. https://aclanthology.org/2022.lrec-1.712

Abro, S., Shaikh, S., Ali, Z., Khan, S., Mujtaba, G., & Khand, Z. H. (2020). Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, *11*(8), 484–491. https://doi.org/10.14569/IJACSA.2020.0110861 admin

Gad, R., Gawali, P., Gite, M., & Pawa. (2022). Hate Speech Detection on Social Media Using Machine Learning Algorithms. *Journal of Cognitive Human-Computer Interaction*, *11*(06), 56–59. https://doi.org/10.54216/jchci.020203

Aldjanabi, W., Dahou, A., Al-qaness, M. A. A., Elaziz, M. A., Helmi, A. M., & Damaševiˇ, R. (2021). *Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model*. 1–13.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

Asians, E. C., Free-, P. R., Approach, P., & Extremism, A. V. (n.d.). *Media and Social Media Analysis on Religious Freedom and Violent Extremism in Central Asia : Cases of Kazakhstan , Tajikistan*.

Assefa, M. (2020). Role of social media in Ethiopia's recent political transition. *Journal of Media and Communication Studies*, *12*, 13–22.

Association, american library. (2017). *Hate Speech and Hate Crime", American Library Association*. https://doi.org/aa35c1c7-f3aa-4b07-964f-30dcf85a503c

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). *Deep Learning for Hate Speech Detection in Tweets*. https://doi.org/10.1145/3041021.3054223

Beltrami, F. (2021). *Ethiopia. Let's stop the hate campaign on social media. Your report counts!* https://www.focusonafrica.info/en/ethiopia-lets-stop-the-hate-campaign-on-social-media-your-report-counts/

Biradar, S., Saumya, S., & chauhan, A. (2022). Fighting hate speech from bilingual hinglish speaker's perspective, a transformer- and translation-based approach. *Social Network Analysis and Mining*, *12*. https://doi.org/10.1007/s13278-022-00920-w

Brüggemann, S., Robert Prosser, A., & Ru, S. (2022). *A scientific basis for a policy fighting fake news and hate speech*. https://doi.org/10.24989//ocg.v.342

Burnap, P., & Williams, M. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making: Machine Classification of Cyber Hate Speech. *Policy & Internet*, *7*. https://doi.org/10.1002/poi3.85

Chetty, Naganna,Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*(March 2017), 108–118. https://doi.org/10.1016/j.avb.2018.05.003

Das, A. K., Asif, A. Al, & Paul, A. (2021).

Bangla hate speech detection on social media using attention - based recurrent neural network. 578–591.

Defersha, Naol Bakala Tune, K. K. (2021). Detection of Hate Speech Text in Afan Oromo Social Media using Machine Learning Approach. *Indian Journal of Science and Technology*, *14*(31), 2567–2578. https://doi.org/10.17485/ijst/v14i31.1019

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). *Hate me, hate me not: Hate speech detection on Facebook*.

Edwards, S. S. M. (2017). *Cyber-Grooming Young Women for Terrorist Activity: Dominant and Subjugated Explanatory Narratives BT - Cybercrime, Organized Crime, and Societal Responses: International Approaches* (E. C. Viano (ed.); pp. 23–46). Springer International Publishing. https://doi.org/10.1007/978-3-319-44501-4_2

Fred Harter. (2022). *Ethiopia Gets Tough on Journalists Since Tigray Conflict*. https://www.voanews.com/a/ethiopia-gets-tough-on-journalists-since-tigray-conflict-/6683980.html

freedomhouse. (2021). *infrastructural limitations restrict access to the internet or the speed and quality of internet connections*. https://freedomhouse.org/country/ethiopia/freedom-net/2021

Ganfure, G. O. (2022). Comparative analysis of deep learning based Afaan Oromo hate speech detection. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00628-w

Getachew, S., & Kakeba, K. (2020). *Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network*. https://doi.org/10.21203/rs.3.rs-114533/v1

Getahun, T. G. (2023). Countering online hate speech through legislative measures: The Ethiopian approach from a comparative perspective. *The Communication Review*, 1–24. https://doi.org/10.1080/10714421.2023.2177487

Ghasiya, P., & Sasahara, K. (2022). Rapid Sharing of Islamophobic Hate on Facebook: The Case of the Tablighi Jamaat Controversy. *Social Media + Society*, *8*(4), 205630512211291. https://doi.org/10.1177/20563051221129151

Hate, N. O. (2020). *Racism, Intolerance, Hate Speech*.

Kiper, J. (2023). Religious Hate Propaganda: Dangerous Accusations and the Meaning of Religious Persecution in Light of the Cognitive Science of Religion. In *Religions* (Vol. 14, Issue 2). https://doi.org/10.3390/rel14020185

Lata Guta kanessaa. (2021). *Department of Computer Science Hate Speech Detection Framework from Social Media Content : The Case of Afaan Oromoo Language Lata Guta kanessaa A Thesis Submitted to the Department of Computer Science in Partial Fulfilment for the Degree of Master of Scie*.

Lumenwaymaker. (2019). *Identify the importance of avoiding hate speech*. https://courses.lumenlearning.com/wm-publicspeaking/chapter/hate-speech/

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, *14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Mahesh, B. (2019). *Machine Learning Algorithms -A Review*. https://doi.org/10.21275/ART20203995

Mnassri, K., Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2022). *BERT-based Ensemble Approaches for Hate Speech*

Detection. http://arxiv.org/abs/2209.06505

Mohtaj, S., Schmitt, V., & Möller, S. (2022). *A Feature Extraction based Model for Hate Speech Identification*.

Morris Kiruga. (2019). *Ethiopia struggles with online hate ahead of telecoms opening*. https://www.theafricareport.com/19569/ethiopia-struggles-with-online-hate-ahead-of-telecoms-opening/

Mossie, Z., & Wang, J.-H. (2018). *Social Network Hate Speech Detection for Amharic Language*. https://doi.org/10.5121/csit.2018.80604

MUNA SHIFA. (2022). *The Interaction of Mass Media and Social Media in Fuelling Ethnic Violence in Ethiopia*. https://www.accord.org.za/conflict-trends/the-interaction-of-mass-media-and-social-media-in-fuelling-ethnic-violence-in-ethiopia/

Nazmine, Manan, K., Tareen, H. K., Noreen, S., & Tariq, M. (2021). Hate Speech and social media: A Systematic Review. *Turkish Online Journal of Qualitative Inquiry*, *12*, 5285–5294.

Njagi, D., Zuping, Z., Hanyurwimfura, D., & Long, J. (2015). A Lexicon-based Approach for Hate Speech Detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*, 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

Nova, F., Rifat, M. R., Saha, P., Ahmed, S. I., & Guha, S. (2019). *Online sexual harassment over anonymous social media in Bangladesh*. https://doi.org/10.1145/3287098.3287107

Poni Alice JameKolok. (2022). *ways to counter hate speech in the Media through Ethics and Self-regulation*. https://en.unesco.org/5-ways-to-counter-hate-speech

Rahman, A., Manaf, A., Ismail, F., Kastriafuddin, T., & Tengku, S. (2018). *THE LEVEL OF SOCIAL MEDIA INFLUENCES ON CYBER HARASSMENT THE LEVEL OF SOCIAL MEDIA INFLUENCES ON CYBER*. *December*.

Raufi, B., & Xhaferri, I. (2018). Application of machine learning techniques for hate speech detection in mobile applications. *2018 International Conference on Information Technologies (InfoTech)*, 1–4.

Runswick-Cole, K. (2014). Disability hate crime. *A Companion to Criminal Justice, Mental Health & Risk*.

Saha, K., Chandrasekharan, E., & De Choudhury, M. (2019). Prevalence and Psychological Effects of Hateful Speech in Online College Communities. *Proceedings of the ... ACM Web Science Conference. ACM Web Science Conference*, *2019*, 255–264. https://doi.org/10.1145/3292522.3326032

Saha, P., Mathew, B., Goyal, P., & Mukherjee, A. (2019). *HateMonitors: Language Agnostic Abuse Detection in Social Media*.

Samuel, M. (2012). *Hate Speech Detection and Classification System in Amharic Text with Deep Learning* (Issue June).

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Seglow, J. (2016). Hate Speech, Dignity and Self-Respect. *Ethical Theory and Moral Practice*, *19*. https://doi.org/10.1007/s10677-016-9744-3

Sękowska-Kozłowska, K., Baranowska, G., & Gliszczyńska-Grabias, A. (2022). Sexist Hate Speech and the International Human Rights Law: Towards Legal

Recognition of the Phenomenon by the United Nations and the Council of Europe. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, *35*(6), 2323–2345. https://doi.org/10.1007/s11196-022-09884-8

Strategic Communications. (2022). *Hate speech poisons societies and fuels conflicts*. https://www.eeas.europa.eu/eeas/hate-speech-poisons-societies-and-fuels-conflicts_en

Tesfaye, S. G., & Tune, K. K. (2020). Automated Amharic Hate Speech Posts and Comments Detection Model Using Recurrent Neural Network. *Research Square*. https://www.researchsquare.com/article/rs-114533/latest?utm_source=researcher_app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound

Thomas, Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 512–515.

UN. (2019). United Nations Strategy and Plan of Action on Hate Speech. *United Nations Report*, *May*, 1–5.

Violence, O. G. (2023). *GENDER-BASED And Its Impact On The Civic Freedoms of Women GENDER-BASED And Its Impact On The Civic Freedoms of Women* (Issue March). https://www.icnl.org/wp-content/uploads/Online-Gender-Based-Violence-report-final.pdf

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *North American Chapter of the Association for Computational Linguistics*.

Yimam, S. M., Ayele, A. A., & Biemann, C. (2019). Analysis of the ethiopic twitter dataset for abusive speech in amharic. *ArXiv*, 1–5.

Zachary Laub. (2019). *Hate Speech on Social Media: Global Comparisons*. https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons