INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Detection of Hate Speech Text in Afan Oromo Social Media using Machine Learning Approach

Naol Bakala Defersha[1*], **Kula Kekeba Tune**[2]

1 Core Member, Center of Excellence for HPC and Big Data Analytics; Ph.D. Student, Assistant Professor, Department of Software Engineering, College of Electrical and Mechanical Engineering, Addis Ababa Science and Technology, Addis Ababa, 16417, Ethiopia
2 Head, Center of Excellence for HPC and Big Data Analytics; Assistant Professor, Department of Software Engineering, College of Electrical and Mechanical Engineering, Addis Ababa Science and Technology University, Adidas Ababa, 16417, Ethiopia

OPEN ACCESS

Received: 04.06.2021
Accepted: 16.08.2021
Published: 22.09.2021

Citation: Defersha NB, Tune KK (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

* **Corresponding author**.

naolbakala@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This study aims to develop a hate speech detection model for Afan Oromo's texts on social networks like Facebook and Twitter using a machine learning algorithm. **Methods:** we collected comments and posts from social media like Facebook and Twitter pages of BBC Afan Oromo, OBN Afan Oromo, Fana Afan Oromo Program, Politicians, Activists, Religious Men, and Oromia Communication Bureau using Face pager tool. The collected data was labelled using Afan Oromo hate speech evaluation system we developed. Text preprocessing tasks applied on data to remove special characters, stop-words, HTML Tags, extra whitespaces, numbers, lemmatization. The n-gram and TF-IDF was applied for feature extraction task to obtain benchmark Afan Oromo hate speech detection dataset. Researchers split dataset into train and test set. Finally, we applied Support Vector Classifier, Multinomial NB, Linear Support Vector Classifier, Logistic Regression decision tree and Random Forest Classifier on 67% of trained data. The performance of proposed model also evaluated using F-score. We also test the performance of developed model by loading test set into it. **Findings:** Hate speech on social media violates the welfare of Ethnic groups and citizens for living together. Many researches have been doing for English, Amharic, and other Languages to detect hate content from social media. This study has focused on developing a prototype for Afan Oromo hate speech detection model using machine learning algorithms and evaluate its performance in which we found Linear Support Vector Classifier scored highest f1-score value is 64%. **Novelty:** Afan Oromo hate speech detection framework proposed and successfully implemented to develop Afan Oromo hate speech detection model. We wrote python script that overcome problems typos in Afan Oromo in addition to designing python scripts that recognized apostrophe "**'**" as important letter for Afan Oromo word formation. Yet, no researchers have used combination of n-gram and TF-IDF for feature extraction. In this study, the n-gram and TF-IDF used for feature extraction approach to build model

that detect Afan Oromo hate speech on Social media.

Keywords: Afan Oromo; Decision tree; Facebook; Hate Speech; Linear Support Vector Classifier; Machine Learning; MultinomialNB; Social Media; Support Vector Classifier; Decision Tree and Random Forest Classifier

1 Introduction

Social media allows users to create, remove and share their ideas freely using the Internet connection. Recently, there have been several feature changes. For example, the maximum number of characters per tweet has recently been increased from 200 to 280, encouraging greater flexibility in interaction. Nowadays, social media allows users to freely communicate and express their ideas using natural language. In addition to this, it also deals with application of data and text mining approach like analyzing social network information retrieval, discovering patterns from a collection of data to investigate the secret form of information, opinion, or sentiments [(1)].

However, the challenges of using natural language over social media is generation of hate speech that violate rights of individuals by disseminating hate speech on various perspectives when users freely express opinion [(2,3)]. Even if there is no universally accepted meaning of hate speech, it is descriptive that propagate the defame related to individuals racial, Ethnic and like [(4)]. Hate speech is an expression that violates the right of people of different perspectives, insulting people those are part of religion, activists, by posting opinions, expressions, emotion and feelings over social media platforms. Hate speech is taken as harmful to social media and therefore, designing their rule and regulation to avoid such hate speech is essential [(3)].

As a result, social media platform does not have system that block hate content displaying and poisoning the social media environment natural language for natural languages used over it. In order to address those problems, some researchers developed hate speech for resource rich language such as Arabic [(1)], English [(5)], Amharic [(6)], Indonesian language [(7)], Indonesian-English [(8)], and others. Among machine learning algorithms, multinomial nave bays (MNB), support vector machine (SVM) algorithms, Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) and Bengali Unicode characters were implemented to build model that detect threat and Abusive Language on social media in Bengali Language in which Support Vector machine scored high accuracy than others algorithms in experiment conducted in [(2)]. After the authors applied preprocessing tasks on data collected from Indonesian Presidential election program in 2019, Latent Dirichlet Allocation was implemented for feature extraction from collecting data [(9)]. Natural Language Processing approach and machine learning algorithms like a Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF) algorithms were implemented to build a system that detects hate speech on the Twitter platform for Arabic Text document [(1)]. It was also confirmed that best result achieved by implementation of document representation and profile-based feature techniques as techniques [(1)]. Machine learning algorithms such as Random forest and Nave Bayes was implemented for learning after dataset resulted from preprocessing tasks such as annotation of collecting data, discourse analysis, content analysis and automated techniques to develop hate speech detection model for Amharic social media [(6)]. Word to vector and document representation (TF-IDF) are employed for feature selection in [(6)].

In Ethiopia, various social media such as Facebook, Pinterest, YouTube, Twitter, Instagram, Linkedin, Tumbir and others are available. The people of Ethiopia used Afan Oromo, Amharic, Tigrigna, Somali and English to on social media [(10)]. Amharic hate speech detection model for Facebook social media [(6)]. The author collected data from Facebook and applied the processing procedure to prepare Amharic hate speech data set [(11)]. The author Automated Amharic Hate speech Posts and Comments Detection

Model using Recurrent Neural Network[11]. The author tried to explore human-bird relationships of Oromo proverbs that associated with the Northern Ground Hornbill in Ethiopia[12]. The researchers fetched data from Facebook to code and categorize the essential ideas, deals with proverb concerned with the bird. Finally, the researchers also confirmed that Hornbill is culturally important bird in Oromo culture. Even if the researchers collected data from Facebook for this study, this work cannot limit to Afan Oromo proverb associated Hornbill[12].

In work of[10], automatic Sentiment Analysis developed for Afan Oromo using Machine Learning. Totally 6670 statements collected from Facebook and Preprocessing activities and normalization were applied to achieve quality dataset[10]. After dataset prepared split into 80% and 20% train and test set respectively, researchers applied Multinomial Naïve Bayes (MNB), Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) in first experiment and applied Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) in second experiments[10]. Finally, Afan Oromo sentiment analysis model with the highest precision performed by LSTM as indicated study from[10].

Even though, users are also using Afan Oromo on social media platforms to express emotions, feelings, and opinion in form of comments and posts that contain hatred ideas which leads to discrimination, social conflict, and even human genocide, yet, no research work attempted to develop hate speech detection prototype for Afan Oromo for any social media platforms. So, it needs to develop model hate speech detection model for Afan Oromo on social media.

Therefore, to fill this gap, we aim to develop automatic hate speech detection system for Afan Oromo for Facebook and Twitter social media by using Support Vector Classifier (SVC), Multinomial, Linear Support Vector Classifier (LSVC), and Random Forest Classifier.

Research Question

- What are preprocessing techniques need to be applied to prepare Quality Afan Oromo hate speech data set?
- What are appropriate software tools for data collection from Facebook and Twitter?
- What are appropriate feature extraction techniques need to be applied to obtain important features from Afan Oromo hate speech data?
- What is a framework to develop hate speech detection model for Afan Oromo social media?
- Which machine learning algorithms is the most performer to build Afan Oromo hate speech detection model?
- To what extent the performance of a system effective when different algorithms combined?

Related Works

The current literature shown that understanding and analyzing social media become a main concern. Today, one of the main concerns about social media is positive and negative impacts that comments and posts in social media platform have either on individual, groups or society. Hence, sentiment analysis, hate speech detection and classification, abusive language detection and/or classification, offensive language detection and/or classification and cyberbullying detection and/or classification become topic of research interest for researchers, Government and Social Media Company. Hate speech detection techniques that used to identify content is displayed on social media platform in the form of comments or posts irrespective of its nature whether the content is hate or normal. It used approaches such as machine learning, natural language processing, statistics and the like to design a model that detects hate speech. Using hate speech detection, natural language processing tasks and machine learning algorithms, comments and posts in social media platforms like Facebook, YouTube and Twitter can analyze and identify either as it is hate or normal.

In this section, we review related work from the perspective of Machine Learning algorithms, hate speech detection and classification, sentiment analysis, and natural language processing. Brief summary of literature review, we discussed in this study is indicated in Table 1. References, year of publications, feature extraction techniques used, the language on which research conducted, social media platforms selected for research, dataset used, availability of a dataset and performance of model developed are briefly summarized in the table.

The researcher collected data and applied pre-processing tasks such as tokenization and cleaning data performed on data collection in the work of[1]. The researcher collected data and applied pre-processing tasks such as tokenization and cleanup data on data collection in[1]'s work. The performance of the system developed was assessed several times and ultimately confirmed that the random forest classifiers obtained an average g of 91.2%[1].

Code-mixed data concept applied for hate speech detection in work of[7]. Data were collected from three sources using the Twitter API and annotated by three Hindi and English experts. In text preprocessing phase URLs, user names, hashtags emoticons punctuation marks unwanted characters eliminated and extra white spaces erased from collected data and finally, each sentence converting all words to lowercase[7]. As feature extraction techniques, the author used fast Text and domain

specific word embedding[7]. Machine learning algorithms such as SVM, SVM-RBF and Random Forest applied to build hate speech detection system in which experiments conducted three times in the work of[7]. In all three experiments, SVM, SVM-RBF and Random Forest algorithms applied their performance compared and finally, SVM-RBF algorithm scored f1-sore 85.80%[7].

In study[13], the author developed apache spark-based hate speech detection to classify Facebook contents into hate and not hate by employing Random field and Naïve Bayes as training algorithms. The author also confirmed that word2vec embedding feature extraction approach scored higher accuracy than TF-IDF techniques[13]. At the end of the experiment, the hate speech detection model resulted with an accuracy of 79.83%[13].

Integrated deep feature extraction resulted from CNN trained on semantic word embedding and n-gram approaches were applied as feature extraction techniques[14]. The experiment was conducted on data collected whose size is 16k and tweets annotated manually. By understanding that accuracy of model to be developed affected by the accuracy of the features extracting during feature extraction stage[14]. The author of[14], used three feature fusion approaches such as early fusion, late fusion, and middle fusion. The classifier algorithms such as Logistic Regression, Random Forest, and Support Vector Machine[14]. Data collected from twitter social media to develop a system that identifies hate speech in Indonesian language[14]. As feature extraction techniques, word embedding technique applied for collecting data to obtain important feature[14]. To develop a hate detection system, the author applied Long-Short Term Memory (LSTM)[14]. The performance of developing a system evaluated by using F1-score performance evaluation parametric and scored f1-score value of 94.5%[14].

The author of[15] attempted to develop offensive and hate speech detection system for Danish language. Dataset named as DKHATE was prepared from comments generated by users over social media platform and data annotated to different language offensive and target. Social media platforms such as Facebook, Reddit and Twitter were used as sources of data[15]. Data collected annotated by using annotation guide line stated in work of Zampieri et al. (2019a) and final data set which divided into train set 80% and test set 20% in which 88% of data labelled as offensive[15]. Classification system executed in subtasks Offensive language identification, Automatic categorization of offensive language types, and Target identification in such way that each subtask contributed to construction of offensive and hate speech detection system in work of[15]. The authors developed offensive detection system English and Danish Language for system performance scored 74% and 70% respectively[15].

In study[16], the authors collected datasets which contain text and image dataset with size 1100 data with hate content and 1100 data with no hate contents. The data collection approach was conducted manually from Facebook, Twitter, Line Today, YouTube comments, and YouTube video transcript; whereas audio data was collected through recording from speakers (19 men and 8 women) in such way that one speaker has to speak 60 up to 110 texts with hate and no hate contents study[16]. After the authors obtained data, researchers employed three experts to annotate data to hate or not hate and then text preprocessing task were applied onto clean collected data[16]. In work of study[16], researchers used textual, acoustic and their combination for feature selection and compared their accuracy. The model with highest accuracy resulted using textual feature whose performance is F1-score 87.98% than acoustic of F1-score 86.98% accuracy. Finally, the authors used deep learning model based on the Long Short-Term Memory to build model study[16].

The author of[17], collected large size dataset from social media, particularly Twitter and annotated dataset. Multimodal data of 150,000 tweets were collected from those contained textual and image data. Series of steps such as collecting tweets, textual image filtering and annotating data, were involved in data collection by researchers in of[17]. The dataset contained textual dataset and Visual data on which experiment was conducted separately and jointly[17]. The performance compared to unimodal data and multimodal dataset in the work of[17]. The author concluded that the hate speech model built depending on image data which made their dataset multimodal that could not result in high accuracy[17]. They proposed an approach of hates speech detection using more than one data with large size, constructing hate speech detection model of unimodal data and multimodal[17].

Afan Oromo Sentiment analysis model for Facebook developed by[10]. The author has applied text preprocessing tasks, text normalization and finally applied conventional machine learning algorithms and neural network to develop the model. In this study, the author approved that the neural network is a higher performer comparing with conventional machine learning algorithms[10].

2 Materials and Methods

The following series of steps were carried out to achieve the objective of this study, that is to detect hate speech from Facebook and Twitter comments and posts effectively. Research design is an approach that enables us to combine various methodologies in research to solve identified problems. Using appropriate research design is essential for researchers to solve selected problems properly. For accomplishment this study, we used the following techniques of research design.

Table 1. Brief summary of literature review

References	Language	Feature extraction techniques	Social Medias	Algorithms	Dataset	Availability	F1-Score
(15)	Danish and English		Facebook, Reddit and Twitter	-	original	No	74%
(17)		-	Twitter	LSTM	Original	no	70.4%
(16)	Indonesian	textual, acoustic and their combination	Facebook, Twitter, Line Today, YouTube	Long Short-Term Memory	Original	No	87.98%
(7)	Indonesian	word n-gram charactern-gram orthography lexicon	Twitter	Vector Machine (SVM), Naive Bayes (NB),and Random Forest Decision Tree (RFDT)	Existing	yes	77.36%
(1)	Arabic	BoW, TF, and TF-IDF	Twitter	Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree(DT) and Random Forest (RF)	Original	no	91.2%

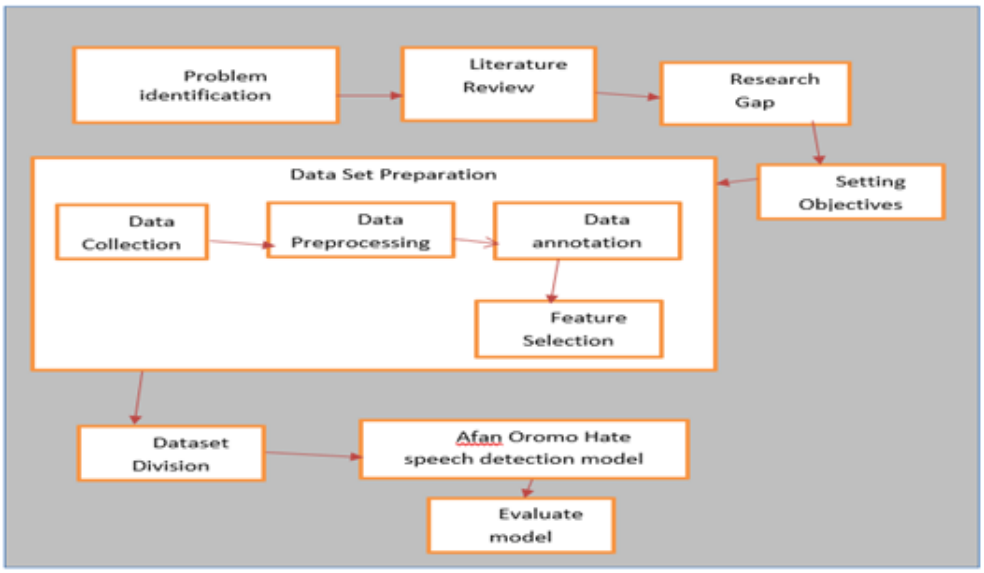

Fig 1. Research flow Diagram for Afan Oromo hate speech detection

2.1 Data set Preparation

2.1.1 Data collection

In the present day, social media is used as a source of data for research. The author of[12], collected Hornbill related Afan Oromo data from Facebook to code and categorize the terms related with proverb concerned. In order to build a sentiment analysis prototype for Afan Oromo, researchers collected data from Facebook[10]. In our research, in addition to Twitter, we also used Facebook as a data source to develop Afan Oromo hate speech detection dataset.

Table 2. Data set sources and Size

Source	Hate	Normal
Facebook	2900	3700
Twitter	3562	3438
Total	6,462	7,138
	13600	

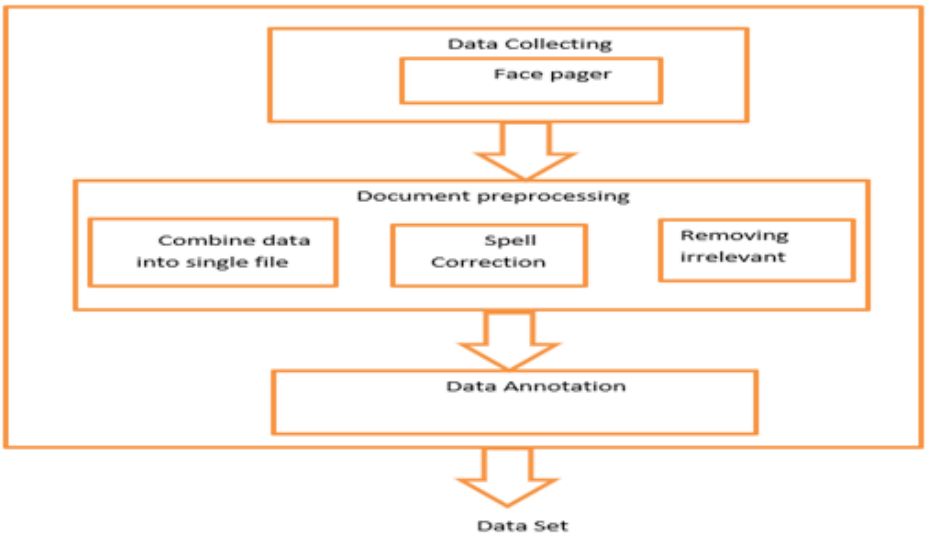

Fig 2. Dataset Preparation Architecture

Our dataset is composed of user written comments and posts from different public pages on Facebook and Twitter of BBC Afan Oromo, OBN Afan Oromo, Fana Afan Oromo Program, Politicians, Activists, Religious Men, and Oromia Communication Bureau.

We collected 13600 comments and posts between September 2019 and 2020 on respective public page using Face pager (https://facepager.software.informer.com/3.6/) in which 7000 and 6600 data were collected from Twitter and Facebook.

Researchers performed word preprocessing tasks to clean up collected data from irrelevant content and consolidated all data into one single file with a comma-separated value (CSV). Data contents also labelled into main classes either hate or normal.

2.1.2 Document preprocessing

Combine collected data into one File: Data collected from various pages of Facebook and Twitter pages. To make collected data ready for text preprocessing, data annotation and data splitting, researchers merged all collected data into single file name with CSV file extension "**AOHSD dataset.CSV**".

Spell correction: most people type Afan Oromo words with the correct spelling, whereas few people type Afaan Oromo words incorrectly. An Afan Oromo word with incorrect spelling changes the meaning of a sentence, paragraph and entire document written in Afan Oromo. To overcome challenges of the misspelled word in Afan Oromo, we identify words with misspelling and try to replace them with correctly spelled words by writing using python scripting.

Removing Irrelevant Contents: The text preprocessing tasks are essential to achieve relevant dataset[7]. In this step, we identified punctuation marks, special symbols, emoji, number, URL and stop words thoroughly. As indicated in the work by[18] irrelevant data has to be eliminated at the text preprocessing phase. To clean punctuation marks, special symbols, emoji, number, URL and stop words, first the plain texts are tokenized into tokens by tokenization process. A second list of stop words in Afan Oromo Languages, punctuation mark, special symbols, html, and URL removed from the data. Finally, researchers wrote python scripts to carry out text preprocessing tasks and removed stop words, punctuation mark, special symbols, convert upper case to lower case, html, and URL. Among all punctuation marks, pseudo code internationally prepared to remove all function marks except, apostrophe "hudhaa" " ' " that helped for word formation in Afan Oromo.

Start:

1. Open the file
2. Read text in dataset;
3. While (! end of text in dataset):
   If the text contains symbol [= <> << >> +! ~] then
   Remove symbol and add space
   If text contain special_char [,!#$@%^*] then
   Remove special_char

If text contains number= [0-9] then
Remove number
If text contains emoji= [EMOJI] then
Remove emoji
If text contains white space, then
Trim text

4. Return corpus;
END:

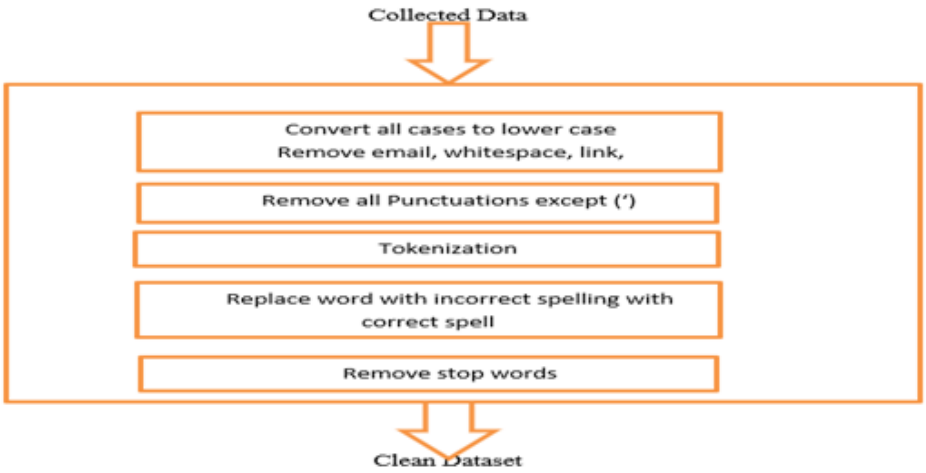

Fig 3. Afan Oromo Text document preprocessing Architecture

2.1.3 Natural Language Processing Tasks in Hate Speech Detection

In this work, we divided hate speech identification tasks into subtasks that were finally combined to support proposed hate speech detection and classification. Each subtask purposely designed to aim of handling all about hate speech. We divided hate speech detection tasks into five tasks. Task 1. Hate speech identification tasks: At hate speech identification tasks, researchers identified whether the given posts and comments are either hating speech or normal speech. Annotations of data carried out based on the specific labels. Task 2. Automatic Hate speech detection tasks: the aim of task in step is to check whether the posts and comments are hating speech or normal based on the label in task 1. Task 3. Automatic Hate speech classification task: in this step the target of the hate speech identified. Task 4. Identifying target: the class identified in tasks 3, the aim of target identification is identifying the target of hate speech based on labeled posts and comments. Task 5, Target of Speech Identification: Analyzing the contents of the text either text content is hating or normal is essential. Therefore, target of speech identification level, the researcher analyzed content, text document content with the help of experts.

2.2 Data Annotation

Nowadays, researchers are using machine learning techniques such as unsupervised, supervised, or semi supervised to conduct experiments. Supervised technique is the most dominated approach in the hate speech detection in social media that require manual annotation of the dataset. In this study, we used manually annotated data used for supervised machine learning algorithms. We used five experts to annotate data depending on the annotation procedure prepared. The number of experts limited to five due to resource scarcity. Experts recruited for data annotation were MA and above MA holders. The selection of experts depends on their interest, computer skill and knowledge in Afan Oromo. The annotators used Afan Oromo hate speech dataset annotation procedures to annotate data whenever they want within a limited period of time.

The researchers generated username and password for experts to login into Afan Oromo hate speech detection dataset evaluation system hosted on the website (https://www.naolinfo.info). After the experts successfully logged into the system, the system allows them to choose Afan Oromo hate speech dataset evaluation page from the displayed list of pages (see figure:

4) on Afan Oromo hate speech dataset evaluation page, the experts read the contents of document to be annotated first and then select hate or normal radio button. Finally, the expert submits his/her options by clicking the "Submit and next" button. After all, five experts submit their selection, the system assigns the class to the data depending on the number of experts select given labels.

Rules employed for Afan Oromo hate speech detection data annotation

If the content of the document contains the following, it has to be labelled as hate, otherwise normal

- if it contains ideas that against gender "Yoo documenting jibe sale when agarsiisuu of cases qabaate"
- If it contains ideas that against groups of persons "Yoo qabiyyeen barreeffamichaa can agree namoota motion faalleessu qabaate".
- if it contains ideas that insults or curse individuals "Yoo qabiyyeen barreeffamichaa nama kan arrabsu tea"
- if it contains ideas against that the specific religion, political party and Ethnicity "Yoo qabiyyeen barreeffamichaa amount, siyaasaafi Saba Tokko can faalleessu tea"
- If it contains ideas that motivate the people or group of people to violence "Yoo qabiyyeen barreeffamichaa hookkaraaf Kana name kakaasuu tea".

In the following Table 3, Sample example of data annotated.

Table 3. Sample Annotated Afan Oromo Text document

Sno	Content/qabiyyee	class/garee
1	Oromo is enemy of Ethiopia "Oromoon diina Itiyoophiyaati"	hate "jibba"
2	Selfish"Abbaa garaa"	hate "jibba
3	struggle you contributed is unforgettable "qabsoon ati giite hin dagatamu"	normal "fayyaalessa"

The meaning of the text document in Oromo is enemy of Ethiopia "Oromoon diina Itiyoophiyaati" is against Oromo Ethnic group and has to labelled as **hate "jibba"**. Like that selfish "abbaa garaa" is insulting somebody, therefore annotators annotate it as **hate "jibba"**. In oromo culture, selfish "abbaa garaa" anybody who is not worrying about others even for his/her brother if they achieve what they want. The contents of "struggle you contributed is unforgettable "qabsoon ati giite hin dagatamu" free and it has to be labelled as normal "fayyaalessa" by annotators.

2.3 Features Extraction

To extract features from all features in the dataset, we used a combination of approaches used in [1] and [7] with combinations for our approach.

N-Gram: n-gram is count of number words n in given Afan Oromo collected data that can be between 1 and N range based on word features. We evaluated the performance of the model, by assigning n value from n=1 to n=5 and the model with highest performance achieved bigram when n=2. In addition to n-gram the TF-IDF is also used.

TF-IDF: TF-IDF is an essential approach for document representation which can be calculated from the number of Terms, and number of documents in corpus. Term Frequency (TF) for given documents in the corpus show the number times a particular word found in the documents whereas the Inverse Document Frequency (IDF) represents the number document in which the term occurs from all available documents in the corpus.

TF-IDF (t, d) as computed as F (t, d) * IDF(t), when IDF (t) = log [n / (DF(t) + 1] such that n = total number of documents in the document set DF(t) = document frequency of t The weighted n-gram is given as: $W(t,d) = NGram(t,d) \times TF\text{-}IDF(t,d)$ [19].

In our approach, we used the combination of Bigram and TFIDF to select important features to obtain relevant Afan Oromo hate speech detection dataset.

2.4 Dataset Division into Train set and Test set

Dataset is used as input to conduct experiments and play a vital role to obtain right output from experiment. Since we used a supervised machine learning approach to the train model, the final annotated dataset is split into a train and test set with a fair distribution of classes and data. Train data contain 67% whereas the remaining part is test set.

2.5 Programming Language and Tool Used

As indicated in the previous section, we collected data from channels of BBC Afan Oromo, OBN Afan Oromo, Fana Afan Oromo Program, Politicians, Activists, Religious Men, and Oromia Communication Bureau. Face pager is a tool that retrieves data from Facebook and Twitter pages and saves retrieved data in csv format on a local machine [15].

Face pager: In this study, we used face pager to collect posts and comments from Facebook and Twitters Pages.

MySQL database: MySQL database server used to develop the Afan Oromo hate speech dataset annotator system alongside using php.

Python programming language: Python programming language is a powerful programming language currently used in various disciplines. In this particular work, python programming language is used for data preprocessing, dataset splitting and model development.

2.6 Afan Oromo Hate Speech Detection

Machine learning algorithms applied to develop Afan Oromo hate speech detection model. The machine learning algorithms, particularly, supervised machine learning require properly annotated dataset to obtain models with highest performance. We annotated a dataset for Afan Oromo hate speech detection depending on the annotation procedure prepared.

2.6.1 Machine learning algorithms

Several activities such as text classification, text categorization, pattern recognition, pattern discovery, decision making and the like, those that need human intelligence are automating by Machine learning. Machine learning is a branch of Artificial Intelligence, which is categorized into supervised, unsupervised. In the machine learning approach for predefined classes, a document that will be classified manually by the user always exists. Therefore, the predefined data sets are used for automatically learning the meaning that the user assigned attributes to the classes due to the existence of available data. It contains two main learning approaches: unsupervised learning and unsupervised learning approaches.

Supervised learning approach needs predefined class and deals with classification techniques; whereas unsupervised learning approach does not predefined data and deals with clustering techniques. Supervised machine learning approach requires human involvement partially for a labelling class of data, to divide a dataset into train and test dataset. Decision tree, support Vector machine and Naïve Bayes are the most known supervised machine learning algorithms.

As we understand from literature review, currently, supervised machine Learning algorithms are also utilized for hate speech detection and classification. In our work, we also used machine learning algorithms listed under for conducting experiments then compared their performance

2.7 Evaluation System

2.7.1 Dataset Labelling Evaluation System

Researchers strategically identified the classes of a hate speech detection dataset into hate and normal. Afan Oromo hates speech detection dataset classes become the name of two radio buttons for row data displayed from the database that holds hate speech detection dataset which was created in the MySQL database we used as back-end software for evaluating Afan Oromo hate speech detection system. Depending on the numbers of experts assigned either hate or normal label (see Figure 4), the system selects classes and assigns them to each in the database. To study, since we used five experts to annotate data, the class of three or more than three experts will assign as a class to the data.

2.7.2 Performance evaluation parameters

In machine learning techniques, accuracy, precision, recall and f-measure are used as the main performance evaluation techniques [9]. Among those performance evaluation parameters, f-measure is the average of Precision and recall. Therefore, in this research work, f-measure/score was used to evaluate the performance of Afan Oromo hate speech detection system [9]. Among those performance evaluation parameters, f-measure is the average of precision and recall. In the confusion matrix, the performance of each machine learning algorithm is evaluated using comparatives of Accuracy, Recall, Precision and F-score. In our study, we evaluated the performance of algorithms we used in experiments. Therefore, in this research work, f-measure/score was used to evaluate the performance of Afan Oromo hate speech detection system.

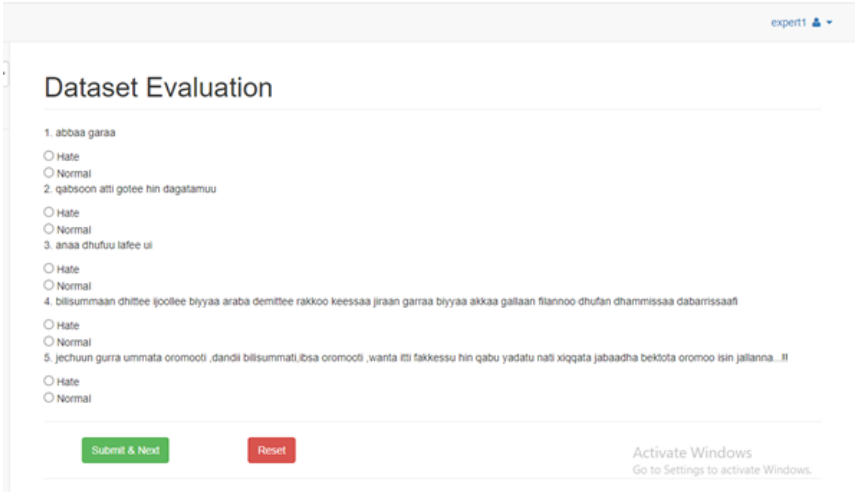

Fig 4. Afan Oromo hate speech detection data Evaluation page

2.7.3 Confusion Matrix

Confusion matrix is a table that visualizes the performance of machine learning algorithms. In the confusion matrix, the variable has positive or negative values such that columns represent the actual values of the variable whereas the row of confusion matrix represent the predicted values.

3 Results and Discussion

The experiment conducted on Afan Oromo hate speech detection dataset using Python 3.6 to develop the proposed model.

Experimental Setup

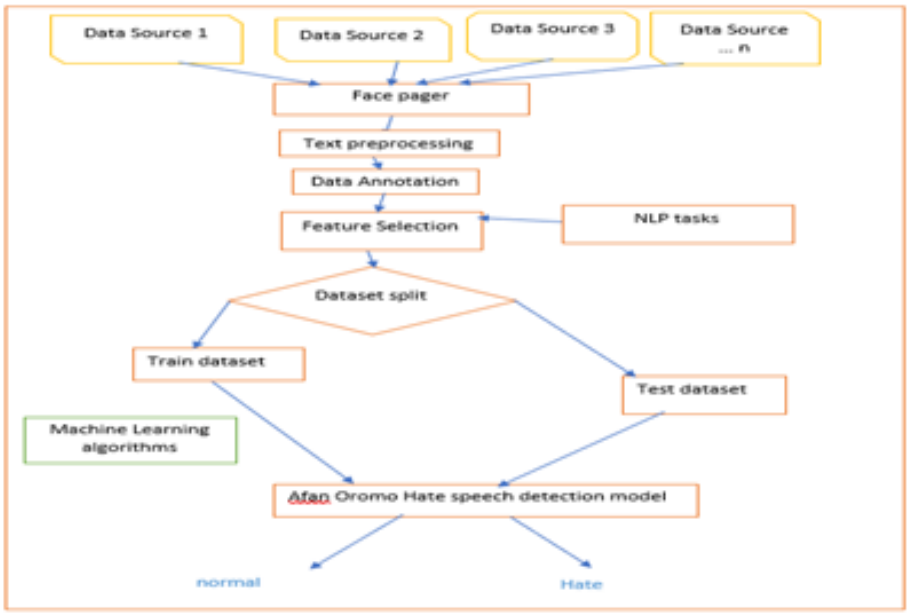

Fig 5. Afan Oromo hate speech detection framework

We present the performance of Linear Support Vector Classifier, Multinomial NB, Random Forest Classifier, Logistic regression, Decision tree and Support vector classifier in Table 4.

Table 4. Machine leering algorithms performance

S.no	Algorithm	Precision	Recall	F-score
1	Linear SVC	66	66	64
2	Multinomial NB	66	65	62
3	Random Forest Classifier	64	64	63
4	Logistic Regression Classifier	65	64	61
5	Support Vector Classifier	66	65	62
6	Decision Tree Classifier	59	59	59

Training set : In order to build the Afan Oromo hate speech detection model, we used 67% of the dataset by applying various machine learning algorithms. The total data of 13600 train sets used for training model is 9112(see Table 4).

Test set : To evaluate the performance of the Afan Oromo Hate speech detection model, we used two ways. First, we loaded 4488 of test data. In the first way, the model tried to predict the class for the loaded test dataset. Second, any comments/posts in text format loaded to the model and able to calculate the class of the loaded text document into either hate or normal.

3.1 Results

Afan Oromo hate speech detection data collected from Facebook and Twitter social media platforms using Face pager. The system we developed using php and MySQL database assigned labels for the loaded data into the database. Generating accounts for experts of the developed system able to annotate the dataset. From Annotated Afan Oromo hate speech dataset, train and test data set obtained after the annotated dataset divided into using python programming language. The important feature selected from the prepared dataset helped to result in a Benchmark Afan Oromo hate speech dataset that contains the train and test set.

We conducted experiments by loading machine learning algorithms turn by turn on the dataset and the performance of each applied algorithm demonstrated in Table 5. The performance of each applied machine learning algorithm is also indicated in Table 5. The performance of each classifier is illustrated by accuracy, precision, recall and F-score measures Table 5. The developed Afan Oromo hate speech detection was able to be tested with the test dataset scored performance of 64%.

3.2 Discussion

The study is centered on developing hate speech detection models for Afan Oromo social media platforms, specifically from Facebook and Twitter. For successful development of the proposed model, we performed a series of activities. First, data collected from selected sources and annotated according to prepared procedures. Then, text preprocessing applied on gathered data to select relevant data and remove irrelevant data. At text preprocessing phase, Afan Oromo stop words, punctuations except', numbers, all none Afan Oromo text document, row with empty space, image, video, audio, link, emoji and email removed. All typos errors tried to replace by word correct spelling. We also applied data normalization. Next to those, feature selection techniques such Bigram and TF-IDF applied for data vectorization. On vectorized Afan Oromo hate speech data, supervised machine learning algorithms were applied. To conduct the experiment, we used linear Support Vector Classifier, Multinomial NB, Random Forest Classifier, Logistic regression and Support vector classifier. From all machine learning algorithms applied to build models, linear support vector classifiers achieved higher accuracy than others and the linear support vector classifiers selected as the highest performer.

Finally, we also tested the performance of developed Afan Oromo hate speech detection using a test data set and model scored f-score 64% Table 4.

Since, the Linear support Vector classifier shows good results to detect hate contents on both training and testing depicts that linear support vector classifier has trained from training data and can also apply the knowledge to new text document with unknown class.

Finally, Afan Oromo hate speech text model from Afan Oromo posts and comments can identify hate speech contents by training by training using dataset collected from Facebook and Twitter in Afan Oromo. This model can be challenged by detecting and alerting the hate contents from Facebook and Twitter. The output of this developed Afan Oromo hate speech detection model can overcome the problems that may the country face due to hate speech if properly implemented by the Ministry of peace in Ethiopia and social media companies.

4 Conclusion

We have outlined that developing hate speech detection for Afan Oromo social media is essential to eradicate the risk of hate speech on social welfare.

Our work has led to the conclusion that machine learning is applicable for the development of hate speech detection models for Afan Oromo on Facebook and Twitter.

We conducted experiments six times by applying machine learning algorithms such as Support Vector Classifier, MultinomialNB, Linear Support Vector Classifier, Logistic Regression and Random Forest Classifier to build hate speech detection prototypes for Facebook and Twitter. To evaluate the performance of each algorithm, researchers used performance metrics such as Accuracy, Precision, Recall and F-score. The feature selection techniques for machine learning, bigram and TF-IDF applied. The result of the study indicated that Support Vector Classifiers achieved Linear support Vector classifier Performance Precision 66%, recall, 66% and F-score 64%. The Multinomial NB achieved performance Precision 60%, recall 65% and F-score 62%. The Random forest classifier achieved performance Precision 64%, recall 64% and F-score 63%. The Logistic Regression classifier achieved the Performance Precision 65%, recall 64% and F-score 61%. The Support Vector Classifier achieved performance Precision 66%, recall 65% and F-score 63%. The result of the experiment shows that the performance of Linear Support Vector Classifier scored f1-score value is 64% and we have confirmed that Linear Support Vector Classifier scored highest performance compared with others. Therefore, the researchers agreed to use Linear support vector classifiers to deploy Afan Oromo hate speech detection model.

Even though we have developed the Afan Oromo hate speech detection model using machine learning algorithms by collecting data from Facebook and Twitter, this study only investigated posts and comments in text documents. The posts and comments in mode of image/photo, audio and video data have not been considered. The most important limitation of this study also lies in applying conventional machine learning algorithms that need manual labelling of dataset.

In this study, experiments conducted on data were of small in size. In future study can also be conducted by collecting data from other Social Media platforms. In addition to collecting data from other social media platforms, the researchers can consider other modes of data for further research to be investigated. Applying beyond conventional machine learning algorithms for experiments can also be the next study.

References

1) Alfawareh IAM, Alfawareh M, Hammo B, Hijazi N. Intelligent detection of hate speech in Arabic social network: A machine learning approach. *Journal of Information Science*. 2020. doi:10.1177/0165551520917651.
2) Chakraborty P, Seddiqui MH. Threat and Abusive Language Detection on Social Media in Bengali Language. *1st International Conference Advances Science Engineering Robotics Technology 2019, ICASERT 2019*. 2019;2019:1–6. doi:10.1109/ICASERT.2019.8934609.
3) Macavaney S, Yao HR, Yang E, Russell K, Goharian N, Frieder O. Hate speech detection: Challenges and solutions. *PLoS One*. 2019;14(8):1–16. doi:10.1371/journal.pone.0221152.
4) George C. Hate Speech Law and Policy. *International Encyclopedia Digitial Communication Society*. 2014;p. 1–10. doi:10.1002/9781118767771.wbiedcs139.
5) Zampieri M. Detecting Hate Speech in Social Media. . Available from: https://www.aclweb.org/anthology/R17-1062/.
6) Mossie Z, Wang JH. Social Network Hate Speech Detection for Amharic Language. 2018. doi:10.5121/csit.2018.80604.
7) Ibrohim MO, Budi I. Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter. 2019. doi:10.18653/v1/w19-3506.
8) Sreelakshmi K, Premjith B, Soman KP. Detection of Hate Speech Text in Hindi-English Code-mixed Data. *Procedia Comput Science*. 2019;171:737–744. doi:10.1016/j.procs.2020.04.080.
9) Febriana T, Budiarto A. Twitter Dataset for Hate Speech and Cyberbullying Detection in Indonesian Language. *Proc 2019 International Conference Information Management and Technology ICIMTech 2019*. 2019;1:379–382. doi:10.1109/ICIMTech.2019.8843722.
10) Kuyu SJ. Developing an Automated Machine Learning Based Sentiment Analysis for Afaan Oromoo. 2021.
11) Tesfaye SG, Tune KK. Automated Amharic Hate speech Posts and Comments Detection Model using Recurrent Neural Network. 2020. doi:10.21203/rs.3.rs-114533/v1.
12) . . Available from: https://ssrn.com/abstract=3770521.
13) Sajjad M, Zulifqar F, Khan MUG, Azeem M. Hate Speech Detection using Fusion Approach. *ICAEM 2019 - Proc*. 2019;p. 251–255. doi:10.1109/ICAEM.2019.8853762.
14) Sazany E, Budi I. Deep Learning-Based Implementation of Hate Speech Identification on Texts in Indonesian: Preliminary Study. *Proc ICAITI 2018 - 1st Information Management and Technologyl Innovation Towar A New Paradig Des Assist Technol Smart Home Care*. 2018;p. 114–117. doi:10.1109/ICAITI.2018.8686725.
15) Sigurbergsson GI, Derczynski L. Offensive language and hate speech detection for danish. *Lr 2020 - 12th International Conference Language Resource Evaluation Confernce Procedia*. 2020;p. 3498–3508. Available from: https://arxiv.org/abs/1908.04531.
16) Sutejo TL, Lestari DP. Indonesia Hate Speech Detection using Deep Learning. *International Conference Asian Language Process*. 2018;p. 39–43. Available from: https://ieeexplore.ieee.org/document/8629154.
17) Gomez R, Gibert J, Gomez L, Karatzas D. 2019. Available from: https://arxiv.org/abs/1910.03814.
18) Mhamdi C, Al-Emran M, Salloum SA. Text mining and analytics: A case study from news channels posts on Facebook. *Stud Computer Intelligent*. 2018;740:399–415. Available from: https://link.springer.com/chapter/10.1007/978-3-319-67056-0_19.
19) Oriola O. Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets. 2020.